VoCa: Unified Autoregressive Modeling for Talking Audio-Video Generation

Zhuofan Zong[1,2], Jiale Yuan[1,2], Yufei Liu[3], Dongzhi Jiang[1], Hao Shao[1], Zimu Lu[1], Ke Wang[1], Yunqiao Yang[1], Mingjie Zhan[2] ✉, and Hongsheng Li[1,4] ✉

[1] CUHK MMLab
zongzhuofan@gmail.com
[2] SenseTime Research
[3] Shanghai Jiao Tong University
[4] CPII under InnoHK

Abstract. Talking-head generation from text typically relies on cascaded pipelines that synthesize audio before video, leading to error accumulation, latency, and audiovisual misalignment. Existing end-to-end models address some of these issues but often lack autoregressive capabilities. In this paper, we present `VoCa`, a unified autoregressive framework that jointly generates speech and talking-head video from text transcripts and a reference image. `VoCa` employs a large language model (LLM) to map text into shared representations, which are concurrently processed by dedicated speech and video decoders to ensure strict synchronization. To bridge the semantic gap between audio-optimized features and human motion, we introduce a speech refiner with a tailored multi-stage training strategy. Furthermore, to enable step-by-step autoregressive co-generation, we propose a unified window partitioning mechanism and an associated improved rolling-window denoising strategy. This adapts a bidirectional diffusion transformer into a causal architecture synchronized with the LLM's autoregressive generation windows. Extensive experiments show that `VoCa` matches or surpasses state-of-the-art baselines in visual quality, improves audio–lip synchronization, and identity preservation. Code and models will be released upon acceptance.

Keywords: Unified Autoregressive Modeling · Talking-Head Generation · Audio-Video Generation

1 Introduction

Talking-head generation [3,14,19,26,38,45] is a core capability for human–computer interaction, telepresence, dubbing, and accessible media production. Conventional talking-head systems are audio-driven: given a reference identity and an input audio track, they animate facial and human motion that follows the acoustics. When no speech audio is available, these systems fall back to a cascaded

✉ Corresponding authors.

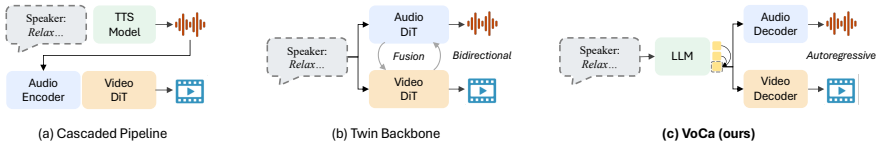

Fig. 1: We compare VoCa with other talking audio-video generation paradigms, including cascaded and end-to-end twin-backbone designs [24, 46]. Cascaded pipelines suffer from error accumulation and latency, while twin-backbone models require training a separate audio branch from scratch and still depend on an external LLM for interactive use. In contrast, VoCa natively integrates a LLM to autoregressively co-generate audio and video, enabling streaming generation while directly inheriting LLM capabilities such as voice cloning.

pipeline that first synthesizes speech from a transcription with text-to-speech (TTS) model and then uses the generated audio to drive a talking head. This cascade compounds errors, introduces extra latency, and often yields misalignment between transcription semantics and human motion because the generative model conditions only on acoustics. End-to-end formulations can address these issues by learning a single mapping from text to synchronized speech and video, yet remain comparatively underexplored. A few recent end-to-end systems [1, 24, 46, 55] generate talking audio-video, but they typically adopt a twin-backbone design that trains a dedicated audio branch entirely from scratch, incurring substantial data and compute cost. Moreover, since these models lack a built-in language model, extending them to real-world interactive applications still requires bolting on an external LLM, effectively reintroducing the very cascade they aim to replace.

In this paper, we present VoCa, a unified autoregressive model that jointly generates speech and talking-head video from text transcripts and reference images. Unlike cascaded pipelines and twin-backbone architectures, VoCa employs a large language model (LLM) to autoregressively map text into shared representations, which are then concurrently decoded by elaborate speech decoder and video decoder to mitigate error accumulation, latency, and audiovisual misalignment. By decoding from a shared autoregressive state, this design ensures that both modalities are jointly synthesized at every step, preserving strict audiovisual synchronization.

Realizing this unified generation paradigm presents two primary challenges: (1) translating the LLM's shared representations into synchronized, high-fidelity audio and natural human motion, and (2) orchestrating these components into a unified framework capable of step-by-step, autoregressive co-generation akin to LLM generation. To address the first challenge, we introduce novel architectural designs and a tailored training paradigm. Specifically, the video decoder in VoCa comprises a speech refiner and a pretrained diffusion transformer. The speech refiner is designed to bridge the semantic gap between the LLM's features optimized purely for audio synthesis and the audio-understanding signals required to drive nuanced lip motions and co-speech gestures. To further enhance this

alignment, we employ an auxiliary audio encoder and elaborate training strategies. By conditioning both the speech and video decoders on the LLM's shared representations, VoCa precisely aligns human motion with both *what* is said and *how* it is spoken. This deep alignment constitutes a core strength of VoCa, enabling the generation of highly lifelike and contextually accurate talking-head videos. To tackle the second challenge, we introduce a unified window partitioning mechanism and adapt the bidirectional diffusion transformer into a causal architecture. During generation, the LLM's output representations and video decoder's video latents are partitioned into two rolling windows: a history window and an active window. The active window encompasses the most recently generated representations or latents. For the video branch, we adopt the rolling-window denoising and distillation training introduced in Rolling Forcing [22]. Combining this strategy with our mechanism, we maintain a rolling key-value cache of past video frames and slide a fixed-size denoising window across the sequence. To ensure synchronized generation, the denoising window's duration strictly matches the unified active window. Within this autoregressive inference process, we regard the reference image and several initial frames as the global temporal context, while latents from the most recently generated window provide a local temporal context. Frames in the active denoising window only attend to themselves, temporal contexts, and the representations in the LLM's active window. On popular benchmarks, VoCa delivers higher visual quality and tighter audio–lip synchronization than state-of-the-art methods. We will release code and models upon acceptance.

