MagicInfinite: Generating Infinite Talking Videos with Your Words and Voice

Hongwei Yi[1*†] **Tian Ye**[1,3*] **Shitong Shao**[1,3*] **Xuancheng Yang**[1*] **Jiantong Zhao**[1*]

Hanzhong Guo[1,4*] **Terrance Wang**[1*] **Qingyu Yin**[1] **Zeke Xie**[3] **Lei Zhu**[3]

Wei Li[1] **Michael Lingelbach**[1] **Daquan Zhou**[2‡]

[1] Hedra Inc. [2] Peking University [3] HKUST(GZ) [4] HKU

Abstract

We present MagicInfinite, a novel diffusion Transformer (DiT) framework that overcomes traditional portrait animation limitations, delivering high-fidelity results across diverse character types—realistic humans, full-body figures, and stylized anime characters. It supports varied facial poses, including back-facing views, and animates single or multiple characters with input masks for precise speaker designation in multi-character scenes. Our approach tackles key challenges with three innovations: (1) 3D full-attention mechanisms with a sliding window denoising strategy, enabling infinite video generation with temporal coherence and visual quality across diverse character styles; (2) a two-stage curriculum learning scheme, integrating audio for lip sync, text for expressive dynamics, and reference images for identity preservation, enabling flexible multi-modal control over long sequences; and (3) region-specific masks with adaptive loss functions to balance global textual control and local audio guidance, supporting speaker-specific animations. Efficiency is enhanced via our innovative unified step and cfg distillation techniques, achieving a 20× inference speed boost over the basemodel—generating a 10-second 540x540p video in 10 seconds or 720x720p in 30 seconds on 8 H100 GPUs—without quality loss. Evaluations on our new benchmark demonstrate MagicInfinite's superiority in audio-lip synchronization, identity preservation, and motion naturalness across diverse scenarios. It is publicly available at `https://www.hedra.com/`, with examples at `https://magicinfinite.github.io/`.

1 INTRODUCTION

Talking avatars—realistic digital characters created from a reference image using audio or text input—are a transformative technology at the intersection of computer vision and human-computer interaction, driving advancements in digital entertainment, education, and AI communication [39, 59, 25, 48, 1, 24], fundamentally reshaping how humans interact with digital content.

Pioneering works [51, 50, 65, 5, 69, 40, 61, 3, 44, 9, 6, 31, 74, 43, 29, 36, 17, 58] use neural networks and rendering like GANs [16], NeRF [42], and Gaussian Splatting [26] to fuse motion and identity features for high-fidelity talking heads. Yet, 3DMM [54] and FLAME [33] struggle with motion accuracy, and rendering limits resolution and quality. Recent LDMs [49] boost video synthesis [4, 19, 7, 62, 60, 64, 73, 46, 28] with better diversity and coherence. We apply T2V diffusion to portrait animation. Studies [53, 57, 8, 63, 23, 71] use LDM priors [49] to model spatial

*Equal contribution, †Project lead, ‡Corresponding author.

Figure 1: Given a portrait image, our model can generate compelling, realistic, and vivid animation videos with control over text and voice, ensuring temporal coherence and perceptual quality even under significant head pose variations and diverse portrait styles.

and temporal dimensions separately, but falter in coherence for large movements, non-frontal faces, or high-resolution cases due to weak frame correspondence [62].

In this work, we propose **MagicInfinite**, a novel zero-shot framework that utilizes 3D full-attention to effectively model video along both temporal and spatial dimensions. **MagicInfinite** demonstrates superior perceptual quality and temporal coherence across various situational motions and actions. We explored the simultaneous use of text and voice for synthesizing portrait animation videos, where the voice primarily provides lip movement dynamics, while the text supplies the speaker's expressions, actions, and background transformations. However, the integration of both audio and textual prompts as motion signals presents certain challenges: The pre-trained T2V model leverages 3D full-attention mechanisms, attending to all tokens from both the video modality and textual prompts, thereby capturing a robust dependency between the entire video content and the associated text. The text plays a significant role in controlling the video content, as each video token is related to the textual tokens. By contrast, the control of character motion by audio primarily focuses on local video regions, such as lip movements, which are only associated with a small subset of video tokens. This creates a conflict between the driving audio and the textual prompt controls. Therefore, when building upon the traditional form of incorporating audio in prior portrait animation works [53], which applies cross-attention [56] between audio features and vision representations within a spatial latent space directly, without further manipulation, the model tends to overlook the driving audio's influence on lip movements. This limitation becomes even more apparent when the face occupies a smaller portion of the image.

To address this challenge, we design a two-stage curriculum learning scheme to effectively enable audio control, guiding the model to progressively learn both textual and audio control. Specifically, in the first training phase, only the portrait image and textual prompt are introduced, allowing the

model to learn text-controlled image-to-video (I2V) generation. In the second phase, both text and audio are incorporated. Here, we utilize a face region mask and an adaptive loss function to guide the conditional cross-attention of audio, specifically targeting local facial movements around the mouth. This training scheme effectively integrates the control provided by textual prompts and driving audio for portrait animation.

