LIA: Latent Image Animator

Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva
https://wyhsirius.github.io/LIA-project/

***Abstract*—Previous animation techniques mainly focus on leveraging explicit structure representations (*e.g.*, meshes or keypoints) for transferring motion from driving videos to source images. However, such methods are challenged with large appearance variations between source and driving data, as well as require complex additional modules to respectively model appearance and motion. Towards addressing these issues, we introduce the Latent Image Animator (LIA), streamlined to animate high-resolution images. LIA is designed as a simple autoencoder that does not rely on explicit representations. Motion transfer in the pixel space is modeled as linear navigation of motion codes in the latent space. Specifically such navigation is represented as an orthogonal motion dictionary learned in a self-supervised manner based on proposed Linear Motion Decomposition (LMD). Extensive experimental results demonstrate that LIA outperforms state-of-the-art on VoxCeleb, TaichiHD, and TED-talk datasets with respect to video quality and spatio-temporal consistency. In addition LIA is well equipped for zero-shot high-resolution image animation.**

***Index Terms*—Generative adversarial networks, image animation, video generation, interpretability, disentanglement.**

I. Introduction

WITH the remarkable success of deep generative models such as Generative Adversarial Networks [19] and Diffusion Models [25], [59], [60], general video generation has achieved tremendous progress. Current methods [5], [17], [24], [57], [77] are able to synthesize photorealistic videos directly from text description. Apart from general video generation, being able to produce videos from a single image, *i.e.*, image animation, has also gained much attention due to its real-world applications, such as filmmaking, digital humans, and art creation. In this work, we focus on *human-centric (e.g., face and human body) image animation*, and we examine the scenario where a *source image* is animated by the motion representations learned from a *driving video*.

Early approaches for image animation are classically related to computer graphics [9], [62], [63], [89] or exploit motion labels [75] and structure representations such as semantic maps [43], [69], [70], human keypoints [11], [30], [52], [68], [69], [82], [85], 3D meshes [12], [39], and optical flows [37], [42]. We note that the ground truth of such structure representations has been computed a-priori for the purpose of supervised training, which poses constraints on applications, where such representations of unseen testing images might be fragmentary or difficult to access.

Recently, self-supervised image animation approaches [52], [55], [79] accept raw videos as input and learn to reconstruct driving images by warping a source image with predicted *dense optical flow fields*. While the need for domain knowledge or labeled ground truth data has been obviated, which improves performance on in-the-wild testing images, such methods entail necessity of explicit structure representations as motion guidance. Prior information such as keypoints [52], [71] or regions [55] are learned in an end-to-end training manner by additional networks as intermediate features, in order to predict target flow fields. Although online prediction of such representations is less tedious than the acquisition of ground truth labels, it still strains the complexity of networks.

Deviating from such approaches, we here aim to fully *eliminate* the need of *explicit structure representations* by directly manipulating the latent space of a deep generative model. To the best of our knowledge, this constitutes a new direction in the context of *image animation*. Our work is motivated by *interpretation of GANs* [18], [29], [50], [67], showcasing that latent spaces of StyleGAN [32], [33] and BigGAN [6] contain rich semantically meaningful directions. Given that walking along such directions, basic visual transformations such as *zooming* and *rotation* can be induced in generated results. As in image animation, we have that motion between source and driving images can be considered as higher-level transformation, a natural question here arises: *can we discover a set of directions in the latent space that induces high-level motion transformations collaboratively?*

Towards answering this question, we introduce Latent Image Animator (LIA), a novel image animation model constituting of an autoencoder for animating still images via latent space navigation. LIA seeks to animate a source image via linearly navigating associated source latent code along a learned path to reach a target latent code, which represents the high-level transformation for animating the source image. We introduce a Linear Motion Decomposition (LMD) approach aiming to represent a latent path via a linear combination of a set of learned motion directions and associated magnitudes. Specifically, we constrain the set as an orthogonal basis, where each vector indicates a basic visual transformation. By describing the whole motion space using such learned basis, LIA eliminates the requirement of explicit structure representations. In addition, we design LIA to disentangle motion and appearance within a single encoder-generator architecture. Deviating from existing methods using separate networks to learn disentangled features, LIA integrates both, latent *motion* code, as well as *appearance* features in a *single encoder*, which highly reduces the model complexity and simplifies training.

We conduct evaluation on multiple datasets including VoxCeleb [13], TaichiHD [52], CelebV-HQ [91] and TED-talk [55]. In addition, we show that LIA outperforms the state-of-the-art in preserving the facial structure in generated videos in the setting of zero-shot image animation on unseen datasets such as FFHQ [32] and GermanPublicTV [61]. As shown in Fig. 1, on high-resolution generation, LIA also achieves high quality results.

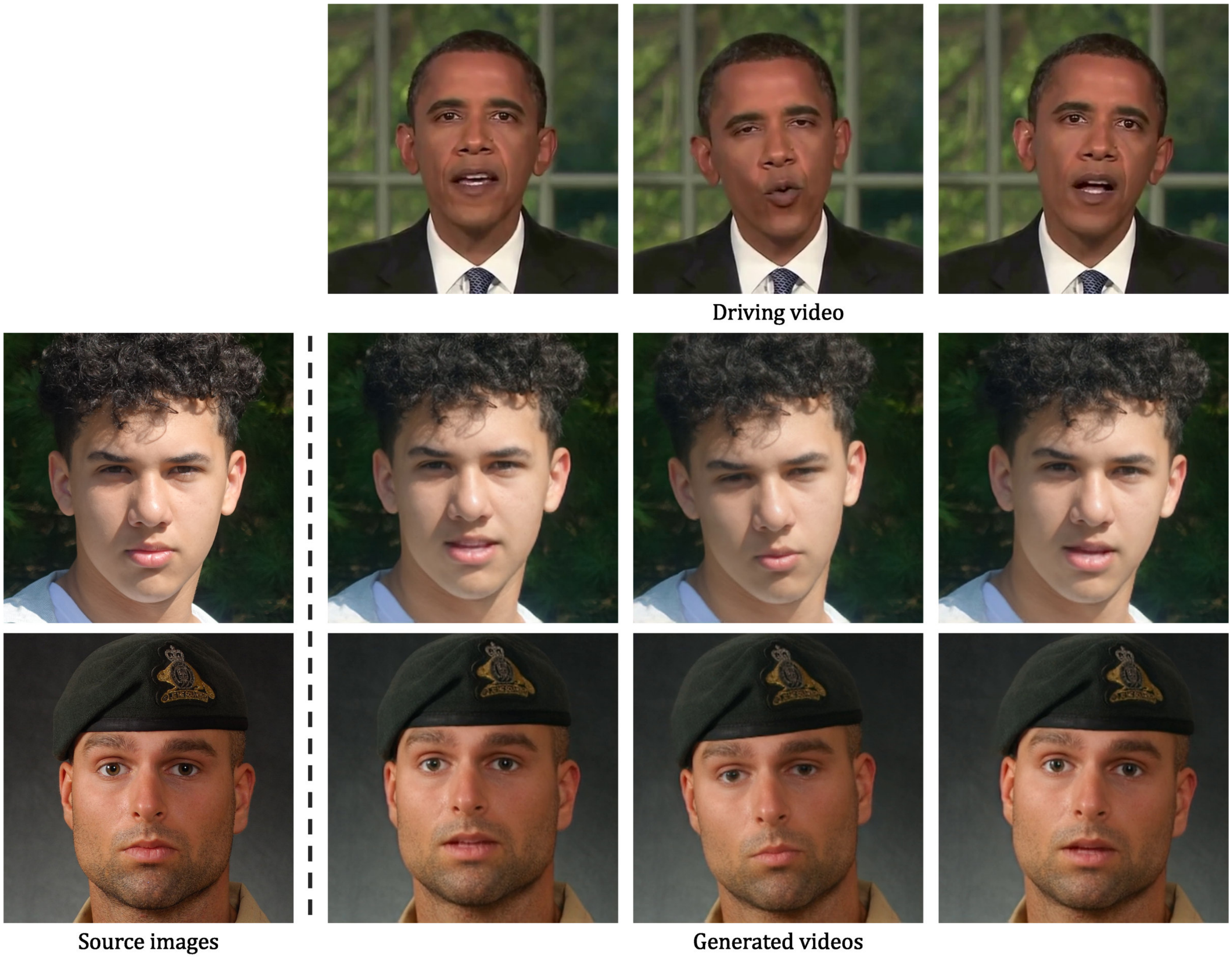


Fig. 1: **LIA animation examples in high-resolution** (512 × 512)**.** The two images from FFHQ [32] are animated by LIA, which transfers motion of a driving video (top row) of Obama onto the still images. LIA is able to successfully animate these two images without relying on any *explicit structure representations*, such as landmarks and region representations.

This work is an extension of our previous work [78]. We here proceed to expand the original work by (a) proposing a *novel high-resolution talking head dataset*, namely VoxCelebHQ which contains 17,995 videos of 512 resolution, aimed towards boosting high-quality talking head animation. We provided detailed statistics on the new dataset. We (b) provide additional details on the design of the improved high-resolution LIA-model and have conducted an ablation study on associated improvement. In particular, we have quantitatively compared the improved LIA with the original LIA on high-resolution generation, and compared high-resolution LIA with MRAA, both trained on VoxCelebHQ. We have qualitatively shown the improvement of high-resolution LIA over the low-resolution model. In addition, we (c) conduct an analysis on the orthogonal basis to study the effectiveness of the orthogonality and show quantitative results. Further, we (d) provide more technical details by presenting additional experiments and an expanded literature review.