The **contributions** of this work are: **(1)** We introduce VoCa, a unified autoregressive framework that jointly generates speech and talking-head video from text transcripts and a reference image. **(2)** We design a speech refiner and a tailored training strategy to bridge the semantic gap, precisely aligning nuanced human motion with the generated audio. **(3)** We introduce a unified window partitioning mechanism and an associated improved rolling-window strategy, adapting a bidirectional diffusion transformer into a causal architecture for synchronized autoregressive co-generation. **(4)** On popular benchmarks, VoCa surpasses state-of-the-art methods in visual quality and audio-lip synchronization.

2 Related Work

2.1 Joint Audio-Video Generation

Early diffusion work [31] shows the ability to generate audio and video together, but suffers from limited generalization. Veo3 [8] raised the bar with short high quality text to audio-video generation results and a few projects [21, 43] in the community have attempted to replicate its performance. UniVerse-1 [43] combines a pretrained text to video diffuser with a separate music model using cross modal adapters and an alignment loss. OmniTalker [46] shows end-to-end text-driven speech and video with diffusion and achieves real-time performance. Ovi [24] adopts identical diffusion transformer backbones for audio and for video and connects them with bidirectional cross attention in every layer,

which improves synchronization but requires training from scratch. In contrast, our method connects a frozen large language model (LLM) to a video diffusion model, where the LLM transfers in-context semantic and acoustic cues into the video generator, ensuring consistent avatar generation.

2.2 Talking-Head Generation

Talking-head methods [4, 13, 36, 39, 45] animate a static portrait into a talking video given a speech audio clip via a diffusion framework [15–17, 25, 33–35, 58, 59]. The common pipeline first encodes the waveform into audio features with a separate audio encoder. A diffusion model then conditions on these features and a reference image to synthesize the speaker's frames. For example, Hallo [48] follows an end-to-end diffusion paradigm for audio-driven portrait animation, and the latest Hallo3 [6] further introduces a video diffusion transformer to better handle dynamic and complex backgrounds. This design requires a real audio track and does not work when only textual transcript is available. A text-to-speech model can supply the audio, but the extra step often adds latency and causes misalignment between voice and motion. Our unified autoregressive framework takes a transcript as input and simultaneously generates the spoken audio and the talking-head video, yields more consistent speaker motion and supports streaming generation for interactive scenarios.

2.3 Autoregressive Video Generation

Recent works [2, 12, 22, 32, 52] have adopted the autoregressive paradigm to address the challenges of long video generation and interactive generation. In this approach, video frames or chunks are generated sequentially. While effective, a critical drawback of purely sequential generation is its susceptibility to error accumulation and drift. To combat this, Rolling Diffusion [32] and related works [22, 37] employ a "rolling" or sliding-window approach that relaxes strict frame-by-frame causality by jointly denoising a small window of adjacent frames at progressively increasing noise levels. Inspired by the success of this rolling paradigm, we extend the approach to the task of talking-head synthesis. Our framework achieves robust joint audio-video generation, enabling long-duration and interactive performance suitable for real-world scenarios.

3 Architecture

3.1 Overall Framework

`VoCa` is a unified autoregressive text-to-audio-video generation model comprising three core components: (1) a large language model (LLM) that autoregressively maps input text into shared representations that comprises discrete audio tokens and hidden states, (2) a speech decoder that transforms these representations into a speech waveform, and (3) a video decoder, consisting of a speech refiner and a pretrained diffusion transformer, that produces the talking-head videos.

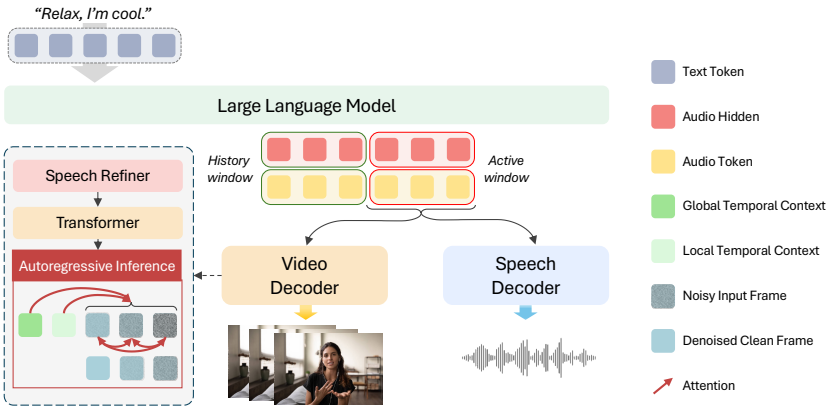

Fig. 2: Architecture of VoCa.

During inference, VoCa converts an input transcription and a reference image into synchronized speech and video through multiple autoregressive steps. As the transcription is fed, the LLM continuously generates shared representations. Once the LLM generates a new audio token, it and its corresponding hidden state are fed into the unified active window. This rolling window's length is fixed, as new representations enter, the oldest representations are eliminated from the active window and placed into the history window. The shared representations within the active window are simultaneously fed into the speech decoder and the video decoder for generation. Within the video branch, the speech refiner first refines the active window's representations to bridge the semantic gap. The causal diffusion transformer also maintains an active denoising window, which consists of adjacent frames that are assigned progressively higher noise levels. The unified active window in both the LLM and the video decoder are strictly time-aligned. To be specific, at each denoising step, it denoises the active window's latents conditioned on the shared representations and attends to KV cache of both global and local contexts to preserve temporal coherence. After a few diffusion denoising steps, the newest video frame in the window is generated and eliminated from the active denoising window to become part of the temporal context. In parallel, the speech decoder synthesizes the speech waveform from the shared representations. Because they decode from a shared autoregressive state, audio and video are produced concurrently and remain strictly locked in time. This end-to-end pipeline avoids the need for intermediate text-to-speech and audio re-encoding stages, reducing error accumulation and ensuring that visual motion is intrinsically aligned with the spoken content.

