



LEO: Generative Latent Image Animator for Human Video Synthesis

Yaohui Wang¹ · Xin Ma^{1,2} · Xinyuan Chen¹ · Cunjian Chen² · Antitza Dantcheva³ · Bo Dai¹ · Yu Qiao¹

Received: 31 January 2024 / Accepted: 20 August 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Spatio-temporal coherency is a major challenge in synthesizing high quality videos, particularly in synthesizing human videos that contain rich global and local deformations. To resolve this challenge, previous approaches have resorted to different features in the generation process aimed at representing appearance and motion. However, in the absence of strict mechanisms to guarantee such disentanglement, a separation of motion from appearance has remained challenging, resulting in spatial distortions and temporal jittering that break the spatio-temporal coherency. Motivated by this, we here propose LEO, a novel framework for human video synthesis, placing emphasis on spatio-temporal coherency. Our key idea is to represent motion as a sequence of flow maps in the generation process, which inherently isolate motion from appearance. We implement this idea via a flow-based image animator and a Latent Motion Diffusion Model (LMDM). The former bridges a space of motion codes with the space of flow maps, and synthesizes video frames in a warp-and-inpaint manner. LMDM learns to capture motion prior in the training data by synthesizing sequences of motion codes. Extensive quantitative and qualitative analysis suggests that LEO significantly improves coherent synthesis of human videos over previous methods on the datasets TaichiHD, FaceForensics and CelebV-HQ. In addition, the effective disentanglement of appearance and motion in LEO allows for two additional tasks, namely infinite-length human video synthesis, as well as content-preserving video editing. Project page: <https://wyhsirius.github.io/LEO-project/>.

Keywords Video generation · Diffusion models · Human analysis · Deep generative models

Communicated by Kwan-Yee Kenneth Wong.

✉ Yaohui Wang
wangyaohui@pjlab.org.cn

Xin Ma
xin.ma1@monash.edu

Xinyuan Chen
chenxinyuan@pjlab.org.cn

Cunjian Chen
cunjian.chen@monash.edu

Antitza Dantcheva
antitza.dantcheva@inria.fr

Bo Dai
daibo@pjlab.org.cn

Yu Qiao
qiaoyu@pjlab.org.cn

¹ Shanghai Artificial Intelligence Laboratory, 701 Yunjin Road, Shanghai 200030, China

² Monash University, Wellington Road, Clayton, VIC 3800, Australia

1 Introduction

Deep generative models such as generative adversarial networks (GANs) (Goodfellow et al., 2014) and Diffusion Models (Ho et al., 2020; Song et al., 2021) have fostered a breakthrough in video synthesis (Ge et al., 2022; Ho et al., 2022a; Singe et al., 2023; Skorokhodov et al., 2022; Saito et al., 2020; Tulyakov et al., 2018; Vondrick et al., 2016; Villegas et al., 2023; Wang et al., 2020, 2021; Yu et al., 2022), elevating tasks such as text-to-video generation (Singe et al., 2023; Villegas et al., 2023), video editing (Bar-Tal et al., 2022), as well as 3D-aware video generation (Bergman et al., 2022). While existing work has demonstrated promising results w.r.t. frame-wise visual quality, synthesizing videos of strong spatio-temporal coherency, tailored to human videos, containing rich global and local deformations, remains challenging.

