GeoDiff4D: Geometry-Aware Diffusion for 4D Head Avatar Reconstruction

Supplementary Material

A. Implementation Details

A.1. Video Generation Model

Noise Schedule During inference, we do not initialize the diffusion process from pure random noise. Instead, we inject reference information directly into the initial latent by adding Gaussian noise to both the reference image and its corresponding normal map, and then use these noisy signals as the starting point for DDIM denoising. This reference-conditioned initialization encourages the model to preserve identity features, fine-grained facial details, and geometric structure more faithfully throughout the iterative sampling process, reducing ambiguity compared to fully noise-based initialization. Concretely, we first construct the two domain-specific latents (image and normal) independently and then concatenate them along the domain dimension to form an initial latent of shape $[B \times D \times C \times T \times H \times W]$, following the formulation described in Section 3. By starting from this structurally coherent latent, the model benefits from stronger reference guidance, resulting in more consistent and identity-preserving outputs across both spatial and temporal dimensions.

Classifier-Free Guidance During training, we randomly drop the expression latent with a probability of 0.1 to facilitate classifier-free guidance, enabling the model to learn to generate outputs with and without explicit expression conditioning. During inference, classifier-free guidance is applied simultaneously to both domains to maintain consistent control over the generated image and normal signals. Specifically, after constructing the initial noisy latents as described previously, we first reshape them into an image–normal latent of shape $[(B\times D)\times C\times T\times H\times W]$. This latent is then duplicated along the batch–domain dimension to form an unconditional–conditional latent pair, which allows the guidance mechanism to differentiate between conditioned and unconditioned signals. All other conditioning inputs, including expression latents, head pose maps, and class labels, are processed in the same manner to maintain alignment across all control signals. Finally, we apply the standard classifier-free guidance procedure with a guidance scale of 2.5 for all experiments, ensuring strong adherence to the desired expressions and head poses while enhancing overall fidelity and consistency across both spatial and temporal dimensions.

Video Synthesis for Reconstruction We use the video generation model to synthesize videos of approximately 200 frames across 12 viewpoints. Specifically, we select 12 out of the 16 available viewpoints from NeRSemblev2 and adopt their camera configuration, keeping the head pose fixed while animating only the facial expressions of the reference image. We then concatenate all video clips into a single sequence, treating it as a monocular video, which is subsequently used for avatar reconstruction.

A.2. 4D Reconstruction

FLAME Refinement To refine the initial FLAME tracking and compensate for errors from monocular fitting, we introduce learnable residuals for all FLAME parameters, including shape, expression, jaw, neck, eye, and eyelid coefficients, as well as global pose (R, t). These parameters are jointly optimized with the Gaussian attributes using a three-phase learning-rate schedule (warmup, stable, decay). We divide the FLAME parameters into two groups: pose- and shape-related parameters are trained with a conservative schedule and a peak learning rate of $1\mathrm{e}{-5}$, while expression-related parameters adopt a higher peak learning rate of $1\mathrm{e}{-4}$ to better capture fine-grained facial dynamics.

All parameters start from an extremely low learning rate of $1\mathrm{e}{-10}$ and linearly ramp up during the first 40K iterations to prevent instability in early training. The learning rate is maintained at its peak from 20K to 80K iterations to allow sufficient exploration and then exponentially decayed from 80K to 100K iterations to ensure convergence. This staged schedule, combined with group-specific learning rates, enables stable joint optimization of tracking parameters and Gaussian attributes while preserving photometric fidelity and temporal coherence.

Topology-Preserving Remeshing To ensure geometric consistency when remeshing the FLAME template into UV space, we adopt a topology-preserving strategy that validates the connectivity of newly generated faces. Standard UV remeshing subdivides each UV grid cell into two triangles to create a dense tessellation, but this approach can inadvertently produce invalid faces that connect distant or topologically unrelated regions of the original mesh. Such invalid connections can lead to visual artifacts and geometric distortions during rendering and deformation.

To prevent such artifacts, we introduce an adjacency-based validation mechanism grounded in the original FLAME topology. We first construct a face-adjacency graph encoding connectivity between all FLAME faces, which is precomputed and cached for efficiency. For each candidate UV face, we retrieve the FLAME face indices of its three vertices through the UV rasterization mapping. A

UV face is retained only if (1) all vertices belong to the same FLAME face, or (2) the vertices span multiple FLAME faces that are mutually connected within a bounded hop distance in the adjacency graph. This multi-hop connectivity check is performed via breadth-first search (BFS) with a maximum hop threshold of 5. This mechanism effectively filters out topologically invalid or distorted triangles while preserving sufficient mesh density for high-quality Gaussian splatting. The resulting UV mesh maintains the original FLAME topology and provides a reliable, deformation-aware surface for attaching Gaussian attributes.