3.2 LLM and Speech Decoder

We initialize the LLM and speech decoder using the publicly available Higgs-Audio-v2 [5] foundation model. The LLM is built upon a 3B-parameter Llama-

[5] https://github.com/boson-ai/higgs-audio/

3.2 [9] backbone, augmented with a lightweight Dual-FFN architecture. This Dual-FFN design enhances the LLM's ability to model acoustic tokens with minimal computational overhead, leveraging its deep language and acoustic understanding to achieve highly expressive generation. During inference, the LLM autoregressively predicts discrete audio tokens that capture fine-grained rhythm, prosody, and articulation cues. These tokens ensure that the generated lip motions are strictly synchronized with the transcript, while the corresponding hidden states encapsulate higher-level semantic features to drive natural nonverbal behaviors, such as emotion, co-speech gestures, and head motion.

To generate the final speech waveform, we employ a speech decoder based on the DAC [18]. Utilizing an improved RVQGAN for high-fidelity audio compression, this decoder seamlessly converts the LLM's predicted discrete audio tokens into a continuous, high-quality time-domain waveform. During inference, the speech decoder offers flexible generation modes: it can either decode the complete sequence of the LLM's audio tokens in a single pass for better quality, or operate in a streaming fashion by decoding exclusively the tokens within the unified active window, thereby maintaining strict parallel synchronization with the video generation process.

3.3 Video Decoder

The video decoder transforms the LLM's shared representations into high-fidelity, temporally coherent talking-head videos under a causal conditioning scheme: while the LLM predicts shared representations token by token, the video decoder consumes the same active window in parallel, ensuring that visual frames are produced in lock-step with the audio stream. To achieve this, the decoder is structured into two components: a speech refiner and a causal diffusion transformer. The speech refiner serves as a crucial bridge across modalities, because the LLM's representations are inherently optimized for audio synthesis, the refiner is necessary to extract, enhance, and temporally align the semantic and acoustic cues required to drive accurate lip synchronization and natural co-speech gestures. Subsequently, the diffusion transformer synthesizes the talking video in an autoregressive manner. Conditioned on the refiner's aligned outputs, this causal architecture ensures that the generated visual motion is both highly realistic and strictly synchronized with the concurrent speech.

Speech Refiner. We begin by aligning temporal rates through linear interpolation of the LLM representations onto the video frame grid. To strengthen audio understanding, the token sequence is then passed through a convolutional network that amplifies rhythm, stress, and other cues essential for precise lip synchronization. This network consists of a stack of one-dimensional convolutional residual blocks. The output of this network is supervised by an auxiliary objective that distills semantic cues from a self-supervised speech model to compensate for the reconstruction-oriented nature of the LLM's audio tokens (described in Sec. 4). Next, the enhanced token features are concatenated with the

corresponding audio hidden states, which provide utterance-level context and acoustics for improved semantics-to-motion alignment. Then, a temporal compression module aggregates the concatenated sequence and downsamples it to match the temporal shape of the video latents. It comprises two temporal convolution layers with stride two, which progressively reduce the sequence length. A multilayer perceptron (MLP) then projects the compressed representation into the diffusion conditioning space. The i-th aligned conditioning vector is applied when denoising the i-th video latent, ensuring temporal alignment between audio and motion.

Diffusion Transformer. We build on Wan's [41] architecture, which comprises a VAE, a umT5 [49] text encoder, an image encoder, and a 3D transformer, and extend it with causal and cross-attention mechanisms suited to autoregressive talking-head generation. The computation is performed in the VAE latent space. In our implementation, each transformer block applies causal self-attention over the current frame and cached past frames, followed by cross-attention to the text prompt embeddings and the reference image embeddings, then a feed-forward network. Residual connections and normalization are used.

Audio Attention. We employ audio attention to drive lip movement and co-speech gestures from the provided audio conditioning. Specifically, the video latents are reshaped into a per-frame token grid, and the cross-attention operation is strictly constrained: the spatial tokens of any given frame attend exclusively to the speech refiner's outputs at the exact corresponding time index. This localized attention strategy serves a dual purpose. First, it prevents information leakage from future audio, preserving the causal nature of the generation and yielding a highly natural, deterministic alignment between acoustic cues and visual motion. Second, by restricting the cross-attention scope to strictly time-aligned features rather than the entire audio sequence, this design significantly reduces the computational complexity and memory overhead of the attention mechanism, ensuring highly efficient generation.

Identity Attention. In addition to audio, the transformer conditions on the reference image representations to ensure the generated video retains the subject's appearance and facial geometry. We encode the reference image through two feature extractors: a DINOv2 [27] vision encoder to capture holistic appearance (e.g. clothing and overall look) and an InsightFace [30] model to capture detailed facial geometry and identity-specific features. These two types of identity features are injected into the transformer blocks via decoupled identity cross-attention layers. In each transformer block, we perform two cross-attention operations from the frame's latent tokens and outputs of these two attentions are then added together. By summing their effects, we preserve both the high-level look and the precise facial identity of the character throughout the video.

4 Training Recipe

The training of VoCa is structured into three distinct stages to progressively align the modalities, enable joint autoregressive generation, and accelerate inference.

