

VLOGGER: Multimodal Diffusion for Embodied Avatar Synthesis

Enric Corona Andrei Zanfir Eduard Gabriel Bazavan Nikos Kolotouros
 Thiemo Alldieck Cristian Sminchisescu

Google DeepMind



Figure 1. VLOGGER is a novel framework to synthesize humans from audio. Given a single input image like the ones shown in the first column and a sample audio input, our method generates photorealistic and temporally coherent videos of the person talking and vividly moving. As seen in the synthesized images in the right columns, we generate head motion, gaze, blinking, lip movement, and, unlike previous methods, upper body and hand gestures, thus taking audio-driven synthesis one step further.

Abstract

We propose VLOGGER, a method for audio-driven human video generation from a single input image of a person, which builds on the success of recent generative diffusion models. Our method consists of 1) a stochastic human-to-3d-motion diffusion model, and 2) a novel diffusion-based architecture that augments text-to-image models with both spatial and temporal controls. This supports the generation of high quality video of variable length, easily controllable through text or speech via high-level representations of human faces and bodies. In contrast to previous work, our method does not require training for each person, does not rely on face detection and cropping, generates the complete image (not just the face or the lips), and considers a broad spectrum of scenarios (e.g. visible torso or diverse subject identities) that are critical to correctly synthesize humans who communicate. We also curate MENTOR, a new and diverse dataset with 3d pose and expression annotations, one order of magnitude larger than previous ones (800,000 identities) and with dynamic gestures, where we train and ablate our main technical contributions. VLOGGER outperforms state-of-the-art methods in three public

benchmarks, considering image quality, identity preservation and temporal consistency while also generating upper-body gestures. We analyze the performance of VLOGGER with respect to multiple diversity metrics, showing that our architectural choices and the use of MENTOR benefit training a fair and unbiased model at scale. Finally we show applications in video editing and personalization.

1. Introduction

We present VLOGGER, a method to automatically generate a video of a talking and moving person, based on text or audio, and given only a single image of that person. Industries like content creation, entertainment, or gaming all have high demand for audio-visual human synthesis. Yet, the creation of realistic videos of humans is still complex and ripe with artifacts. This requires significant manual intervention for realistic results. Full automation, however, would not only ease creative processes, but also unlock entirely new use cases, such as enhanced online communication, education, or personalized virtual assistants, to name a few. The latter is especially relevant, given recent success of chat agents [46, 55]. Research has shown that such solutions

are not perceived as sufficiently natural to develop empathy [111] and some [39] argue that anthropomorphism and behavioral realism (*e.g.* gaze, facial expressions, whole-body movements, *etc.*) are critical in creating social presence and eliciting empathetic user responses. Such features could result in the wide adoption of agents [49], in areas like customer service [1, 58], telemedicine [68], education [67], or human-robot interaction [63]. It is precisely automation and behavioral realism that we aim for in this work: VLOGGER is a multi-modal interface to an *embodied conversational agent* [80], featuring audio, complex facial expressions and gestures, designed to support natural conversations with a human user. VLOGGER can be used as a stand-alone solution for presentations, education, narration, low-bandwidth online communication, and as an interface for text-only HCI [3, 108]. In this paper, we additionally illustrate its potential for video editing tasks.

Multimodal realistic human synthesis is complex due to challenges like data acquisition, enacting natural facial expressions, image and audio synchronization, occlusion, or representing full-body movements — especially given a single input image. Many attempts focused exclusively on lip sync [59, 81, 88], by editing the mouth region of a driving video. Recently, [101, 103] relied on extensive advances in face reenactment [9, 20, 31, 53, 75, 95, 104] to generate talking heads from a single image by predicting facial motion from audio. Temporal consistency is usually achieved with a single-frame image generation network by relying on a smooth guiding motion from face keypoints. However, this might cause blurriness and does not ensure temporal coherency in areas other than the face. Consequently, most methods require detecting and cropping the head, whenever a significant part of the body is visible. For example, more recently, photorealistic methods such as [18, 54, 89, 93] are not applicable to images where the body is visible. In this paper, we argue that communication is more than “just” audio combined with lip and face motion – humans communicate using their body via gestures, gaze, blinks, or pose. MODA [43] recently started exploring the animation of face and body, however in limited scenarios, and without generalization to new identities. In contrast, we aim for a *general, person agnostic synthesis solution*, focusing on realism and diversity of motion, including both head and hand gestures. Our objective is to bridge the gap between recent video synthesis [2, 6, 38, 70], that generate dynamic videos with no control over identity or pose, and controllable image generation [9, 20, 65]. Towards that goal, we propose a two-step approach where first a generative diffusion network predicts body motion and facial expressions according to an audio input signal. This stochastic approach is necessary to model the nuanced (one-to-many) mapping between speech and pose, gaze, and expression. Second, we propose and ablate a novel image diffusion architecture, which

Audio Control	Face Control	Body Control	Stochastic	Photorealistic	Generalizes to new subjects	Can edit videos	
✗	✓	✗	✗	✓	✓	✗	Face Reenactment [75, 85]
✓	✓	✗	✓	✗	✓	✗	Audio-to-Motion [19, 74]
✓	✗	✗	✗	✓	✓	✓	Lip Sync [22, 59]
✓	✓	✗	✓	✓	✓	✗	SadTalker [103]
✓	✓	✗	✗	✓	✓	✗	Styletalk [45]
✓	✓	✓	✓	✓	✓	✓	VLOGGER (Ours)

Table 1. **Key properties of VLOGGER compared to related work.** Face reenactment [9, 20, 31, 53, 75, 95, 104] generally does not consider driving using audio or text. Works on audio-to-motion [14, 19, 62, 71, 74, 91, 98] shares components by encoding audio into 3d face motion, however lack photorealism. Lip sync [22, 59] consider input videos of different subjects, but only model mouth motion. Given their generalisation capacity, SadTalker [103] and Styletalk [45] are the closest to us, but require cropped images of faces, lack body control, and cannot edit videos.

offer spatio-temporal control. By additionally relying on generative human priors, acquired during pre-training, we show how this combined architecture improves the capacity of image diffusion models, which often struggle to generate consistent human images or video (*e.g.* eyes). VLOGGER consists of a base model followed by a super-resolution diffusion model for high quality video. We condition video generation on dense motion controls that capture the full body, including facial expressions as in previous work, but also body and hands. To generate videos of arbitrary length, we temporally outpaint and condition based on previous frames. Finally, the flexibility of VLOGGER enables editing only specific regions, like lips or the face region.

