VISA: 3D-AWARE VIDEO SHADING FOR REAL-TIME UPPER-BODY AVATAR CREATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating high-fidelity upper-body 3D avatars from a single image remains challenging. Large reconstruction models provide fast inference and stable body structures but suffer from blurry textures and stiff motion. Generative video models produce photorealistic dynamics but exhibit structural errors and identity drift. We propose ViSA, which combines both paradigms: a 3D reconstruction model supplies structural and appearance priors to a real-time autoregressive video diffusion renderer. This combination synthesizes photorealistic, high-frequency details and fluid, temporally coherent motion while reducing texture blur, motion stiffness, and structural inconsistencies. Experiments demonstrate substantially improved visual quality and fewer artifacts than leading methods, supporting robust and efficient real-time avatar generation for applications such as gaming and virtual reality.



Figure 1: **Photorealistic, consistent, and controllable character animation from a single reference image**. Our method enables photorealistic upper-body avatar generation that preserves appearance fidelity across diverse poses and expressions while maintaining temporal coherence in real-time video synthesis. Input images are synthesized using Gemini (Team et al., 2023).

1 INTRODUCTION

Significant progress has been made in generating expressive digital avatars in real time from a single image. 3D reconstruction methods, leveraging representations like Neural Radiance Fields or 3D Gaussian Splatting (Hong et al., 2023; Qiu et al., 2025a; Tang et al., 2024; He et al., 2025; Zhang et al., 2025a), excel at providing a stable body structure. By binding the reconstruction to parametric human templates (Pavlakos et al., 2019), they can produce dynamic results. Yet, these methods

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.

struggle to achieve true photorealism. The inherent ill-posed nature of single-view reconstruction often leads to artifacts such as blurry textures and incomplete geometry that persist in the rendering pipeline, degrading the final output and making the avatars appear unrealistic.

On the other hand, video diffusion models (Wan et al., 2025; Kong et al., 2024; Gao et al., 2025) have demonstrated promising capabilities in synthesizing photorealistic videos with rich details and natural motion. Existing methods adapt pretrained video diffusion models to generate video avatars with different conditioning signals, such as keypoints (Hu, 2024), DensePose (Güler et al., 2018; Xu et al., 2024), and rendered SMPL meshes (Zhu et al., 2024). While these signals improve structural accuracy, existing methods still lose fine-grained details such as texture and clothing, making identity preservation difficult. Moreover, the slow inference speed of these methods also makes them impractical for generating long avatar sequences for interactive applications.

To address these problems, we explore a way of generating realistic avatars by combining an explicit 3D reconstruction process with an implicit 3D-aware video rendering process, capitalizing on the complementary strengths of both methods. The core motivation is that explicit 3D reconstruction largely preserves the identity, shape, and texture of the reference avatar, providing a high-fidelity prior for the subsequent generative stage, which alleviates generation uncertainty and enables lightweight, few-step autoregressive models for real-time controlled video synthesis. Specifically, our method first extracts multi-level 2D features from the input avatar image. These features are then lifted into a 3D feature space via a lightweight reconstruction network, animated and projected back onto the image plane to guide the rendering process. To achieve real-time and long-sequence rendering, we propose to build the implicit rendering process with a pre-distilled, few-step causal video diffusion model (Huang et al., 2025). Specifically, we repurpose it as a powerful neural shader by finetuning it with 3D-aware features rendered from our reconstruction model. This is non-trivial, as naively finetuning the distilled model with a standard denoising objective leads to blurry outputs, which is caused by the high variance of the optimization target (Luo et al., 2025a;b). To prevent the model from deviating from its learned distribution, we propose an adversarial distribution preservation loss that enhances generation fidelity, leading to significantly sharper outputs with realistic high-frequency details.

To train our model, we curate a dataset from SpeakerVid-5M (Zhang et al., 2025b), a large-scale dataset of human speaking videos. We process this data by annotating it with SMPL-X and FLAME parameters to support our 3D-aware framework. Extensive experiments demonstrate that our method significantly outperforms existing approaches, generating photorealistic and temporally coherent animation results from a single image in real time. Our contributions can be summarized as follows: i) We propose a novel framework that integrates one-shot 3D avatar reconstruction with a real-time autoregressive video shader, unifying geometric stability with generative rendering power. ii) We introduce a novel adversarial distribution preservation loss, which enables the efficient finetuning of a pre-distilled, few-step video diffusion model to act as a powerful neural renderer while mitigating the typical quality degradation. iii) Our method establishes a new state-of-the-art in real-time upper-body avatar generation, achieving superior visual quality and motion realism, validated by comprehensive qualitative and quantitative comparisons.

2 Related Work

2.1 One-shot 3D Human Reconstruction

Reconstructing animatable avatars from one-shot input remains a difficult task. Early approaches typically relied on parametric body models (Alldieck et al., 2018; Jiang et al., 2025a). Recent feed-forward methods (Qiu et al., 2025a; Zhuang et al., 2025; Qiu et al., 2025b; He et al., 2025; Qiu et al., 2025c) use scalable transformer architectures to directly regress human avatars in a predefined canonical space, typically conditioned on the parametric head model FLAME (Li et al., 2017) or the full-body model SMPL-X (Pavlakos et al., 2019). GUAVA (Zhang et al., 2025a) leverages inverse texture mapping and projection sampling to generate animatable upper-body avatars with an expressive human parametric model. Although these approaches have advanced human reconstruction, the scarcity of large-scale, high-quality 3D supervision or precisely aligned training annotations often leads to blurred outputs and artifacts in fine-grained regions, thereby degrading the overall realism and perceptual quality.