4.1 Alignment Pretraining

The first stage optimizes the video decoder while keeping the large language model (LLM) and the speech decoder frozen. The primary objective of this stage is to bridge the semantic and modality gap between the LLM's audio-centric representations and the visual motion space. We adopt the Flow Matching [20] training objective to train the video decoder. Given a clean video latent x_0 and Gaussian noise $x_1 \sim \mathcal{N}(0, I)$, we sample a timestep $t \in [0, 1]$ and construct the interpolated noisy latent $x_t = (1-t)x_0 + tx_1$. The network is trained to predict the velocity vector $v = x_1 - x_0$, conditioned on the LLM's shared representations c_{llm} and the reference image features c_{ref}. The Flow Matching loss is defined as:

$$\mathcal{L}_{flow} = \mathbb{E}_{t,x_0,x_1,c_{llm},c_{ref}} \left[\| v_\theta(x_t, c_{llm}, c_{ref}, t) - (x_1 - x_0) \|_2^2 \right]$$

To strengthen the audio-visual semantic alignment, we introduce an auxiliary distillation loss applied to the intermediate tokens of the speech refiner. The motivation is that the LLM's discrete audio tokens are primarily optimized for audio waveform reconstruction during tokenizer training, so they encode rich acoustic details that benefit audio quality but carry relatively little semantic information. This imbalance limits precise audio–visual alignment in the video branch. To bridge this gap, while the Flow Matching loss provides global visual supervision, our auxiliary loss explicitly distills latent representations from HuBERT [11] into the speech refiner's output tokens $h_{refiner}$, encouraging them to capture the semantic and temporal cues necessary for precise audio–visual alignment. Specifically, layer-wise HuBERT representations are extracted from the ground-truth audio and average-pooled to provide robust phonetic and linguistic supervision. We minimize the mean squared error between these pooled HuBERT features h_{target} and the refiner's outputs $h_{refiner}$:

$$\mathcal{L}_{aux} = \mathbb{E} \left[\| h_{refiner} - h_{target} \|_2^2 \right] \tag{1}$$

The final pretraining objective combines these two terms:

$$\mathcal{L}_{pre} = \mathcal{L}_{flow} + \lambda_1 \mathcal{L}_{aux}$$

where λ_1 is a hyperparameter balancing the visual generation and auxiliary alignment losses. Following this stage, the video decoder is capable of producing lip-synced video frames conditioned on the LLM's shared representations.

4.2 Joint Finetuning

The second stage involves joint finetuning. Here, we optimize both the LLM and the video decoder while keeping the speech decoder frozen. This stage aims to: (1) achieve strong content consistency between the jointly generated audio and video streams, and (2) ensure that the LLM's shared representations capture both the acoustic details and the necessary semantic cues required to drive nuanced facial dynamics in the video decoder.

The overall loss in this stage combines the visual Flow Matching loss $\mathcal{L}_{flow}$ with an audio generation loss. The audio loss is formulated as a standard autoregressive cross-entropy objective, which forces the LLM to predict the discrete audio tokens a_i given the text prompt and previous tokens:

$$\mathcal{L}_{ce} = -\mathbb{E}\left[\sum_i \log P_\phi(a_i \mid a_{<i}, \text{text})\right]$$

To enhance audio generation quality, we apply the delay pattern to the LLM's token prediction during training. For the video decoder, we employ teacher forcing and apply a causal attention mask across all self-attention layers. Specifically, we partition the whole frame sequence into multiple non-overlapping chunks and each chunk consists of several consecutive frames. The causal mask ensures that frames within the current chunk can only attend to themselves and frames in the preceding history chunks, strictly preventing future information leakage. The combined objective for this stage is:

$$\mathcal{L}_{ar} = \mathcal{L}_{flow} + \lambda_2 \mathcal{L}_{ce}$$

where λ_2 controls the relative weight of the autoregressive audio loss.

4.3 Distribution Matching Distillation

The third stage distills the multi-step video decoder into a few-step causal generator, so that video chunks can be emitted at the same autoregressive rate as the LLM audio stream. We follow the rolling-window distillation strategy of Rolling Forcing [22], and adapt it to image-conditioned talking-head generation. During the student-generator update, the LLM and speech decoder remain frozen and gradients are applied only to the video decoder.

During rollout, the noisy latent sequence is partitioned into non-overlapping frame chunks $\{B_i\}_{i=1}^{N}$ of K latent frames each. The reference image is encoded by the VAE and inserted as an initial chunk B_0 with timestep 0. This clean chunk is cached before generation and excluded from the distillation gradient. Given Gaussian noise z, the student generator G_θ is unrolled with the causal sampler used at inference. At each rolling step, an active window of up to M adjacent chunks is assigned progressively higher noise levels from the earliest to the newest chunk following the few-step denoising schedule. The newest chunk enters the window from pure noise, older chunks are reused from the noisy cache,

and the earliest chunk is finalized and written to the clean KV cache. The DiT self-attention attends to three sources of context: the active denoising window, a local history cache of previously generated clean chunks, and a global context cache. Unlike the original Rolling Forcing setting, our global context is image-conditioned: it is initialized with B_0 and expanded to include both B_0 and B_1 once the first generated chunk is clean-cached, anchoring identity and appearance to the reference image while using the first motion chunk as a long-range temporal anchor. To avoid full backpropagation through the rollout, we sample a window phase and retain the computation graph only for windows of that phase within the gradient horizon; all other windows and cache updates run under stop-gradient.