For robustness and generalization, we curate a large-scale dataset featuring a considerably larger diversity than previously available, in terms of skin tone, body pose, viewpoint, speech and body visibility. In contrast to previous work, the dataset also contains video with dynamic hand gestures, which are important in capturing more complex human communication. VLOGGER outperforms previous work across different diversity metrics, and obtains state-of-the-art image quality and diversity results on the previous HDTF [105] and TalkingHead-1KH [85] datasets. Our method considers a larger range of scenarios than baselines, by generating high resolution video of head and upper-body motion, and by featuring considerably diverse facial expressions and gestures. To summarize, the main contributions of VLOGGER are: 1) an approach to generate talking and moving humans given speech inputs; (2) leveraging a diverse, curated dataset, called MENTOR, one order of magnitude larger than existing ones, for training and testing our models; (3) outperforms SOTA in large quantitative comparisons on three public benchmarks; (4) Diversity analysis showing low bias and outperforming baselines on different perceived human attributes.

2. Related Work

Audio-Driven Talking Face Generation. There has been significant work in talking face generation, which can be categorized according to the driving inputs, intermediate representations and output formats. We provide a high-level overview and positioning with respect to our work in Tab. 1. There exists a body of work in animation of 3D morphable faces [14, 19, 62, 71, 74, 91] or full bodies [42, 98] based on audio segments. These can generate diverse 3d talking heads with temporally coherent pose and shape parameters of various statistical head or body models [5, 7, 40, 57, 92]. We consider a similar network to guide the motion but, in this paper, we instead aim to generate photorealistic talking humans with diversity in expression and head or body motion, given a single image (real or synthesized) of a target subject. We consider challenges such as temporal consistency, subject diversity, hair coverage, and gaze detail in the production of the output video.

For images and video, incipient work has focused on mouth editing [11, 13, 33, 59, 79, 105], by predicting lip motion synchronized with the input audio. Follow up work added extended features such as head motion, gaze and blinking [34, 44, 61, 73, 106, 110], using intermediate 2d or 3d landmarks, or flow based representations. To increase photorealism, many methods have extensively used discriminator losses [8, 9, 17, 60, 86, 100], and some recent methods relied on diffusion models [54, 71, 72, 93, 101]. However, it is hard to ensure proper disentanglement between body, head motions, blinking, gaze and facial expressions when operating in the latent space of GANs [21, 36] or generic diffusion models. Our method does not have to rely on custom perceptual losses for gaze, identity preservation or lip syncing. Body motion and gestures have not been considered because of lack of data and the difficulty of generating coherent video. We curate a large-scale dataset and propose a complete pipeline to address this problem. VLOGGER can generate coherent face and upper-body motion with a variety of expressions, head and body motion, gaze, eye blinking and accurate lip movement. Moreover, we show that our method is more expressive and robust across different diversity axes.

Face Reenactment. Video-based talking face generation aims to transfer the motion of a source video to a target person, and has been widely explored in the past [9, 18, 25, 30, 31, 53, 75, 87, 89, 95, 104, 107, 109]. Most methods rely on an intermediate representation, such as sparse or dense landmarks, semantic masks, dense 3d representations or warped features. In the 3d domain, several have taken advantage of NeRF [4, 47] based solutions [23, 41, 96, 97]. However, this requires a significant number of image samples of a target person talking, to support training and animation. This task is closely related to ours, and some previous work adapt intermediate representations

for audio input. We aim to extend controllable video synthesis towards the full body with *e.g.* gestures, hair, or eye movement coverage.

Video Generation. Also related to our work is generic video synthesis. This task has been widely explored recently, hence we here cover the most related directions. With the success of text-to-image diffusion models [16], many have also explored extensions to video [2, 6, 24, 26, 28, 37, 38, 52, 64, 70, 78, 90]. However, such methods are still limited in the length of the video they can generate, resolution, or fine grain control, which would be necessary *e.g.* for audio-visual human synthesis. We extend the current state-of-the-art image diffusion models to the temporal domain by adding spatio-temporal control and propose an iterative out-painting procedure to generate video of variable length. While concurrent work explore similar network architectures [2, 70] for more generic scenarios, our goal is fine grained control for the realistic animation of talking humans. We condition each frame with 1) dense motion controls resulting from rendering a posed 3D human body model, and 2) warped reference images. These inputs make the generative process both controllable and more stable, as ablated in the experimental section.

3. Method

Our goal is to generate a photorealistic video V of variable length synthesizing a human talking, with realistic head motion and gestures. The input is an audio and a reference image of the person, which can be either real or synthesized. Our framework, which we call VLOGGER, is illustrated in Fig. 2. VLOGGER is a two-stage pipeline based on stochastic diffusion models to represent the one-to-many mapping from speech to video. The first network takes as input an audio waveform $a \in \mathbb{R}^{NS}$ at sample rate S to generate intermediate body motion controls C , which are responsible for gaze, facial expressions and 3D pose over the target video length N . The second network is a temporal image-to-image translation model that extends large image diffusion models, taking the predicted body controls as input, to generate the corresponding video frames. To condition the process on a particular identity, the network also takes a reference image of a person. We train VLOGGER on our newly introduced MENTOR dataset (§3.3). We describe both networks next.

3.1. Audio-Driven Motion Generation

Architecture. The first network of our pipeline M is designed to predict the driving motion based on an input speech. We also consider input text through a text-to-speech model to convert inputs to waveforms [76], and represent the resulting audio as standard Mel-Spectrograms. M is based on a transformer architecture [77] with four multi-head attention layers on the temporal dimension. We in-

