FREE360HEAD: CONTINUOUS-VIEW FULL-HEAD GENERATION VIA 3D-AWARE GUIDANCE

Anonymous authors
Paper under double-blind review

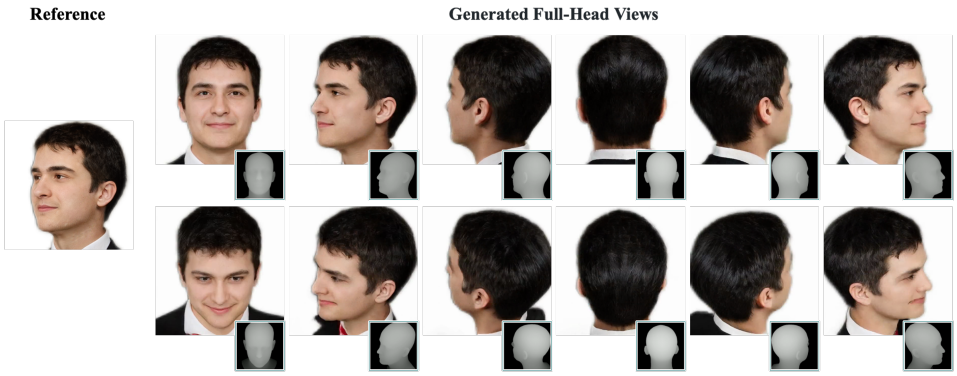

Figure 1: Free360Head generates depth-guided full-head videos from one portrait (left): a zero-pitch video (top) and different nonzero-pitch videos (bottom). Depth maps are shown in the insets.

ABSTRACT

Generating a continuous free-viewpoint video of a complete human head from a single portrait requires plausible completion of unseen regions, identity preservation, precise trajectory control, and cross-view consistency. Existing 3D-based approaches provide consistent rendering but are limited by the quality of single-image reconstruction, whereas generative approaches produce realistic details but may drift in geometry and identity under large viewpoint changes. We present FREE360HEAD, to the best of our knowledge the first reference-independent 3D-aware trajectory control framework, combining the strengths of both routes. Our method explicitly disentangles identity from trajectory: an arbitrary-view reference portrait specifies identity without defining any output viewpoint, while an identity-agnostic depth sequence independently controls the target trajectory. Compared with camera-pose conditioning, the depth sequence provides dense, frame-aligned cues for projected silhouette, visible surface, and depth ordering, enabling fine-grained geometric control. Built on the pretrained Open-Sora video diffusion Transformer, FREE360HEAD uses a reference-prefixed latent injection scheme and an auxiliary reference reconstruction objective to propagate identity throughout depth-guided video generation. Experiments demonstrate improved target-view fidelity, identity preservation, and trajectory control over existing approaches. approaches.

1 INTRODUCTION

Generating continuous free-viewpoint videos of a complete human head from a single portrait supports digital avatars, immersive communication, and portrait-based content creation (An et al., 2023b; Li et al., 2024; 2026; Lyu et al., 2025; Gu et al., 2025). Given a reference portrait and a target camera trajectory, the goal is to synthesize an identity-consistent video that follows the prescribed viewpoints. Unlike narrow-baseline novel-view synthesis, this task spans wide azimuth and elevation changes, revealing unobserved regions such as the ears, hair, and back of the head. Inferring these regions from partial evidence is inherently ambiguous: they must be plausible, consistent with the reference identity, and stable across views and time. This requires jointly addressing generative completion, identity preservation, camera control, and spatiotemporal consistency.

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.

Existing solutions broadly follow two routes. The first reconstructs a shared 3D representation, such as a neural radiance field (Mildenhall et al., 2020), 3D Gaussians (Kerbl et al., 2023b), or a 3D-aware feature field (Chan et al., 2022a; An et al., 2023a), and renders it from target cameras. A shared representation naturally offers explicit viewpoint control and strong cross-view consistency. Its output quality, however, is bounded by the difficult single-image reconstruction step. Geometry and appearance in unobserved regions must be committed to before rendering, and reconstruction or inversion errors persist across all views, often compromising subject-specific details and photo-realism (Lyu et al., 2025). In contrast, the second route uses image or video generative priors to synthesize unobserved appearance. Diffusion models produce high-quality details and have been extended to portrait novel-view and 360° synthesis (Rombach et al., 2022; Gu et al., 2024; 2025). Video diffusion further models dependencies among ordered frames (Ho et al., 2022b; Guo et al., 2024b; Sun et al., 2025a), and spatial or camera conditions provide viewpoint control (Zhang et al., 2023; He et al., 2025; Chen et al., 2026). Nevertheless, high per-frame quality does not guarantee 3D consistency under large full-head rotations: unseen appearance may drift, geometry may change across frames, and identity can degrade. Moreover, camera-pose conditioning specifies *where* the camera moves but leaves the network to infer the corresponding image-space head structure. Such a compact control signal can be insufficient for precise full-head motion, especially when large rotations change the silhouette and visible surface. Conventional image-to-video conditioning also tends to treat the reference as an output frame, coupling identity conditioning and the reference viewpoint with the generated trajectory.

These complementary strengths motivate a different formulation: use 3D information to control the trajectory, but avoid reconstructing a subject-specific 3D representation; use a video foundation model to synthesize appearance, but constrain all frames with a shared, dense geometric sequence. This formulation must explicitly separate *who* is generated from *how* the head is viewed. Identity should come from the reference portrait regardless of its pose, whereas trajectory should be controlled by an independent signal that provides more detailed geometric guidance than a camera-pose sequence alone.