We apply Distribution Matching Distillation (DMD) [51] on the student outputs. For a generated latent sequence $x = G_\theta(z, c)$, where $c = \{c_{llm}, c_{ref}\}$ denotes the LLM audio representations and reference conditioning, we sample a score timestep τ, perturb x to x_τ, and contrast a frozen real-score model s_{real} (initialized from the fully trained multi-step decoder) with a trainable fake-score model s_{fake} fit on the student distribution. The loss is implemented in its stop-gradient form:

$$\mathcal{L}_{DMD} = \frac{1}{2} \left\| x - \text{sg}\big[x - w(\tau)\big(s_{\text{fake}}(x_\tau, \tau, c) - s_{\text{real}}(x_\tau, \tau, c)\big)\big] \right\|_2^2, \tag{2}$$

where $w(\tau)$ is the DMD normalization term. We mask out the clean context chunks so the student is optimized only on newly generated motion latents. The fake-score network is updated in alternation on the same student samples with the standard denoising objective, while the real-score network remains frozen. This distillation preserves the causal rolling-window behavior from the previous stage, reduces the number of denoising evaluations per chunk, and enables `VoCa` to synthesize speech and video in a unified streaming autoregressive pipeline.

5 Experiments

5.1 Implementation Details

Training. Our training corpus combines Hallo3 [6], Casual Conversations v2 (CCv2) [28], CelebV-Text [53], MEAD [44], and SpeakerVid [55]. We exclude clips with multiple speakers because the method only focuses on single-speaker scenarios. Video captions are generated with Qwen2.5-VL-7B [50], and speech transcriptions are obtained with Whisper-Large-v3 [29]. We instantiate the video decoder with Wan2.1-1.3B [41] and Wan2.2-5B. Models are trained on 64 NVIDIA H800 GPUs with a global batch size of 64. During training we employ a multi-scale strategy. Frames are sampled at resolutions up to 832 $\times$ 480 pixels, and clip durations are drawn uniformly between 2 and 7 seconds. The loss coefficient λ_1 is set 1 and λ_2 is set 0.25. To improve robustness, individual conditioning inputs are randomly dropped with probability 0.05, and all conditioning signals are simultaneously dropped with probability 0.05. For distribution matching distillation, we follow the settings of the prior work [22]. We use a unified active window of 5 chunks, and each frame chunk contains 3 latent frames.

Table 1: Quantitative comparisons with talking-head generation baselines. The best score is in blue, with the second-best score in green.

Methods	HDTF				CelebV-HQ				RAVDESS			
	IDC↑	Sync-C↑	FID↓	FVD↓	IDC↑	Sync-C↑	FID↓	FVD↓	IDC↑	Sync-C↑	FID↓	FVD↓
Diffusion Models												
Hallo [48]	0.981	7.312	35.830	37.060	0.934	4.417	48.410	122.510	0.948	3.536	27.700	418.980
AniPortrait [47]	0.978	3.607	36.100	71.270	0.938	1.804	50.000	110.600	0.935	2.364	23.800	505.040
DreamTalk [54]	0.874	5.777	87.150	207.660	0.805	3.804	93.520	288.460	0.859	4.464	109.370	224.000
V-Express [42]	0.972	7.744	40.990	122.270	0.878	4.029	61.330	180.110	0.939	4.444	29.100	601.920
EchoMimic [4]	0.935	5.858	54.270	113.860	0.896	3.284	58.030	135.330	0.893	4.174	73.240	158.060
Hallo3 [6]	0.976	6.898	35.550	42.850	0.930	4.173	44.750	91.410	0.949	4.948	24.520	54.330
EchoMimicV3 [26]	0.976	2.573	35.570	58.380	0.930	2.134	49.880	92.540	0.965	2.494	15.090	44.360
FantasyTalking [45]	0.969	3.836	37.360	46.600	0.930	2.797	47.220	106.810	0.952	3.803	29.350	57.820
HunyuanAvatar [3]	0.970	7.584	35.570	106.270	0.915	4.494	46.550	97.800	0.957	5.214	19.760	99.490
OmniAvatar [7]	0.959	3.534	37.710	119.470	0.910	2.428	51.960	93.580	0.952	2.893	23.510	84.890
Autoregressive Models												
VoCa-1.3B	0.981	6.082	33.620	39.630	0.927	3.620	48.040	95.870	0.970	4.585	14.070	55.190
VoCa	0.982	8.054	33.020	36.930	0.936	4.769	44.480	91.170	0.970	5.636	13.160	48.710

Evaluation. We evaluate VoCa across two primary tasks: audio-driven video generation and text-driven audio-video generation. For the audio-driven task, we conduct evaluations on the HDTF [56], CelebV-HQ [57], and RAVDESS [23] datasets, randomly sampling 300 videos from each. Visual fidelity is assessed using Fréchet Inception Distance (FID) [10] and Fréchet Video Distance (FVD) [40]. While FID measures the spatial distributional gap between generated and real individual frames, FVD extends this comparison to the temporal domain to evaluate sequence-level coherence. Additionally, we report Identity Consistency (IDC) following the protocol established in FantasyTalking [45]. To measure audio-visual alignment, we compute the SyncNet-based Sync-C score [5], which quantifies the synchronization between the generated lip movements and the driving speech. For audio-video generation, we adopt the Verse-Bench [43] evaluation framework, strictly following its original experimental settings. Finally, we conduct a user study to subjectively compare VoCa against existing methods.

5.2 Quantitative Results

Audio-Driven Generation. We compare VoCa with state-of-the-art audio-driven methods in Tab. 1. For fast and fair evaluation, we use a fixed prompt "the person is talking" and the ground-truth audio track as input reference audio for all samples. This plain text prompt creates a mismatch between inference and training because our training captions are more descriptive, which can reduce performance at test time. Even under this setting, VoCa attains the best IDC, FID, and FVD scores, indicating strong identity preservation, visual fidelity, and temporal coherence. VoCa with Wan2.2-5B also surpasses FantasyTalking with a stronger base model Wan2.1-14B. These results suggest that an autoregressive end-to-end design can match or exceed the state-of-the-art specialist models.