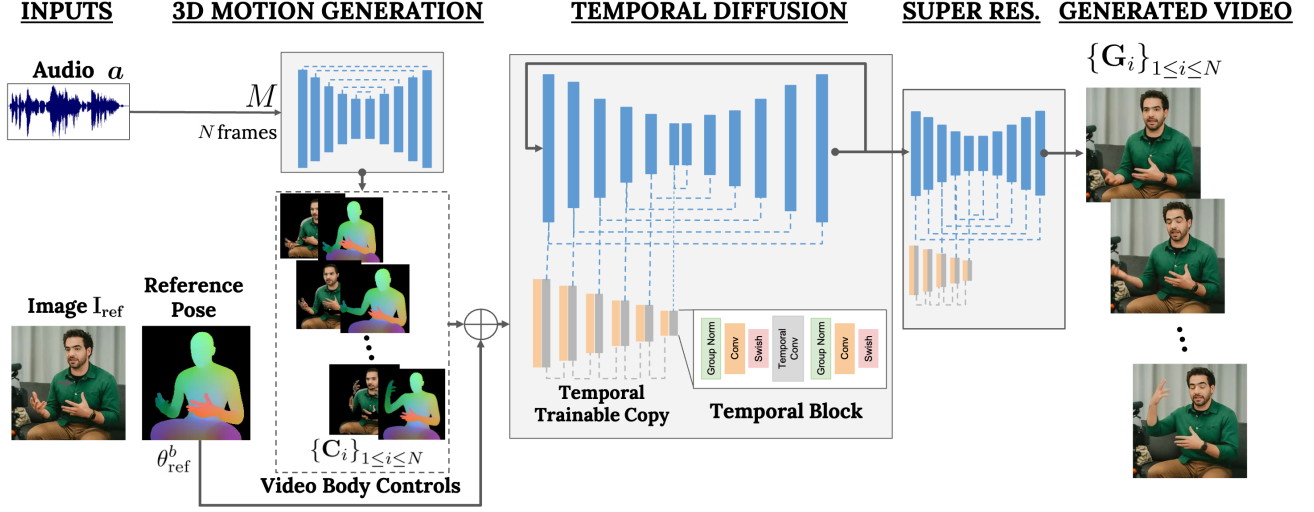


Figure 2. **High-level overview.** VLOGGER conditions the video generation process using a statistical 3D body model. Given an input image I_{ref} (left), the predicted shape parameters encode the geometric properties of the target identity. First, a network M takes the Mel-Spectrogram a of an input speech and generates a sequence of 3D facial expressions $\{\theta_i^e\}_{1 \leq i \leq N}$ and body poses $\{\theta_i^b\}_{1 \leq i \leq N}$ for N frames. We render dense representations of the moving 3D body to act as 2D controls $\{C_i\}_{1 \leq i \leq N}$ in the video generation stage. Together with the reference image of the subject, these are given as input to a temporal diffusion model and a super-resolution module, which are trained to generate a sequence of photorealistic reenactments $\{G_i\}_{1 \leq i \leq N}$ of the target identity.

clude positional encoding on the number of frames and the diffusion step, and an embedding MLP for the input audio and the diffusion step. At each frame, we use a causal mask to make the model attend only to previous frames. The model is trained using variable length video to enable generation of very long sequences, as *e.g.* in the TalkingHead-1KH Dataset [85] (see §4).

We rely on the estimated parameters of a statistical and expressive 3D body model [35, 56, 69, 92], fitted to the reference image of the person¹, to produce intermediate control representations for the synthesized video. These models consider both facial expressions and body motion, opening possibilities for human synthesis with expressive and dynamic gestures. We task the motion generation network to predict face and body parameters $M(a_i) = \{\theta_i^e, \Delta\theta_i^b\}$ based on the input audio a_i in frame i . In particular, the model generates expression θ_i^e and residuals over body pose θ_i^b . By predicting displacements, *i.e.* $\Delta\theta_i^b$, we enable the model to take an input image with reference pose θ_{ref}^b for the target subject, and animate the person relatively with $\theta_i^b = \theta_{\text{ref}}^b + \Delta\theta_i^b$, for frames $1 \leq i \leq N$. The identity of the person in the geometric domain is modeled by the body shape code. During both training and testing, we use the estimated 3D shape parameters obtained by fitting the parametric body model to the input image. In order to leverage

¹During training, the statistical 3D human model is fitted to all the human video data, whereas at run-time the model is only fitted to the reference image of the person.

the 2D/3D predictions with CNN-based architectures, we pose the model using the predicted expression and pose parameters and rasterize the template vertex positions of the posed body as dense representations to obtain dense masks $\{C_i^d\}_{1 \leq i \leq N} \in \mathbb{R}^{H \times W \times 3}$. We also rasterize the semantic regions of the body, $\{C_i^m\}_{1 \leq i \leq N} \in \{0, 1\}^{H \times W \times N_c}$, for N_c different semantic classes.

Previous face reenactment work often relies on warped images [20, 82, 103, 107], however these have been so far overlooked in diffusion-based architectures for human animation [10, 32, 84]. We propose bridging the gap between these two representations and use warped images to guide the generative process. This facilitates the task of the network and helps preserving subject identity (see Tab. 3). We assign a pixel color to each body vertex visible in the reference image, and render the body in each new frame, obtaining partial warps $\{C_i^w\}_{1 \leq i \leq N} \in \mathbb{R}^{H \times W \times 3}$. For all renders, rasterization assumes a full-perspective camera, with a diagonal field-of-view inferred from either the training video, or the reference image. For illustration, see Fig. 2. We describe the temporal image diffusion model in the next section. We also ablate the use of dense representations and warped images in the experimental section.

Loss functions. This model follows a diffusion framework which progressively adds Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ to ground-truth samples $x_0 = \{\{\theta_i^e, \Delta\theta_i^b\}\}_{1 \leq i \leq N}$, with a conditional audio input a . The goal is to model the motion distribution of real heads and bodies, $x_0 \sim q(x_0|a)$,

by training a denoising network ϵ_ϕ that predicts the added noise from the noisy input x_t , where t is an arbitrary diffusion step. In our case, we obtained better performance by directly predicting the ground-truth distribution, x_ϕ^0

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0, t, \mathbf{a}, \epsilon \sim \mathcal{N}(0, 1)} \left[\|x_0 - x_\phi^0(x_t, t, \mathbf{a})\|_2^2 \right]. \quad (1)$$

We also include an additional temporal loss to penalize the prediction difference at consecutive frames, $\mathcal{L}_{\text{temp}} = \|x_\phi^0(x_t, t, \mathbf{a})_{i+1} - x_\phi^0(x_t, t, \mathbf{a})_i\|_2^2$, for any given frame $i \in N$, and train the full model using a linear combination of both losses, *i.e.* $\mathcal{L}_{\text{diff}} + \lambda_{\text{temp}} \mathcal{L}_{\text{temp}}$. In practice, we use different temporal loss weights for expression and body pose to ensure smoother motion for the head and hands while allowing larger dynamism for facial expressions.

3.2. Generating Dynamic & Photorealistic Humans

Architecture. Our next goal is to animate an input reference image $\mathbf{I}_{\text{ref}}$ of a person given audio, such that it follows the predicted body and face motion, which is represented in terms of semantic, sparse and dense masks $\mathbf{C}$. Based on these image-based controls, we propose a temporally-aware extension of state-of-the-art diffusion models [66]. Inspired by ControlNet [102], we freeze the initial trained model and make a zero-initialized trainable copy of its encoding layers, which take as input temporal controls $\mathbf{C}$. We interleave 1d convolutional layers in the temporal domain, after the first layer of each downsampling block and before the second GroupNorm activation, as shown in Fig. 2. The network is trained by taking N consecutive frames and controls, and tasked to generate short clips of the reference person animated according to input controls.