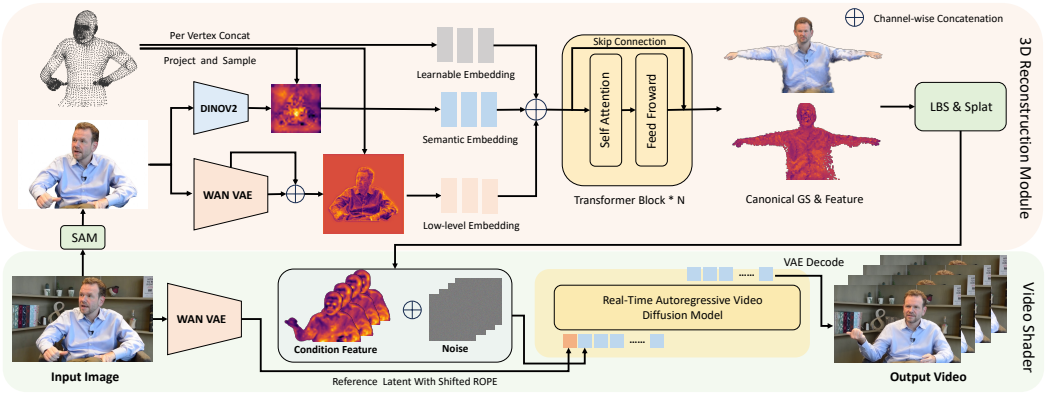

Figure 2: **Overview of ViSA.** The reconstruction module uses a feed-forward transformer to regress a canonical 3D Gaussian avatar from geometric, semantic, and low-level embeddings. The video shader conditions an autoregressive video model on the rendered 3D-aware features to generate photorealistic results in real time.

2.2 Controllable Human-Centric Video Generation

Fueled by large-scale datasets and advances in diffusion architectures (Rombach et al., 2022; Peebles & Xie, 2023), diffusion-based human-centric video generation approaches (Wan et al., 2025; Kong et al., 2024; Gao et al., 2025; Team et al., 2025; Ali et al., 2025) have progressed substantially, especially in generalization and high-fidelity visual quality. Animate Anyone (Hu, 2024) uses a reference network to preserve identity and a lightweight pose-conditioning module for guidance. EMO (Tian et al., 2024) generates expressive portrait videos from audio and a reference image. Subsequent works, such as Champ (Zhu et al., 2024), MIMO (Men et al., 2024), and Human4DiT (Shao et al., 2024), replace sparse 2D keypoints with 3D shape guidance (Pavlakos et al., 2019; Loper et al., 2015), enabling more accurate and controllable generation. Realisdance (Zhou et al., 2024) utilizes a variety of control signals, such as sparse keypoints and rendered images from the MANO (Romero et al., 2022) and SMPL (Loper et al., 2015) models, to achieve more controllable generation effects. Although they can produce vivid human animation videos, they struggle to generate temporally consistent long videos with stable body structures. Additionally, these methods require several minutes to generate a video sequence given the human poses (Zhu et al., 2024; Shao et al., 2024), limiting their practicality in interactive applications.

2.3 Autoregressive Video Diffusion Model

Autoregressive diffusion models (Chen et al., 2024; Jin et al., 2024; Chen et al., 2025; Gu et al., 2025) generate video frames sequentially, offering a promising way to reduce generation latency. Recent advances such as CausVid (Yin et al., 2025) and Self-Forcing (Huang et al., 2025) enable real-time, high-fidelity video generation by combining few-step diffusion distillation with causal inference. We adapt an autoregressive video diffusion model into a real-time video shader. Conditioned on dense 3D-aware features, our method achieves photorealistic rendering in real time.

3 Method

3.1 Overview

ViSA generates photorealistic avatars from a single image through two stages: 3D reconstruction (Section 3.2) provides geometric stability, and a real-time autoregressive video renderer (Section 3.3) supplies generative detail. Section 3.5 describes progressive training; Appendix A details dataset curation.

3.2 3D RECONSTRUCTION

To construct a comprehensive guide for the reconstruction and rendering, we extract and fuse three complementary types of features: i) *Semantic Features (F_{sem}).* We employ a pretrained DINOv2 encoder (Oquab et al., 2023), denoted as E_{dino}, to capture high-level semantic information about the subject. The semantic feature is extracted as

$$F_{sem} = E_{dino}(I_{dino}), \tag{1}$$

where I_{dino} is the input image I resized to the input resolution of E_{dino}.

ii) *Low-Level Visual Features (F_{vis}).* DINOv2 is not optimized for reconstruction, as it tends to discard fine-grained local details and is constrained to relatively low input resolutions. To address this problem, we use the pretrained VAE encoder from a video diffusion model (Wan et al., 2025), denoted as E_{vae}, to provide complementary information. As this encoder is trained for reconstruction, it excels at preserving high-fidelity local information and can process much higher-resolution inputs. We extract multi-scale feature maps $\{f^l_{vae}\}_{l\in L}$ from a set of intermediate layers L of the encoder. These features are then aggregated by a pyramid encoding network, Net_{pyr}, to produce a fused visual feature map:

$$F_{vis} = Net_{pyr}(\{f^l_{vae}\}_{l\in L}). \tag{2}$$

iii) *General Human Prior (F_{prior}).* While the features above provide person-specific information, we also introduce a person-invariant prior to encode general human structure. To achieve this, we create a learnable embedding table for the vertices of the SMPL-X template. For each of the N_v vertices in the canonical SMPL-X model, we assign a learnable d-dimensional feature vector. This collection of embeddings forms a human prior feature table, which is independent of the input image and shared across all subjects:

$$F_{prior} = \{e_i\}^{N_v}_{i=1}, \quad \text{where } e_i \in \mathbb{R}^d. \tag{3}$$