We therefore present FREE360HEAD, a reference-independent 3D-aware trajectory control framework for continuous free-viewpoint full-head generation. Here, reference-independent means that trajectory specification is independent of the reference viewpoint, while generation remains conditioned on the reference identity. FREE360HEAD builds on the pretrained Open-Sora video diffusion Transformer (Zheng et al., 2024b) to exploit its spatial–temporal generative prior. Our key design is to factorize identity and trajectory conditioning. An arbitrary-view reference portrait specifies *who* to generate, but contributes no target-view constraint. Independently, we render an identity-agnostic 3D head proxy along the requested camera path to obtain a sequence of depth maps that specifies *how* the head should be viewed. Unlike camera poses, this 3D-aware sequence supplies dense, frame-aligned cues for projected silhouette, visible surface, and depth ordering, translating the camera path into direct image-space geometric targets for fine-grained trajectory control. To inject identity without tying the reference viewpoint to any frame of the final video, we construct a reference-prefixed latent condition and jointly learn auxiliary reference reconstruction and depth-guided video synthesis under a rectified-flow objective (Liu et al., 2023b). Spatial and temporal attention then propagate identity and geometric cues across the complete viewpoint sequence, supporting detailed generation and consistent trajectory control without subject-specific 3D reconstruction.

Our contributions are summarized as follows.

- To the best of our knowledge, we introduce the reference-independent 3D-aware trajectory control framework for continuous free-viewpoint full-head generation. Instead of conditioning only on camera poses, our reusable depth sequence provides dense, frame-aligned geometric targets for fine-grained trajectory control without requiring subject-specific 3D reconstruction.
- We cast the task as conditional video generation and adapt the pretrained Open-Sora backbone, leveraging its spatial–temporal prior to jointly synthesize detailed and coherent views, including head regions unobserved in the input portrait.
- We explicitly disentangle identity from trajectory: an arbitrary-view reference controls identity without defining any output viewpoint, while the 3D-aware depth sequence independently controls the trajectory. A reference-prefixed latent injection scheme and auxiliary reconstruction objective preserve identity throughout depth-guided video synthesis.

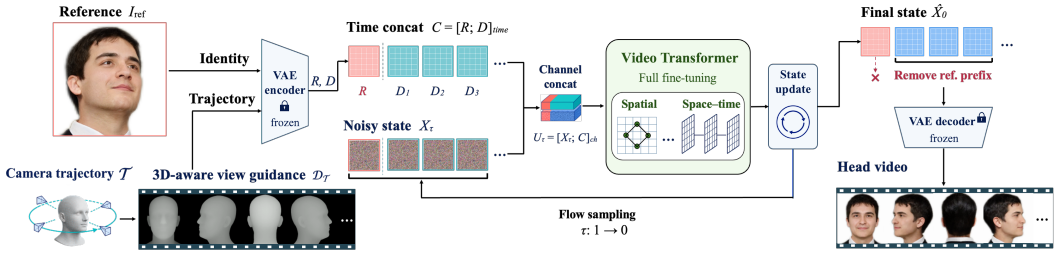

Figure 2: Overview of FREE360HEAD. The reference portrait and the 3D-aware depth sequence independently specify identity and camera trajectory, respectively. A frozen video VAE encodes the reference, depth controls, and target video into latent representations. We construct a clean condition $C = [R; D]_{\text{time}}$ and a reference-prefixed generation target $X = [R; V]_{\text{time}}$. The Open-Sora backbone jointly denoises the complete latent sequence.

2 METHOD

2.1 OVERVIEW

Given a single reference portrait $I_{\text{ref}} \in \mathbb{R}^{3\times H\times W}$ and a camera trajectory $\mathcal{T} = \{\pi_t\}_{t=1}^{T}$, our goal is to synthesize a full-head video $\widehat{\mathbf{Y}} = \{\widehat{Y}_t\}_{t=1}^{T}$ that preserves the identity in I_{ref} while following $\mathcal{T}$. Importantly, I_{ref} may be captured from an arbitrary viewpoint and is *not* required to coincide with any frame on the target trajectory: it only specifies identity, while the viewpoint of every output frame is determined by $\mathcal{T}$. We formulate this task as

$$\widehat{\mathbf{Y}} \sim p_\theta(\mathbf{Y} \mid I_{\text{ref}}, \mathbf{D}_\mathcal{T}), \tag{1}$$

where $\mathbf{D}_\mathcal{T} = \{D_t\}_{t=1}^{T}$ is a 3D-aware depth sequence rendered along $\mathcal{T}$.

Built on the pretrained Open-Sora video generator (Zheng et al., 2024b), FREE360HEAD separates identity conditioning from trajectory control. A reference-prefixed latent provides identity without constraining the output viewpoints, while an identity-agnostic depth sequence specifies the target trajectory. Open-Sora's spatial–temporal prior then generates coherent target views, including head regions unseen in the reference (Figure 2).

We build FREE360HEAD upon the Open-Sora spatial–temporal diffusion Transformer (STDiT) and its frozen video VAE. The VAE maps an RGB video to and from the latent space using its encoder $\mathcal{E}$ and decoder $\mathcal{G}$, respectively. Specifically, it encodes $\mathbf{Y} \in \mathbb{R}^{3\times T\times H\times W}$ as $V = \mathcal{E}(\mathbf{Y}) \in \mathbb{R}^{c\times N\times h\times w}$, where N is the temporally compressed latent length. STDiT operates in this space to model within-frame structure and cross-view correspondence, and $\mathcal{G}$ maps the generated latents back to RGB video.

2.2 REFERENCE-INDEPENDENT 3D-AWARE TRAJECTORY CONTROL

3D-aware control representation. A camera pose alone specifies where to render but does not expose the dense image-space geometry expected at that pose. We therefore convert the requested trajectory $\mathcal{T} = \{\pi_t\}_{t=1}^{T}$ into a sequence of depth maps by rendering a canonical 3D head proxy:

$$D_t = \mathcal{R}_{\text{depth}}(\mathcal{M}, \pi_t), \qquad \mathbf{D}_\mathcal{T} = (D_1, \ldots, D_T), \tag{2}$$

where $\mathcal{M}$ denotes the 3D head proxy and $\mathcal{R}_{\text{depth}}$ is a depth renderer. Each D_t encodes the projected silhouette, visible surface, and relative depth ordering at viewpoint π_t; the ordered sequence additionally describes their continuous temporal evolution. We normalize the depth sequence and encode it using the same frozen video VAE as the RGB data.