Training. We train our method on the MENTOR dataset, which consists of full-length videos of unique human subjects. Because, during training, the network takes a sequence of consecutive frames and an arbitrary reference image $\mathbf{I}_{\text{ref}}$ of the person, we theoretically can assign any video frame as reference. In practice, we sample the reference to be farther away (temporally) from the target clip, as closer examples trivialize training and provide less potential for generalization. The network is trained in two stages by first learning the new control layers [102] on single frames, and later training on video by adding the temporal components. This enables using a large batch size in the first stage and learning the head associated to the reenactment task faster. We train the image models with learning rate $5e-5$, for 400k steps with batch size 128 in both stages. We ablate the effect of this training schedule in Table 3.

Loss functions. Similarly to the previous section and the loss described in Eq. (1), we follow a diffusion process where we add noise ϵ^I to the ground-truth images $\mathbf{I}$. We base our model on a version of Imagen [66] trained on internal data sources, which predicts the added noise ϵ^I

$$\mathcal{L}_{\text{diff}}^I = \mathbb{E}_{x_0^I, t, \mathbf{C}, \epsilon^I \sim \mathcal{N}(0, 1)} \left[\|\epsilon^I - \epsilon_\phi^I(x_t^I, t, \mathbf{C})\|_2^2 \right]. \quad (2)$$

Super Resolution. While the previous approach is resolution independent, we generate base video at 128×128 resolution, and use a cascaded diffusion approach to extend the temporal conditioning in two super-resolution variants for higher quality video at 256×256 or 512×512 . The generated images are denoted as $\{\mathbf{G}_i\}_{1 \leq i \leq N}$. High resolution examples are shown in Fig. 1 and Fig. 4.

Temporal outpainting during inference. The proposed temporal diffusion model is trained to generate only a fixed number of frames N , so it is not obvious how to extend it to variable length video. Most previous diffusion-based video generation methods are limited to short clips [29, 37, 90] or rely on smoothly generated intermediate token representations [78], but without guarantees of smooth changes in the image domain. Here, we explore the idea of temporal outpainting: we first generate N frames, and then iteratively outpaint $N' < N$ frames based on the previous $N - N'$. The extent of overlap between two consecutive clips, *i.e.* $N - N'$ is chosen as a trade-off between quality and running time. We use DDIM to generate each video clip, and show this approach can scale to thousands of frames. For details, see the ablation in Tab. 2, where we validate the main design choices and show that our final network can generate realistic and temporally coherent video of humans.

3.3. MENTOR Dataset

We curate the MENTOR Dataset from a large repository of videos that contain a single speaker, mostly facing the camera, from the torso up, communicating mostly in English. The videos contain 240 frames at 24 fps (10 seconds clips), with audio at 16 kHz.

With the goal of modeling full-body communicating humans, we estimate 3D body joints and hands and fit a statistical articulated 3D body model by minimizing the projection error and temporal difference between consecutive frames. We filter out videos where the background changes meaningfully, the face or body have been only partially detected or their estimations are *jittery*, where hands are completely undetected (*e.g.* in cases of humans grasping and manipulating objects), or the audio is of low quality. This process resulted in a training set of more than 8M seconds (2.2K hours) and 800K identities, and a test set of 120 hours and $\sim 4K$ identities, making it the largest dataset used to date in terms of identities and length, at higher resolution.

4. Experiments

Data and Training. We train VLOGGER on the MENTOR dataset as described in Sec. 3.3, at a base resolution of 128×128 and cascade resolutions at 256×256 and

512 × 512. Evaluation is performed on the test sets of the HDTF [105], TalkingHead-1KH [85] and MENTOR. We also ablate the performance of our method in different scenarios on the MENTOR dataset and report its performance against baselines across several diversity metrics, such as age, perceived gender, or skin tone.

Baselines. We compare against several state-of-the-art methods, *i.e.* [45, 82, 83, 103, 112]. Note that, unlike our method, all baselines require cropping the face region, as they can detect and animate only the head.

Metrics. We rely on a combination of metrics to evaluate image quality, lip sync, temporal consistency, and identity preservation of the generated videos. For image quality, the FID score [27] measures the distance between ground-truth and generated image distributions, while the Cumulative Probability of Blur Detection (CPBD) [50, 51] and Natural Image Quality Evaluator (NIQE) [48] validate the quality of generated images. Following the literature in talking face generation, we next estimate face landmark coordinates and report the difference in mouth vertex position (LME) to measure lip sync quality. We also report the LSE-D [12] score. Similarly, we report the jitter (or *jerk*) error following [99] to measure the temporal smoothness in the generated video. We also provide the standard deviation of the expression parameters predicted from generated videos, to assess diversity in terms of expression and gaze, given that speech-to-video is not always a one-to-one mapping and it is important to generate a distribution of realistic videos. Regarding diversity of body and hand motion, VLOGGER is the first model to consider gestures, and we assess this qualitatively.

4.1. Ablation Study

We ablate our main design choices extensively in Tables 2 and 3. Tab. 2 summarizes the most representative metrics for the complete method (last row) and each row represents the effect of changing one feature (*e.g.* not using a temporal loss when training the motion predictor). Tab. 3 validates the importance of the 2d controls used to generate videos. We discuss the results next.

Motion generation. In the upper-part of Tab. 2 we show the drop in temporal consistency when not using temporal loss or not predicting Δ (see §3.1). The network gains in smoothness and stability when predicting a residual over body motion, resulting in overall higher image quality. We also show the positive use of classifier-free guidance regarding LME and FID [27].