B. More Ablation Results

B.1. VGM Ablation

Portrait Generation Qualitative results are presented in Fig. 1. Our full model shows clear improvements in fine facial details, such as wrinkles, eyelashes, and teeth, compared to the ablated versions. When joint representation learning is removed, the model's ability to transfer expressions is noticeably weakened. We believe this is due to the absence of 3D consistency provided by the surface normal domain, which increases the entanglement between identity, head pose, and expression. Similarly, removing the domain attention module degrades both expression transfer and the fidelity of facial details, emphasizing the importance of cross-domain information exchange between image and normal features. Ablating the cross-view pairing strategy leads to obvious identity leakage and degraded driving quality, where the generated results inappropriately retain characteristics from the driving sequence rather than faithfully following the reference identity and driving expressions. The results obtained without synthetic data are closer to the complete model, indicating that while the inclusion of synthetic data provides additional gains in generation quality, the contributions of the core modules—joint representation, domain attention, and cross-view pairing—are essential for robust and high-fidelity video generation.

Normal Generation We further evaluate the impact of synthetic data on the quality of generated surface normals. As shown in Fig. 2, removing synthetic data results in a modest but noticeable reduction in high-frequency facial details, such as wrinkles and fine contours. Incorporating synthetic data provides an overall improvement in visual fidelity, particularly in regions that are challenging to reconstruct from pseudo-ground-truth normals alone. We believe this improvement is primarily due to the high accuracy and fine-grained detail of the synthetic normals, which offer stronger geometric supervision compared to the limited-fidelity pseudo-ground-truth normals present in other datasets. This effect is further illustrated in Fig. 3, where synthetic normals help preserve subtle facial geometry that is otherwise lost.

B.2. Head Avatar Ablation

Qualitative results in Fig. 4 show that removing hierarchical refinement leads to noticeably degraded reconstruction quality, particularly in sequences with large head rotations and exaggerated expressions, where errors from monocular FLAME tracking propagate more severely without hierarchical correction. Ablating normal regularization further highlights its critical role in providing geometric guidance for challenging regions such as the mouth and teeth, and in mitigating artifacts under extreme head poses. Fig. 5 further demonstrates the contribution of both modules to free-view rendering quality, where the full model significantly reduces artifacts and produces more faithful novel-view synthesis. As shown in Fig. 6, normals generated by our model capture finer facial details compared to monocular estimated normals, enabling more accurate geometric guidance for Gaussian splatting optimization. Overall, these results confirm that hierarchical refinement, normal regularization, and high-quality generated normals are all essential for producing high-fidelity 3D head avatars.

C. More Results

C.1. Comparisons with More Baselines

We provide additional comparisons with more baselines in Tab. 1 and Fig. 7, including diffusion-based generative models X-NeMo and Wan-Animate, as well as LivePortrait and VoodooXP. Both qualitative and quantitative results show that our method achieves the best performance on most metrics, delivering more accurate expression transfer and remaining robust under large head pose variations.

C.2. Cross-View Animation

The results illustrated in Fig. 8 demonstrate that our pose-free expression encoder exhibits strong cross-view consistency, producing highly consistent facial expressions across a wide range of camera viewpoints. Even when the head pose is held fixed, the encoder effectively disentangles expression from pose, maintaining detailed and coherent facial dynamics regardless of the viewing angle. This highlights the robustness of our encoder in preserving expression fidelity across diverse perspectives.

C.3. Computational Efficiency Discussion

As computational efficiency is a practical concern for diffusion-based models, we evaluate generation quality against inference cost across different sampling steps (Tab. 2). We adopt 25 steps as our default, striking a favorable balance between quality and speed.

Figure 1. Ablation of portrait generation. We ablate joint representation learning, the Domain-Spatial attention module, cross-view pairing, and synthetic data.

Figure 2. Ablation of normal generation. We ablate synthetic data to evaluate its contribution to normal generation quality.

Figure 3. Comparison of ground-truth and pseudo-ground-truth normals. Ground-truth normals from synthetic data are more accurate and capture finer geometric details, thereby improving the quality of generated surface normals.