II. Related Work

Video generation is aimed at generating videos by simultaneously modeling spatio-temporal distribution. A number of deep generative models have been proposed, such as GANs [7], [14], [49], [58], [64]–[66], [73]–[76], [84], VAEs [3], [15], [38], [80], and VQ-based models [16], [81]. Most recently, Diffusion Models (DMs) [25], [41], [59] have contributed to a remarkable progress in image synthesis [46]–[48], as well as video generation. Building upon this success, numerous recent works [22], [26] have explored the application of DMs for video generation. These works showcase the promising capability of DMs to model complex video distributions by integrating spatio-temporal operations into image-based models, surpassing previous approaches in terms of video quality. Unlike these approaches, which generate general videos based on noise input in an unconditional manner, in this paper, we focus on conditionally creating human-centric videos by transferring motion from driving videos to input

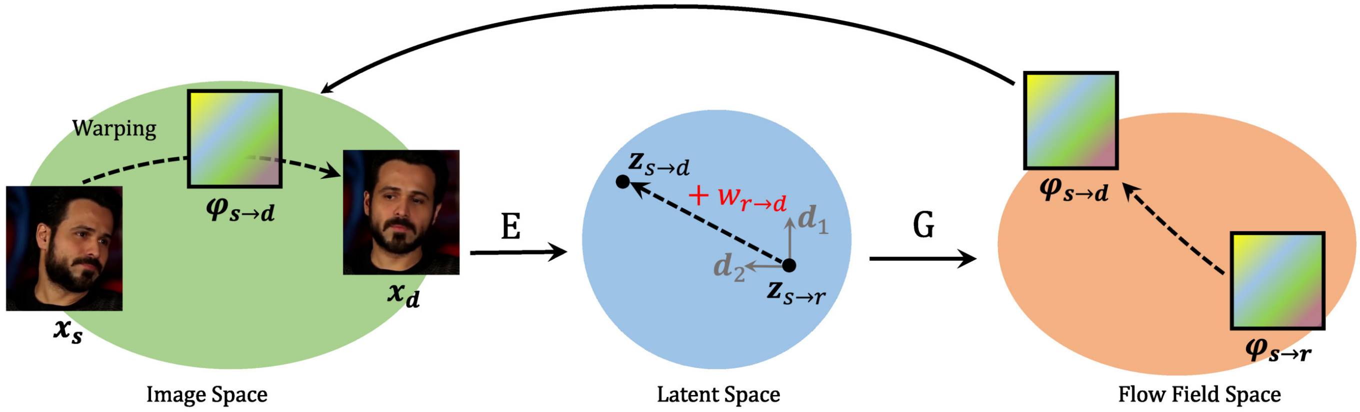


Fig. 2: **General pipeline.** Our objective is to transfer motion via latent space navigation. The entire training pipeline consists of two steps. Firstly, we encode a source image x_s into a latent code $z_{s\to r}$. By linearly navigating $z_{s\to r}$ along a path $w_{r\to d}$, we reach a target latent code $z_{s\to d}$. The latent paths are represented by a linear combination between a set of learned motion directions (*e.g.*, d_1 and d_2), which is an orthogonal basis, and associated magnitudes. In the second step, we decode $z_{s\to d}$ to a target dense optical flow field $\phi_{s\to d}$, which is used to warp x_s into the driving image x_d. While we train our model using images from the same video sequence, in the testing phase, x_s and x_d generally pertain to different identities.

source images.

Latent space editing. In an effort to control generated images, recent works explored the discovery of semantically meaningful directions in the latent space of pre-trained GANs, where linear navigation corresponds to desired image manipulation. Supervised [18], [29], [50] and unsupervised [45], [51], [67] approaches were proposed to edit semantics such as facial attributes, colors and basic visual transformations (*e.g.,* rotation and zooming) in generated or inverted real images [1], [92]. In this work, as opposed to finding directions corresponding to individual visual transformations, we seek to learn a set of directions that cooperatively allows for high-level visual transformations that can be beneficial in image animation.

Image animation. Related approaches [11], [69], [70], [83], [85], [88] in image animation required strong prior structure labels as motion guidance. In particular, [11], [83] and [70] proposed to map representations such as human keypoints and facial landmarks to videos in the setting of image-to-image translation proposed by [28]. However, such approaches were only able to learn an individual model for a single identity. Transferring motion on new appearances requires retraining the entire model from scratch by using videos of target identities. Several recent works [69], [85] explored meta learning in fine-tuning models on target identities. While only few images of target identities were required during inference time, it was still compulsory to input pre-computed structure representations in those approaches, which usually are hard to access in many real-world scenarios. Towards addressing this issue, very recent works [52], [55], [72], [79], [88] proposed to learn image animation in self-supervised manner, only relying on RGB videos for both, training and testing without any priors. They firstly predicted dense flow fields from input images, which were then utilized to warp source images, in order to obtain final generated results. Inference only required one image of a target identity without any fine-tuning step on pre-trained models. While no priors were required, state-of-the-art methods still followed the idea of using explicit structure representations. FOMM [52] proposed a first order motion approach to predict keypoints and local transformations online to generate flow fields. MRAA [55] developed this idea to model articulated objects by replacing a keypoints predictor by a PCA-based region prediction module. Similar to FOMM, Zhao *et al.* [88] proposed a thin-plate spline motion model towards learning more precise flow fields between source and target images. To enable 3D-aware animation, Wang *et al.* [72] extended FOMM by predicting 3D keypoints for view-free generation. Siarohin *et al.* [54] introduced a unsupervised method to learn 3D structure and dynamics solely from single-view RGB videos. Li *et al.* [36] proposed to leverage deformable nerf towards improving fidelity in 3D talking-head animation. Recently, DMs-based approaches also achieve tremendous progress in both human-centric [21], [27] and general [4], [87] image animation, where powerful stable diffusion is leveraged to produce more photorealistic results.

In contrast to such approaches, our method does not require any explicit structure representations. We dive into the latent space of the generator and self-learn to navigate motion codes in certain directions with the goal to reach target codes, which are then decoded to flow fields for warping.

III. Method

Self-supervised image animation aims at learning to transfer motion from a subject of a driving video to a subject in a source image based on training with a large-scale video dataset. In this work, we propose to model such motion transformation via latent space navigation. The general pipeline is illustrated in Fig. 2. Specifically, in training time, our model takes in a pair of source and driving frames, randomly sampled from one video sequence. These two images are encoded into a latent code which is used to represent motion transformation in the image space. The training objective is to reconstruct

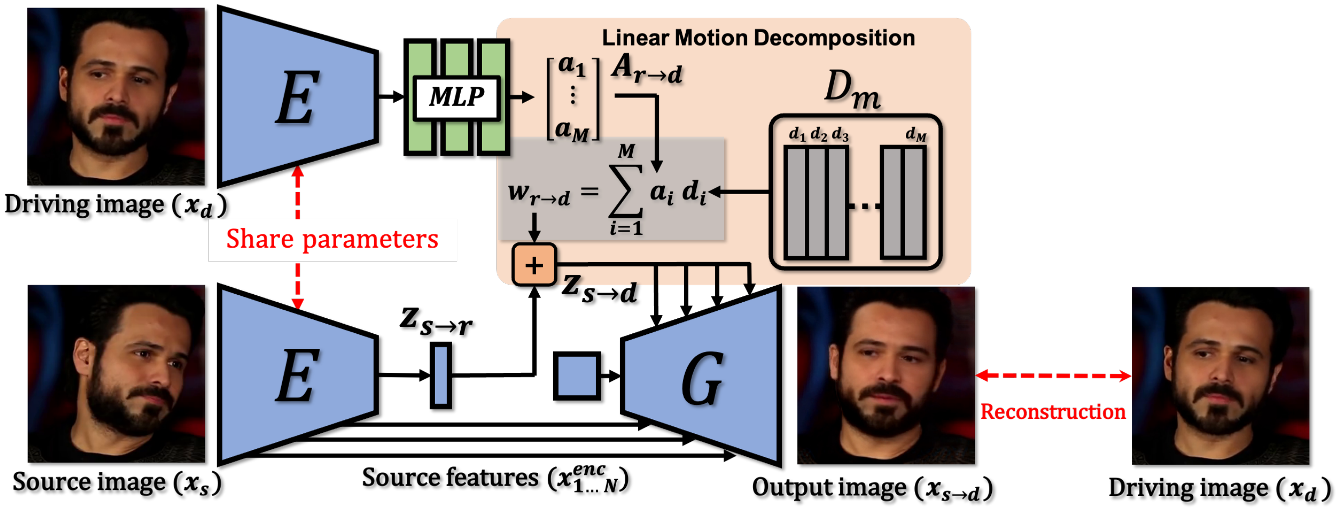


Fig. 3: **Overview of LIA.** LIA is an autoencoder consisting of two networks, an encoder E and a generator G. In the latent space, we apply Linear Motion Decomposition (LMD) towards learning a motion dictionary D_m, which is an orthogonal basis where each vector represents a basic visual transformation. LIA takes two frames sampled from the same video sequence as source image x_s and driving image x_d respectively during training. Firstly, it encodes x_s into a source latent code $z_{s\to r}$ and x_d into a magnitude vector $A_{r\to d}$. Then, it linearly combines $A_{r\to d}$ and a trainable D_m using LMD to obtain a latent path $w_{r\to d}$, which is used to navigate $z_{s\to r}$ to a target code $z_{s\to d}$. Finally, G decodes $z_{s\to d}$ into a target dense flow field and warps x_s to an output image $x_{s\to d}$. The training objective is to reconstruct x_d using $x_{s\to d}$.

the driving image by combining source image with learned motion transformation. In testing, frames of a driving video are sequentially processed with the source image to animate the source subject.

We provide an overview of the proposed model in Fig. 3. Our model is an autoencoder, consisting of two main networks, an encoder E and a generator G. In general, our model requires two steps to transfer motion. In the first step, E encodes source and driving images $x_s, x_d \sim \mathcal{X} \in \mathbb{R}^{3\times H\times W}$ into latent codes in the latent space. The source code is then navigated into a target code, which is used to represent target motion transformation, along a learned latent path. Based on proposed Linear Motion Decomposition (LMD), we represent such a path as a linear combination of a set of learned motion directions and associated magnitudes, which are learned from x_d. In the second step, once the target latent code is obtained, G decodes it as a dense flow field $\phi_{s\to d} \sim \Phi \in \mathbb{R}^{2\times H\times W}$ and uses $\phi_{s\to d}$ to warp x_s and then to obtain the output image. In the following, we proceed to discuss the two steps in detail.

A. Latent motion representation

Given a source image x_s and a driving image x_d, our first step constitutes of learning a *latent code* $z_{s\to d} \sim \mathcal{Z} \in \mathbb{R}^N$ to represent the motion transformation from x_s to x_d. Due to the uncertainty of two images, directly learning $z_{s\to d}$ puts forward a high requirement on the model to capture a complex distribution of motion. Mathematically, it requires modeling directions and norms of the vector $z_{s\to d}$ simultaneously, which is challenging. Therefore, instead of modeling motion transformation $x_s \to x_d$, we assume there exists a reference image x_r and motion transfer can be modeled as $x_s \to x_r \to x_d$, where $z_{s\to d}$ is learned in an indirect manner. Such reference space can be considered to represent a 'normalized' image in an implicit manner, which enables the model to transfer large deformations from source image to target image. In particular, during inference time, directly transferring motion from different identities with different poses may be difficult. With such reference space, the model is able to process one input at a time, reducing mismatches when different identities are used at inference time. We show qualitative justification of reference space in Sec. IV-E. We model $z_{s\to d}$ as a target point in the latent space, which can be reached by taking linear walks from a starting point $z_{s\to r}$ along a linear path $w_{r\to d}$ (see Fig. 2), denoted by

$$z_{s\to d} = z_{s\to r} + w_{r\to d}, \tag{1}$$

where $z_{s\to r}$ and $w_{r\to d}$ indicate the transformation $x_s \to x_r$ and $x_r \to x_d$ respectively. Both $z_{s\to r}$ and $w_{r\to d}$ are learned independently and $z_{s\to r}$ is obtained by passing x_s through E.