Unlike an identity-specific 3D reconstruction, the proxy only defines a geometric control coordinate system. The same depth sequence can be paired with different subjects:

$$\widehat{\mathbf{Y}}^{(i,j)} \sim p_\theta\left(\mathbf{Y} \mid I_{\text{ref}}^{(i)}, \mathbf{D}_{\mathcal{T}^{(j)}}\right). \tag{3}$$

Consequently, changing $I_{\text{ref}}^{(i)}$ while fixing $\mathbf{D}_{\mathcal{T}^{(j)}}$ changes the identity but preserves the rendering trajectory; changing the depth sequence while fixing the reference changes the trajectory without

redefining identity. This factorized control interface is the basis of our reference-independent trajectory control.

Trajectory condition encoding. The depth sequence is spatially and temporally aligned with the target RGB video. During training, target frame Y_t and depth frame D_t correspond to the same camera pose and undergo the same deterministic resize-and-crop transformation. Both sequences are sampled with identical temporal indices. We encode the complete control sequence using the frozen VAE:

$$D = \mathcal{E}(\mathbf{D}_{\mathcal{T}}) \in \mathbb{R}^{c\times N\times h\times w}. \tag{4}$$

Thus, each temporally compressed depth latent remains aligned with the corresponding target-video latent. The dense control preserves spatial correspondence throughout the network, unlike a global camera embedding that must implicitly infer the projected head structure.

2.3 Arbitrary-View Reference Identity Conditioning

Reference as identity context. The central requirement of our setting is that the reference viewpoint must not constrain the generated trajectory. We therefore do not insert I_{ref} as a clean frame of the output video, nor do we require its pose to match π_1 or any other pose in $\mathcal{T}$. Instead, the frozen VAE independently encodes the portrait as

$$R = \mathcal{E}(I_{\text{ref}}) \in \mathbb{R}^{c\times 1\times h\times w}. \tag{5}$$

The single latent R is used as an auxiliary prefix that communicates identity through STDiT's temporal attention. Since the reference is encoded independently of $\mathbf{D}_{\mathcal{T}}$, it may depict any available view of the subject.

Reference-prefixed joint modeling. Let $V = \mathcal{E}(\mathbf{Y})$ be the ground-truth RGB video latent. We construct the clean visual condition and the denoising target as

$$C = [R; D]_{\text{time}}, \qquad X = [R; V]_{\text{time}}, \tag{6}$$

where $[\cdot;\cdot]_{\text{time}}$ denotes concatenation along the latent temporal axis. Both tensors have shape $c \times (N+1) \times h \times w$, but their semantics are deliberately separated. In the condition C, the first position carries identity appearance and the remaining N positions carry trajectory geometry. In the target X, the first position is an auxiliary reference reconstruction target and the remaining positions are the output video. Ground-truth video latents V never appear in C.

At noise level τ, we concatenate the noisy generation state X_τ and the clean visual condition channel-wise:

$$U_\tau = [X_\tau; C]_{\text{channel}} \in \mathbb{R}^{2c\times(N+1)\times h\times w}. \tag{7}$$

The first STDiT projection jointly embeds each noisy latent and its aligned clean condition. We expand its patch embedding from c to $2c$ input channels to accommodate this concatenation, while retaining the subsequent Open-Sora Transformer blocks. Spatial attention incorporates dense depth structure within each target view, whereas temporal attention propagates reference identity to all trajectory positions and exchanges appearance information among generated views. This design keeps the reference available throughout generation without assigning its viewpoint to an output timestamp.

Including R in both C and X defines an *auxiliary reference reconstruction objective*. If R appeared only in C, identity would be supervised indirectly through the target views. Reconstructing its noisy counterpart in X instead provides a direct appearance-preservation signal, while the video positions are jointly optimized to satisfy this signal and the depth sequence. Identity and trajectory therefore enter through separate condition slots but are reconciled by the shared video Transformer.

2.4 Training

We train Free360Head with the rectified-flow objective used by Open-Sora (Liu et al., 2023b). For a clean reference-prefixed target X and Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, we sample $\tau \in [0, 1]$ and form

$$X_\tau = (1-\tau)X + \tau\epsilon, \tag{8}$$

where $\tau = 0$ denotes clean data and $\tau = 1$ denotes noise. The condition C remains clean at every noise level. The network predicts the denoising velocity $X - \epsilon$:

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}_{X,C,\tau,\epsilon}\left[\|v_\theta(X_\tau, \tau, C) - (X - \epsilon)\|_2^2\right]. \tag{9}$$

The noise level is sampled with Open-Sora's logit-normal schedule and its spatial- and temporal-scale-dependent timestep transformation.

Because X consists of one reference latent followed by N video latents, the implemented mean loss can be written as

$$\mathcal{L}_{\mathrm{RF}} = \frac{\mathcal{L}_{\mathrm{ref}} + N\mathcal{L}_{\mathrm{video}}}{N+1}, \tag{10}$$

where $\mathcal{L}_{\mathrm{ref}}$ and $\mathcal{L}_{\mathrm{video}}$ are the mean velocity errors over the reference prefix and target-video positions, respectively. This is the ordinary element-wise mean over the complete sequence: the reference term directly supervises identity reconstruction, while the video term learns identity-preserving synthesis under the requested trajectory.

3 Experiments

3.1 Experimental Setup

Evaluation Datasets. We use 400 synthetic head instances generated with PanoHead (An et al., 2023a) and 182 instances derived from RenderMe-360 (Pan et al., 2023). For the latter, we estimate and refine camera parameters from 60 images per instance, reconstruct heads using 3D Gaussian Splatting (Kerbl et al., 2023b), and render supervision videos under the target camera convention. The depth sequence for each video is obtained by volume rendering in PanoHead (An et al., 2023a) using the corresponding camera parameters. Each instance provides ten reference images randomly sampled within yaw and pitch ranges of $[-45°, +45°]$ and $[-40°, +40°]$, respectively. Target videos follow nine full-head trajectories spanning pitch angles from $-40°$ to $+40°$ in $10°$ increments, with depth-control sequences shared across instances. Each video contains 49 frames at 512×512 resolution and 24 FPS. We split data by head instance, keeping its references and trajectories together. The training set contains 320 PanoHead and 146 RenderMe-360-derived instances, while the test set contains 80 and 36, respectively. Pairing all references and trajectories yields 41,940 training pairs and 10,440 test pairs. Quantitative comparisons use 25 PanoHead and 5 RenderMe-360-derived instances randomly sampled from the test sets. All methods use references from the same set and matching target viewpoints, with method-specific camera conversion and preprocessing as needed.