Figure 4. Ablation on 4D reconstruction. We ablate hierarchical refinement and normal regularization to evaluate their contributions to reconstruction quality.

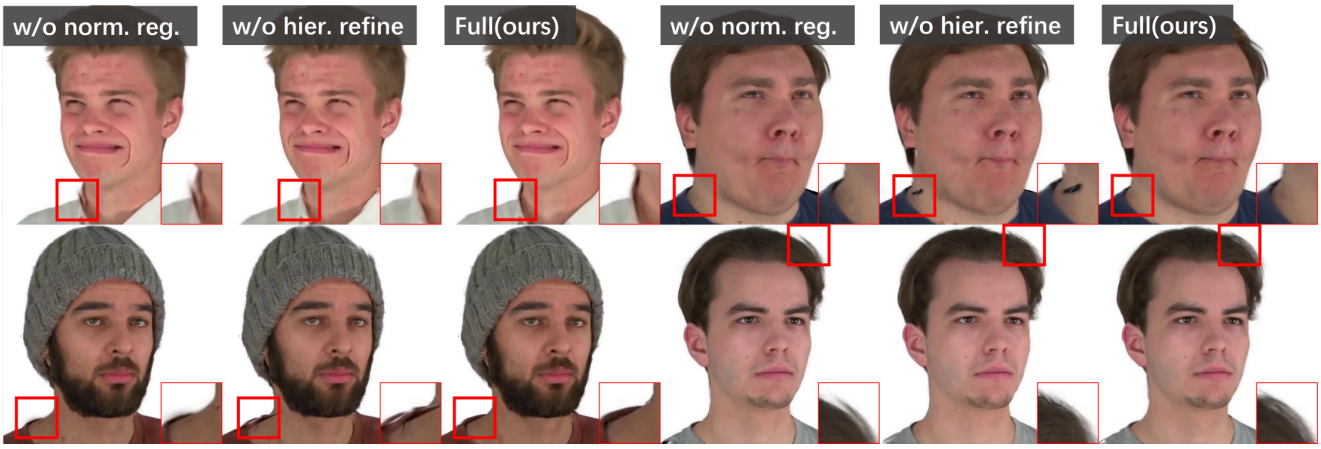

Figure 5. Ablations on hierarchical refinement and normal regularization for free-view rendering.

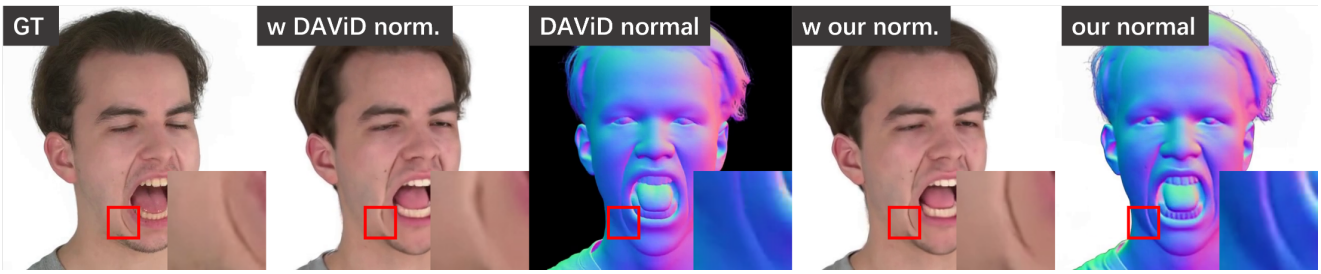

Figure 6. Ablations on our normals versus DAViD normals.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	JOD ↑
VOODOOXP	13.111	0.714	0.276	0.598	5.116
LivePortrait	18.047	0.787	0.214	0.736	6.230
Wan-Animate	12.805	0.655	0.354	0.712	4.597
X-NeMo	14.202	0.683	0.311	0.732	4.987
Our VGM	21.586	0.831	0.174	0.754	7.127
GeoDiff4D	19.953	0.822	0.195	0.737	6.780

Table 1. Quantitative comparison with more baselines (best, second-best).

Figure 7. Qualitative comparison with more baselines.