Owing to the substantial increase in the number of function evaluations (NFEs) caused by multi-step sampling and classifier-free guidance (CFG) [22, 10], diffusion models often suffer from sluggish inference speeds—an issue that has been largely neglected in earlier studies. Previous work [18] obtained an avatar-based few-step generator by employing latent consistency model (LCM) [38], which effectively condensed multiple sampling steps into a reduced number, thereby minimizing the sampling process. However, we found that utilizing LCM for MagicInfinite step distillation inevitably led to noticeable degradation in visual quality, often manifesting as significant blurring. To address this pressing concern, we turn to DMD2 [66], which has been rigorously validated across numerous image diffusion models [45] and has consistently demonstrated superior performance compared to LCM, to enable accelerated sampling. Our approach achieves a groundbreaking advancement by simultaneously eliminating **MagicInfinite**'s reliance on CFG [41] during the step distillation process in DMD2, directly reducing NFEs from 50 to just 4. To circumvent the GPU memory constraints posed by importing three models simultaneously in vanilla DMD2, we adopt LoRA to update the parameters of the fake data distribution estimator, thereby ensuring an efficient and resource-aware training process.

We conducted a thorough evaluation of **MagicInfinite** using a rigorous benchmark we developed, called the **MagicInfinite**-Benchmark. This benchmark consists of 25 driving audio clips spanning diverse speaking scenarios—such as singing, speeches, and rapping—paired with 20 textual prompts describing the speaker's emotions, actions, and background variations. Additionally, the **MagicInfinite**-Benchmark includes a varied collection of portrait images featuring different styles, face orientations relative to the camera, and face sizes. Under this benchmark, **MagicInfinite** exhibits robust performance across a wide range of portrait images and motion scenarios, effectively driven by both audio and textual prompts, while maintaining efficient inference speeds.

The key contributions of our work are as follows:

- A novel zero-shot talking avatar framework using 3D full-attention in a pre-trained video DiT with a sliding window denoising strategy, enabling infinite video generation with strong temporal coherence and visual quality across diverse portrait styles and speaking scenarios.
- A two-stage curriculum learning approach, incorporating face region-guided cross-attention and adaptive loss weighting, which effectively balances global text-based control with local audio-driven control, enabling high-fidelity joint conditioning from both modalities.
- An innovative strategy for synergistic step and classifier-free guidance (CFG) distillation, combined with a sliding window mechanism, allowing seamless sampling of arbitrarily long videos. This advancement yields a 20× increase in inference speed while preserving competitive quality metrics, overcoming a key limitation in large-scale diffusion models for portrait animation.

2 RELATED WORK

2.1 GAN-based Portrait Animation

The goal of portrait animation is to generate talking videos driven by motion signals. Traditional methods typically employ neural networks to extract motion features from motion signals. These motion features are then transformed into intermediate representations such as landmarks [51, 50], 3D head parameters [65, 5, 69, 40, 61, 3] (3DMM [54]), faces learned with an articulated model and expressions [44] (FLAME [33], 3D Gaussian parameters [9, 6, 31, 74], 3D tri-plane hash representations [32], or latent representations [43, 29, 36, 17, 58]. Rendering techniques, such as GANs [16], NeRF [42], and Gaussian splatting [26] are employed to project these intermediate representations into dynamic portrait animations. Despite notable successes, limitations in rendering methods and feature extractors hinder the generation of realistic portrait animation videos.

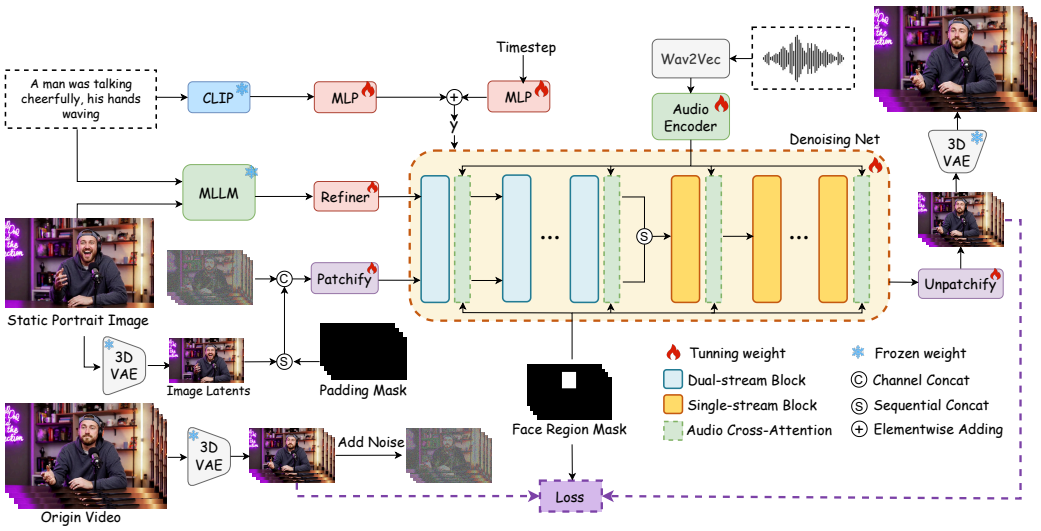

Figure 2: **Overview of MagicInfinite.** MagicInfinite employs a hybrid dual-to-single-stream denoising network with Audio Cross-Attention in final blocks. MLLM encodes static portrait and text into tokens, concatenated for T2V, refined, and denoised. Wav2Vec encodes audio, resampled by an Audio Encoder, and guided by a Face Region Mask for precise lip sync and adaptive loss.