We learn $w_{r\to d}$ via Linear Motion Decomposition (LMD). Our idea is to learn a set of motion directions $D_m = \{\mathbf{d_1}, ..., \mathbf{d_M}\}$ to represent any path in the latent space. We constrain D_m as an orthogonal basis, where each vector indicates a motion direction $\mathbf{d_i}$. We then combine each vector in the basis with a vector $A_{r\to d} = \{a_1, ..., a_M\}$, where a_i represents the magnitude of $\mathbf{d_i}$. Hence, any linear path in the latent space can be represented using a linear combination

$$w_{r\to d} = \sum_{i=1}^{M} a_i \mathbf{d_i}, \tag{2}$$

where $\mathbf{d_i} \in \mathbb{R}^N$ and $a_i \in \mathbb{R}$ for all $i \in \{1, ..., M\}$. Semantically, each $\mathbf{d_i}$ should represent a basic visual transformation and a_i indicates the required steps to walk in $\mathbf{d_i}$ towards achieving $w_{r\rightarrow d}$. Due to D_m entailing an orthogonal basis, any two directions $\mathbf{d_i}, \mathbf{d_j}$ follow the constrain

$$< \mathbf{d_i}, \mathbf{d_j} >= \begin{cases} 0 & i \neq j \\ 1 & i = j. \end{cases} \tag{3}$$

We implement D_m as a learnable matrix and apply the Gram-Schmidt process during each forward pass, in order to meet the requirement of orthogonality. $A_{r\rightarrow d}$ is obtained by mapping $z_{d\rightarrow r}$, which is the output of x_d after E, through a 5-layer MLP. The final formulation of latent motion representation for each x_s and x_d is thus given as

$$z_{s\rightarrow d} = z_{s\rightarrow r} + \sum_{i=1}^{M} a_i \mathbf{d_i}. \tag{4}$$

B. Latent code driven image animation

Once we obtain $z_{s\rightarrow d}$, in our second step, we use G to decode a flow field $\phi_{s\rightarrow d}$ and warp x_s. Our G consists of two components, a flow field generator G_f and a refinement network G_r.

Towards learning multi-scale features, G is designed as a residual network containing N models to produce a pyramid of flow fields $\phi_{s\rightarrow d} = \{\phi_i\}_1^N$ in different layers of G_f. Multi-scale source features $x_s^{enc} = \{x_i^{enc}\}_1^N$ are obtained from E and are warped in G_f.

However, as pointed out by Siarohin *et al.* [52], only relying on $\phi_{s\rightarrow d}$ to warp source features is insufficient to precisely reconstruct driving images due to the existing occlusions in some positions of x_s. In order to predict pixels in those positions, the network is required to inpaint the warped feature maps. Therefore, we predict multi-scale masks $\{m_i\}_1^N$ along with $\{\phi_i\}_1^N$ in G_f to mask out the regions required to be inpainted. In each residual module, we have

$$x_i' = \mathcal{T}(\phi_i, x_i^{enc}) \odot m_i, \tag{5}$$

where $\odot$ denotes the Hadamard product and $\mathcal{T}$ denotes warping operation, whereas x_i' signifies the masked features. We generate both dense flow fields, as well as masks by letting each residual module output a 3-channel feature map in which the first two channels represent ϕ_i and the last channel m_i. Based on an inpainted feature map $f(x_i')$, as well as an upsampled image $g(x_{i-1})$ provided by the previous module in G_r, the RGB image from each module is given by

$$o_i = f(x_i') + g(o_{i-1}), \tag{6}$$

where f and g denote the inpainting and upsampling layers, respectively. The output image o_N from the last module constitutes the final generated image $x_{s\rightarrow d} = o_N$.

C. Learning

We train LIA in a self-supervised manner to reconstruct x_d using three losses, *i.e.*, a reconstruction loss $\mathcal{L}_{recon}$, a perceptual loss $\mathcal{L}_{vgg}$ [31] and an adversarial loss $\mathcal{L}_{adv}$. We use $\mathcal{L}_{recon}$ to minimize the pixel-wise L_1 distance between x_d and $x_{s\rightarrow d}$, calculated as

$$\mathcal{L}_{recon}(x_{s\rightarrow d}, x_d) = \mathbb{E}[\|x_d - x_{s\rightarrow d}\|_1]. \tag{7}$$

Towards minimizing the perceptual distance, we apply a VGG19-based $\mathcal{L}_{vgg}$ on multi-scale feature maps between real and generated images, written as

$$\mathcal{L}_{vgg}(x_{s\rightarrow d}, x_d) = \mathbb{E}[\sum_{n}^{N} \|F_n(x_d) - F_n(x_{s\rightarrow d})\|_1], \tag{8}$$

where F_n denotes the n^{th} layer in a pre-trained VGG19 [56]. In practice, towards penalizing multi-scale real and generated images, we use a pyramid containing four resolutions. For the low-resolution model, the four resolutions include 256×256, 128×128, 64×64 and 32×32. For the high-resolution LIA-model, we use a pyramid consisting of 512×512, 256×256, 128×128 and 64×64 resolutions. The final perceptual loss is the addition of perceptual losses in four resolutions.

Further, towards generating photo-realistic results, we incorporate a non-saturating adversarial loss $\mathcal{L}_{adv}$ on $x_{s\rightarrow d}$, which is calculated as

$$\mathcal{L}_{adv}(x_{s\rightarrow d}) = \mathbb{E}_{x_{s\rightarrow d} \sim p_{rec}}[-log(D(x_{s\rightarrow d}))], \tag{9}$$

where D is a discriminator, aimed at distinguishing reconstructed images from the original ones. Our full loss function is the combination of three losses with λ as a balanced hyperparameter

$$\begin{aligned} \mathcal{L}(x_{s\rightarrow d}, x_d) = \mathcal{L}_{recon}(x_{s\rightarrow d}, x_d) + \lambda \mathcal{L}_{vgg}(x_{s\rightarrow d}, x_d) \\ + \mathcal{L}_{adv}(x_{s\rightarrow d}). \end{aligned} \tag{10}$$

D. Inference

During the inference stage, given a driving video sequence $V_d = \{x_t\}_1^T$, we aim to transfer motion from V_d to x_s, in order to generate a novel video $V_{d\rightarrow s} = \{x_{t\rightarrow s}\}_1^T$. In case that V_d and x_s stem from the same video sequence, *i.e.*, $x_s = x_1$, our task comprises of reconstructing the entire original video sequence. Therefore, we construct the latent motion representation of each frame using *absolute transfer*, which follows the training process, given as

$$z_{s\rightarrow t} = z_{s\rightarrow r} + w_{r\rightarrow t},\ t \in \{1, ..., T\}. \tag{11}$$

However, in real world applications, interest is naturally rather placed on the scenario, where motion transfer between x_s and V_d, the latter stemming from different identities, *i.e.*, $x_s \neq x_1$. Taking a *talking head* video as an example, in this setting, beyond identity, x_1 and x_s might also differ in pose and expression. Therefore, we propose *relative transfer* to eliminate the motion impact of $w_{r\rightarrow 1}$ and involve motion of $w_{r\rightarrow s}$ in the full generated video sequence. Owing to a linear representation of the latent path, we can easily represent $z_{s\rightarrow t}$ for each frame as

$$\begin{aligned} z_{s\rightarrow t} &= (z_{s\rightarrow r} + w_{r\rightarrow s}) + (w_{r\rightarrow t} - w_{r\rightarrow 1}) \\ &= z_{s\rightarrow s} + (w_{r\rightarrow t} - w_{r\rightarrow 1}),\ t \in \{1, ..., T\}. \end{aligned} \tag{12}$$

The first term in Eq. (12), $z_{s\rightarrow s}$ indicates the reconstruction of x_s, while the second term $(w_{r\rightarrow t} - w_{r\rightarrow 1})$ represents the motion from x_1 to x_t. This equation indicates that the original pose is preserved in x_s, at the same time motion is transferred from V_d. We note that in order to completely replicate the position and pose in V_d, it requires x_s and x_1 to contain similar poses in relative motion transfer.

E. Model architecture

We show architecture details of E and G for 256px model in Fig. 4 and Fig. 5 respectively. E contains several ResBlocks to downsample the input images into multi-scale feature maps and latent motion codes. We take feature maps of spatial size from 8×8 to 256×256 as our appearance features x_i^{enc}. We use MLPs to obtain the magnitude vector $A_{d\rightarrow r}$ from $z_{d\rightarrow r}$. G consists of two components, a flow field generator G_f and a refinement network G_r. We adapt style blocks (modulation, convolution, and normalization) proposed in StyleGAN2 in G_f. Latent code $z_{s\rightarrow t}$ works as the style code in convolutional layer. The refinement network G_r upsamples and refines inpainted feature maps to target resolution. Flow field ϕ_i and corresponding mask m_i are generated from each G Block.

For high-resolution generation, we keep the model architecture, as well as the amount of directions the same as in the low-resolution model. However, we add an additional layer in both, E and G, respectively. We observe that for high-resolution generation, using the original design of LIA requires longer time to reconstruct details of the original target image during training, and hardly achieves high-quality photorealistic results. Therefore, we propose to reuse the information from the original source image, employing it as an additional feature map. Towards this, we involve additional layers to warp and in-paint the original source image in G. We find that this modification enables a more efficient learning, as well as an improved visual quality. We have quantitatively and qualitatively evaluated the high-resolution model in Sec. IV-D.

IV. Experiments

In this section, we firstly describe our experimental setup including implementation and dataset details. Secondly, we qualitatively demonstrate generated results on both 256×256 and 512×512 resolutions on VoxCeleb datasets. Then, we provide quantitative evaluation *w.r.t.* image quality on (a) same-identity reconstruction, (b) cross-video motion transfer, presenting (c) a user study. Next, we conduct an ablation study to demonstrate (d) the effectiveness of our proposed motion dictionary, as well as (e) the chosen dictionary size. Finally, we provide an in-depth analysis of our (f) latent codes and (g) motion dictionary.

A. Experimental setup

Datasets. We conduct experiments on four datasets, VoxCeleb [40], CelebV-HQ [91], TaichiHD [52] and TED-talk [55]. Following FOMM [52], we first download the original videos based on the provided YouTube-IDs and then process each frame into 256×256 resolution images for quantitative evaluation. In addition, we re-download and process a high-resolution version of VoxCeleb to create a new dataset, namely VoxCelebHQ towards qualitatively and quantitatively demonstrating the capacity of LIA on high resolution (512×512) image animation.