We merge these multi-modal features in 3D space to predict the attributes of the 3D Gaussian Splatting (GS) representation as well as the 3D-aware features to guide the following generation process. Specifically, we first lift the 2D features, F_{sem} and F_{vis}, into the 3D canonical space of the SMPL-X model with the estimated pose and shape parameters of the input image. Denoting these lifted features for the i-th vertex as $f_{sem,i}$ and $f_{vis,i}$, they are then concatenated with the per-vertex learnable prior e_i to formulate the final feature token T_i:

$$T_i = \text{concat}(f_{sem,i}, f_{vis,i}, e_i). \tag{4}$$

The tokens $\{T_i\}^{N_v}_{i=1}$ are then processed by a lightweight reconstruction network, which consists of five transformer layers, to predict the 3D Gaussian attributes and 3D-aware features.

The overall structure of our reconstruction model is illustrated in Fig. 2. During training, our model is supervised by a composite loss function, combining RGB and mask supervision:

$$\begin{aligned}\mathcal{L}_{recon} = \lambda_{L1}\|I_{ren} - I_{gt}\|_1 + \lambda_p\mathcal{L}_{LPIPS}(I_{ren}, I_{gt}) \\ + \lambda_{mask}\|M_{ren} - M_{gt}\|^2_2,\end{aligned} \tag{5}$$

where λ_{L1}, λ_p, and λ_{mask} are weighting coefficients. This 3D reconstruction model provides the following video shader with 3D-aware guidance, which is essential for stable avatar generation.

3.3 REAL-TIME VIDEO SHADER

We build our video shader upon Self-Forcing (Huang et al., 2025), a state-of-the-art few-step autoregressive video diffusion model for high-quality, real-time video synthesis. We introduce several key architectural modifications to adapt it into a powerful, 3D-aware neural renderer.

3.3.1 DYNAMIC CONDITIONING WITH 3D-AWARE FEATURES.

Traditional controllable generation methods (Hu, 2024; Xu et al., 2024; Zhu et al., 2024; Zhou et al., 2024) typically rely on 2D guidance signals, such as keypoints or rendered SMPL(-X) images. These

RGB-based conditions are encoded into the latent space using a pretrained VAE encoder. We argue that this paradigm has two fundamental limitations: **Information Bottleneck**: 2D guidance is either sparse (keypoints) or semantically simplistic (SMPL). These signals primarily convey motion, offering weak priors for subject-specific shape and texture, which often leads to identity drift. **Latency Bottleneck**: In real-time settings, the necessity of a forward pass through the VAE encoder for each conditioning frame introduces significant latency, hindering performance. In contrast, we propose conditioning the video model directly on the 3D-aware feature maps from our reconstruction model. These dense features encode rich, multi-level information about identity, shape, and texture, providing a strong prior that mitigates identity drift. We concatenate the 3D-aware features and noisy latents along the channel dimension before feeding them into the diffusion model, adding minimal parameters. Our approach not only supplies rich semantic priors for generation but also bypasses VAE encoding, yielding a more informative and efficient solution.

3.3.2 Static Conditioning with a Reference Image.

To further anchor the avatar's identity, we condition the generation on the original input image. The Self-Forcing model, which is distilled from WAN2.1 1.3B (Wan et al., 2025), is designed for text-to-video generation and lacks a native mechanism for image conditioning. However, its causal autoregressive structure offers an elegant solution. Specifically, we treat the VAE-encoded reference image as the first frame of the generation sequence. We perform a single forward pass to compute its Key-Value (KV) cache within the model's attention layers. This KV cache is then held constant and reused throughout the subsequent autoregressive generation process, serving as a persistent identity reference. To address the spatial misalignment between the static reference frame and the animated subsequent frames, we adopt the shifted Rotary Position Embedding (RoPE) (Su et al., 2024; Zhou et al., 2024), which ensures coherent feature integration despite significant motion, effectively injecting the reference identity into the entire generated video.

3.4 Supervision

Our finetuning process is designed to effectively adapt the pre-distilled video diffusion model into a high-fidelity, 3D-aware video renderer. This is achieved through a combination of a paired conditioning loss to improve alignment with control signals and a novel adversarial loss to maintain sharpness.

Paired Condition Supervision. Given a training triplet $(I_{ref}, V_{gt}, \Theta_{gt})$, which includes a reference image, ground-truth video frames, and the corresponding SMPL-X parameters, our pipeline proceeds as follows: First, the 3D reconstruction module generates a canonical Gaussian avatar from I_{ref}. This avatar is then animated using the target poses from Θ_{gt}. For each frame, we render a 3D-aware feature map F_{cond} to serve as the dynamic condition and adopt V_{gt} to provide paired supervision. During finetuning, we randomly sample a timestep from the adapted few-step schedule in Appendix D and then generate video frames $\hat{V}$ with a self-rollout training strategy.