2.2 Diffusion-based Portrait Animation

Latent Diffusion Models [49] (LDMs) have demonstrated strong capabilities in image and video generation [4, 19, 7, 62, 60]. Models such as EMO [53] and V-Express [57] leverage audio cues for lip synchronization and weak visual signals for head motion, enabling the generation of natural portrait animation videos. Echomimic [8] and MegActor-Sigma [63] combine both audio and video to offer finer control over the animation process. Loopy [23] and MEMO [71] enhance temporal consistency and emotional expressiveness by integrating modules for audio-to-emotion conversion and past-frame fusion during long video generation. However, these methods face limitations in video fidelity and narrative coherence due to separate attention computations in spatial and temporal domains. DiT-based diffusion methods, such as CogVideoX [64], Allegro [73] Movie Gen [46], and HunyuanVideo [28], employ 3D Full-Attention mechanisms, yielding high-quality video generation. Inspired by previous works, we applied the 3D Full-Attention architecture to portrait animation video generation and achieved remarkable results. Our concurrent work, Hallo3 [12], was outperformed by MagicInfinite in terms of video length and resolution.

2.3 Efficiency Inference

Diffusion model inference is inherently limited by its multi-step sampling process, which impacts computational efficiency. To address this challenge, step distillation techniques have emerged as essential approaches for accelerating inference by reducing the number of function evaluations (NFEs). The current research landscape in this domain can be categorized into two primary strategies: aligning few steps with many steps and distribution matching distillation.

In the first category, LCM [38] has established itself as a prominent framework that effectively aligns few-step models with their multi-step counterparts. This approach has been widely adopted, leading to the development of several open-source few-step VDMs, including MCM [68], the T2V-turbo series [30], and FastVideo [20].

Meanwhile, the distribution matching paradigm has demonstrated superior performance in image synthesis compared to LCM-based methods, as evidenced by techniques such as DMD [67], DMD2 [66], and SiD-LSG [72]. Despite its promising results in image generation, the application of distribution matching to video synthesis remains relatively underexplored, with limited implementations such as [66].

3 METHOD

Given a single static portrait, our objective is to generate a head animation video that is conditioned on both driving audio and a textual prompt. The resulting video should preserve the identity of the subject and the background content depicted in the static portrait while accurately reflecting lip movements, as well as head and facial gestures, and background variations in the portrait image, as dictated by the driving audio and textual prompt. In this work, we leverage our internal powerful generative prior of the production-ready DiT-based video generative model, i.e., the text-to-video (T2V) model, integrating control from the static portrait, audio, and textual prompt 3.1. We introduce a cross-attention layer into the model to enable the injection of driving audio 3.2. Building upon this, we propose a novel stepwise training strategy to enable the model to learn control from both voice and text, using the face region mask and adaptive loss function to guide the conditional cross-attention for audio 3.3. Finally, we detail our video model distillation techniques, and inference strategies for efficient long video inference 3.4.

3.1 Preliminaries

3.1.1 Flow Matching Model

Flow Matching [34] transforms a complex probability distribution into a simpler one via the probability density function, enabling the generation of new data samples through inverse transformations. Stable Diffusion V3 [14] is a subclass of Flow Matching models that operate in the latent space, leveraging a pre-trained AutoEncoder [27] to facilitate this process. Unlike standard Text-to-Image (T2I) models, which are conditioned solely on textual inputs T_s, MagicInfinite derives scene context and motion constraints from the textual prompt T_s, the driving audio A and a static portrait image I_s. The model trained to learn the reverse transformations of Flow Matching with the objective,

$$L_{base} = \mathbb{E}_{z_0, z_1, t\sim[0,1]}\left[\left\| v_t - u_\theta\big(z_t, t, T_s, I_s, A\big)\right\|_2^2\right], \tag{1}$$

where u_θ is a trainable denoising net. z_1 and z_0 notes the latent embedding of the training sample and the initialized noise sample drawn from the Gaussian distribution $\mathcal{N}(0, 1)$. z_t is the training sample constructed using a linear interpolation. $\mathbf{v}_t = d\mathbf{z}_t/dt = z_1 - z_0$ is the velocity which is the target of the model prediction.

3.1.2 T2V model

The naive T2V model consists of a Causal 3D VAE [27] and a diffusion backbone (denoising net). The 3D VAE compresses pixel-space videos and images into a compact latent space, reducing the token count for the subsequent diffusion transformer model. The diffusion backbone follows a "dual-stream to single-stream" hybrid model design, similar to [14]. In the dual-stream phase, video and text tokens are processed independently, and multimodal information is fused in the single-stream phase. Both phases employ a unified Full-Attention mechanism, which models both spatial and temporal dimensions, ensuring strong alignment between text and video content, and guaranteeing high visual quality, motion dynamics, and text-video alignment.

3.2 Model Architecture

To achieve control over the static portrait, we extend the T2V model to an image-to-video (I2V) model, as shown in Fig. 2. Specifically, we treat the first frame of a video as a static portrait image and apply zero-padding to create a tensor of the same shape as the latent input. We utilize a binary mask to encode temporal position information and adjust the parameters of the first convolutional module by zero-initialization.