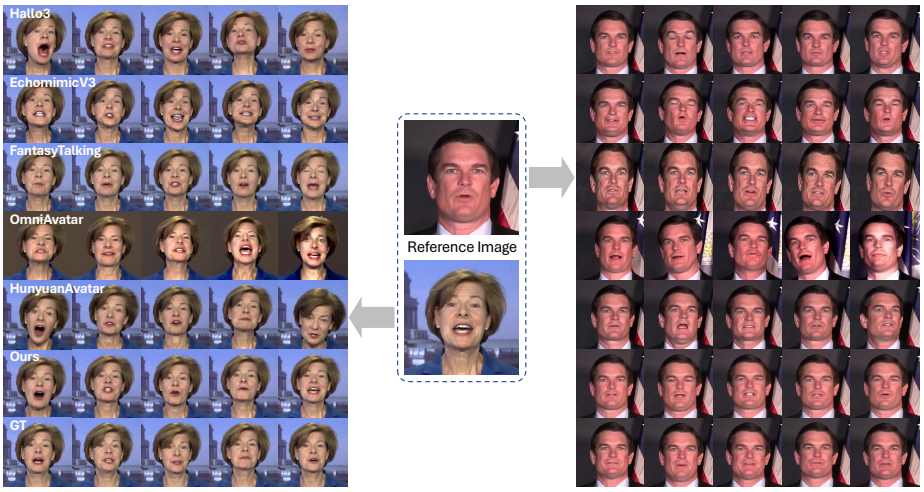

Fig. 3: Comparison of talking-face generation results.

Fig. 4: Characters with dynamic background.

Text-Driven Audio-Visual Generation. We also evaluate the audio-visual generation performance of `VoCa` against several competitive counterparts on the Verse-Bench. The detailed results are available in the Supplementary Materials.

User Study. To subjectively evaluate the naturalness of the generated co-speech human motion, we conducted a user study involving 40 participants. On a rating scale from 0 to 5, our proposed method, `VoCa`, achieved the highest average score of 4.46. This result demonstrates a clear user preference over existing baselines, outperforming EchoMimicV3 (4.02), Hallo3 (3.64), and FantasyTalking (3.45), thereby highlighting `VoCa`'s superior capability in synthesizing lifelike human dynamics.

5.3 Qualitative Results

Talking Face Generation. As shown in Fig. 3, we qualitatively compare `VoCa` with five state-of-the-art methods: Hallo3, EchoMimicV3, FantasyTalking, Om-

Fig. 5: Half-body animation.

Fig. 6: Diverse character styles.

niAvatar, and HunyuanAvatar. Benefiting from its unified framework and the speech refiner that aligns semantic and acoustic cues, `VoCa` produces coherent facial motion tightly synchronized with generated speech. Unlike cascaded pipelines that often exhibit lip jitter or motion drift, `VoCa` maintains stable head movement and natural co-speech gestures while preserving the subject's identity and expression consistency across frames. The results highlight the model's strong audiovisual alignment and temporal stability in open-domain settings.

Character with Dynamic Background. As shown in Fig. 4, `VoCa` remains robust under dynamic backgrounds, including forest fires, lightning, and rainy scenes with umbrellas. The model preserves the identity of the character while avoiding background bleeding or temporal flickering, and maintains stable human motion even when the background content changes significantly over time.

Half-body Animation. Fig. 5 illustrates half-body animation results produced by `VoCa`. In addition to accurate audio–lip synchronization, the model generates coherent upper-body and shoulder motion, as well as co-speech hand gestures that correlate with the spoken content and prosodic emphasis. The resulting motion appears fluid and temporally consistent, rather than restricted to rigid head-only movement.

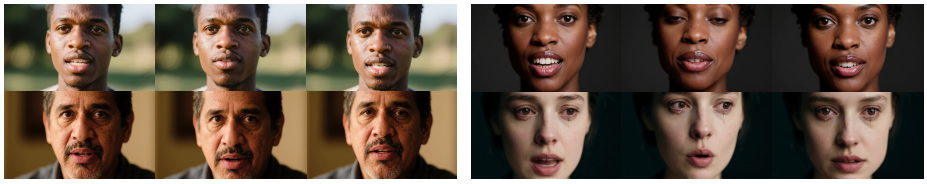

Fig. 7: High-fidelity close-up animation.

Fig. 8: Head-to-head comparisons.

Diverse Character Styles. As depicted in Fig. 6, `VoCa` generalizes to diverse character styles, including sketch portraits, cosmic-themed characters, cartoon boys, and non-human cartoon avatars. Across these styles, the model preserves the original artistic appearance while still producing plausible lip motion and head dynamics that remain synchronized with the generated audio.

High-Fidelity Close-ups. As demonstrated in Fig. 7, `VoCa` exhibits robust performance in extreme close-up scenarios. Our method successfully preserves and synthesizes high-fidelity fine details, including beard, skin textures, and realistic teeth structures during speech. Furthermore, the model accurately captures subtle micro-expressions and facial muscle dynamics.

Head-to-Head Comparisons. We conduct head-to-head comparisons in Fig. 8 between `VoCa` and leading methods including HunyuanAvatar, EchoMimicV3, and OmniAvatar. While `VoCa` robustly maintains dynamic backgrounds and realistic human-object interactions, the competing methods suffer from inaccurate lip-sync, spatial distortions, and a lack of co-speech gestures, respectively.

5.4 Ablation Studies

Shared Representations of LLM. We ablate the LLM's shared representations by comparing three variants: using only LLM audio tokens, only LLM hidden states, and both in Tab. 2. Using only tokens lacks contextual semantics, leading to less expressive nonverbal behavior. Using only hidden states provides rich semantic context but loses detailed acoustic cues, causing severe audio–lip misalignment. Leveraging both tokens and hidden states yields the best overall performance.

Table 2: Shared representations of LLM.

Method	IDC↑	Sync-C↑	FID↓	FVD↓
Token	0.970	5.799	37.066	59.961
Hidden	0.973	3.773	36.124	56.256
Both	0.981	6.082	33.620	39.630

Table 3: Auxiliary distillation loss.