Video Generation. The lower-part of Tab. 2 ablates the design choices on the temporal video generation model. First, it validates the effectiveness of the proposed outpainting procedure, which not only supports variable-length video generation, but also ensures smoothness and low jitter. Our final model has an overlap of 50% between generated and

	FID [27] ↓	LME [mm] ↓	Jitter [mm/s ³] ↓
Metrics in the final video			
Motion Generation			
No Δ over body pose	52.27	4.22	6.56
No temporal loss	16.56	3.18	4.64
No CFG	16.54	3.32	3.49
Temporal Diffusion Model			
No body controls	16.95	3.10	4.45
No temporal outpainting	15.88	3.25	3.70
25% outpainting	15.90	3.23	3.61
Full model	15.36	3.06	3.58

Table 2. **Ablation of the main design choices in VLOGGER** evaluated on the MENTOR Dataset, where we report metrics that measure image quality through the FID [27] score, expressiveness and lip sync quality via landmark error (LME), and temporal consistency through vertex jitter. The first part shows that the temporal loss and classifier-free guidance lead to the best performance in image quality and LME. The second part summarizes improvements for design choices in the temporal diffusion model. The final pipeline benefits from taking body controls, and the proposed temporal outpainting (50% overlap in the full model) results in the best temporal consistency. We noticed the model plateaus with more overlap.

	Face		Body		Hands		Full Image			
	PSNR ↑	L1 ↓	PSNR ↑	L1 ↓	PSNR ↑	L1 ↓	PSNR ↑	SSIM ↑	LPIPS ↓	L1 ↓
w/ 2D keypoints	20.5	.0591	17.9	.0778	17.8	.0763	19.8	.702	0.138	.0564
w/ dense control	20.4	.0604	18.3	.0750	18.2	.0744	20.1	.719	0.128	.0548
+ warping	21.6	.0517	19.3	.0668	19.1	.0680	20.7	.722	0.113	.0496
+ train schedule	22.2	.0468	20.2	.0594	20.0	.058	21.6	.76	.095	.0447

Table 3. **Ablation of 2d controls in video generation**, in the MENTOR Dataset. We ablate different 2d controls considered in concurrent works, such as driving 2d skeleton [32, 84], dense body representations [94] or our proposed controls which include dense representations and warped images.

given frames, and plateaus at larger values, but obtains a noticeable improvement with respect to a smaller overlap (25%), or no outpainting. The model also performs better with body pose control.

Effect of 2d controls in video generation. We ablate the representation designs used to guide the video generation, in Tab. 3, by reenacting test set samples with their GT motion and reporting image reconstruction metrics. We explore 2d landmarks, dense representations and our controls, which combine dense body representations with a warped reference image. The latter eases the task of the network significantly and leads to the best results. Moreover, we obtain an additional boost in performance with the training schedule described in §3, of first training in single images and later finetuning the temporal layers in videos.

4.2. Quantitative Results

Talking Head Generation. Tab. 4 summarizes the performance of VLOGGER against previous state-of-the-art methods on the task of audio-driven video generation.

HDTF Dataset [105]									
	Photorealism			Lip Sync		Diversity	Identity Preserv.		Temp. Consist.
	FID [27] ↓	CPBD [51] ↑	NIQE [48] ↓	LSE-D [12] ↓	LME [mm] ↓	Expression ↑	Head Err. ↓	ArcFace [15] ↓	Jitter [mm/s ³] ↓
Groundtruth	0.00	0.562	6.31	7.79	0.0	0.401	0.00	0.00	5.19
MakeItTalk [112]	22.63	0.428	6.65	8.30	3.26	0.364	0.911	0.828	6.21
Audio2Head [82]	19.58	0.512	<u>6.41</u>	7.55	3.08	<u>0.415</u>	0.896	1.92	6.15
Wang <i>et al.</i> [83]	21.23	0.428	7.71	8.04	4.48	0.365	1.37	2.52	6.46
SadTalker [103]	<u>19.44</u>	<u>0.520</u>	6.48	<u>7.73</u>	3.01	0.287	<u>0.880</u>	0.874	5.51
StyleTalk [45]	34.16	0.472	6.47	7.87	3.79	0.416	1.14	0.692	4.34
Ours	18.98	0.621	5.92	8.10	<u>3.05</u>	0.397	0.877	<u>0.759</u>	<u>5.05</u>
Ours (Best of 3)	-	0.628	5.64	7.43	2.95	0.425	0.829	0.706	4.75
Ours (Best of 5)	-	0.631	5.53	7.22	2.91	0.436	0.814	0.687	4.67
Ours (Best of 8)	-	0.634	5.44	7.04	2.84	0.448	0.800	0.677	<u>4.58</u>

TalkingHead-1KH Dataset [85]									
	Photorealism			Lip Sync		Diversity	Identity Preserv.		Temp. Consist.
	FID [27] ↓	CPBD [51] ↑	NIQE [48] ↓	LSE-D [12] ↓	LME [mm] ↓	Expression ↑	Head Err. ↓	ArcFace [15] ↓	Jitter [mm/s ³] ↓
Groundtruth	0.00	0.512	7.27	8.70	0.0	0.452	0.00	0.00	3.91
MakeItTalk [112]	34.84	<u>0.493</u>	7.86	10.48	3.50	0.382	<u>1.20</u>	0.909	4.69
Audio2Head [82]	46.49	0.475	7.55	9.38	4.33	0.494	1.47	2.01	4.66
Wang <i>et al.</i> [83]	34.52	0.440	8.61	10.18	3.49	0.338	1.48	2.93	4.70
SadTalker [103]	<u>31.45</u>	0.482	<u>7.46</u>	8.17	3.10	0.347	1.21	0.961	4.26
StyleTalk [45]	38.98	0.468	7.96	9.46	3.44	0.421	1.29	0.663	3.19
Ours	28.94	0.575	6.91	<u>8.50</u>	<u>3.33</u>	<u>0.436</u>	1.05	<u>0.881</u>	<u>4.16</u>
Ours (Best of 3)	-	0.582	6.33	8.43	3.07	0.448	1.03	0.853	3.68
Ours (Best of 5)	-	0.585	6.21	8.35	2.96	0.455	1.01	0.833	3.57
Ours (Best of 8)	-	0.589	6.08	<u>8.20</u>	2.94	<u>0.469</u>	0.99	<u>0.813</u>	<u>3.56</u>

Table 4. **Quantitative evaluation on the HDTF and TalkingHead-1KH Datasets.** We measure the capacity of our model to generate realistic talking heads in multiple metrics. VLOGGER achieves the highest visual quality with highest identity preservation summarized in several metrics, while obtaining expression diversity and temporal consistency close to the groundtruth videos. Regarding lip sync quality, all methods obtain comparable scores. To demonstrate the diversity generated by VLOGGER, we also report the improvement in performance when generating 3, 5 or 8 videos (Except for FID which measures a similarity within an image distribution). Results are consistent for all metrics on both datasets.