Baselines. We compare our method with six baselines from three categories: (i) 3D-aware GANs, including PanoHead (An et al., 2023a), SphereHead (Li et al., 2024), and HyPlaneHead (Li et al., 2026). To adapt these pretrained generators to each reference image, we apply Pivotal Tuning Inversion (PTI) (Roich et al., 2022) and render the resulting 3D-aware representations along the target camera trajectory; (ii) the diffusion-based multi-view generation method DiffPortrait360 (Gu et al., 2025); and (iii) single-image 3D reconstruction methods, Unique3D (Wu et al., 2024) and FaceLift (Lyu et al., 2025). For methods that recover explicit 3D representations, we render their outputs along the same target camera trajectories for evaluation. We use publicly available implementations and pretrained weights with method-specific camera conversion and preprocessing. RGB outputs are evaluated at corresponding target viewpoints under a unified resolution and cropping protocol. PercHead (Oroz et al., 2026) is a closely related concurrent CVPR 2026 method for single-image 3D head reconstruction. We do not include it in the quantitative comparison because its repository did not provide executable training or inference code, or pretrained weights, at the time of our evaluation.

3.2 Comparisons with Existing Methods

We compare Free360Head with existing methods on the PanoHead synthetic and RenderMe-360-derived test sets, focusing on target-view fidelity, identity preservation, and image quality. The evaluation covers frontal views, full-head rotation at zero pitch, and trajectories with varying pitch.



Figure 3: Qualitative comparisons with existing methods on samples from the PanoHead test set. All target trajectories are rendered at zero pitch. Our method produces higher-quality, more identity-faithful frontal and rear full-head views.

Table 1: **Quantitative comparison on the PanoHead synthetic test set.** Higher PSNR, SSIM, and CSIM indicate better performance, while lower LPIPS and FID are preferred. Best results are shown in **bold**, and second-best results are underlined.

Method	Input	Frontal View (0°)				Average over 8 Views (Pitch = 0°)			
		PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
SphereHead	10 views	18.8167	0.540234	0.238663	0.362129	16.1010	0.420772	0.343405	37.7552
Unique3D	10 views	12.2819	0.413880	0.433095	0.357781	10.9735	0.352622	0.507756	115.1995
DiffPortrait360	10 views	18.1241	0.474150	0.229620	0.562269	15.9982	0.357061	0.329623	49.5373
FaceLift	10 views	12.2686	0.376660	0.472877	0.318592	9.9594	0.326830	0.542061	96.0240
HyPlaneHead	10 views	19.2902	0.546255	0.223499	0.395723	17.3090	0.442712	0.309282	34.8966
Ours	10 views	**22.0160**	**0.638075**	**0.204164**	**0.670004**	**20.1936**	**0.530491**	**0.281705**	**30.9758**

Quantitative comparisons. Tables 1 and 2 show that FREE360HEAD achieves the strongest overall multiview performance among the compared methods on both test sets. The improvements span image reconstruction, structural similarity, perceptual similarity, and distributional image quality, supporting more faithful synthesis across the evaluated head views. On PanoHead, this advantage also extends to the frontal-view evaluation, including identity similarity to the reference.

On the RenderMe-360-derived test set, our method achieves the best frontal structural and perceptual similarity, although the advantage is less uniform across frontal-view metrics. DiffPortrait360 retains higher frontal PSNR, while DiffPortrait360 and Unique3D achieve higher frontal CSIM. Our strongest advantage on this test set therefore lies in overall multiview synthesis rather than uniformly superior frontal identity preservation.

Table 3 extends the comparison to results aggregated over trajectories with varying pitch. FREE360HEAD maintains the leading reconstruction, perceptual, and distributional quality among the reported methods on both test sets. These results extend the evidence for improved full-head synthesis beyond the zero-pitch evaluation.

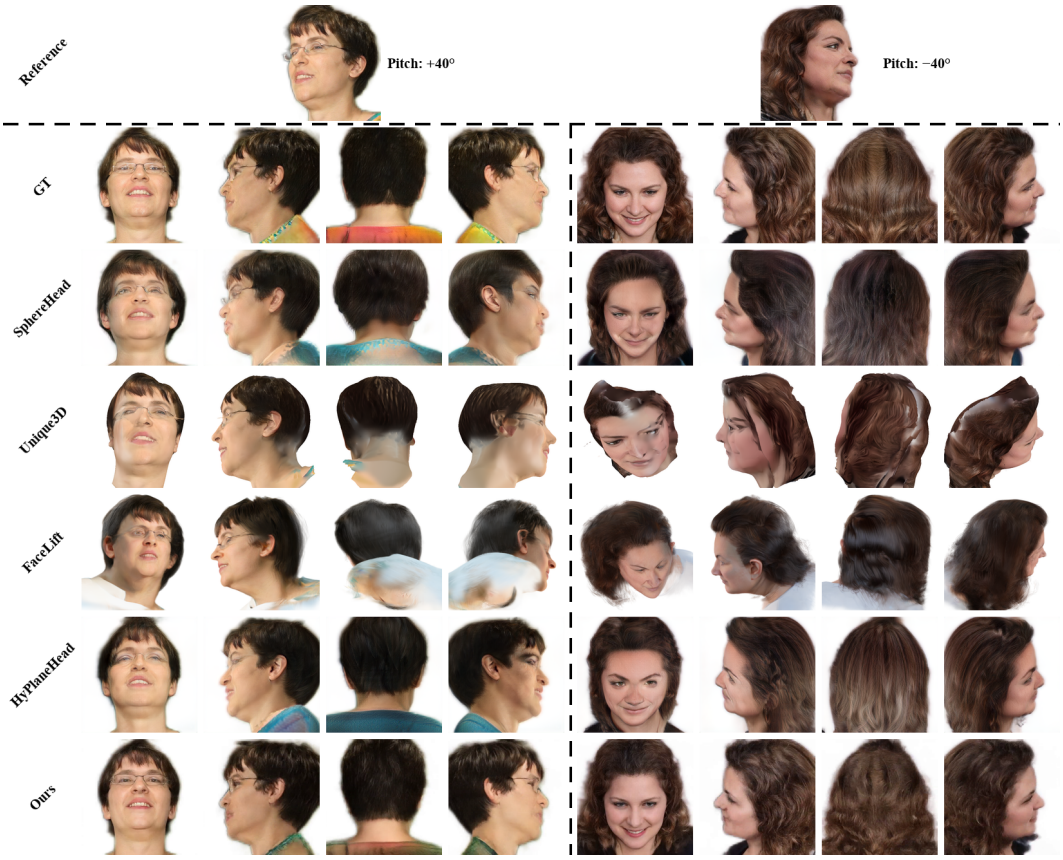

Figure 4: Qualitative comparisons with existing methods on the PanoHead test set under target camera trajectories at pitch angles of $\pm 40°$. Our method better preserves identity and appearance across frontal, side, and rear views.