To guide the generation, we apply supervision in both the latent and pixel spaces. For latent-space supervision, we adopt MSE loss to enforce content consistency between generated frames and ground-truth frames. For pixel-space supervision, which ensures the preservation of fine-grained details, we use a combination of pixel-wise, perceptual, and style losses. To mitigate the computational cost of decoding for this pixel-level comparison, we employ a tiny decoder (Boer Bohan, 2025), which is distilled from the WAN VAE (Wan et al., 2025) with minimal performance degradation, as our VAE decoder. The final paired conditioning loss $\mathcal{L}_{\text{pair}}$ is formulated as a weighted sum of these components:

$$\begin{aligned}\mathcal{L}_{\text{pair}} = \lambda_{\text{latent}} \|Z_{\hat{V}} - Z_{V_{\text{gt}}}\|_2^2 + \lambda_{L2} \|\hat{V} - V_{\text{gt}}\|_2^2 \\ + \lambda_p \mathcal{L}_{\text{LPIPS}}(\hat{V}, V_{\text{gt}}) + \lambda_s \mathcal{L}_{\text{style}}(\hat{V}, V_{\text{gt}}),\end{aligned} \tag{6}$$

where λ_{latent}, λ_{L2}, λ_p, and λ_s are loss weights.

Adversarial Distribution Preservation Loss. During our experiments, we found that naively finetuning the distilled video diffusion model with the denoising objective ($\mathcal{L}_{pair}$), while capable of incorporating conditioning signals, results in blurry outputs. This is caused by the high variance between randomly sampled noise and fixed ground-truth targets. As a result, the model tends to

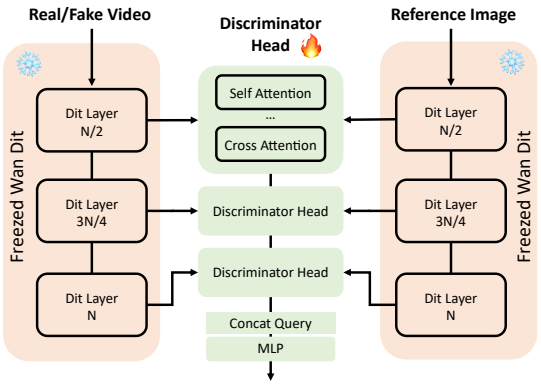

Figure 3: **Discriminator architecture.** A pretrained WAN backbone and a trainable classification head predict logits conditioned on reference-image features.

predict an average of multiple plausible images (Luo et al., 2025a;b). To counteract this and enforce the sharpness and realism of the generation results, we introduce an adversarial training scheme to keep its outputs aligned with the distribution of real-world videos.

We formulate the finetuning as an adversarial training process, where our video shader acts as the generator and we introduce a powerful critic network as the discriminator. Specifically, we build our discriminator with the frozen backbone of the pretrained WAN video diffusion model. We argue that this large model, trained on vast amounts of video data, already possesses a strong implicit prior of natural video statistics, making it an ideal critic. A small, trainable classification head is added to this backbone. To make the discriminator conditional, we integrate the reference image I_{ref} into the training process by feeding features extracted from real or generated videos and the reference image into the classification head. The architecture of our discriminator is shown in Fig. 3.

For the training objective, we employ a relativistic loss (Jolicoeur-Martineau, 2018) for more stable training. The training objectives for the discriminator D (with parameters ψ) and the generator G (with parameters θ) are defined as:

$$\mathcal{L}_D(\psi) = -\mathbb{E}\big[\log(\text{sigmoid}(D(V_{gt}, I_{ref}) - D(\hat{V}, I_{ref})))\big] + \lambda_{reg}\mathcal{L}_{reg}, \tag{7}$$

$$\mathcal{L}_G(\theta) = -\mathbb{E}\big[\log(\text{sigmoid}(D(\hat{V}, I_{ref}) - D(V_{gt}, I_{ref})))\big]. \tag{8}$$

We also adopt approximate R1 and R2 regularization (Huang et al., 2025; Lin et al., 2025b;a) to stabilize training.

3.5 Progressive Training Strategy

To facilitate stable end-to-end training and enhance temporal consistency, we adopt a two-stage progressive training strategy.

Stage 1: End-to-End Alignment. We first initialize the reconstruction model through a brief warm-up phase and then train the entire framework, i.e., the reconstruction model and video shader, end-to-end on short video clips (e.g., 9 frames). This stage focuses on aligning the 3D-aware features with the autoregressive video diffusion model. The total training objective is:

$$\mathcal{L}_{stage1} = \mathcal{L}_{recon} + \mathcal{L}_{pair} + \mathcal{L}_{adv}. \tag{9}$$

Stage 2: Temporal Extension. After the initial alignment, we freeze the parameters of the 3D reconstruction model and continue training the video shader on longer sequences. This stage enhances the model's ability to maintain temporal consistency over extended periods. The training objective in this stage is:

$$\mathcal{L}_{stage2} = \mathcal{L}_{pair} + \mathcal{L}_{adv}. \tag{10}$$

During finetuning, we employ the same self-rollout strategy as Self-Forcing (Huang et al., 2025) to prevent train-test mismatch. To better preserve the rich knowledge of the pretrained model and



Input Image Champ MimicMotion VACE GUAVA Ours

Figure 4: **Qualitative comparisons on self-reenactment**. Compared with existing approaches, our method generates more accurate and clearer results. Baselines: Champ (Zhu et al., 2024), MimicMotion (Zhang et al., 2024), VACE (Jiang et al., 2025b), and GUAVA (Zhang et al., 2025a).