To further preserve the identity of the subject and the background content depicted in the static portrait image, we use a pre-trained Multimodal Large Language Model (MLLM) [35] to encode the static portrait image. The encoded representations of both the static portrait image and the textual prompt are concatenated along the sequence dimension and used as text tokens.

The clip of driving audio is passed through the Wav2Vec [2] feature extraction module to obtain audio features. The Audio Encoder then processes these features for resampling, producing latent

audio features, which are injected into the denoising net via cross-attention. We selectively insert cross-attention layers at the end of both the single and double blocks, denoted as audio cross-attention, for the controlling of driving audio. The video tokens undergo cross-attention with the latent audio features, performed independently between frames in the latent space.

3.3 Curriculum Learning Scheme

Training with both textual prompts and audio often leads to the model neglecting audio control, resulting in incorrect lip synchronization. This occurs because the pre-trained T2V model has already established a strong association between video and text tokens. However, the driving audio primarily focuses on fine details of lip movement, which occupies only a small portion of the video frame. During fine-tuning on new data with supervision via MSE loss, the model tends to further strengthen the association between text and video, particularly focusing on head movements and background changes, while neglecting the alignment of small lip regions. Thus, the key to effectively achieving audio control over lip movements is to enhance the influence of driving audio. To address this, we design a novel two-stage curriculum learning scheme.

In the first stage, we feed driving image and textual prompts into the model through channel-wise concatenation with noise latents and Full-Attention. At this stage, the denoising net remains the same as the T2V model, without audio cross-attention blocks. The training objective is to predict target head animation video in the latent space of 3D VAE, guided by the static portrait image and textual prompt.

In the second stage, we introduce driving audio into the model. Define $M_{face} \in \mathbb{R}^{H \times W}$ as the face region mask (the union of facial movement regions across all frames), a binary mask, where H, and W represent the height and width of the compressed video frame. To emphasize the control of the driving audio, face region mask is applied to the output of the audio cross-attention layer at the pixel level, which can be described as:

$$h_i = h_i + Attn_{audio}(h_i, A_e, A_e) \times M'_{face} \tag{2}$$

h_i is the hidden states of the video after i-th single or double block in denoising net. A_e is the feature representation encoded by the audio encoder. M'_{face}represents M_{face} reformatted into tokens, specifying which tokens within the ensemble of video modalities are to be aligned with the corresponding A_e. $*$ stands for element-wise multiplication. This forces the audio cross-attention layer to focus on the correlation between driving audio and local facial movements, especially lip movements. Additionally, we design an adaptive loss function that amplifies the loss weight in the facial region, based on the size of the face area. Denote the target latents are of shape $z_1 \in \mathbb{R}^{T \times C \times H \times W}$, where T, C, H, and W represent the frame, channel, height, and width of the compressed video, respectively. The adaptive loss can be written as:

$$L_{adap} = L_{base} \frac{H \times W}{\sum_{i=0}^{H-1} \sum_{j=0}^{W-1} M_{face}^{ij}}, \tag{3}$$

. Thus, the total loss function can be written as:

$$L_{total} = L_{base} + \lambda_{\text{adap}} L_{adap} \tag{4}$$

while λ_{adap} represents the weight of the adaptive loss.

This ensures that when the facial region is smaller, the loss weight for the facial area is higher, causing the model to focus more on the correlation between driving audio and local facial movements. Conversely, when the facial region is larger, the model focuses more on the correlation between audio and the overall motion. This approach allows the model to better learn the fine-grained control of audio over lip movements.

3.4 Model Acceleration

The multi-step sampling process during diffusion model inference severely limits the model's inference speed, a problem that becomes even more pronounced in MagicInfinite, a 13B diffusion model, due to the slow inference in each sampling step. As illustrated in Fig. 3, we achieve high-quality accelerated sampling through the collaborative distillation of DMD and CFG. Our baseline algorithm is DMD2. DMD2 is a cutting-edge algorithm inspired by score distillation sampling (SDS) [47]