Table 2: **Quantitative comparison on the RenderMe-360-derived test set.** Higher PSNR, SSIM, and CSIM indicate better performance, while lower LPIPS and FID are preferred. Best results are shown in **bold**, and second-best results are underlined.

Method	Input	Frontal View ($0°$)				Average over 8 Views (Pitch $= 0°$)			
		PSNR ↑	SSIM ↑	LPIPS ↓	CSIM ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
PanoHead	10 views	14.9893	0.551023	0.320975	0.251904	13.5042	0.514948	0.367256	104.7034
SphereHead	10 views	14.3864	0.526899	0.358972	0.220655	13.1728	0.510721	0.381476	110.6829
Unique3D	10 views	11.8088	0.486468	0.390034	0.532731	9.9581	0.457726	0.457697	177.3816
DiffPortrait360	10 views	**15.3150**	0.497566	0.355898	**0.543424**	13.7865	0.512460	0.367717	106.6136
FaceLift	10 views	11.5762	0.444478	0.425071	0.388536	9.0840	0.408068	0.502084	148.3266
HyPlaneHead	10 views	14.4406	0.522851	0.343222	0.239916	13.2195	0.494215	0.389046	107.8217
Ours	10 views	15.2793	**0.565336**	**0.286236**	0.481531	**14.5710**	**0.583218**	**0.328441**	**98.9775**

Table 3: **Quantitative comparison over all trajectories on the PanoHead synthetic and RenderMe-360-derived test sets.** Higher PSNR and SSIM indicate better performance, while lower LPIPS and FID are preferred. Best and second-best results are shown in **bold** and underlined, respectively.

Method	Input	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
SphereHead	10 views	16.0071	0.415696	0.348149	24.3754
Unique3D	10 views	10.8418	0.347784	0.513407	99.5295
FaceLift	10 views	9.8395	0.323540	0.545676	79.1152
HyPlaneHead	10 views	17.2236	0.438905	0.313566	22.4741
Ours	10 views	**20.1099**	**0.524669**	**0.284672**	**21.6434**

(a) PanoHead synthetic test set

Method	Input	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
PanoHead	10 views	13.5516	0.515558	0.363296	83.2223
SphereHead	10 views	13.2764	0.511585	0.377124	86.8884
Unique3D	10 views	9.9386	0.456961	0.455985	152.8734
FaceLift	10 views	9.0211	0.408971	0.498672	122.6948
HyPlaneHead	10 views	13.3150	0.496944	0.383835	86.8678
Ours	10 views	**14.4673**	**0.582186**	**0.328257**	**77.8106**

(b) RenderMe-360-derived test set

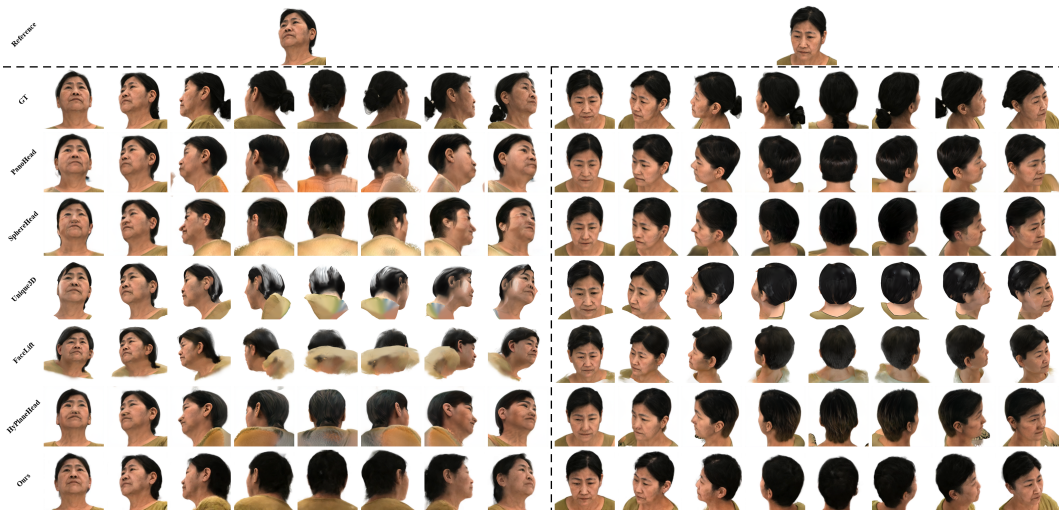

Figure 5: Qualitative comparisons with existing methods on the RenderMe-360-derived test set under target camera trajectories at pitch angles of $\pm 40°$. Our method better preserves identity and appearance across frontal, side, and rear views.



Figure 6: **Qualitative trajectory comparison.** Results on (a) PanoHead synthetic and (b) RenderMe-derived heads at pitch-offset settings $+20$ and $+40$. Each head combines a generated left half with its GT right half.

Qualitative comparisons. Figure 3 compares frontal and rear views at zero pitch on the PanoHead test set. SphereHead and HyPlaneHead generate plausible rear views, but their frontal views exhibit discrepancies in facial appearance relative to the ground truth. DiffPortrait360 synthesizes plausible frontal and rear views, although the generated hairstyles do not consistently match the ground truth. Unique3D exhibits discontinuities around the neck and collar, while FaceLift shows blurred hair and deviations in head orientation in the displayed examples. Our method more closely matches the target facial details, hair color, and overall hairstyle across these views.