Input Reference Champ MimicMotion VACE GUAVA Ours

Figure 5: **Qualitative results on cross-reenactment against existing methods**. Our approach outperforms alternatives in maintaining identity consistency across a range of poses while more faithfully capturing facial expressions and fine-grained details.

enhance training efficiency, only the input layer for the 3D-aware features and the self-attention blocks in the video diffusion model are finetuned, while other weights are frozen.

4 Experiments

4.1 Experimental Setup

Implementation Details. Using the curation procedure described in Appendix A, we construct a dataset of 46,000 video clips, with approximately 150 frames per clip on average (corresponding to a typical duration of 3–10 seconds). From this collection, we randomly select 100 subject IDs, each contributing one video clip, as the test set, with the remaining clips forming the training set. Our



2D-Keypoint 3D-RGB 3D-Feature

(a) **Conditioning inputs.**



Input Image w/o ADP w/ ADP

(b) **Distribution preservation.**

Figure 6: **Ablation studies.** (a) Conditioning inputs (Sec. 3.3.1); red boxes highlight differences. (b) Adversarial Distribution Preservation (ADP) reduces blur and artifacts.

Table 1: Quantitative comparison of avatar generation methods: VACE (Jiang et al., 2025b), Champ (Zhu et al., 2024), MimicMotion (Zhang et al., 2024), and GUAVA (Zhang et al., 2025a).

Method	PSNR ↑	SSIM ↑	LPIPS ↓	IPS-self ↓	IPS-cross ↓	Speed (fps) ↑
VACE	14.83	0.78	0.110	0.053	0.066	0.28
Champ	15.57	0.79	0.106	0.049	0.074	1.90
MimicMotion	17.91	0.83	0.087	0.045	0.065	0.72
GUAVA	18.59	0.86	0.072	0.040	0.061	**89.12**
ViSA (Ours)	**22.09**	**0.87**	**0.043**	**0.037**	**0.060**	15.15

ViSA model is initialized from the Self-Forcing 1.3B model (Huang et al., 2025) and trained on this training set using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a learning rate of 5×10^{-5}. Stage one is trained on 32 NVIDIA H20 GPUs with a total batch size of 64 for 4,000 steps. Stage two is then trained for 6,000 steps using the same batch size and number of GPUs. The entire training process takes approximately five days.

Metrics. *Self-reenactment.* We use the first frame of each video as the reference image and animate it using the full motion sequence from the same video. To evaluate image quality, we compute PSNR, SSIM, and LPIPS (Zhang et al., 2018) between all generated frames and their corresponding ground-truth frames in the original video. Additionally, we measure Identity Preservation Score (IPS-self) by computing the cosine distance between ArcFace (Deng et al., 2019) features extracted from the face regions of the generated and their corresponding original frames. *Cross-reenactment.* Since no ground-truth output exists when animating a source identity with motion from a different subject, we assess identity consistency using the Identity Preservation Score (IPS-cross), defined as the average cosine similarity between ArcFace features of the reference image and all frames in the generated video.

4.2 Comparison with State-of-the-art Methods

We compare our ViSA with state-of-the-art *2D-based methods*, including MimicMotion (Zhang et al., 2024), Champ (Zhu et al., 2024), and VACE (Jiang et al., 2025b), and the *3D-based method* GUAVA (Zhang et al., 2025a).

As shown in Fig. 4 and Fig. 5, video-based methods often suffer from identity drift: the generated avatar fails to maintain consistent appearance with the input reference image, and its identity gradually shifts over time. Additionally, pose details are frequently unstable. For instance, hands ap-

Table 2: Component-wise latency (ms per frame) for different conditioning methods, measured on a single A100 GPU.

Condition	GS Render	Encode	Denoise	VAE	All
Feature Condition	2	-	62	2	**66**
RGB Condition	1	36	62	2	101

pear blurry or adopt anatomically implausible configurations. Backgrounds are also poorly handled; VACE tends to replace them with arbitrary scenes, while Champ hallucinates unrealistic elements such as a floating microphone. These issues are reflected in Table 1, where these methods exhibit inferior PSNR, SSIM, and LPIPS scores, as well as higher IPS-self and IPS-cross distances indicating poor identity preservation. Moreover, their reliance on multi-step diffusion models incurs high computational cost, requiring several minutes to generate a short 10–20-second clip. In contrast, the 3D-based method GUAVA leverages efficient 3D Gaussian Splatting (3DGS) rendering to achieve real-time synthesis. By attaching Gaussians to an SMPL-X template, it ensures motion controllability and stable identity. However, this rigidity limits expressiveness: motions appear stiff, fine details are lost, and backgrounds cannot be synthesized.

Our approach bridges the gap between these paradigms. By integrating 3D-aware avatar features with a lightweight, few-step autoregressive video diffusion model, ViSA achieves both high fidelity and real-time performance. The 3D prior ensures identity consistency, pose accuracy, and coherent background integration, while the generative renderer enriches details and produces fluid motion. This combination improves visual quality while enabling generation at 15.15 fps on a single A100 GPU (Table 1).

4.3 Ablation Study

Different Conditioning Inputs. Figure 6a shows the ablation study for different guidance inputs. For human keypoint inputs, the skeletal representation provides only coarse positional information and lacks pixel-level guidance for image synthesis. As a result, it produces severe artifacts in high-frequency regions, such as the fingers. In contrast, our model conditions on pixel-aligned guidance generated from a 3D reconstruction model, yielding more accurate and consistent animation results. Figure 6a further shows that the model conditioned on 3D-aware features can capture more precise semantic information and produces more natural, photorealistic animations than that conditioned on 3D-aware RGB. The component-wise speed analysis in Table 2 further compares the efficiency of feature-based condition injection versus traditional RGB-based injection. Our feature-based conditioning bypasses the VAE encoder, reducing per-frame latency from 101 to 66 ms, or by approximately 35%, compared with the RGB-conditioned pipeline.