Steps	PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	JOD ↑	AKD ↓	AED ↓	Speed (s/frame) ↓
10	21.177	0.8286	0.1779	0.7419	7.033	4.215	2.503	1.16
25	21.505	0.8294	0.1746	0.7419	7.093	4.242	2.541	2.74
50	21.529	0.8285	0.1740	0.7421	7.097	4.283	2.573	5.39
100	21.545	0.8280	0.1736	0.7412	7.100	4.261	2.550	10.66

Table 2. Ablation on sampling steps (best, second-best).

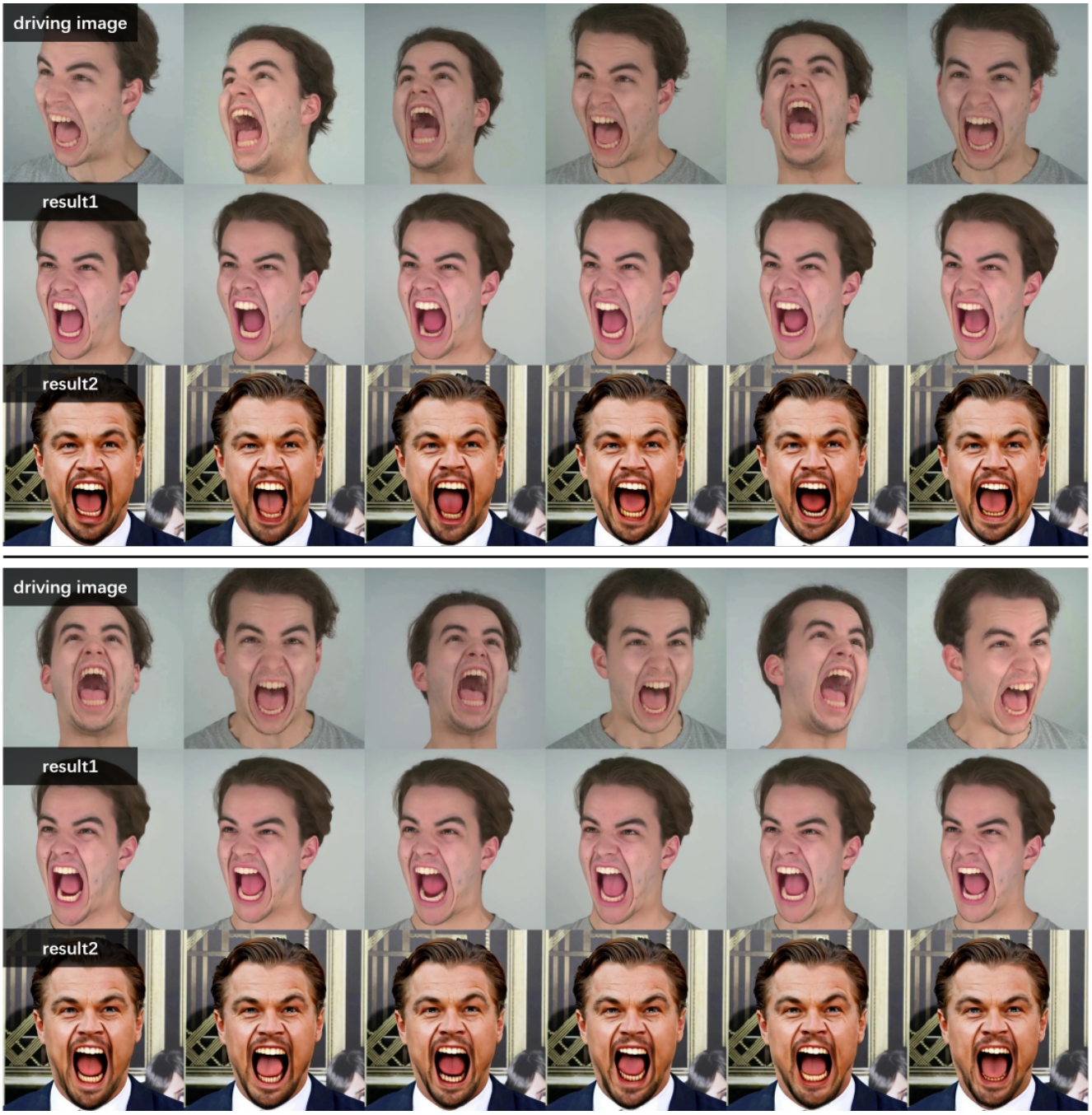

Figure 8. Cross-view animation results. We generate images using expression sequences from 12 different camera viewpoints while fixing the head pose, demonstrating the view consistency of our pose-free expression encoder.