- **VoxCeleb** consists of a large-scale interview videos of different celebrities. In total, it contains 17928 and 495 videos in training and testing set respectively.

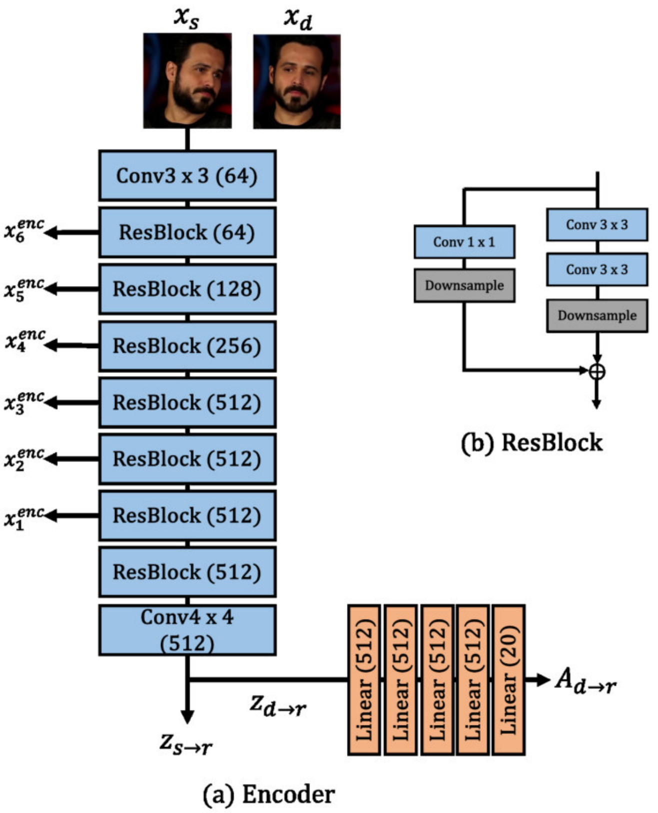
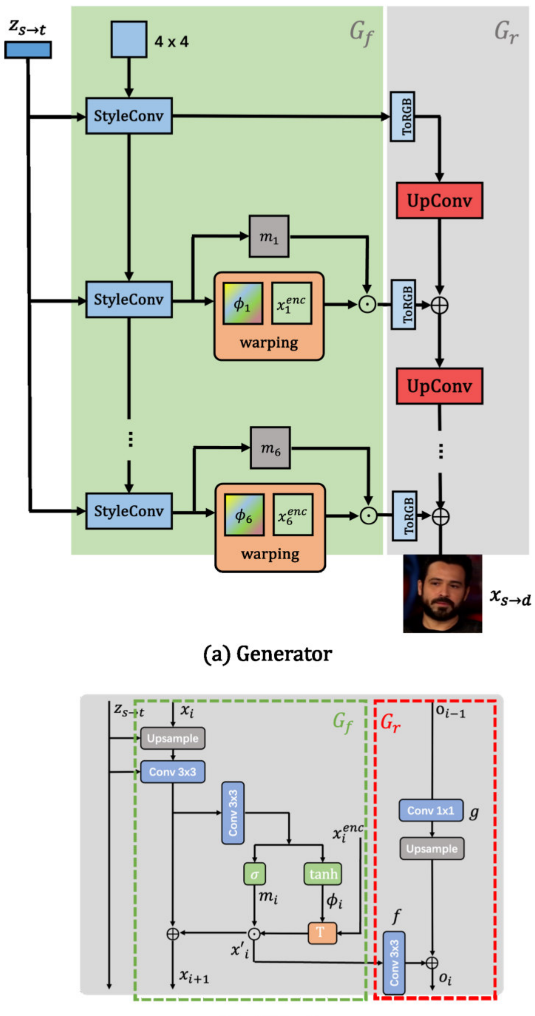


Fig. 4: **Encoder architecture.** (a) E contains 7 ResBlocks to downsample the input images into multi-scale feature maps and latent motion codes. We take feature maps of spatial size from 8×8 to 256×256 as our appearance features x_i^{enc}. We use a 5-layer MLP to obtain a magnitude vector $A_{d\rightarrow r}$ from $z_{d\rightarrow r}$. (b) Each ResBlock downsamples the resolution of previous feature maps to half resolution.

- **CelebV-HQ** contains 35,666 video clips involving 15,653 identities and 83 manually labeled facial attributes covering appearance, action, and emotion. The resolution of all the videos are 512×512.
- **TaiChiHD** consists of videos of full human bodies performing Tai Chi action. It contains 1096 training videos and 115 testing videos.
- **TED-talk** is proposed in MRAA which comprises TED-talk videos. The speakers are cropped out based on the detected bounding boxes. It includes 1124 training videos and 130 testing videos.
- **VoxCelebHQ** is a new dataset based on the YouTube-IDs provided by FOMM. We first filter out the videos whose resolutions are lower than 1080p. Faces are then cropped out in each frame based on the detected bounding boxes. In total, there are around 15000 cropped video clips with a resolution higher than 512×512. Towards adding more diversity in the training set, we keep the videos with a resolution larger than 400×400, disregarding the rest, which results in 17995 videos consisting of 434 identities in total. The video duration ranges from 2s to 80s. On average, the duration is around 8s. All cropped clips are



(a) Generator

(b) G Block

Fig. 5: **Generator architecture.** (a) G consists of two components, a flow field generator G_f and a refinement network G_r. We apply style blocks (modulation, convolution, and normalization) proposed in StyleGAN2 in G_f. Latent code $z_{s\rightarrow t}$ works as the style code in convolutional layer. The refinement network G_r upsamples and refines inpainted feature maps to target resolution. (b) Flow field ϕ_i and corresponding mask m_i are generated from each G Block.

resized to 512×512 resolution. Therefore, the training set contains 17512 videos, whereas the testing set comprises 483 videos.

Implementation details. We implement our model using PyTorch [44]. All models are trained on 4 NVIDIA V100 GPUs. For generating videos with 256 × 256 resolution, the total batch size is 32 with 8 images per GPU. We use a learning rate of 0.002 to train our model with the Adam optimizer [34]. The dimension of latent codes, as well as directions in D_m is set to be 512. In final loss function, we set $\lambda = 10$ in order to penalize more on the perceptual loss. It takes approximate 150 hours to fully train the model. For our high-resolution model, we train LIA on the same number of GPUs but with smaller batch size, *i.e.,* 16 with 4 images per GPU. It takes around 250 hours for the model to fully converge.

Evaluation metrics. We evaluate our model *w.r.t.* (i) reconstruction faithfulness using $\mathcal{L}_1$, LPIPS, (ii) generated video quality using video FID, as well as (iii) semantic consistency using average keypoint distance (AKD), missing keypoint rate (MKR) and average euclidean distance (AED).

- $\mathcal{L}_1$ represents the mean absolute pixel difference between reconstructed and real videos.
- **LPIPS** [86] aims at measuring the perceptual similarity between reconstructed and real images by leveraging the deep features from AlexNet [35].
- **FID** is the video version from original FID [23]. We here follow the same implementation as [74] and utilize a pre-trained 3D ResNext101 [20] to extract spatio-temporal features to compute the distance between real and generated videos distributions. We take the first 100 frames of each video as input of the feature-extractor to compute the scores.
- **Average Keypoint Distance (AKD) and Missing Keypoint Rate (MKR)** evaluate the difference between keypoints of reconstructed and ground truth videos. We extract landmarks using the face alignment approach [8] and extract body poses for both TaiChiHD and TED-talks using OpenPose [10]. AKD is computed as the average distance between corresponding keypoints, whereas MKR is the proportion of keypoints present in the ground-truth that are missing in a reconstructed video.
- **Average Euclidean distance (AED)** measures the ability of preserving identity in reconstructed video. We use a person re-identification pretrained model [90] for measuring human bodies (TaichiHD and TED-talk) and OpenFace [2] for faces to extract identity embeddings from reconstructed and ground truth frame pairs, then we compute MSE of their difference for all pairs.

B. Qualitative results

Firstly, as shown in Fig. 6, we qualitatively evaluate the ability of LIA to generate realistic videos and compare related results with two state-of-the-art, FOMM and MRAA. For VoxCeleb dataset, we conduct a cross-dataset generation experiment, where we transfer motion from videos in VoxCeleb to images in FFHQ dataset. We observe that our method outperforms FOMM and MRAA *w.r.t.* image quality, as both approaches visibly deform the shape of the original faces. This is specifically notable in the case that source and driving images entail large pose variations. At the same time, LIA is able to successfully tackle this challenge and no similar deformations are visible. For TaichiHD and TED-talk datasets, we conduct experiments related to cross-video generation. Corresponding results confirm that our method is able to correctly transfer motion on articulated human bodies, in the absence of explicit structure representations.

Secondly, for high-resolution face animation, we pretrain LIA on VoxCelebHQ and then train the model on CelebV-HQ. We found that the pre-trained process enables LIA to obtain strong prior knowledge of faces and help the fine-tuning process to quickly obtain good visual quality and increase the

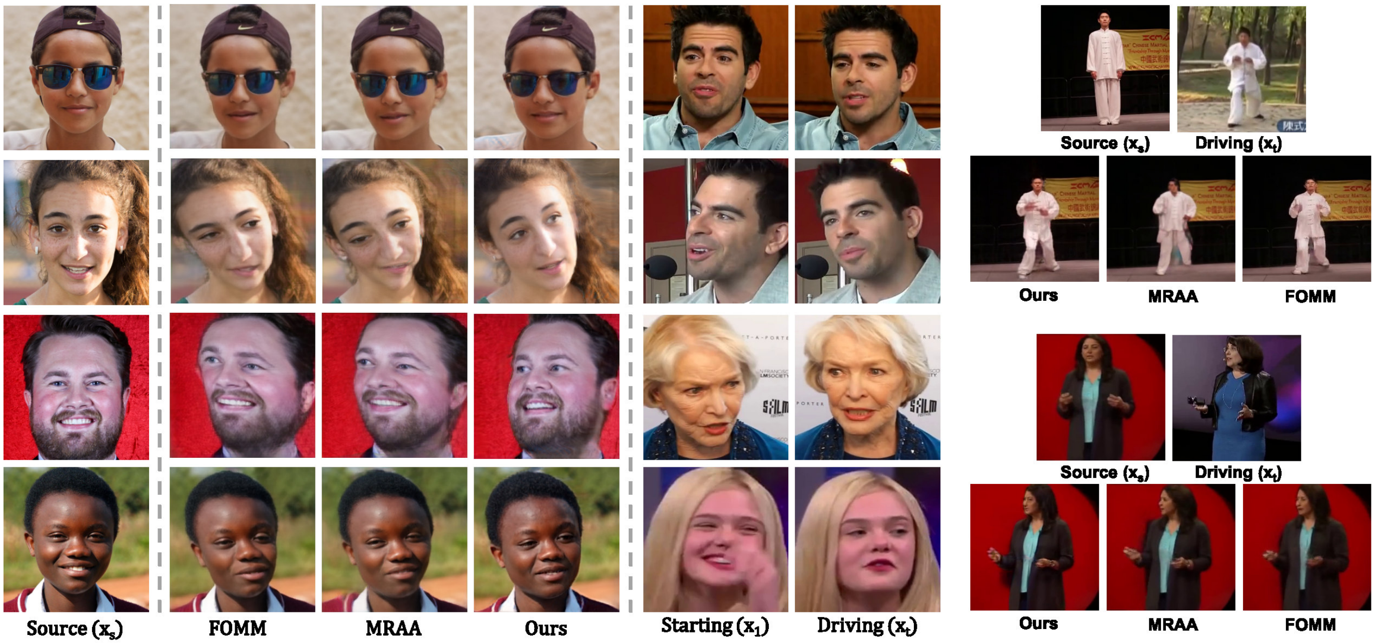


Fig. 6: **Qualitative results.** Examples for same-dataset *absolute motion transfer* on TaichiHD (top-right) and TED-talk (bottom-right). On VoxCeleb (left), we demonstrate cross-dataset *relative motion transfer*. We successfully transfer motion between x_1 and x_t from videos in VoxCeleb to x_s from FFHQ, the latter not being used for training.