The Effects of Adversarial Distribution Preservation. Figure 6b presents an ablation study of our Adversarial Distribution Preservation loss (ADP). The results show that ADP effectively improves overall visual quality compared to training without it, particularly in high-frequency regions, and yields more natural renderings.

Additional user-study results are provided in Appendix C, and limitations are discussed in Appendix B.

5 Conclusion

We presented ViSA, a real-time video shading framework for photorealistic 3D avatar generation that effectively bridges geometric reconstruction and generative video synthesis. By leveraging a 3D-aware reconstruction module to provide structural and appearance priors, our method guides a distilled autoregressive video diffusion model to render high-fidelity, temporally coherent avatars with rich details and natural motion. Experiments demonstrate that ViSA improves visual quality over the evaluated methods while maintaining real-time inference.

AI USE STATEMENT

Generative AI tools were used to assist with document formatting and language polishing.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018.

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024.

Ollin Boer Bohan. Taehv: Tiny autoencoder for hunyuan video. `https://github.com/madebyollin/taehv`, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=gVM2AZ5xA6`.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7297–7306, 2018.

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 *SIGGRAPH*, 2025.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Yiming Jiang, Wenfeng Song, Shuai Li, and Aimin Hao. Hfhuman: High-fidelity human reconstruction from single image with multi-modality fusion. *IEEE Transactions on Visualization and Computer Graphics*, 2025a.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *ICCV*, 2025b.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.

A Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *SIGGRAPHA*, 2017.

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025a.

Shanchuan Lin, Ceyuan Yang, Hao He, Jianwen Jiang, Yuxi Ren, Xin Xia, Yang Zhao, Xuefeng Xiao, and Lu Jiang. Autoregressive adversarial post-training for real-time interactive video generation. *arXiv preprint arXiv:2506.09350*, 2025b.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *TOG*, 2015.

Yihong Luo, Tianyang Hu, Yifan Song, Jiacheng Sun, Zhenguo Li, and Jing Tang. Adding additional control to one-step diffusion with joint distribution matching. *arXiv preprint arXiv:2503.06652*, 2025a.

Yihong Luo, Shuchen Xue, Tianyang Hu, and Jing Tang. Noise consistency training: A native approach for one-step generator in learning additional controls. *arXiv preprint arXiv:2506.19741*, 2025b.

Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. *arXiv preprint arXiv:2409.16160*, 2024.

Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *ECCV*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. Lhm: Large animatable human reconstruction model from a single image in seconds. In *ICCV*, 2025a.

Lingteng Qiu, Peihao Li, Qi Zuo, Xiaodong Gu, Yuan Dong, Weihao Yuan, Siyu Zhu, Xiaoguang Han, Guanying Chen, and Zilong Dong. Pf-lhm: 3d animatable avatar reconstruction from pose-free articulated human images. *arXiv preprint arXiv:2506.13766*, 2025b.

Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In *CVPR*, 2025c.

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

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4dit: 360-degree human video generation with 4d diffusion transformer. *TOG*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *ECCV*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260. Springer, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22963–22974, 2025.

Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Yang Li, Minghan Qin, Yu Li, and Haoqian Wang. Guava: Generalizable upper body 3d gaussian avatar. *arXiv preprint arXiv:2505.03351*, 2025a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. *arXiv preprint arXiv:2507.09862*, 2025b.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

Jingkai Zhou, Benzhi Wang, Weihua Chen, Jingqi Bai, Dongyang Li, Aixi Zhang, Hao Xu, Mingyang Yang, and Fan Wang. Realisdance: Equip controllable character animation with realistic hands. *arXiv preprint arXiv:2409.06202*, 2024.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint arXiv:2403.14781*, 2024.

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. In *CVPR*, 2025.

SUPPLEMENTARY MATERIAL

A DATASET CURATION

We construct our training dataset from SpeakerVid-5M (Zhang et al., 2025b), a large-scale dataset of high-quality human portrait videos. Our data processing pipeline involves several key stages to ensure high-quality annotations suitable for 3D-aware avatar generation.

Initial Filtering. The pipeline begins by discarding static clips, identified by analyzing the temporal variance of 2D keypoint locations. As our focus is on upper-body avatars, we further select clips where the face and hands are clearly visible, retaining only those with corresponding keypoint confidence scores above a predefined threshold.

Parameter Estimation and Refinement. We employ Multi-HMR (Baradel et al., 2024) and GAGATracker (Chu & Harada, 2024) to obtain initial estimates for SMPL-X and FLAME parameters, respectively. However, direct feed-forward estimation for in-the-wild, upper-body videos is notoriously challenging, often yielding inaccurate poses (particularly for hands), shapes, and expressions. To refine these estimates, we adopt a per-video optimization strategy similar to recent works (Moon et al., 2024; Zhang et al., 2025a). This involves minimizing the reprojection error between the 3D body/facial keypoints and their corresponding 2D detections. Following this optimization, a second filtering pass is performed to discard any clips where the final reprojection error remains above a specified threshold, ensuring a high-quality ground-truth dataset.