Figure 4 shows results at pitch angles of $\pm 40°$ on the PanoHead test set. Unique3D exhibits distorted or incomplete regions around the face and neck, while FaceLift produces conspicuous head and shoulder artifacts. SphereHead and HyPlaneHead generate more complete head outlines, but do not faithfully preserve the subjects' facial appearance. In these examples, our method more closely matches the target facial appearance and hairstyle, with fewer conspicuous head and neck artifacts across the displayed viewpoints.

Figure 5 presents qualitative comparisons on the RenderMe-360-derived test set at pitch angles of $\pm 40°$. Unique3D produces patchy hair textures and irregular head and neck boundaries, while FaceLift distorts the lower head and shoulder regions, particularly in the rear views of the left group. Our method produces more complete head outlines with fewer conspicuous surface and neck artifacts in these examples. However, the bun visible in the ground truth is not fully recovered in our rear and oblique views, revealing remaining errors in the completion of occluded hair.

3.3 Ablation Study

3D-Aware Guidance for Trajectory Control. We compare Free360Head with Camera-only using identical reference images and target trajectories. Camera-only uses numerical camera parameters, whereas our method provides dense, frame-aligned depth guidance. Table 4(a) shows closer agreement with target head orientations and adjacent-frame rotation changes on both test sets.

Table 4: **Comparisons of trajectory guidance and reference conditioning.** Results on 25 PanoHead and 5 RenderMe-derived test identities. (a) Trajectory agreement over nine 49-frame trajectories per identity using one reference; errors are in degrees. (b) Reconstruction quality averaged over eight target views at zero pitch, with ten references evaluated independently per identity; PSNR is in dB. Arrows indicate the preferred direction. Better reported values are shown in **bold**.

(a) Trajectory control

	PanoHead		RenderMe-derived	
Method	PGE ↓	RRE ↓	PGE ↓	RRE ↓
Camera-only	7.66	4.61	13.10	7.25
Ours	**4.54**	**3.90**	**8.64**	**5.16**

(b) Reference conditioning

	PanoHead			RenderMe-derived		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Spatial concat	14.5274	0.3565	0.4446	12.6785	0.4769	0.4291
Ours	**20.1936**	**0.5305**	**0.2817**	**14.5710**	**0.5832**	**0.3284**

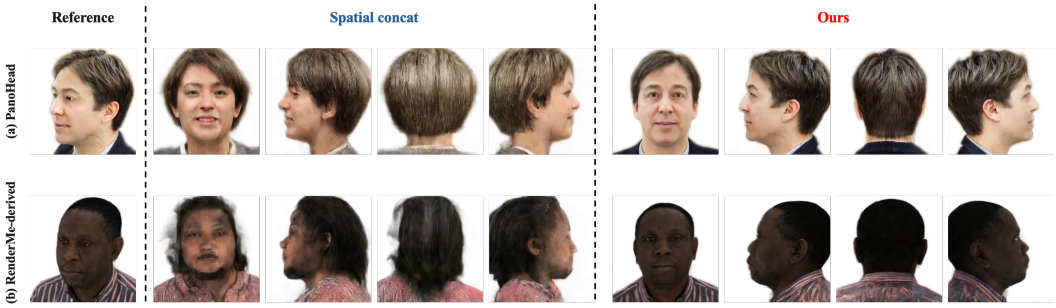

Figure 7: **Qualitative comparison of reference conditioning layouts.** Results on (a) PanoHead synthetic head and (b) RenderMe-derived head. Each row shows the reference image, four views generated by Spatial concat, and the corresponding four views generated by FREE360HEAD. Both methods receive the same reference image and depth-control sequence.

Figure 6 visualizes alignment by combining generated left halves with ground-truth right halves. Camera-only exhibits visible orientation and facial misalignment, whereas our method more closely matches the target head orientation and facial structure. These results support the effectiveness of depth guidance for trajectory following.

Reference Conditioning Layout. We compare our reference-prefixed layout with Spatial concat using the same reference images and depth-control sequences. Spatial concat concatenates temporally repeated reference latents with zero-prefixed depth latents along the width dimension, then resizes the result to the original latent resolution. Our layout instead places the reference in a dedicated temporal slot. Table 4(b) shows improved reconstruction and perceptual quality on both test sets. As illustrated in Figure 7, Spatial concat alters facial appearance and hairstyle, whereas our method better preserves reference-visible facial features, hair, and clothing. These results favor reference-prefixed conditioning over the spatial concatenation and resizing scheme evaluated here.

4 CONCLUSIONS AND DISCUSSION

We presented FREE360HEAD, a framework for continuous-view full-head generation from a single portrait. Built on a pretrained video diffusion Transformer, it combines reusable depth guidance with reference-prefixed latent conditioning and auxiliary reference reconstruction to specify target viewpoints independently of the reference pose, without subject-specific 3D reconstruction. Experiments on PanoHead and RenderMe-360-derived test subsets demonstrate improved multiview fidelity and distributional image quality over the evaluated baselines. The RenderMe-360-derived training data contain floating artifacts introduced by 3D Gaussian reconstruction and rendering, which may affect generation quality. Future work will focus on improving the quality of these data to mitigate such artifacts.

AI USE STATEMENT

We used generative AI tools only for language polishing, grammar checking, translation assistance, LaTeX formatting, readability improvement, and consistency checks of terminology, equations, tables, figures, and references. We did not use generative AI tools to generate research ideas, design experiments, implement code, process data, produce proofs, conduct experiments, analyze results, or generate reported numbers. The authors reviewed all AI-assisted edits against the manuscript, source files, and experimental outputs, and take full responsibility for the accuracy, originality, and compliance of the final content.

REFERENCES

Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. PanoHead: Geometry-aware 3D full-head synthesis in 360°. In *CVPR*, pp. 20950–20959, 2023a.

Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360°. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20950–20959. IEEE, 2023b.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22875–22889. IEEE, 2025.

Volker Blanz and Thomas Vetter. A morphable model for thesynthesis of 3d faces. In *SIGGRAPH'99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 187–194, 1999.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5543–5552, 2016.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, pp. 16123–16133, 2022a.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022b.