TABLE I: **Same-identity reconstruction.** Comparison with state-of-the-art methods on three datasets for same-identity reconstruction.

	VoxCeleb				TaichiHD				TED-talks			
Method	$\mathcal{L}_1$	AKD	AED	LPIPS	$\mathcal{L}_1$	(AKD, MKR)	AED	LPIPS	$\mathcal{L}_1$	(AKD, MKR)	AED	LPIPS
X2Face	0.078	7.687	0.405	-	0.080	(17.654, 0.109)	-	-	-	-	-	-
Monkey-Net	0.049	1.878	0.199	-	0.077	(10.798, 0.059)	-	-	-	-	-	-
FOMM	0.046	1.395	0.141	0.136	0.063	(6.472, 0.032)	0.495	0.191	0.030	(3.759, 0.0090)	0.428	0.13
MRAA *w/o* bg	0.043	1.307	0.140	0.127	0.063	(5.626, 0.025)	0.460	0.189	0.029	(**3.126**, **0.0092**)	**0.396**	0.12
MRAA	**0.041**	**1.303**	**0.135**	0.124	**0.045**	(5.551, 0.025)	0.431	**0.178**	**0.027**	(**3.107**, **0.0093**)	**0.379**	**0.11**
TPS	0.042	1.325	0.141	0.126	0.062	(5.512, 0.027)	0.448	0.019	0.029	(3.203, 0.0096)	0.400	0.12
LIA (256px)	**0.041**	1.353	0.138	**0.123**	0.057	(**4.823**, **0.020**)	**0.431**	0.180	**0.027**	(3.141, 0.0095)	0.399	**0.11**

TABLE II: **Same-identity reconstruction on high-resolution generation.** Comparison with MRAA on VoxCelebHQ for same-identity reconstruction.

	$\mathcal{L}_1$	AKD	AED	LPIPS
MRAA (512px)	0.051	1.385	0.141	0.130
LIA (512px)	**0.033**	**1.211**	**0.125**	**0.115**

TABLE III: **Cross-video generation.** We report video FID for both inner- and cross-dataset tasks on VoxCeleb, VoxCelebHQ and GermanPublicTV.

	VoxCeleb	GermanPublicTV	VoxCelebHQ
FOMM	0.323	0.456	-
MRAA (256px)	0.308	0.454	-
LIA (256px)	**0.161**	**0.406**	-
MRAA (512px)	-	0.932	0.811
LIA (512px)	-	**0.425**	**0.231**

TABLE IV: **User study.** We ask 30 human raters to conduct a subjective video quality evaluation. Results show that LIA outperforms other methods by a large margin.

	VoxCeleb(%)	TaichiHD(%)	TED-talk(%)
Ours/FOMM	**92.9**/7.1	**64.5**/35.5	**71.4**/28.6
Ours/MRAA	**89.7**/10.3	**60.7**/39.9	**54.8**/45.2

generalizability of the entire model. Fig. 1 shows the results from the first stage and Fig. 7 and Fig. 8 show the results from the second stage. As can be clearly seen, our model performs well on high-resolution data and motion can be well transferred from target videos to source images. In addition, after the fine-tuning stage, LIA is able to conduct zero-shot animation for different types of images such as cartoon, black and white film and even oil painting.

We also compare the results between high-resolution model and low-resolution model using the same source image and driving video. As illustrated in Fig. 9, our improved high-resolution LIA is able to better preserve the identity from original source image, as we utilize the original source image. Even with large deformation, the improved LIA outperforms

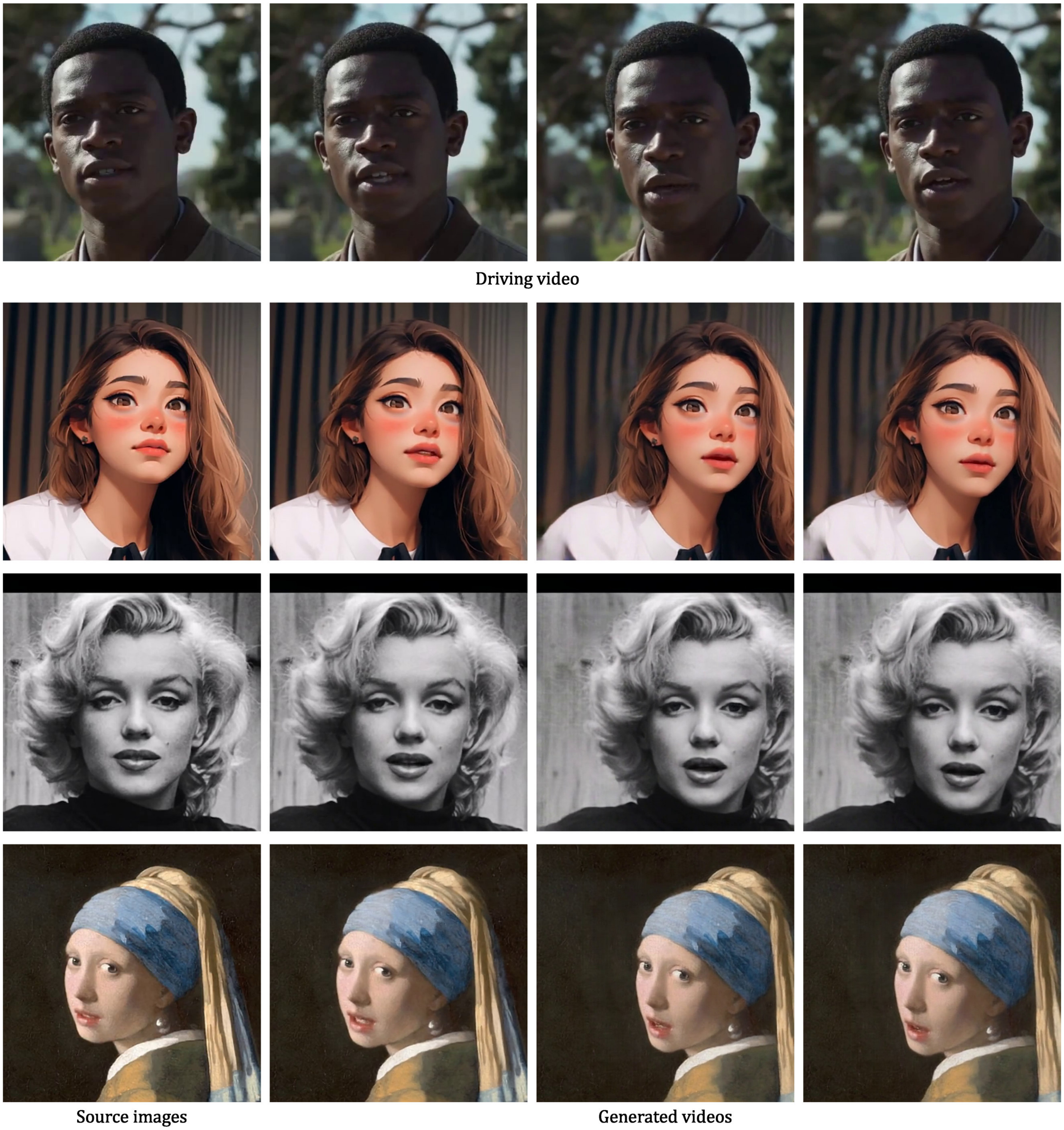


Fig. 7: Zero-shot high-resolution (512×512) image animation. LIA is fine-tuned on CelebV-HQ.

the low-resolution model *w.r.t.* visual quality.

C. Quantitative evaluation

We quantitatively compare low-resolution LIA with the state-of-the-art approaches, namely X2Face [79], MonkeyNet [53], FOMM [52], MRAA [55] and TPS [88] on two tasks, (a) same-identity reconstruction and (b) cross-video motion transfer. Additionally, we conduct a (c) user study towards subjectively analyzing results of our generated results.