Seamless Head-Body Fusion. To achieve enhanced facial expressiveness, we replace the head of the SMPL-X model with the more detailed FLAME model (Zhang et al., 2025a). A naive replacement, however, often results in a noticeable discontinuity artifact at the neck seam due to geometric inconsistencies. To resolve this, we propose a refined fusion technique. Instead of swapping the entire head region, we selectively graft only the facial region of FLAME onto the SMPL-X model, while retaining the original SMPL-X neck.

Specifically, we identify the neck vertices of the FLAME model by leveraging its skinning weights, $W_{flame} \in \mathbb{R}^{N_{flame} \times J_{flame}}$. A vertex v_i is classified as part of the neck if its largest skinning weight corresponds to the neck joint:

$$\underset{j}{\text{argmax}} \left(W_{flame}[i, j]\right) = id_{neck} \tag{11}$$

By excluding these identified neck vertices from the FLAME model during the fusion, we create a seamless transition to the original SMPL-X neck. As shown in Fig. 7, this simple yet effective strategy successfully eliminates the neck artifact, ensuring a coherent and natural-looking full-body template for our training.

B LIMITATIONS

While ViSA demonstrates promising results in synthesizing photorealistic, real-time upper-body avatars, its performance is still bounded by the capacity and quality of the underlying base model. Another primary bottleneck is the inference speed, which is currently constrained by the 4-step denoising process. Reducing the number of denoising steps often leads to a degradation in quality, particularly in high-frequency details. Therefore, a key future direction is to develop methods that can achieve high-fidelity synthesis with fewer denoising steps, thereby further accelerating inference. Moreover, the current work focuses mainly on reconstructing and conditioning on the human subject, omitting the background. Integrating the background into the 3D reconstruction process to ensure stable and consistent background generation presents another important and interesting avenue for future research.

C USER STUDY

To further evaluate the practical performance of our method, we conducted a comprehensive user study focusing on the cross-ID animation task. Our evaluation set consists of diverse data pairs, each comprising a driving video and a distinct character image. We invited 22 participants for a

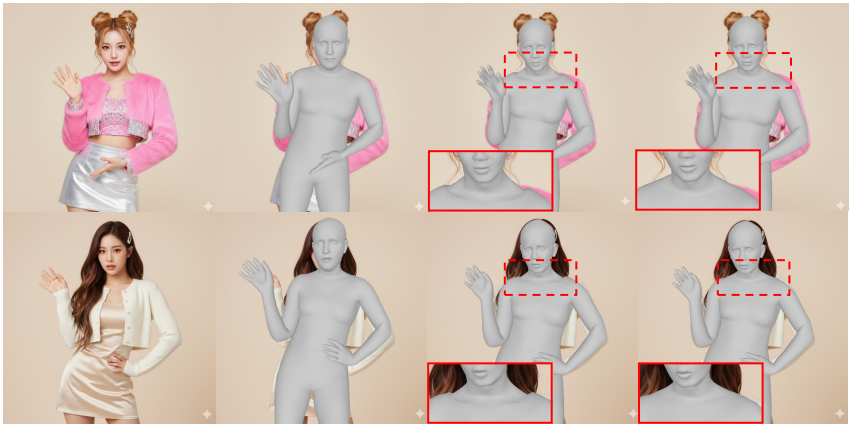

Figure 7: **Effects of our neck-vertex removal strategy**. This simple yet effective strategy eliminates neck artifacts, yielding a coherent, natural-looking full-body template for training. The red boxes highlight the differences.

subjective evaluation, where generated videos from our method and competing baselines, including MimicMotion (Zhang et al., 2024), VACE (Jiang et al., 2025b), and GUAVA (Zhang et al., 2025a), were presented side-by-side in an anonymous fashion. Participants were asked to select their preferred result based on a holistic consideration of video generation quality and identity (ID) consistency. As summarized in Table 3, our method receives the highest preference rates for both generation quality and ID preservation.

Table 3: User study on generation quality and ID consistency. Baselines: MimicMotion (Zhang et al., 2024), VACE (Jiang et al., 2025b), and GUAVA (Zhang et al., 2025a).

Method	MimicMotion	VACE	GUAVA	ViSA (Ours)
Generation Quality	0.05	0.22	0.18	**0.55**
ID consistency	0.12	0.08	0.19	**0.61**

D More Implementation Details

Our video diffusion shader is built upon the 1.3B-parameter, 4-step Self-Forcing model (Huang et al., 2025), which is distilled from the WAN2.1 base model (Wan et al., 2025). To adapt this model for our task, we modify its initial convolutional layer to accept our 3D-aware features as conditioning input. As the 3D-aware features already provide coarse shape information, we replace the original uniform timestep schedule from Self-Forcing (where $[t_4, t_3, t_2, t_1] = [1000, 750, 500, 250]$) with a non-uniform one, where $[t_4, t_3, t_2, t_1] = [1000, 250, 125, 75]$. This new schedule concentrates the denoising steps in the low-noise region, which is crucial for refining high-frequency details. For autoregressive synthesis, we adopt a similar chunk-wise strategy to Self-Forcing, where each new chunk consists of three latent frames. The 3D Gaussian features, with a dimension of [N, 32, C, H, W], are reshaped to [N/4, 128, C, H, W] before being input to the shader. This aligns their spatio-temporal structure with that of the noisy latents. During the second training stage, we train on video clips of varying lengths, processed chunk-by-chunk. We employ a truncated backpropagation strategy to reduce the high memory cost caused by supervision in the image space: for each sampled video, we detach the previously generated chunks from the computation graph and compute the loss only on the final chunk. This approach reduces memory use and enhances the model's ability to generate long, coherent videos during inference.