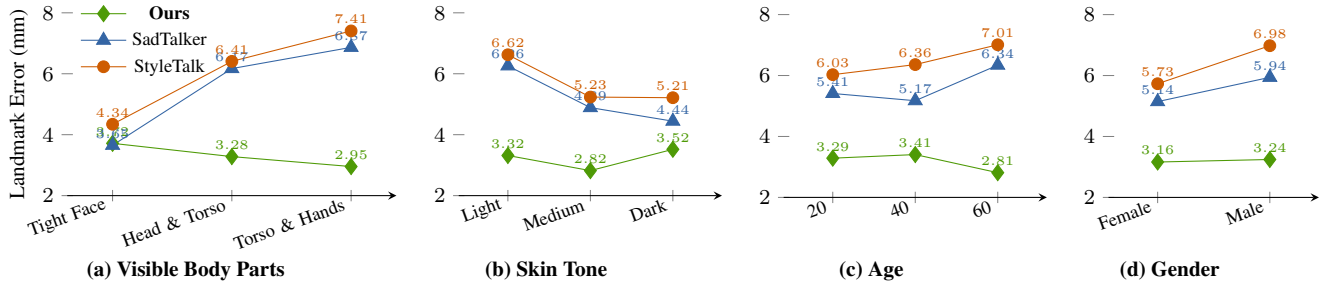


Figure 3. Our model and closest competitors across **different perceived attributes**, such as skin tone, gender and age, on the test set of MENTOR dataset. Our model leverages priors from large pre-trained diffusion models and our proposed large-scale dataset. Thus, in contrast to other methods, it manages to perform consistently across all categories, showing little to no bias. We also show in (a) that our model is capable of animating humans in images at a wide range of viewpoints, instead of cropping tight bounding boxes around the face.

We report results on the HDTF Dataset [105] and on the TalkingHead-1KH Dataset [85]. Talking head generation is a challenging task with several desirable properties, assessed by different metrics. Noticeably, there is a trade-off between image quality, diversity and identity preservation. VLOGGER comes close to the amount of expression diversity present in real video while achieving the highest image quality and identity preservation, with second lowest motion jitter after StyleTalk [45], which introduces very little face motion (see Fig. 4). The temporal consistency vali-

dates the contribution of our temporal layer and the out-painting procedure, while still leveraging the high-quality image generation capabilities of state-of-the-art diffusion models. All methods obtain comparable Lip Sync scores, and results are consistent for all metrics on both datasets evaluated. While no other baseline consider body or gestures, we ablate our design choices in this regard in Tables 2 and 3. In Fig. 3, we showcase our fairness and generalization capabilities (in part benefiting also from the training set scale), by running comparisons to other methods across



Figure 4. **Qualitative comparison** showing input images (left) and generated frames. Baselines typically maintain the expression along the whole sequence, and require cropping the head [45, 83, 103]. In contrast, VLOGGER generates changes in the visible areas when considering faces (third row) but also visible upper-body (fifth row). This figure shows animated faces, but examples with gestures are shown in Fig. 1.

several perceived attributes. Previous work exhibits a clear performance degradation for different classes (*e.g.* light vs dark skin, young vs old, *etc.*), and do not generalize to video with visible torsos or hands. In contrast, VLOGGER exhibits fairly low bias along all evaluation axes. We hope that the release of MENTOR will enable the community to address critical generalization and fairness issues and further advance the state-of-the-art.

4.3. Qualitative Results

We show qualitative results in Fig. 4 against SOTA baselines on images in-the-wild. Many of the existing methods have somewhat limited generative capacity, which makes it difficult to synthesize any parts occluded in the reference image (*e.g.* if the teeth were obscuring the mouth interior, this could persist across the generated video). In contrast, our model is able to generate more diverse expressions and correctly inpaint occluded regions of the moving head.

5. Conclusions

We have presented VLOGGER, a methodology for audio-visual human video synthesis, including both face and body, from a single input image, conditioned by audio or text. VLOGGER is built as a temporal extension of control-based diffusion models, with underlying scaffolding based on 3D human head and body pose representations. Such representations support consistency and generalization and the production of high quality animations of variable length. We introduce a diverse and large scale dataset, and validate the performance of VLOGGER on this and multiple other repositories. We show that VLOGGER outperforms previous state-of-the-art on the task of talking face generation, and that it is more robust on different diversity axes and is able to generate gestures consistent with the spoken content.

Acknowledgements: We gratefully acknowledge Alonso Martinez, Anja Hauth, Sergi Caelles, Hernan Moraldo, Erik Frey, Krishna Somandepalli and Brendan Jou for their careful collection and analysis of a large and diverse repository of videos from which we curated MENTOR.

References

- [1] Adam, M., Wessel, M., Benlian, A.: Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets* **31**(2), 427–445 (2021) [2](#)
- [2] Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Li, Y., Michaeli, T., et al.: Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945* (2024) [2](#), [3](#)
- [3] Bard: Bard: A large language model from google ai (2023), <https://blog.google/technology/ai/bard-google-ai-search-updates/> [2](#)
- [4] Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR* (2022) [3](#)
- [5] Bazavan, E.G., Zanfir, A., Szente, T.A., Zanfir, M., Alldieck, T., Sminchisescu, C.: Sphear: Spherical head registration for complete statistical 3d modeling. *3DV* (2024) [3](#)
- [6] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023) [2](#), [3](#)
- [7] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In: *ECCV* (2016) [3](#)
- [8] Bounareli, S., Argyriou, V., Tzimiropoulos, G.: Finding directions in gan’s latent space for neural face reenactment. *BMVC* (2022) [3](#)
- [9] Bounareli, S., Tzelepis, C., Argyriou, V., Patras, I., Tzimiropoulos, G.: Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces. In: *ICCV*. pp. 7149–7159 (2023) [2](#), [3](#)
- [10] Chang, D., Shi, Y., Gao, Q., Fu, J., Xu, H., Song, G., Yan, Q., Yang, X., Soleymani, M.: Magicdance: Realistic human dance video generation with motions & facial expressions transfer. *arXiv preprint arXiv:2311.12052* (2023) [4](#)
- [11] Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: *CVPR*. pp. 7832–7841 (2019) [3](#)
- [12] Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: *ACCV-W* (2016) [6](#), [7](#)
- [13] Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? (2017) [3](#)
- [14] Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: *CVPR* (2019) [2](#), [3](#)
- [15] Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: *CVPR*. pp. 4690–4699 (2019) [7](#)
- [16] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. *NeurIPS* **34**, 8780–8794 (2021) [3](#)
- [17] Doukas, M.C., Zafeiriou, S., Sharmanska, V.: Headgan: One-shot neural head synthesis and editing. In: *ICCV*. pp. 14398–14407 (October 2021) [3](#)
- [18] Drobyshev, N., Casademunt, A.B., Vougioukas, K., Landgraf, Z., Petridis, S., Pantic, M.: Emoportraits: Emotion-enhanced multimodal one-shot head avatars (2024) [2](#), [3](#)
- [19] Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: *CVPR*. pp. 18770–18780 (2022) [2](#), [3](#)
- [20] Gao, Y., Zhou, Y., Wang, J., Li, X., Ming, X., Lu, Y.: High-fidelity and freely controllable talking head video generation. In: *CVPR*. pp. 5609–5619 (2023) [2](#), [4](#)
- [21] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *NeurIPS* (2014) [3](#)
- [22] Guan, J., Zhang, Z., Zhou, H., Hu, T., Wang, K., He, D., Feng, H., Liu, J., Ding, E., Liu, Z., et al.: Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In: *CVPR*. pp. 1505–1515 (2023) [2](#)
- [23] Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In: *ICCV*. pp. 5784–5794 (2021) [3](#)
- [24] Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animate-diff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023) [3](#)
- [25] Ha, S., Kersner, M., Kim, B., Seo, S., Kim, D.: Marionette: Few-shot face reenactment preserving identity of unseen targets. In: *AAAI* (2020) [3](#)
- [26] He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221* (2022) [3](#)
- [27] 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. *NeurIPS* **30** (2017) [6](#), [7](#)