(a) Same-identity reconstruction. We firstly evaluate the reconstruction ability of our method. Specifically, we reconstruct each testing video by using the first frame as x_s and the remaining frames as x_d. Results of low-resolution LIA on three datasets are reported in Table I. Our method outperforms other approaches *w.r.t.* all metrics. While we do not explicitly

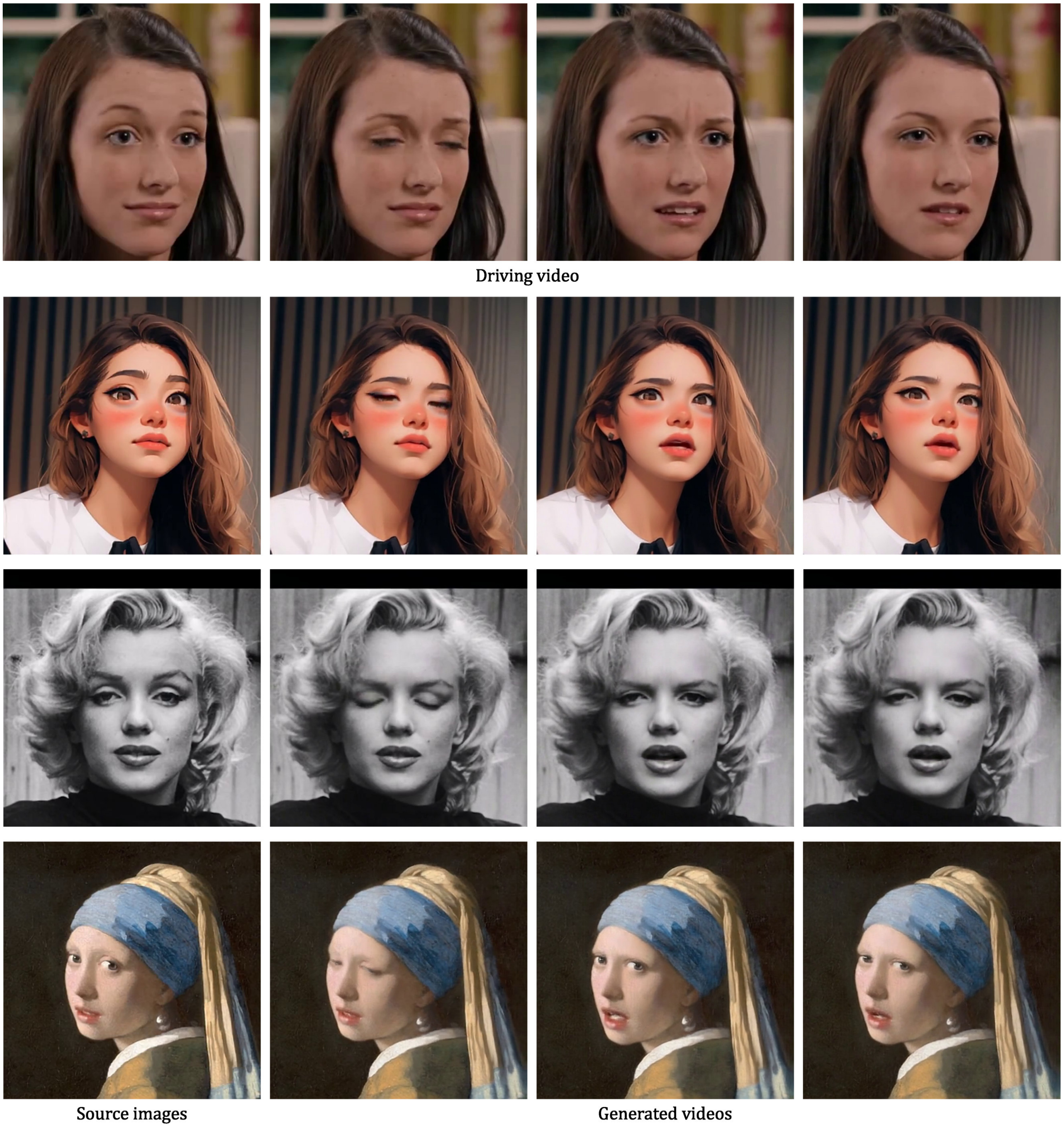


Fig. 8: Zero-shot high-resolution (512×512) image animation. LIA is fine-tuned on CelebV-HQ.

predict keypoints, *w.r.t.* the TaichiHD dataset, interestingly we outperform MRAA in both, AKD and MKR. Such results showcase the effectiveness of our proposed method on modeling articulated human structures. However, we notice that current evaluation metrics cannot provide a completely fair comparison on how well the main subjects (*e.g.,* faces and human bodies) are generated in videos. This is in particular the case for TaichiHD and TED-talk, where reconstructing backgrounds has large contributions to the final scores. We also report comparison results with MRAA on VoxCelebHQ in Table II. Results show that our proposed method outperforms MRAA by a large margin, which demonstrates that the improved LIA is effective in high-resolution generation.

(b) Cross-video motion transfer. Next, we conduct experiments, where source images and driving videos stem from different video sequences. In this context, we mainly focus

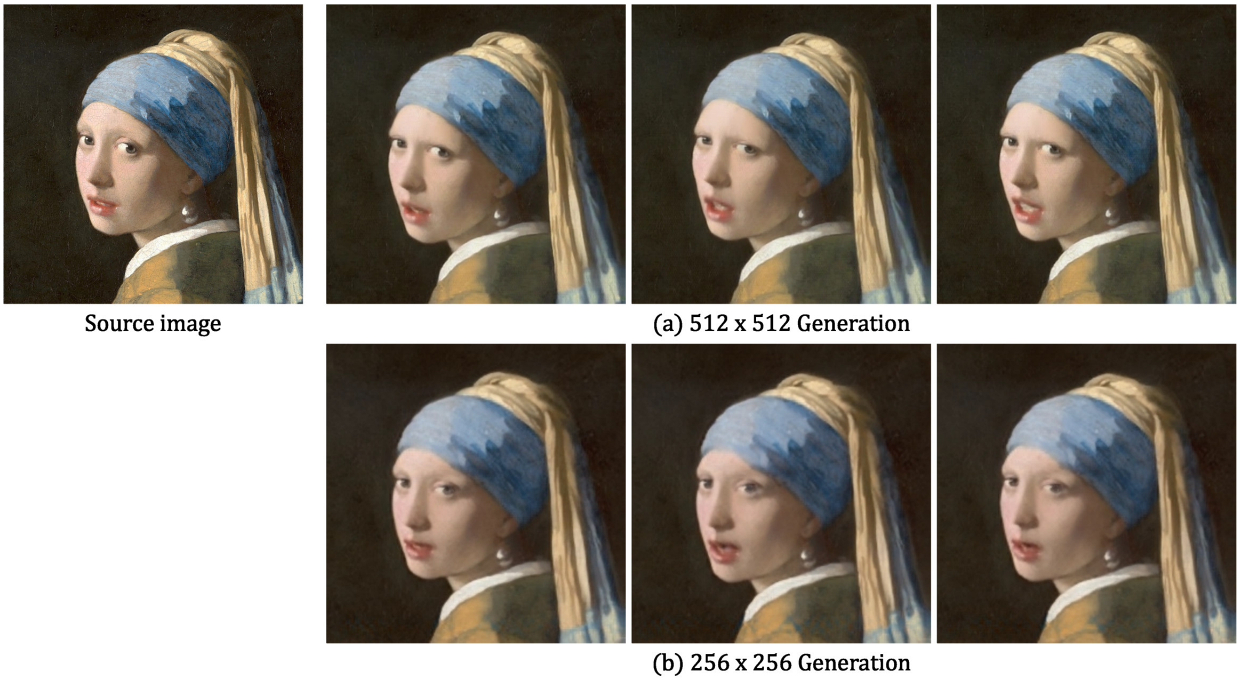


Fig. 9: We show results using the same source image and driving video from high-resolution model (up row) and low-resolution model (bottom row) respectively.

on evaluating *talking head* videos and explore two different tasks. In the first task, we generate videos using the VoxCeleb testing set to conduct inner-dataset *cross-identity motion transfer*. In the second task, source images are from an unseen dataset, namely the GermanPublicTV dataset, as we conduct *cross-dataset motion transfer*. We note that the model we use in both cases are *only* trained on VoxCeleb dataset. In both experiments, we randomly construct source and driving image pairs and transfer motion from driving videos to source images to generate a novel manipulated dataset. We adopt this experimental protocol for evaluating our high-resolution model, however utilizing VoxCelebHQ instead of VoxCeleb. Since there are no ground truth videos for our generated results in both tasks, we choose to use video FID (as initialized by [74]) to measure the distance between generated and real data distributions. As shown in Tab. III, our method outperforms all other approaches *w.r.t.* video FID, indicating the best generated video quality.

(c) User study. We finally conduct a user study to evaluate generated video quality. In particular, we focus on evaluating the quality of cross-video motion transfer, as it represents the real-world application. Towards achieving fair evaluation, we displayed paired videos generated by different approaches and asked 30 human raters the same question ‘which clip is more realistic?’. Each video-pair contains a generated video from our method, as well as a video generated from FOMM or MRAA. Results suggest that our results are more realistic in comparison to FOMM and MRAA across all three datasets (see Tab. IV). Hence, the obtained human preference is in accordance with our quantitative evaluation.

D. Ablation study

We here analyze our proposed motion dictionary and focus on answering following two questions.

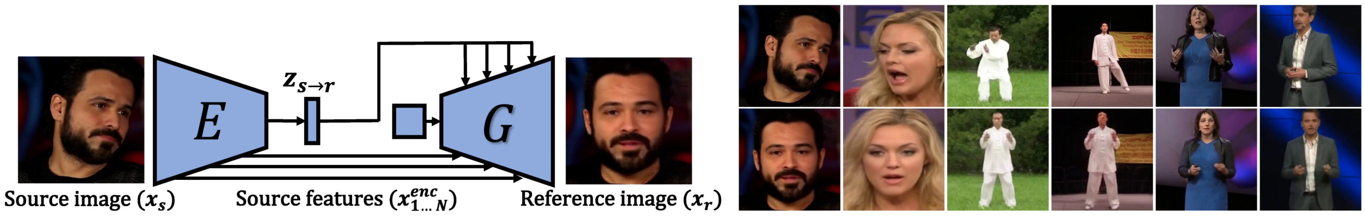
(d) Is the motion dictionary D_m beneficial? We explore the impact of proposed D_m, by training our model without D_m. Specifically, we output $w_{r\rightarrow d}$ directly from MLP, without using LMD to learn an orthogonal basis. From the evaluation results reported in Tab. VI and qualitative results in Fig. 12, we observe that in the absence of D_m, appearance information is undesirably transferred from driving videos to generated videos, which proves the effectiveness of D_m, consistently on all datasets.

(e) How many directions are required in D_m? Towards finding an effective size of D_m, we empirically test three different size, viz. 5, 10, 20, 40 and 100. Quantitative results in Tab. VII show that when using 20 directions, LIA achieves the best reconstruction results on all three datasets.

However, we note that when employing more than 20 motion directions, the difference of perceptual errors becomes minor for all datasets. We note that adding additional motion directions in the latent space will contribute to an improvement up to a certain level, after which, adding more components in the latent space will not bring to the fore difference in the results.

We postulate that the number of directions in the motion dictionary has strong correlation with the motion complexity occurring in the training datasets. However, to fully analyze such a phenomenon, the design of the dataset necessitates the inclusion of different motion complexity. We will explore this in future work.

(f) Effectiveness of using the source image as last feature map. Towards demonstrating the effectiveness of utilizing



(a) (b)

Fig. 10: **Visualization of *reference images*.** (a) Given a source image x_s, without using any driving video, LIA is able to obtained a reference image x_r of the same identity. (b) Example source (top) and reference (down) images from VoxCeleb, TaichiHD and TED-talk datasets. Our network learns *reference images* of a consistently frontal pose, systematically for all input images of each dataset.