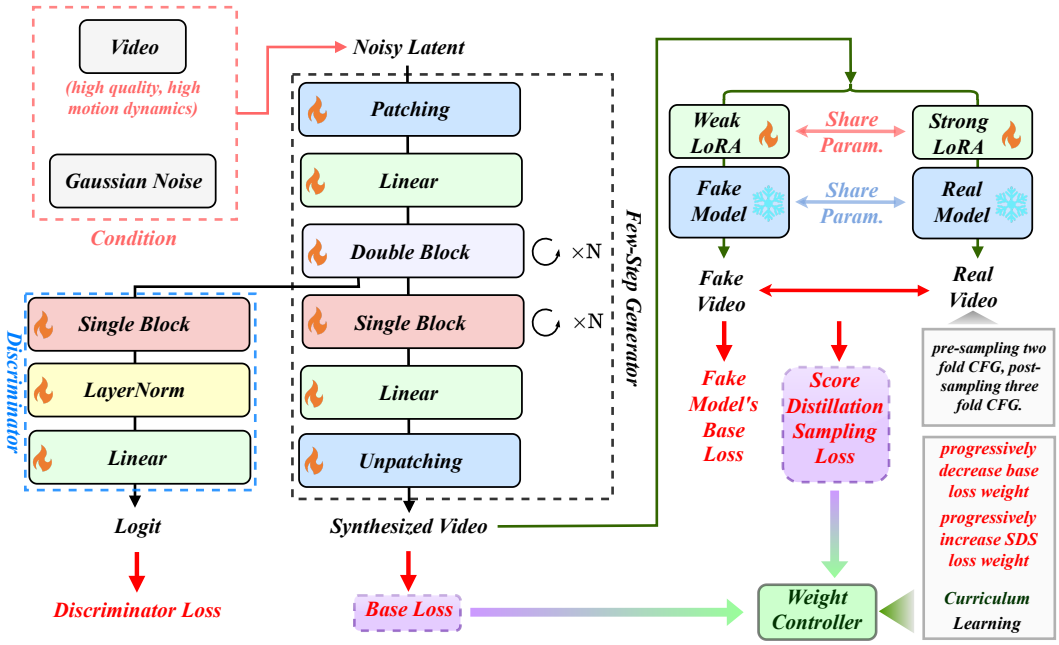

Figure 3: The overview of our modified DMD2. We employed a curriculum learning strategy to gradually reduce the weight of the base loss while progressively increasing the weight of the SDS loss, effectively avoiding abrupt shifts in learning objectives. Furthermore, we adopted a two-fold to three-fold CFG attenuation strategy in the calculation of the real data distribution, which significantly enhances the motion dynamics of the generated video.

designed to facilitate efficient distribution matching and achieve step distillation. Specifically, the training process of DMD2 involves the collaboration of three models: the one/four-step generator G_ϕ, which undergoes parameter updates; the real MagicInfinite u_θ^{real}, responsible for estimating the real data distribution p_{real}; and the fake MagicInfinite u_θ^{fake}, which estimates the fake data distribution p_{fake}. Notably, all three models are initialized from the pre-trained MagicInfinite to ensure consistency and efficiency throughout the process. Its distribution matching loss can be written as

$$\begin{aligned}\nabla_\phi L_{\text{SDS}} \triangleq \mathbb{E}_t \nabla_\phi \mathcal{D}_{\text{KL}}\left(p_{\text{fake}} \| p_{\text{real}}\right) \approx -\mathbb{E}_t \int_\epsilon \Big([z_t - \sigma_t \\ u_\theta^{\text{real}}(z_t, t, T_s, I_s, A) - z_t + \sigma_t u_\theta^{\text{fake}}(z_t, t, T_s, I_s, A)] \frac{\partial G_\phi(\epsilon)}{\partial \phi} d\epsilon \Big),\end{aligned} \tag{5}$$

where $z_t = \sigma_t z_1 + (1 - \sigma_t)\hat{z}_0$ and $\hat{z}_0$ is synthesized from the few-step generator. σ_t stands for the noise schedule. This equation transforms the original distribution matching on the score function (i.e., vanilla DMD2) into a novel distribution matching at $t = 0$, aligning with the training paradigm of MagicInfinite. Furthermore, DMD2 must update u_θ^{fake} in real time to ensure accurate estimation of p_{fake}:

$$L_{\text{fake}} = \mathbb{E}_{z_0, z_1, t \sim [0,1]} \left[\left\| z_1 - \hat{z}_0 - u_\theta^{\text{fake}}\left(z_t, t, T_s, I_s, A\right) \right\|_2^2 \right]. \tag{6}$$

The key difference between DMD2 and DMD lies in the addition of adversarial training, a technique employed to enhance and refine visual quality. In practice, we train our distillation model with 16 NVIDIA H100 GPUs (each with 80GB of memory). However, we found that even with ZeRO3 optimization, loading all three models simultaneously for standard DMD2 training was infeasible. To overcome this, we adopted an alternative approach: we leveraged the real MagicInfinite u_θ^{real}, augmented with low-rank adaptation (LoRA), to implement the fake MagicInfinite u_θ^{fake}. This adjustment enabled efficient minimization of Eq. 6 using LoRA, offering a computationally viable solution to the resource constraints encountered.

Furthermore, directly applying L_{SDS} does not guarantee the quality of the synthesized video or its motion dynamics. Moreover, directly incorporating a pre-trained video model for distribution

Figure 4: Qualitative experimental results of MagicInfinite

matching can lead to training instability or result in an uncontrollable few-step generator producing low-quality videos. To address these challenges, we first gather real data with the highest possible quality and motion dynamics. Subsequently, we employ a curriculum learning strategy to gradually reduce the weight of L_{base} while progressively increasing the weight of L_{SDS}, ensuring stable model training.