- [28] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022) 3
- [29] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. *arXiv:2204.03458* (2022) 5
- [30] Hong, F.T., Zhang, L., Shen, L., Xu, D.: Depth-aware generative adversarial network for talking head video generation. In: *CVPR*. pp. 3397–3406 (2022) 3
- [31] Hsu, G.S., Tsai, C.H., Wu, H.Y.: Dual-generator face reenactment. In: *CVPR*. pp. 642–650 (2022) 2, 3
- [32] Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117* (2023) 4, 6
- [33] Jamaludin, A., Chung, J.S., Zisserman, A.: You said that?: Synthesising talking faces from audio. *IJCV* **127**, 1767–1779 (2019) 3
- [34] Ji, X., Zhou, H., Wang, K., Wu, Q., Wu, W., Xu, F., Cao, X.: Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In: *SIGGRAPH. SIGGRAPH '22* (2022). <https://doi.org/10.1145/3528233.3530745>, <https://doi.org/10.1145/3528233.3530745> 3
- [35] Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 8320–8329 (2018) 4
- [36] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: *CVPR*. pp. 8110–8119 (2020) 3
- [37] Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439* (2023) 3, 5
- [38] Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Hornung, R., Adam, H., Akbari, H., Alon, Y., Birodkar, V., et al.: Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125* (2023) 2, 3
- [39] Kyrilitsias, C., Michael-Grigoriou, D.: Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality* **2**, 786665 (2022) 2
- [40] Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. *TOG* **36**(6), 194–1 (2017) 3
- [41] Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., Zhou, B.: Semantic-aware implicit neural audio-driven video portrait generation. In: *ECCV*. pp. 106–125. Springer (2022) 3
- [42] Liu, Y., Cao, Q., Wen, Y., Jiang, H., Ding, C.: Towards variable and coordinated holistic co-speech motion generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1566–1576 (2024) 3
- [43] Liu, Y., Lin, L., Yu, F., Zhou, C., Li, Y.: Moda: Mapping-once audio-driven portrait animation with dual attentions. In: *ICCV* (2023) 2
- [44] Lu, Y., Chai, J., Cao, X.: Live Speech Portraits: Real-time photorealistic talking-head animation. *ACM Transactions on Graphics* **40**(6) (December 2021). <https://doi.org/10.1145/3478513.3480484> 3
- [45] Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., Yu, X.: Styletalk: One-shot talking head generation with controllable speaking styles. *arXiv preprint arXiv:2301.01081* (2023) 2, 6, 7, 8
- [46] Manyika, J.: An overview of bard: an early experiment with generative ai. *AI. Google Static Documents* (2023) 1
- [47] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: *ECCV* (2020) 3
- [48] Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *Signal Processing Letters* **20**(3), 209–212 (2012) 6, 7
- [49] Moussawi, S., Koufaris, M., Benbunan-Fich, R.: How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. *Electronic Markets* **31**, 343–364 (2021) 2
- [50] Narvekar, N.D., Karam, L.J.: A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In: *International Workshop on Quality of Multimedia Experience*. pp. 87–91. IEEE (2009) 6
- [51] Narvekar, N.D., Karam, L.J.: A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). *Image Processing* **20**(9), 2678–2683 (2011) 6, 7
- [52] Ni, H., Shi, C., Li, K., Huang, S.X., Min, M.R.: Conditional image-to-video generation with latent flow diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 18444–18455 (2023) 3
- [53] Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: *ICCV*. pp. 7184–7193 (2019) 2, 3