Dong-Yu Chen, Yixin Guo, Shuojin Yang, Tai-Jiang Mu, and Shi-Min Hu. Beyond inpainting: Unleash 3D understanding for precise camera-controlled video generation. *arXiv preprint arXiv:2601.10214*, 2026.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. *arXiv preprint arXiv:2305.13840*, 2023.

Xi Chen, Zhiheng Liu, Mengting Chen, Yutong Feng, Yu Liu, Yujun Shen, and Hengshuang Zhao. Livephoto: Real image animation with text-guided motion control. In *European Conference on Computer Vision*, pp. 475–491. Springer, 2024.

Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20311–20322, 2022.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

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

Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. DiffPortrait3D: Controllable diffusion for zero-shot portrait view synthesis. In *CVPR*, pp. 10456–10465, 2024.

Yuming Gu, Phong Tran, Yujian Zheng, Hongyi Xu, Heyuan Li, Adilbek Karmanov, and Hao Li. DiffPortrait360: Consistent portrait diffusion for 360 view synthesis. In *CVPR*, pp. 26263–26273, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *ICLR*, 2024a.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024b.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. Toward robust and unconstrained full range of rotation head pose estimation. *IEEE Transactions on Image Processing*, 33: 2377–2387, 2024. doi: 10.1109/TIP.2024.3378180.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *NeurIPS*, volume 35, pp. 8633–8646, 2022b.

Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. *arXiv preprint arXiv:2310.01406*, 2023.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025.

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

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *TOG*, 42(4), 2023b.

Tobias Kirschstein, Simon Giebenhain, Jiapeng Tang, Markos Georgopoulos, and Matthias Nießner. Gghead: Fast and generalizable 3d gaussian heads. In *SIGGRAPH Asia 2024 conference papers*, pp. 1–11, 2024.

Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *NeuIPS*, 2023.

Heyuan Li, Ce Chen, Tianhao Shi, Yuda Qiu, Sizhe An, Guanying Chen, and Xiaoguang Han. SphereHead: Stable 3D full-head synthesis with spherical tri-plane representation. In *European Conference on Computer Vision*, pp. 324–341. Springer, 2024.

Heyuan Li, Kenkun Liu, Lingteng Qiu, Qi Zuo, Keru Zheng, Zilong Dong, and Xiaoguang Han. Hyplanehead: Rethinking tri-plane-like representations in full-head image synthesis. *Advances in Neural Information Processing Systems*, 38:95995–96024, 2026.

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, 2017a.

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

Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *2020 IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, pp. 5890–5899. IEEE, 2020.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14951–14961. IEEE, 2025.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9264–9275. IEEE, 2023a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023b.

Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. FaceLift: Learning generalizable single image 3D face reconstruction from synthetic heads. In *ICCV*, pp. 12691–12701, 2025.

Weijie Lyu, Ming-Hsuan Yang, and Zhixin Shu. Facecam: Portrait video camera control via scale-aware conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30607–30617, 2026.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision – ECCV 2020*, pp. 405–421. Springer, 2020. doi: 10.1007/978-3-030-58452-8_24.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. *CVPR2022*, 2022.

Antonio Oroz, Matthias Nießner, and Tobias Kirschstein. Perchead: Perceptual head model for single-image 3d head reconstruction & editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4097–4108, 2026.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7993–8005, 2023.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR 2023*, 2022.

Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 fourth international conference on 3D vision (3DV)*, pp. 460–469. IEEE, 2016.

Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020.

Liao Shen, Wentao Jiang, Yiran Zhu, Jiahe Li, Tiezheng Ge, Zhiguo Cao, and Bo Zheng. Identity-preserving image-to-video generation via reward-guided optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27281–27290, 2026.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 573–580. IEEE, 2012. doi: 10.1109/IROS.2012.6385773.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhu, Jun Zhang, and Yikai Wang. DimensionX: Create any 3D and 4D scenes from a single image with decoupled video diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13695–13706, 2025a.

Xiaokun Sun, Zeyu Cai, Ying Tai, Jian Yang, and Zhenyu Zhang. Strandhead: Text to hair-disentangled 3d head avatars using human-centric priors. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13393–13404. IEEE, 2025b.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2024.

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.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3D: High-quality and efficient 3D mesh generation from a single image. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125116–125141, 2024.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38027–38037, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

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.

Yabo Zhang, Yuxiang Wei, XIAOPENG ZHANG, Wangmeng Zuo, Qi Tian, et al. Controlvideo: Training-free controllable text-to-video generation. In *International Conference on Learning Representations*, volume 2024, pp. 54441–54461, 2024.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024a.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024b.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5745–5753, June 2019.

Zijian Zhou, Shikun Liu, Haozhe Liu, Haonan Qiu, Zhaochong An, Weiming Ren, Zhiheng Liu, Xiaoke Huang, Kam-Woh Ng, Tian Xie, et al. Scaling zero-shot reference-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9253–9262, 2026.

A Appendix

A.1 Related Work

3D Face Generation and Reconstruction. Single-image 3D face reconstruction is inherently ill-posed, as geometry and appearance must be inferred from limited visual evidence. Classical methods build on 3D Morphable Models (3DMMs) Blanz & Vetter (1999) and articulated variants such as FLAME Li et al. (2017a), which represent facial shape and appearance in low-dimensional spaces learned from scans. Subsequent 3DMM-based methods Booth et al. (2016); Li et al. (2017b); Lin et al. (2020); Richardson et al. (2016); Feng et al. (2021); Daněček et al. (2022) regress these parameters from a single image to recover identity, expression, pose, and appearance. Although controllable, their restricted priors often yield overly smooth geometry and limited high-frequency detail. Beyond parametric models, 3D-aware GANs Schwarz et al. (2020); Chan et al. (2022b); Or-El et al. (2022) learn data-driven 3D priors from in-the-wild image collections through neural radiance fields Mildenhall et al. (2020) and hybrid explicit–implicit representations; recent work further adopts 3D Gaussian primitives for efficient explicit rendering Kirschstein et al. (2024); Kerbl et al. (2023a). For faces, these models have progressed from predominantly frontal portrait synthesis Chan et al. (2022b) to view-consistent 360° full-head generation An et al. (2023b); Li et al. (2026; 2024). However, reconstructing real faces with GAN priors typically requires per-instance latent inversion, which can compromise reconstruction fidelity. Pretrained diffusion models Rombach et al. (2022) offer a complementary generative prior for 3D face reconstruction and generation. Existing methods either distill frozen 2D diffusion priors into optimizable 3D representations via score distillation Poole et al. (2022); Kolotouros et al. (2023); Huang et al. (2023); Sun et al. (2025b), or perform amortized novel-view synthesis with image-conditioned multi-view diffusion models Liu et al. (2023a); Voleti et al. (2024); Gu et al. (2024; 2025). The latter can be coupled with 3D Gaussian fitting or feed-forward prediction of explicit 3D representations Lyu et al. (2025); Oroz et al. (2026). Nevertheless, image-space novel-view synthesis does not guarantee cross-view consistency, while subsequent 3D reconstruction may lose input-specific identity and fine-grained appearance.