Method	IDC↑	Sync-C↑	FID↓	FVD↓
$\lambda_1 = 0$	0.975	4.782	34.808	45.082
$\lambda_1 = 1$	0.981	6.082	33.620	39.630
$\lambda_1 = 2$	0.963	6.003	36.916	64.509

Table 4: Refiner depth.

Method	IDC↑	Sync-C↑	FID↓	FVD↓
0 layers	0.973	4.754	34.996	47.177
2 layers	0.969	5.743	35.322	54.948
4 layers	0.976	5.919	34.915	46.671
6 layers	0.981	6.082	33.620	39.630
8 layers	0.982	6.054	33.506	39.269

Auxiliary Distillation Loss. We ablate the auxiliary distillation loss weight and report its effect on generation performance in Tab. 3. The baseline setting achieves the best overall performance, yielding better visual quality and more accurate lip motion. Without the loss, we observe degraded lip–audio alignment, while an overly large weight slightly harms visual fidelity. This suggests our default choice is effective and does not require elaborate tuning.

Speech Refiner Design. We ablate the depth of the convolutional network in the speech refiner in Tab. 4. We compare the baseline configuration (6 layers) with shallower and deeper variants using 0, 2, 4, and 8 layers. Reducing the number of layers leads to progressively worse audio representations, reflected in lower audio quality and less precise audio–video alignment. The 6-layer refiner achieves the best overall performance, suggesting that a deeper refinement network is beneficial for capturing richer acoustic cues that are important for talking-head generation.

6 Conclusion

We have presented `VoCa`, a unified autoregressive framework that jointly generates synchronized speech and high-fidelity talking-head video directly from text transcripts and reference images. Unlike cascaded pipelines that suffer from error accumulation and bidirectional twin-backbone models, `VoCa` utilizes a large language model to autoregressively predict shared representations that concurrently drive dedicated speech and video decoders. To bridge the inherent semantic gap between audio-optimized LLM outputs and visual motion, we introduced a speech refiner and tailored training strategies, ensuring that the generated human motion aligns precisely with both the linguistic content and the acoustic realization of the utterance. Furthermore, we proposed a unified window partitioning mechanism coupled with a rolling-window denoising strategy, effectively adapting a bidirectional diffusion transformer into a causal architecture for step-by-step, synchronized audiovisual co-generation. Evaluations on popular benchmarks demonstrate that `VoCa` surpasses state-of-the-art methods in visual quality and audio-lip synchronization, offering a highly lifelike, efficient, and contextually accurate solution for end-to-end talking-head synthesis.

References

1. Chatziagapi, A., Morency, L.P., Gong, H., Zollhöfer, M., Samaras, D., Richard, A.: Av-flow: Transforming text to audio-visual human-like interactions. arXiv preprint arXiv:2502.13133 (2025) 2
2. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2024) 4
3. Chen, Y., Liang, S., Zhou, Z., Huang, Z., Ma, Y., Tang, J., Lin, Q., Zhou, Y., Lu, Q.: Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. arXiv preprint arXiv:2505.20156 (2025) 1, 11
4. Chen, Z., Cao, J., Chen, Z., Li, Y., Ma, C.: Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2403–2410 (2025) 4, 11
5. Chung, J.S., Zisserman, A.: 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. pp. 251–263. Springer (2017) 11
6. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. arXiv preprint arXiv:2412.00733 (2024) 4, 10, 11
7. Gan, Q., Yang, R., Zhu, J., Xue, S., Hoi, S.: Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. arXiv preprint arXiv:2506.18866 (2025) 11
8. Google DeepMind: Veo: A text-to-video generation system. Tech. rep., Google (2024), https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf, accessed: 2025-09-24 3
9. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 6
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 11
11. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing **29**, 3451–3460 (2021) 8
12. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025) 4
13. Huang, Y., Guo, H., Wu, F., Wang, W., Zhang, S., Huang, S., Gan, Q., Liu, L., Zhao, S., Chen, E., et al.: Live avatar: Streaming real-time audio-driven avatar generation with infinite length. arXiv preprint arXiv:2512.04677 (2025) 4
14. Ji, X., Hu, X., Xu, Z., Zhu, J., Lin, C., He, Q., Zhang, J., Luo, D., Chen, Y., Lin, Q., et al.: Sonic: Shifting focus to global audio perception in portrait animation. arXiv preprint arXiv:2411.16331 (2024) 1
15. Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703 (2025) 4

16. Jiang, D., Song, G., Wu, X., Zhang, R., Shen, D., Zong, Z., Liu, Y., Li, H.: Comat: Aligning text-to-image diffusion model with image-to-text concept matching. arXiv preprint arXiv:2404.03653 (2024) 4
17. Jiang, D., Zhang, R., Li, H., Zong, Z., Guo, Z., He, J., Guo, C., Ye, J., Fang, R., Li, W., et al.: Draco: Draft as cot for text-to-image preview and rare concept generation. arXiv preprint arXiv:2512.05112 (2025) 4
18. Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., Kumar, K.: High-fidelity audio compression with improved rvqgan. Advances in Neural Information Processing Systems **36**, 27980–27993 (2023) 6
19. Lin, G., Jiang, J., Yang, J., Zheng, Z., Liang, C.: Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. arXiv preprint arXiv:2502.01061 (2025) 1
20. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022) 8
21. Liu, K., Li, W., Chen, L., Wu, S., Zheng, Y., Ji, J., Zhou, F., Jiang, R., Luo, J., Fei, H., et al.: Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. arXiv preprint arXiv:2503.23377 (2025) 3
22. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025) 3, 4, 9, 10
23. Livingstone, S.R., Russo, F.A.: Ravdess emotional speech audio (2019). https://doi.org/10.34740/KAGGLE/DSV/256618, https://www.kaggle.com/dsv/256618 11
24. Low, C., Wang, W., Katyal, C.: Ovi: Twin backbone cross-modal fusion for audio-video generation. arXiv preprint arXiv:2510.01284 (2025) 2, 3
25. Ma, B., Zong, Z., Song, G., Li, H., Liu, Y.: Exploring the role of large language models in prompt encoding for diffusion models. arXiv preprint arXiv:2406.11831 (2024) 4
26. Meng, R., Wang, Y., Wu, W., Zheng, R., Li, Y., Ma, C.: Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. arXiv preprint arXiv:2507.03905 (2025) 1, 11
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 7
28. Porgali, B., Albiero, V., Ryda, J., Ferrer, C.C., Hazirbas, C.: The casual conversations v2 dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10–17 (2023) 10
29. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518. PMLR (2023) 10
30. Ren, X., Lattas, A., Gecer, B., Deng, J., Ma, C., Yang, X.: Facial geometric detail recovery via implicit representation. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–8. IEEE (2023) 7
31. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q., Guo, B.: Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10219–10228 (2023) 3
32. Ruhe, D., Heek, J., Salimans, T., Hoogeboom, E.: Rolling diffusion models. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 42818–42835. PMLR (21–27 Jul 2024) 4