- [54] Niu, M., Cun, X., Wang, X., Zhang, Y., Shan, Y., Zheng, Y.: Mofa-video: Controllable image animation via generative motion field adaptations in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222* (2024) 2, 3
- [55] OpenAI: Gpt-4 technical report (2023) 1
- [56] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10975–10985 (2019) 4
- [57] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: *International Conference on Advanced Video and Signal Based Surveillance*. pp. 296–301. Ieee (2009) 3
- [58] Pizzi, G., Vannucci, V., Mazzoli, V., Donvito, R.: I, chatbot! the impact of anthropomorphism and gaze direction on willingness to disclose personal information and behavioral intentions. *Psychology & Marketing* **40**(7), 1372–1387 (2023) 2
- [59] Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: *ACM International Conference on Multimedia*. pp. 484–492 (2020) 2, 3
- [60] Pumarola, A., Agudo, A., Martinez, A., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: One-shot anatomically consistent facial animation. In: *IJCV* (2019) 3
- [61] Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: *ICCV*. pp. 13759–13768 (2021) 3
- [62] Richard, A., Zollhöfer, M., Wen, Y., De la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: *ICCV*. pp. 1173–1182 (2021) 2, 3
- [63] Roesler, E., Manzey, D., Onnasch, L.: A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics* **6**(58), eabj5425 (2021) 2
- [64] 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
- [65] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arxiv:2208.12242* (2022) 2
- [66] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS* **35**, 36479–36494 (2022) 5
- [67] Seeger, A.M., Pfeiffer, J., Heinzl, A.: Texting with humanlike conversational agents: Designing for anthropomorphism. *Journal of the Association for Information Systems* **22**(4), 8 (2021) 2
- [68] Seitz, L., Bekmeier-Feuerhahn, S., Gohil, K.: Can we trust a chatbot like a physician? a qualitative study on understanding the emergence of trust toward diagnostic chatbots. *International Journal of Human-Computer Studies* **165**, 102848 (2022) 2
- [69] Shen, K., Guo, C., Kaufmann, M., Zarate, J.J., Valentin, J., Song, J., Hilliges, O.: X-avatar: Expressive human avatars. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16911–16921 (2023) 4
- [70] Sora. <https://openai.com/sora> (2024) 2, 3
- [71] Stan, S., Haque, K.I., Yumak, Z.: Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. *arXiv preprint arXiv:2309.11306* (2023) 2, 3
- [72] Styplukowski, M., Vougioukas, K., He, S., Zieba, M., Petridis, S., Pantic, M.: Diffused heads: Diffusion models beat gans on talking-face generation. *arXiv preprint arXiv:2301.03396* (2023) 3
- [73] Suzhen, W., Lincheng, L., Yu, D., Changjie, F., Xin, Y.: Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In: *IJCAI* (2021) 3
- [74] Thambiraja, B., Habibie, I., Aliakbarian, S., Cosker, D., Theobalt, C., Thies, J.: Imitator: Personalized speech-driven 3d facial animation. In: *ICCV*. pp. 20621–20631 (2023) 2, 3
- [75] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: *CVPR*. pp. 2387–2395 (2016) 2, 3
- [76] Text-to-speech - google cloud. <https://cloud.google.com/text-to-speech> (2019) 3
- [77] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *NeurIPS* (2017) 3
- [78] Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399* (2022) 3, 5
- [79] Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with gans. *IJCV* **128**, 1398–1413 (2020) 3

- [80] Wahde, M., Virgolin, M.: Conversational agents: Theory and applications. In: *HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation*, pp. 497–544. World Scientific (2022) [2](#)
- [81] Wang, J., Qian, X., Zhang, M., Tan, R.T., Li, H.: Seeing what you said: Talking face generation guided by a lip reading expert. In: *CVPR*. pp. 14653–14662 (2023) [2](#)
- [82] Wang, S., Li, L., Ding, Y., Fan, C., Yu, X.: Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293* (2021) [4](#), [6](#), [7](#)
- [83] Wang, S., Li, L., Ding, Y., Yu, X.: One-shot talking face generation from single-speaker audio-visual correlation learning. In: *AAAI* (2022) [6](#), [7](#), [8](#)
- [84] Wang, T., Li, L., Lin, K., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. *arXiv e-prints* pp. *arXiv–2307* (2023) [4](#), [6](#)
- [85] Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: *CVPR* (2021) [2](#), [4](#), [6](#), [7](#)
- [86] Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: *CVPR*. pp. 10039–10049 (2021) [3](#)
- [87] Wiles, O., Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: *ECCV*. pp. 670–686 (2018) [3](#)
- [88] Wu, X., Hu, P., Wu, Y., Lyu, X., Cao, Y.P., Shan, Y., Yang, W., Sun, Z., Qi, X.: Speech2lip: High-fidelity speech to lip generation by learning from a short video. In: *ICCV*. pp. 22168–22177 (2023) [2](#)
- [89] Xie, Y., Xu, H., Song, G., Wang, C., Shi, Y., Luo, L.: X-portrait: Expressive portrait animation with hierarchical motion attention. In: *ACM SIGGRAPH 2024 Conference Papers*. pp. 1–11 (2024) [2](#), [3](#)
- [90] Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. *arXiv preprint arXiv:2306.00943* (2023) [3](#), [5](#)
- [91] Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In: *CVPR*. pp. 12780–12790 (2023) [2](#), [3](#)
- [92] Xu, H., Bazavan, E.G., Zafir, A., Freeman, B., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: Generative 3D human shape and articulated pose models. *CVPR* (2020) [3](#), [4](#)
- [93] Xu, S., Chen, G., Guo, Y.X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., Guo, B.: Vasa-1: Life-like audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667* (2024) [2](#), [3](#)
- [94] Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498* (2023) [6](#)
- [95] Yang, K., Chen, K., Guo, D., Zhang, S.H., Guo, Y.C., Zhang, W.: Face2face ρ : Real-time high-resolution one-shot face reenactment. In: *ECCV*. pp. 55–71. Springer (2022) [2](#), [3](#)
- [96] Yao, S., Zhong, R., Yan, Y., Zhai, G., Yang, X.: Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering. *arXiv preprint arXiv:2201.00791* (2022) [3](#)
- [97] Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., Zhao, Z.: Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430* (2023) [3](#)
- [98] Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: *CVPR*. pp. 469–480 (June 2023) [2](#), [3](#)
- [99] Yi, X., Zhou, Y., Xu, F.: Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)* **40**(4), 1–13 (2021) [6](#)
- [100] Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., Bai, Q., Wu, B., Wang, J., Yang, Y.: Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In: *ECCV*. pp. 85–101. Springer (2022) [3](#)
- [101] Yu, Z., Yin, Z., Zhou, D., Wang, D., Wong, F., Wang, B.: Talking head generation with probabilistic audio-to-visual diffusion priors. In: *ICCV*. pp. 7645–7655 (2023) [2](#), [3](#)
- [102] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023) [5](#)
- [103] Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: *CVPR*. pp. 8652–8661 (2023) [2](#), [4](#), [6](#), [7](#), [8](#)
- [104] Zhang, Y., Zhang, S., He, Y., Li, C., Loy, C.C., Liu, Z.: One-shot face reenactment. *arXiv preprint arXiv:1908.03251* (2019) [2](#), [3](#)
- [105] Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: *CVPR*. pp. 3661–3670 (2021) [2](#), [3](#), [6](#), [7](#)

- [106] Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: CVPR. pp. 3661–3670 (2021) [3](#)
- [107] Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: CVPR. pp. 3657–3666 (2022) [3](#), [4](#)
- [108] Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., et al.: A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arXiv preprint arXiv:2302.09419 (2023) [2](#)
- [109] Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: AAI (2019) [3](#)
- [110] Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: CVPR. pp. 4176–4186 (2021) [3](#)
- [111] Zhou, Q., Li, B., Han, L., Jou, M.: Talking to a bot or a wall? how chatbots vs. human agents affect anticipated communication quality. *Computers in Human Behavior* **143**, 107674 (2023) [2](#)
- [112] Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makeltalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)* **39**(6), 1–15 (2020) [6](#), [7](#)