Finally, to achieve the collaborative distillation of CFG and steps, we leverage CFG during inference to estimate the real data distribution. This strategy was derived through empirical exploration, where a two-fold CFG is applied for timesteps between 0.75 and 1, while a three-fold CFG is utilized for timesteps ranging from 0 to 0.75. Specifically:

$$\begin{aligned} \hat{u_\theta}^{\text{two-fold}} &= (1+\omega_{\text{audio}})u_\theta(\mathbf{z}_t, t, T_s, I_s, A) - \omega_{\text{audio}}u_\theta(\mathbf{z}_t, t, \emptyset, \emptyset, \emptyset), \\ \hat{u_\theta}^{\text{three-fold}} &= (1+\omega_{\text{audio}})u_\theta(\mathbf{z}_t, t, T_s, I_s, A) - \omega_{\text{audio}}u_\theta(\mathbf{z}_t, t, T_s, I_s, \emptyset) \\ &\quad + (1+\omega_{\text{text}})u_\theta(\mathbf{z}_t, t, T_s, I_s, \emptyset) - \omega_{\text{text}}u_\theta(\mathbf{z}_t, t, \emptyset, \emptyset, \emptyset), \end{aligned} \tag{7}$$

where ω_{audio} and ω_{text} represent the CFG scales specifically designed for the audio condition and text condition, respectively. In addition, our CFG scale is dynamic and will sample from a Gaussian distribution with mean values ω_{audio} and ω_{text}. We found that this approach enhances the robustness of the few-step generator in synthesizing diverse characters and varying backgrounds.

4 EXPERIMENT

4.1 Implementation Details

4.1.1 Datasets

We train the model using internally collected data. All data were processed with a cropped resolution of 720×1280 and a video length of 129 frames, with the first frame serving as the portrait image during training. To address performance degradation due to variations in video frame rates (fps), we train the MagicInfinite model exclusively on videos resampled to a consistent 25 fps. We use MediaPipe[37] for human face detection and tracking to extract face mask annotations, also filtering out videos with low-resolution faces, multiple faces or static content. Videos with speakers occluded by subtitles or with audios of multiple speakers are also filtered using off-the-shelf tools. Training

textual instructions were derived from video descriptions extracted using LLAVA-34B [35], and rewritten by Gemini [52]. A total of 1.85 million video clips are used for training. To remove redundant computation during the training process, 3D VAE video features temporally aligned with Wav2Vec audio features, as well as reference image features, are also extracted during data preprocessing.

4.1.2 Training

We trained the base model of MagicInfinite on a GPU server equipped with 200 NVIDIA H100 GPUs, using the T2V model as the generative backbone. The training process for MagicInfinite was conducted in two stages. In the first stage, we trained the model for 80,000 steps to generate videos guided by a static portrait image and textual features, excluding audio-related layers. In the second stage, we introduced audio cross-attention blocks—inserted at the end of each single and double block—and trained the model for an additional 30,000 steps. This enabled the model to animate the portrait image using both driving audio and textual prompts.

For the dual-stream phase, an audio cross-attention block was added after every double block, while in the single-stream phase, one was inserted after every two single blocks. We used the AdamW optimizer [13] with a learning rate of 10^{-5} for all modules across both stages. To emphasize facial local motion in the second stage, the adaptive loss weight w was set to 10. Accelerated sampling was achieved through steps distillation and CFG distillation.

4.1.3 Efficient Long Video Inference via Sliding Window Denoising

For long-form video generation, we introduce a sliding window denoising approach that enables the synthesis of temporally coherent videos of arbitrary duration. Rather than generating the entire video at once—which would be computationally prohibitive—or generating independent segments that would result in temporal discontinuities, our method processes overlapping batches of frames while maintaining global consistency.

Given a long driving audio signal, our algorithm processes it through a multi-stage pipeline. First, the audio is preprocessed and padded to ensure clean transitions between segments. The temporal audio features are then resampled and aligned with the intended video frames, with careful handling of frame boundaries to prevent artifacts.

The core of our approach involves a progressive denoising diffusion process where a latent representation of the entire sequence is initialized and systematically denoised through a fixed number of diffusion steps. At each denoising timestep, we process overlapping batches of frames (33 frames per batch with a configurable overlap). For each batch, we predict noise conditioned on both the audio features and optional text prompts, applying classifier-free guidance separately for audio and text conditions. We implement an adaptive blending mechanism at the overlapping regions that ensures smooth transitions between consecutive batches.

This weighted blending strategy is particularly important, as it distributes the influence of each prediction according to a frame's position within the batch overlap. For frames in the overlapping regions, we compute:

$$w_{\text{orig}} = \frac{(w_{\text{overlap}} - 1) - i}{w_{\text{overlap}} - 2}, \quad w_{\text{new}} = 1 - w_{\text{orig}}$$

where i represents the frame's index position in the overlap region and w_{overlap} is the overlap width. Here, w_{orig} is the weight applied to the previously computed frame and w_{new} is the weight applied to the newly computed frame in the current batch. Our approach reduces memory requirements while maintaining global temporal coherence, addressing the "infinite context" problem in long-form video generation. The model can maintain consistent identity and motion patterns throughout videos of arbitrary duration, without the need for explicit keyframe planning or segment-level supervision.

To further accelerate inference, we implemented Sequence Parallelism for Full-Attention computations using USP [15]. This approach enables the efficient generation of a 1-minute animation (540×540 resolution, four-step sampling) within 60 seconds on 8 NVIDIA H100 GPUs.