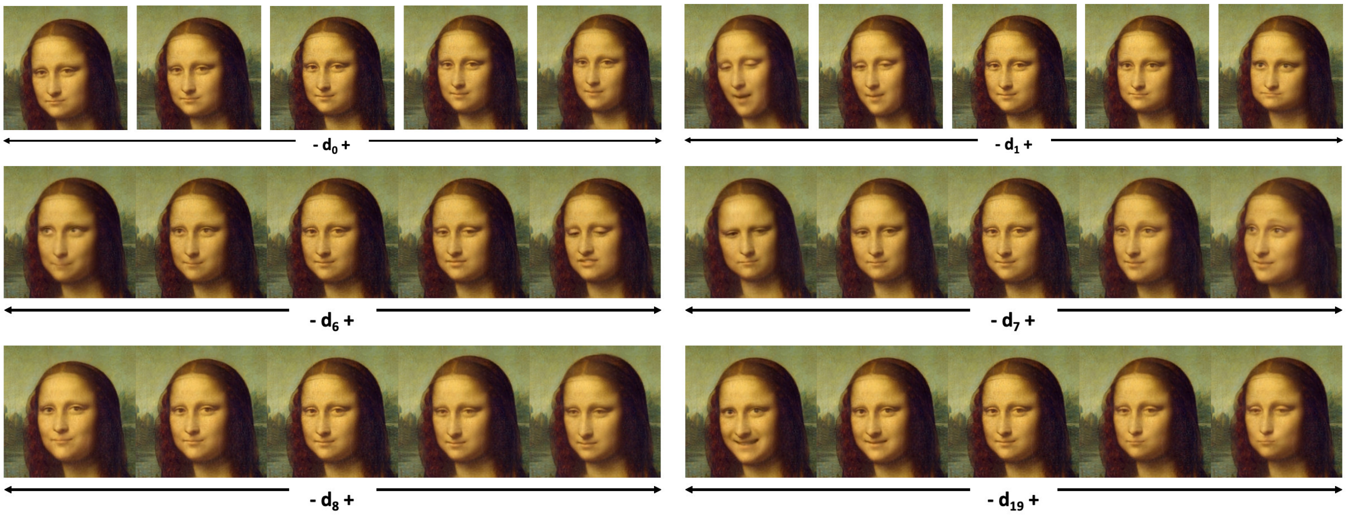


Fig. 11: **Linear manipulation of four motion directions on the painting of Mona Lisa.** Manipulated results indicate that d_6 represents *eye movement*, d_8 represents *head nodding*, whereas d_{19} and d_7 represent facial expressions.

the original source image as last feature map, we conduct an ablation study, in order to compare the original LIA design with the improved high-resolution model. Table V shows the quantitative comparison on VoxCelebHQ dataset for same-identity reconstruction. As can be seen, for high-resolution generation, directly warping source image is able to achieve a significantly improved reconstruction results leading to improved visual quality.

TABLE V: **Ablation study on high-resolution models.** We compare LIA using original and improved design on VoxCelebHQ dataset.

	$\mathcal{L}_1$	AKD	AED	LPIPS
LIA (original)	0.041	1.311	0.131	0.120
LIA (improved)	**0.033**	**1.211**	**0.125**	**0.115**

E. Further analysis

(f) Reference space analysis. While our method successfully transfers motion via latent space navigation, we here aim at answering the question — *what does x_r represent*? Towards answering this question, we proceed to visualize x_r.

TABLE VI: **Ablation study on motion dictionary.** We conduct experiments on three datasets with and without D_m and show reconstruction results.

	VoxCeleb		TaichiHD		TED-talks	
Method	$\mathcal{L}_1$	LPIPS	$\mathcal{L}_1$	LPIPS	$\mathcal{L}_1$	LPIPS
w/o D_m	0.049	0.165	0.062	0.186	0.031	0.12
Full	**0.041**	**0.123**	**0.057**	**0.180**	**0.028**	**0.11**

TABLE VII: **Ablation study on D_m size.** We conduct experiments on three datasets with 5 different D_m size and show reconstruction results.

	VoxCeleb		TaichiHD		TED-talks	
M	$\mathcal{L}_1$	LPIPS	$\mathcal{L}_1$	LPIPS	$\mathcal{L}_1$	LPIPS
5	0.051	0.15	0.070	0.22	0.037	0.15
10	0.043	0.13	0.065	0.20	0.036	0.13
20	**0.041**	**0.12**	**0.057**	**0.18**	**0.028**	**0.11**
40	0.042	**0.12**	0.060	0.19	0.030	0.12
100	**0.041**	**0.12**	0.058	**0.18**	**0.028**	**0.11**

As shown in Fig. 10, we use G to decode $z_{s\rightarrow r}$ into the flow field $\phi_{s\rightarrow r}$, which is used to warped x_s into x_r. Fig. 10 shows examples of x_s and x_r. Interestingly, we observe that x_r represents the *canonical pose* of x_s, regardless of original poses of the subjects. And for all datasets, reference images resemble each other *w.r.t.* pose and scale. As such reference

TABLE VIII: **Comparison between models using orthogonal and non-orthogonal basises.** Results show that using orthogonal basis slightly outperforms the other model *w.r.t.* image quality.

	VoxCeleb		TaichiHD		TED-talks	
	$\mathcal{L}_1$	LPIPS	$\mathcal{L}_1$	LPIPS	$\mathcal{L}_1$	LPIPS
Random basis	0.044	0.13	0.060	**0.18**	0.030	0.12
Orthogonal basis	**0.041**	**0.12**	**0.057**	**0.18**	**0.028**	**0.11**

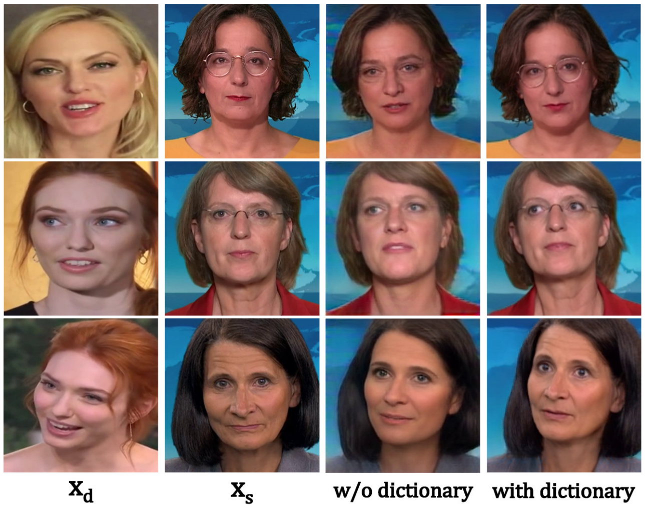


Fig. 12: **Generated results with and without D_m.** We illustrate results by transferring motion from VoxCeleb (x_d) to GermanPublicTV (x_s) with and without motion dictionary. We observe that the disentanglement of appearance and motion is much better by using D_m.

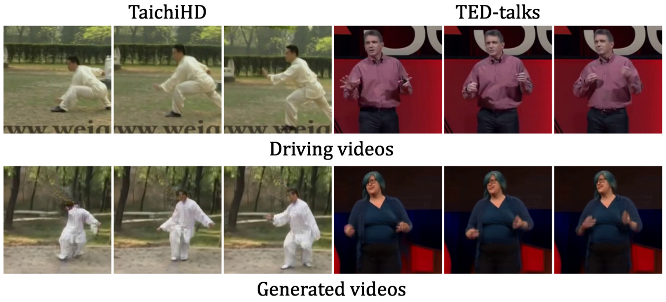


Fig. 13: **Failure cases.** We observed that it is still challenging for LIA to handle arm-leg occlusion (Taichi) and hand motion (TED-talk).

images can be considered as a normalized form of x_s, learning transformations between x_s and x_d using $x_s \to x_r \to x_d$ is considerably more efficient than $x_s \to x_d$, once x_r is fixed.

Noteworthy, we found the similar idea of learning a 'reference image' has also been explored by FOMM [52] and X2Face [79]. However, deviating from our *visualized* 'reference image', the 'reference image' in FOMM refers to a non-visualized and abstract concept. In addition, LIA only requires a latent code $z_{s \to r}$, rather than the 'reference image' for both, training and testing, which is contrast to X2Face as well.

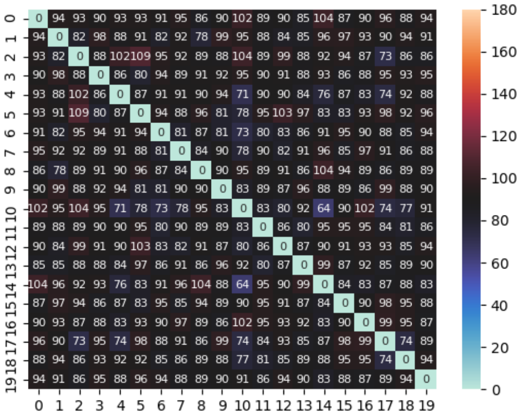


Fig. 14: **Analysis on non-orthogonal basis.** Cross-direction angle (degree) analysis on VoxCeleb.

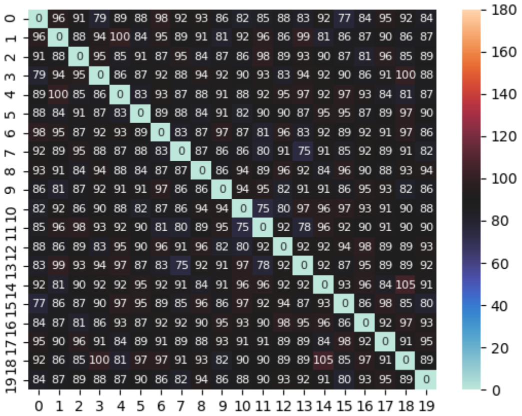


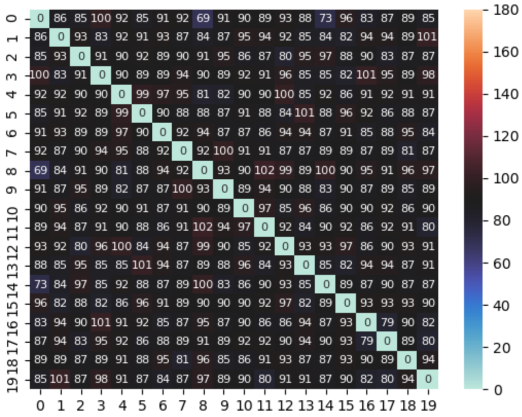
Fig. 15: **Analysis on non-orthogonal basis.** Cross-direction angle (degree) analysis on TaichiHD.

(g) Motion dictionary interpretation. Towards further interpretation of directions in D_m, we conduct linear manipulations on each d_i. Images pertained to manipulating four motion directions are depicted in Fig. 11. The results suggest that the directions in D_m are semantically meaningful, as they represent basic visual transformations such as head nodding (d_8), eye movement (d_6) and facial expressions (d_{19} and d_7). More results can be found on our project webpage[1].