Controllable Video Diffusion Models. Video diffusion models extend denoising diffusion to video sequences Ho et al. (2022a), with latent-space formulations enabling scalable high-resolution synthesis Blattmann et al. (2023); Zheng et al. (2024a); Guo et al. (2024a). Beyond text conditioning, reference images guide subject appearance and identity Zhou et al. (2026); Chen et al. (2024); Shen et al. (2026); Liu et al. (2025), while structural cues such as depth and pose provide finer spatial control Jiang et al. (2025); Chen et al. (2023); Zhang et al. (2024). For viewpoint control, camera trajectories are encoded as per-frame poses within temporal modules Wang et al. (2024) or pixel-aligned Plücker ray maps He et al. (2025). Further advances selectively inject camera guidance into early denoising stages and specific Transformer layers Bahmani et al. (2025), or unify camera pose, intrinsics, and lens distortion through relative ray encoding Zhang et al. (2026). For portraits, FaceCam Lyu et al. (2026) uses scale-aware facial landmark maps rendered from a generic head proxy to re-render monocular videos along target trajectories. Despite these advances, preserving identity across continuous 360° full-head views from a single portrait remains insufficiently explored, as existing approaches primarily target generic scenes or rely on multi-frame portrait inputs.

A.2 DETAILS

Training. We initialize our model from pretrained Open-Sora 1.3 and fine-tune the diffusion backbone while keeping the video VAE frozen. Training uses 49-frame clips at 512×512 resolution and 24 FPS, taking approximately one week on four NVIDIA A800 GPUs with 80 GB memory each. We train for 10 epochs using AdamW with a global batch size of 72, a learning rate of 10^{-4}, and zero weight decay. The learning rate increases linearly over the first 100 updates and remains constant thereafter. We optimize the rectified-flow velocity-matching objective with logit-normal timestep sampling and the original Open-Sora timestep transformation.

Latent Encoding. The reference portrait, target RGB video, and depth sequence are encoded separately using the same pretrained VAE, which produces 16-channel latents with $8\times$ spatial downsampling. Each 49-frame video is encoded in 17-frame chunks with one-frame overlap, yielding 15 latent frames at 64×64 resolution. The reference portrait contributes one additional latent frame, forming a 16-frame sequence. We concatenate the clean conditioning latents with the noisy target latents along the channel dimension and expand the input patch embedding from 16 to 32 channels, initializing both channel groups with the pretrained embedding weights.

Inference. Given an arbitrarily posed reference portrait and a user-specified depth trajectory, we construct the clean condition $C = [R; D]_{\text{time}}$. All $N + 1$ target latent positions, including the auxiliary reference position, are initialized from independent Gaussian noise. We integrate the rectified-flow velocity field using $K = 50$ Euler steps without classifier-free guidance or semantic text prompts. Under the descending schedule $1 = \tau_0 > \tau_1 > \cdots > \tau_K = 0$, each update is

$$\widehat{X}_{\tau_{k+1}} = \widehat{X}_{\tau_k} + (\tau_k - \tau_{k+1}) v_\theta(\widehat{X}_{\tau_k}, \tau_k, C). \tag{11}$$

We then discard the auxiliary reference prefix and decode only the trajectory latents:

$$\widehat{V} = \Pi_{\text{video}}(\widehat{X}_0), \qquad \widehat{\mathbf{Y}} = \mathcal{G}(\widehat{V}), \tag{12}$$

where Π_{video} removes the first latent position and $\mathcal{G}$ denotes the VAE decoder. The reference thus provides identity conditioning throughout denoising without being imposed as a fixed frame in the output video, while the depth trajectory specifies the requested viewpoints.

Evaluation Metrics. We measure target-view fidelity using PSNR (↑), SSIM (↑) (Wang et al., 2004), and LPIPS (↓) (Zhang et al., 2018) at matched cameras and resolutions. For paired image metrics, we apply the ground-truth foreground mask to both the generated image and the corresponding target image before evaluation. We report the frontal view ($0°$), the average over eight yaw views at pitch $0°$, and the average over all 72 evaluation views covering nine pitch angles and eight yaw angles. The image quality is measured by Fréchet Inception Distance (FID, ↓) (Heusel et al., 2017) using generated and target frame sets with matched viewpoint distributions and equal sample counts for each FID computation (2000 samples). Identity preservation is assessed by cosine similarity (CSIM, ↑) between normalized ArcFace embeddings (Deng et al., 2019) extracted from the reference image and generated face crops. Since back-head views lack detectable facial regions, CSIM cannot be computed for these views and is therefore reported only for the frontal view. For trajectory control, we estimate head rotations from generated and ground-truth frames using frozen 6DRepNet360 (Hempel et al., 2024). Pose geodesic error (PGE, ↓) measures the framewise $\mathrm{SO}(3)$ distance between the estimated rotations (Zhou et al., 2019). Relative rotation error (RRE, ↓) measures the geodesic discrepancy between adjacent-frame rotation increments, adapting rotational relative pose evaluation (Sturm et al., 2012). Both metrics are evaluated over all frames of the nine 49-frame trajectories per identity, using one reference image. Errors are reported in degrees and averaged within each identity before averaging across identities.

A.3 MORE RESULTS

Please refer to our demo video for additional visual comparisons and video results.