4.2 Evaluations and Comparisons

4.2.1 Qualitative analysis

To more comprehensively assess the model's ability to animate free-style portraits and its applicability in various scenarios, we introduce a new benchmark, MagicInfinite-Benchmark. Specifically, we collect 30 portraits from different domains, including various styles(e.g. anime, sculpture and realistic styles), which are generated using text-to-image models. These images featured adults, young people, teenagers, and infants, with backgrounds ranging from indoor settings (e.g., bedrooms, living rooms, classrooms) to outdoor environments (e.g., beaches, forests, streets).

The MagicInfinite-Benchmark also includes 20 audio clips representing different speaking scenarios such as singing, speech, and rapping, as well as 20 textual prompts that describe different emotions (e.g., happiness, anger, sadness) and actions (e.g., hand raising, head shaking) during speaking. Some textual prompts included descriptions of background changes specific to certain portrait images, such as leaves rustling in the wind or waves crashing on the beach. This benchmark, with its diverse styles and scenarios, will benefit the development of the community.

We evaluate MagicInfinite on the MagicInfinite-Benchmark. As shown in Figure 1, MagicInfinite demonstrates the superior perceptual quality and motion smoothness, even after model acceleration. MagicInfinite exhibits superior domain generalization capabilities to style portraits, despite being trained exclusively on real portrait animations.

IDs	Questions
1	Which video shows the best match between the character's lip movements and the audio?
2	Which video has the character that looks most like the person in the image?
3	Which video feels the smoothest and least choppy to you?
4	Which video has character movements that seem most realistic or similar to those in movies and animations?
5	Which video has scene changes that seem most realistic or similar to those in movies and animations?

Table 1: The user study comprises a list of five questions. Participants in the user study will answer these questions based on the videos synthesized by Hallo3, SadTalker, and MagicInfinite, as well as the corresponding portrait images.

We compared MagicInfinite with prior portrait animation works, including state-of-the-art (SOTA) GAN-based methods (SadTalker [69]) and recent diffusion-based approaches (Hallo3 [12]). To elucidate the distinctions between MagicInfinite and the SOTA, we conducted a user study using the MagicInfinite-Benchmark. Specifically, for each portrait image in the benchmark, we randomly selected a textual prompt and an audio clip to form a test case. We invited 30 participants from various regions worldwide to evaluate the quality of videos synthesized by MagicInfinite and those generated by publicly available implementations of SOTA methods across all selected test cases. Each participant was asked to complete five questions, as illustrated in Tab. 1. They were instructed to select the synthesized videos that appeared more realistic and vivid. Among the 150 responses collected, **137 out of 150 (91.33%)** participants confirmed that MagicInfinite outperformed SadTalker [69] and Hallo3 [12] in terms of video quality.

Method	Sync-C (HDTF) ↑	Sync-C (Internal) ↑	Sync-D (HDTF) ↓	Sync-D (Internal) ↓
SadTalker	6.7526	4.4568	8.0753	9.9851
Hallo3	6.7997	5.6112	8.6029	9.4386
MagicInfinite	**7.2777**	**6.6943**	**7.9670**	**8.4012**

Table 2: Quantitative comparison of synchronization metrics on HDTF and internal data.

4.2.2 Quantitative analysis

We evaluate the quality of generated portrait videos under different methods using widely adopted metrics: FID [21], FVD [55] and Sync [11] (Sync-C and Sync-D). We randomly sample 100 video clips from HDTF [70] and internally collected data (which were not used in the model's training) as test videos. For each test video, we use the first frame as the static portrait image and generate the entire video, where the audio of the test video serves as the driving audio. The test video is used as the ground truth, and the textual prompt is extracted from the test video using LLAVA-34B [35], rewritten by Gemini [52], same to 4.1.1. As shown in Table 2, MagicInfinite consistently demonstrates superior image quality, motion accuracy, and lip synchronization accuracy compared to these baselines. Furthermore, with limited sampling steps and without CFG (after distillation), MagicInfinite achieves a $20\times$ speedup, outperforming both SadTalker [69] and Hallo3 [12], while delivering competitive results with the base model.

5 Conclusion

We present MagicInfinite, a novel DiT-based framework for portrait animation with precise audio-lip synchronization. It features (1) 3D Full-Attention with sliding window denoising for coherent, high-quality video synthesis; (2) a curriculum learning scheme integrating audio, text, and images for multi-modal control; (3) region-specific masks with adaptive loss for enhanced lip accuracy; and (4) distillation techniques for 20× faster inference. MagicInfinite excels in sync and animation across diverse portraits, voices, and prompts.

References

[1] DeepBrain AI. https://www.prnewswire.com/news-releases/deepbrain-ai-delivers-ai-avatar-to empower-people-with-disabilities-302026965.html. In *Online*, 2024.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[3] Ziqian Bai, Feitong Tan, Sean Fanello, Rohit Pandey, Mingsong Dou, Shichen Liu, Ping Tan, and Yinda Zhang. Efficient 3d implicit head avatar with mesh-anchored hash table blendshapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1975–1984, 2024.

[4] 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.

[5] Aggelina Chatziagapi, ShahRukh Athar, Abhinav Jain, Rohith Mysore Vijaya Kumar, Vimal Bhat, and Dimitris Samaras. Lipnerf: What is the right feature space to lip-sync a nerf. In *International Conference on Automatic Face and Gesture Recognition 2023*, 2023.