E STREAMING LONG VIDEO GENERATION

During inference, we employ a sliding window strategy for the KV cache to enable efficient generation of long sequences. Unlike the Self-Forcing baseline, which uses continuously increasing positional indices, our approach treats the attention window as a self-contained unit. Specifically, the position index for any query is capped at $\min(\text{current_pos}, \text{window_size})$, where $\text{window_size} = 4$ here. For the keys and values within the window, their positional embeddings are recalculated relative to the start of the current window before each attention computation. This relative, window-based RoPE mechanism breaks the dependency on absolute sequence length, allowing our model to synthesize videos of arbitrary length. Fig. 8 shows long-video generation results of our model.

F MORE VISUALIZATION RESULTS

We provide more visualization results in Fig. 9. We also provide the generation results alongside the corresponding 3D-aware features in Fig. 10.

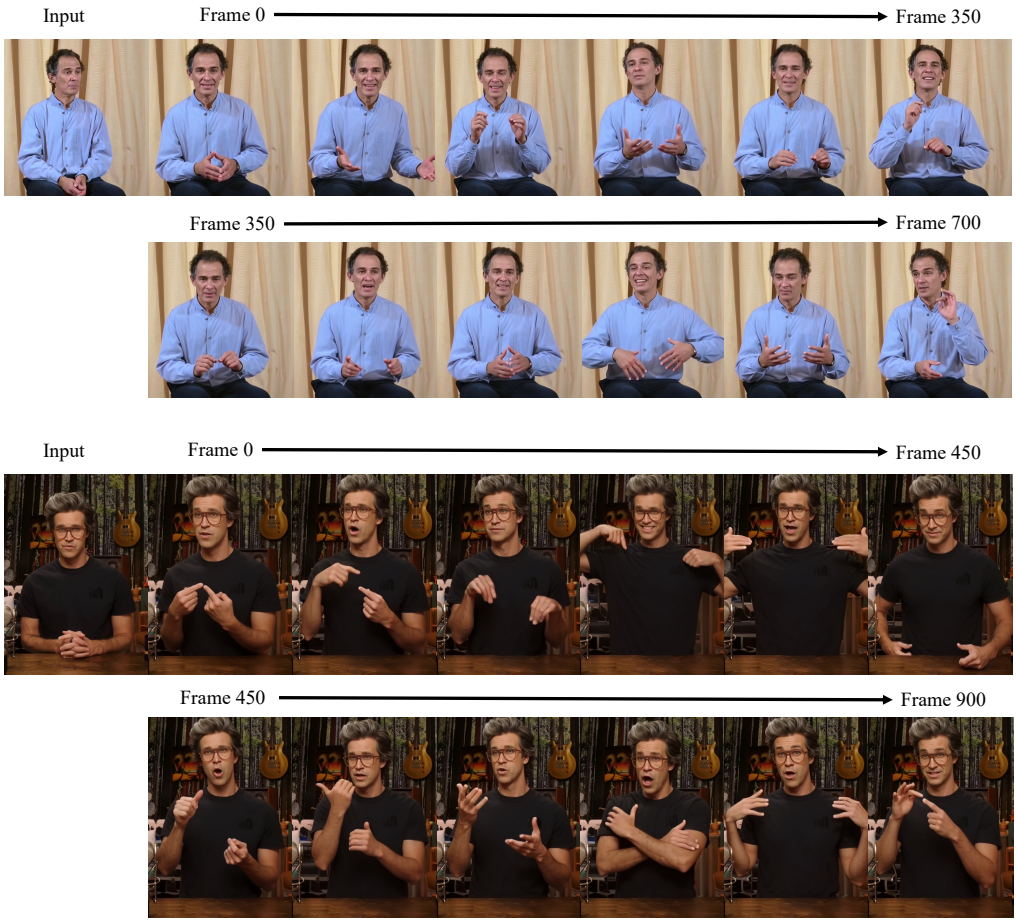

Figure 8: Long video generation results of our method. Our method is capable of generating long video sequences with consistent identity and clear movements.



Figure 9: More generation results of our method. Our method enables photorealistic upper-body avatar generation that preserves appearance fidelity across diverse poses and expressions while maintaining temporal coherence in real-time video synthesis.

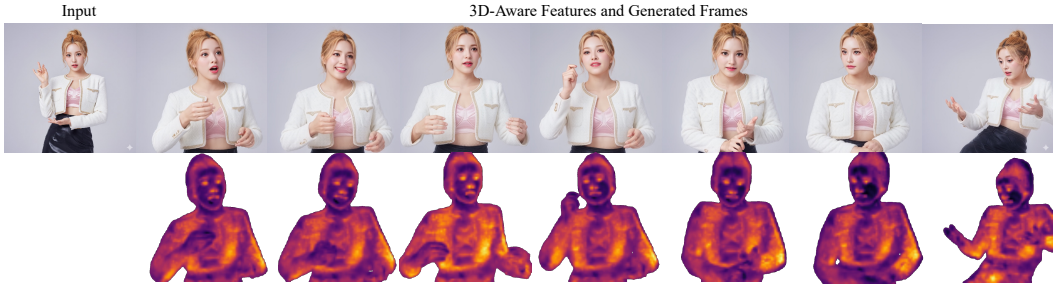

Figure 10: Generated results and corresponding 3D-aware features. Our method produces photorealistic outputs that remain highly consistent with the conditioning features, accurately rendering complex details and expressions.