(h) Orthogonality analysis in motion dictionary. We provided comparison in Tab. VIII to analyze the effectiveness of *orthogonality* in the motion dictionary. In non-orthogonality experiments, we randomly initialize 20 directions without using Gram-Schmidt process.

Tab. VIII indicates that models using an orthogonal basis perform better than the ones using a non-orthogonal basis, which suggests that such constrain enables a model to converge faster and facilitates the decomposition of different

[1] https://wyhsirius.github.io/LIA-project/



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	86	85	100	92	85	91	92	69	91	90	89	93	88	73	96	83	87	89	85
1	86	0	93	83	92	91	93	87	84	87	95	94	92	85	84	82	94	94	89	101
2	85	93	0	91	90	92	89	90	91	95	86	87	80	95	97	88	90	83	87	87
3	100	83	91	0	90	89	89	94	90	89	92	91	96	85	85	82	101	95	89	98
4	92	92	90	90	0	99	97	95	81	82	90	90	100	85	92	86	91	92	91	91
5	85	91	92	89	99	0	90	88	88	87	91	88	84	101	88	96	92	86	88	87
6	91	93	89	89	97	90	0	92	94	87	87	86	94	94	87	91	85	88	95	84
7	92	87	90	94	95	88	92	0	92	100	91	91	87	87	89	89	87	89	81	87
8	69	84	91	90	81	88	94	92	0	93	90	102	99	89	100	90	95	91	96	97
9	91	87	95	89	82	87	87	100	93	0	89	94	90	88	83	90	87	89	85	89
10	90	95	86	92	90	91	87	91	90	89	0	97	85	96	86	90	90	92	86	90
11	89	94	87	91	90	88	86	91	102	94	97	0	92	84	90	92	86	92	91	80
12	93	92	80	96	100	84	94	87	99	90	85	92	0	93	93	97	86	90	93	91
13	88	85	95	85	85	101	94	87	89	88	96	84	93	0	85	82	94	94	87	91
14	73	84	97	85	92	88	87	89	100	83	86	90	93	85	0	89	87	90	87	87
15	96	82	88	82	86	96	91	89	90	90	90	92	97	82	89	0	93	93	93	90
16	83	94	90	101	91	92	85	87	95	87	90	86	86	94	87	93	0	79	90	82
17	87	94	83	95	92	86	88	89	91	89	92	92	90	94	90	93	79	0	89	80
18	89	89	87	89	91	88	95	81	96	85	86	91	93	87	87	93	90	89	0	94
19	85	101	87	98	91	87	84	87	97	89	90	80	91	91	87	90	82	80	94	0

Fig. 16: **Analysis on non-orthogonal basis.** Cross-direction angle (degree) analysis on TED-talks.

motion modes in the latent space. By manipulating motion directions (see Fig. 11), we have that different directions tend to represent different motion mode. To further analyze the correlation between each direction after training, we also computed angles between direction pairs in the motion dictionary for three datasets and show results in Fig. 14, Fig. 15 and Fig. 16.

We observe that most of the directions are nearly orthogonal after training and angles are 90 ± 10. Therefore we consider on current three training datasets that employing an orthogonal basis is effective. However, there are other factors that may affect such choice, e.g., complexity of the datasets and latent code dimension. It remains an open question whether orthogonality is a general solution for all cases or just in our context.

F. Limitations

For human body, one limitation of our method has to do with body occlusion. As shown in Fig. 13, in taichi videos and in case of occlusion by legs and arms, motion is not transferred successfully. In addition, in TED-talks, transferring hand motion is challenging, as hands are of small size, articulated and sometimes occluded by human bodies. Involving implicit 3D reprsentation might be a solution for such limitations. We leave related investigation for our future work.

V. Conclusions

In this paper, we introduced a novel Latent Image Animator (LIA), aimed at animating images via latent space navigation. By the proposed Linear Motion Decomposition (LMD), we were able to formulate the task of transferring motion from driving videos to source images as learning linear transformations in the latent space. We evaluated proposed method on real-world videos and demonstrated that our approach is able to successfully animate still images, while eliminating the necessity of *explicit structure representations*. We extended LIA for high resolution image animation and showcased that for zero-shot image animation, LIA is able to achieve high-quality results. In addition, we showed that the incorporated motion dictionary is interpretable and contains directions pertaining to basic visual transformations. Both quantitative and qualitative evaluations showed that LIA outperforms state-of-art algorithms on all benchmarks. We postulate that LIA opens a new door in design of interpretable generative models for video generation.

Reproducibility Statement

We assure that all results shown in the paper and supplemental materials can be reproduced. We intend to open-source our code, as well as trained models.

Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocation AD011011627R1. It was supported by the French Government, by the National Research Agency (ANR) under Grant ANR-18-CE92-0024, project RESPECT and through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[2] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016. 7

[3] Sarthak Bhagat, Shagun Uppal, Zhuyun Yin, and Nengli Lim. Disentangling multiple features in video sequences using gaussian processes in variational autoencoders. In *ECCV*, 2020. 2

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 1

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 1

[7] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos of dynamic scenes. 2022. 2

[8] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 7

[9] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on graphics (TOG)*, 33(4):1–10, 2014. 1

[10] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 7

[11] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019. 1, 3

[12] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *CVPR*, 2021. 1

[13] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 1

[14] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019. 2

[15] Emily L Denton and vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NeurIPS*, 2017. 2

[16] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, 2022. 2

[17] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *ICCV*, 2023. 1

[18] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *ICCV*, pages 5744–5753, 2019. 1, 3

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

[20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018. 7

[21] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024. 3

[22] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022. 2

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 7

[24] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 2

[27] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 3

[28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*, 2017. 3

[29] Ali Jahanian, Lucy Chai, and Phillip Isola. On the ”steerability” of generative adversarial networks. In *ICLR*, 2020. 1, 3

[30] Yunseok Jang, Gunhee Kim, and Yale Song. Video Prediction with Appearance and Motion Conditions. In *ICML*, 2018. 1

[31] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1, 2

[33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, 2012. 7

[36] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. 2023. 3

[37] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *ECCV*, 2018. 1

[38] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. *ICML*, 2018. 2

[39] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *CVPR*, 2019. 1

[40] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Science and Language*, 2019. 6

[41] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[42] Katsunori Ohnishi, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Hierarchical video generation from orthogonal information: Optical flow and texture. In *AAAI*, 2018. 1

[43] Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Video generation from single semantic label map. *arXiv preprint arXiv:1903.04480*, 2019. 1

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 7

[45] William Peebles, John Peebles, Jun-Yan Zhu, Alexei A. Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 3

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[49] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017. 2

[50] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 1, 3

[51] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 3

[52] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 1, 3, 5, 6, 9, 13

[53] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2019. 9

[54] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Jian Ren, Hsin-Ying Lee, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. In *CVPR*, 2023. 3

[55] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 1, 3, 6, 9

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5

[57] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023. 1

[58] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022. 2

[59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1, 2

[60] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLRs*, 2021. 1

[61] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *ECCV*, 2020. 1

[62] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 1

[63] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *CVPR*, 2016. 1

[64] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *ICLR*, 2021. 2

[65] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In *CVPR*, 2018. 2

[66] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NIPS*, 2016. 2

[67] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. *arXiv preprint arXiv:2002.03754*, 2020. 1, 3

[68] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017. 1

[69] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. In *NeurIPS*, 2019. 1, 3

[70] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018. 1, 3

[71] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 1

[72] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 3

[73] Yaohui Wang. *Learning to Generate Human Videos*. Theses, Inria - Sophia Antipolis ; Université Cote d'Azur, Sept. 2021. 2

[74] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3AN: Disentangling appearance and motion for video generation. In *CVPR*, 2020. 2, 7, 11

[75] Yaohui Wang, Piotr Bilinski, Francois F Bremond, and Antitza Dantcheva. ImaGINator: Conditional Spatio-Temporal GAN for Video Generation. In *WACV*, 2020. 1, 2

[76] Yaohui Wang, Francois Bremond, and Antitza Dantcheva. Inmodegan: Interpretable motion decomposition generative adversarial network for video generation. *arXiv preprint arXiv:2101.03049*, 2021. 2

[77] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 1

[78] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *ICLR*, 2022. 2

[79] Olivia Wiles, A Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686, 2018. 1, 3, 9, 13

[80] Jianwen Xie, Ruiqi Gao, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Motion-based generator model: Unsupervised disentanglement of appearance, trackable and intrackable motions in dynamic patterns. In *AAAI*, 2020. 2

[81] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[82] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *ECCV*, 2018. 1

[83] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. Transmomo: Invariance-driven unsupervised video motion retargeting. In *CVPR*, 2020. 3

[84] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *ICLR*, 2022. 2

[85] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. 1, 3

[86] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

[87] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qing, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. 2023. 3

[88] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, 2022. 3, 9

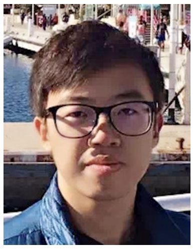
[89] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *ECCV*, 2018. 1

[90] Zhedong Zheng, Tao Ruan, Yunchao Wei, Yi Yang, and Tao Mei. Vehiclenet: Learning robust visual representation for vehicle re-identification. *IEEE Transaction on Multimedia (TMM)*, 2020. 7

[91] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022. 1, 6

[92] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 3

Yaohui Wang is Research Scientist at Shanghai AI Laboratory. He completed his Ph.D. from Inria Sophia Antipolis, STARS Team in 2021. Prior, he received his M.S. degree from Université Paris-Saclay and M.Eng. from ENSIIE in 2017. He obtained his B.S. degree from Xidian University in 2015. His research focuses on generative modeling, video synthesis and representation learning.



Di Yang is a Ph.D. student in Computer Vision at Inria-Sophia Antipolis, France, supervised by Dr. François Brémond. He received his M.Eng. in Computer Vision in 2019 from Télécom Saint-Étienne, France. He earned his B.S. in Communication Engineering in 2017 from Xidian University, China. His research interests include computer vision and deep learning with special interest in video understanding and action recognition.

Francois Bremond is Research Director at Inria Sophia Antipolis, where he has been leading the STARS team since 2012. His research interest includes video understanding, activity recognition and representation learning. He has authored or coauthored more than 140 peer-reviewed publications in the top machine vision and artificial intelligence conferences and journals.

Antitza Dantcheva is Research Director (DR2) with the STARS team of INRIA Sophia Antipolis, France. Previously, she was Marie Curie fellow at Inria and Postdoctoral Fellow at the Michigan State University and the West Virginia University, USA. She received her Ph.D. degree from Télécom ParisTech/Eurecom in image processing and biometrics in 2011. Her research is in computer vision and specifically in designing algorithms that seek to learn suitable representations of the human face in interpretation and generation.