[6] Bo Chen, Shoukang Hu, Qi Chen, Chenpeng Du, Ran Yi, Yanmin Qian, and Xie Chen. Gstalker: Real-time audio-driven talking face generation via deformable gaussian splatting. *arXiv preprint arXiv:2404.19040*, 2024.

[7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[8] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024.

[9] Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time high-fidelity talking head synthesis with audio-driven 3d gaussian splatting. *arXiv preprint arXiv:2404.16012*, 2024.

[10] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024.

[11] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017.

[12] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. *arXiv preprint arXiv:2412.00733*, 2024.

[13] P Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014.

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[15] Jiarui Fang and Shangchun Zhao. A unified sequence parallelism approach for long context generative ai. *arXiv preprint arXiv:2405.07719*, 2024.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[17] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1515, 2023.

[18] Hanzhong Guo, Hongwei Yi, Daquan Zhou, Alexander William Bergman, Michael Lingelbach, and Yizhou Yu. Real-time one-step diffusion-based expressive portrait videos generation. *arXiv preprint arXiv:2412.13479*, 2024.

[19] 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. *arXiv preprint arXiv:2307.04725*, 2023.

[20] Hao-AI-Lab. Fastvideo. `https://github.com/hao-ai-lab/FastVideo/tree/main`, 2025.

[21] 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. *Advances in neural information processing systems*, 30, 2017.

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[23] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024.

[24] Esperanza Johnson, Ramón Hervás, Carlos Gutiérrez López de la Franca, Tania Mondéjar, Sergio F Ochoa, and Jesús Favela. Assessing empathy and managing emotions through interactions with an affective avatar. *Health informatics journal*, 24(2):182–193, 2018.

[25] Oytun Kal and Yavuz Samur. Educational virtual reality game design for film and animation. *Encyclopedia of Computer Graphics and Games*, pages 621–636, 2024.

[26] 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, 2023.

[27] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[28] 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.

[29] Dongze Li, Kang Zhao, Wei Wang, Bo Peng, Yingya Zhang, Jing Dong, and Tieniu Tan. Ae-nerf: Audio enhanced neural radiance field for few shot talking head synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3037–3045, 2024.

[30] Jiachen Li, Qian Long, Jian Zheng, Xiaofeng Gao, Robinson Piramuthu, Wenhu Chen, and William Yang Wang. T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. *arXiv preprint arXiv:2410.05677*, 2024.

[31] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. *arXiv preprint arXiv:2404.15264*, 2024.

[32] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7568–7578, 2023.

[33] 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, 2017.

[34] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[36] Yunfei Liu, Lijian Lin, Fei Yu, Changyin Zhou, and Yu Li. Moda: Mapping-once audio-driven portrait animation with dual attentions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23020–23029, 2023.

[37] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019, 2019.

[38] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[39] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021.

[40] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1896–1904, 2023.

[41] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

[42] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[43] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024.

[44] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20687–20697, 2023.

[45] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*.

[46] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[47] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, kigali, rwanda, May 2023. OpenReview.net.

[48] Imogen C Rehm, Emily Foenander, Klaire Wallace, Jo-Anne M Abbott, Michael Kyrios, and Neil Thomas. What role can avatars play in e-mental health interventions? exploring new models of client–therapist interaction. *Frontiers in Psychiatry*, 7:186, 2016.

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[50] Linsen Song, Wayne Wu, Chaoyou Fu, Chen Change Loy, and Ran He. Audio-driven dubbing for user generated contents via style-aware semi-parametric synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1247–1261, 2022.

[51] Jiacheng Su, Kunhong Liu, Liyan Chen, Junfeng Yao, Qingsong Liu, and Dongdong Lv. Audio-driven high-resolution seamless talking head video editing via stylegan. *arXiv preprint arXiv:2407.05577*, 2024.

[52] 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.

[53] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024.

[54] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018.

[55] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[56] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[57] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024.

[58] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662, 2023.

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

[60] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024.

[61] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023.

[62] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024.

[63] Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Jin Wang. Megactor-sigma: Unlocking flexible mixed-modal control in portrait animation with diffusion transformer. *arXiv preprint arXiv:2408.14975*, 2024.

[64] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[65] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024.

[66] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024.

[67] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6613–6623, 2024.

[68] Yuanhao Zhai, Kevin Lin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Chung-Ching Lin, David Doermann, Junsong Yuan, and Lijuan Wang. Motion consistency model: Accelerating video diffusion with disentangled motion-appearance distillation. *arXiv preprint arXiv:2406.06890*, 2024.

[69] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023.

[70] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.

[71] Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. Memo: Memory-guided diffusion for expressive talking video generation. *arXiv preprint arXiv:2412.04448*, 2024.

[72] Mingyuan Zhou, Zhendong Wang, Huangjie Zheng, and Hai Huang. Long and short guidance in score identity distillation for one-step text-to-image generation. *ArXiv 2406.01561*, 2024.

[73] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024.

[74] Yixiang Zhuang, Baoping Cheng, Yao Cheng, Yuntao Jin, Renshuai Liu, Chengyang Li, Xuan Cheng, Jing Liao, and Juncong Lin. Learn2talk: 3d talking face learns from 2d talking face. *arXiv preprint arXiv:2404.12888*, 2024.