33. Shao, H., Liu, L., Luo, Z., Zong, Z., Li, H.: Mining real-world image relations for large-scale controllable generation and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3893–3902 (2026) 4
34. Shao, H., Wang, S., Zhou, Y., Song, G., He, D., Qin, S., Zong, Z., Ma, B., Liu, Y., Li, H.: Vividface: A diffusion-based hybrid framework for high-fidelity video face swapping. arXiv preprint arXiv:2412.11279 (2024) 4
35. Shen, D., Song, G., Zhang, Y., Ma, B., Li, L., Jiang, D., Zong, Z., Liu, Y.: Adt: Tuning diffusion models with adversarial supervision. arXiv preprint arXiv:2504.11423 (2025) 4
36. Shen, L., Qiao, Q., Yu, T., Zhou, K., Yu, T., Zhan, Y., Wang, Z., Tao, M., Yin, S., Liu, S.: SoulX-FlashTalk: Real-time infinite streaming of audio-driven avatars via self-correcting bidirectional distillation (2025). https://doi.org/10.48550/arXiv.2512.23379, https://arxiv.org/abs/2512.23379 4
37. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025) 4
38. Tian, L., Hu, S., Wang, Q., Zhang, B., Bo, L.: Emo2: End-effector guided audio-driven avatar video generation. arXiv preprint arXiv:2501.10687 (2025) 1
39. Tian, L., Wang, Q., Zhang, B., Bo, L.: 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) 4
40. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019) 11
41. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 7, 10
42. Wang, C., Tian, K., Zhang, J., Guan, Y., Luo, F., Shen, F., Jiang, Z., Gu, Q., Han, X., Yang, W.: V-express: Conditional dropout for progressive training of portrait video generation. arXiv preprint arXiv:2406.02511 (2024) 11
43. Wang, D., Zuo, W., Li, A., Chen, L.H., Liao, X., Zhou, D., Yin, Z., Dai, X., Jiang, D., Yu, G.: Universe-1: Unified audio-video generation via stitching of experts. arXiv preprint arXiv:2509.06155 (2025) 3, 11
44. Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., Loy, C.C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In: European conference on computer vision. pp. 700–717. Springer (2020) 10
45. Wang, M., Wang, Q., Jiang, F., Fan, Y., Zhang, Y., Qi, Y., Zhao, K., Xu, M.: Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. arXiv preprint arXiv:2504.04842 (2025) 1, 4, 11
46. Wang, Z., Zhang, P., Qi, J., Wang, G., Ji, C., Xu, S., Zhang, B., Bo, L.: Omnitalker: One-shot real-time text-driven talking audio-video generation with multimodal style mimicking. arXiv preprint arXiv:2504.02433 (2025) 2, 3
47. Wei, H., Yang, Z., Wang, Z.: Aniportrait: Audio-driven synthesis of photorealistic portrait animation. arXiv preprint arXiv:2403.17694 (2024) 11
48. Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., Zhu, S.: Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv preprint arXiv:2406.08801 (2024) 4, 11
49. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mt5: A massively multilingual pre-trained text-to-text transformer.

In: Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies. pp. 483–498 (2021) 7
50. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115 (2024) 10
51. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024) 10
52. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22963–22974 (2025) 4
53. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: Celebv-text: A large-scale facial text-video dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14805–14814 (2023) 10
54. Zhang, C., Wang, C., Zhang, J., Xu, H., Song, G., Xie, Y., Luo, L., Tian, Y., Guo, X., Feng, J.: Dream-talk: Diffusion-based realistic emotional audio-driven method for single image talking face generation. arXiv preprint arXiv:2312.13578 (2023) 11
55. Zhang, Y., Li, Z., Wang, D., Zhang, J., Zhou, D., Yin, Z., Dai, X., Yu, G., Li, X.: Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. arXiv preprint arXiv:2507.09862 (2025) 2, 10
56. Zhang, Z., Li, L., Ding, Y., Fan, C.: 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. pp. 3661–3670 (2021) 11
57. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: Celebv-hq: A large-scale video facial attributes dataset. In: European conference on computer vision. pp. 650–667. Springer (2022) 11
58. Zong, Z., Jiang, D., Ma, B., Song, G., Shao, H., Shen, D., Liu, Y., Li, H.: Easyref: Omni-generalized group image reference for diffusion models via multimodal llm. In: Forty-second International Conference on Machine Learning (2024) 4
59. Zong, Z., Ma, B., Shen, D., Song, G., Shao, H., Jiang, D., Li, H., Liu, Y.: Mova: Adapting mixture of vision experts to multimodal context. arXiv preprint arXiv:2404.13046 (2024) 4