³ Inria, Université Côte d'Azur, 2004 Rte des Lucioles, Valbonne 06902, France

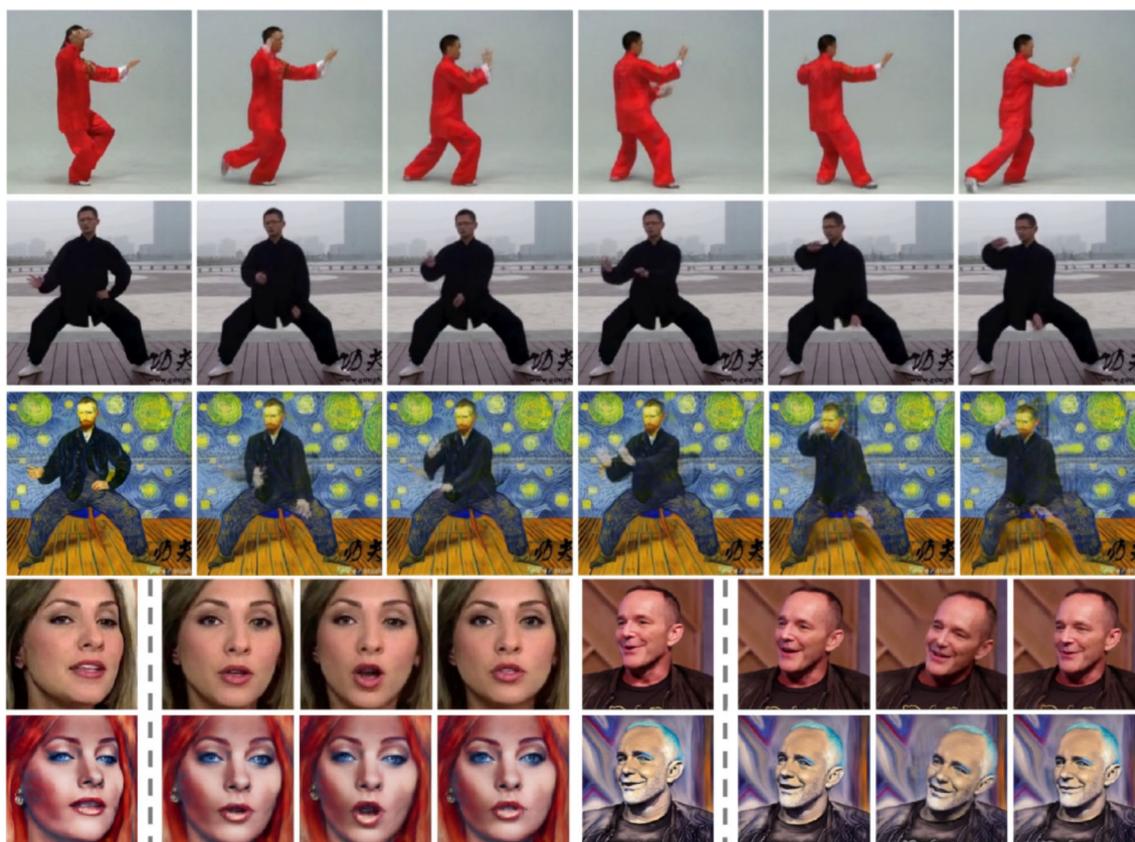


Fig. 1 Our framework caters a set of video synthesis tasks including (i) unconditional video generation (first and second row), (ii) conditional generation based on one single image (fourth row) and (iii) video

editing from the starting image (third and fifth row). Results pertain to our model being trained on the datasets TaichiHD, FaceForensics and CelebV-HQ

Motivated by this, we here propose an effective generative framework, placing emphasis on spatio-temporal coherency in human video synthesis. Having this in mind, a fundamental step has to do with the disentanglement of videos w.r.t. appearance and motion. Previous approaches have tackled such disentanglement by two jointly trained distinct networks, respectively providing appearance and motion features (Tulyakov et al., 2018; Wang et al., 2020, 2021, 2020; Yu et al., 2022), as well as by a two-phase generation pipeline that firstly aims at training an image generator, and then at training a temporal network to generate videos in the image generator's latent space (Ge et al., 2022; Tian et al., 2021; Yan et al., 2021). Nevertheless, such approaches encompass limitations related to spatial artifacts (e.g., distortions of body structures and facial identities in the same sequence), as well as temporal artifacts (e.g., inter-frame semantic jittering), even in short generated videos of 16 frames. We argue that such limitations stem from incomplete disentanglement of appearance and motion in the generation process. Specifically, without predominant mechanisms or hard constraints to guarantee disentanglement, even a minor perturbation in

the high-level semantics will be amplified and will lead to significant changes in the pixel space.

Deviating from the above and towards disentangling videos w.r.t. appearance and motion, in this paper we propose a novel framework for human video generation, referred to as LEO, streamlined to ensure strong spatio-temporal coherency. At the core of this framework is a sequence of flow maps, representing motion semantics, which inherently isolate motion from appearance. Specifically, LEO incorporates a latent motion diffusion module (LMDM), as well as a flow-based image animator. In order to synthesize a video, an initial frame is either provided externally for conditional generation, or obtained by a generative module for unconditional generation. Given such initial frame and a sequence of motion codes sampled from the LMDM, the flow-based image animator generates a sequence of flow maps, and proceeds to synthesize the corresponding sequence of frames in a warp-and-inpaint manner (Fig. 2).

The training of LEO is decomposed into two phases. Firstly, we train the flow-based image animator to encode input images into low-dimensional latent motion codes, and map such codes to flow maps, which are used for recon-

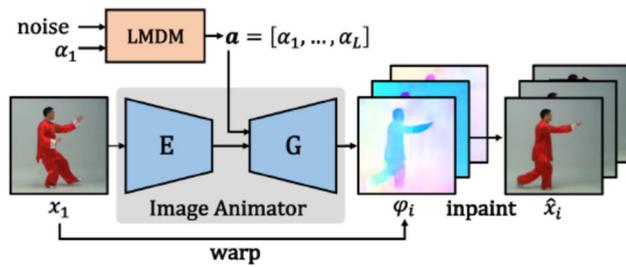


Fig. 2 Inference stage. At the inference stage, LMDM firstly accepts a starting motion code α_1 and a sequence of noise-vectors as input, in order to generate a sequence of motion codes \mathbf{a} , further utilized to synthesize a sequence of flow maps ϕ_i by the pre-trained image animator. The output video is obtained in a warp-and-inpaint manner based on x_1 and ϕ_i

struction via warp-and-inpaint. Therefore, once trained, the flow-based image animator naturally provides a space of motion codes that are strictly constrained to only containing motion-related information. At the second stage, upon the space provided by the image animator, we train the LMDM to synthesize sequences of motion codes and capture motion prior in the training data. To endow LEO with the ability to synthesize videos of arbitrary length beyond the short training videos, we adopt a Linear Motion Condition (LMC) mechanism in LMDM. As opposed to directly synthesizing sequences of motion codes, LMC enables LMDM to synthesize sequences of residuals w.r.t. a starting motion code, in order for longer videos to be easily obtained by concatenating additional sequences of residuals.

To evaluate LEO, we conduct extensive experiments pertained to three human video datasets, including TaichiHD (Siarohin et al., 2019), FaceForensics (Rössler et al., 2018), and CelebV-HQ (Zhu et al., 2022). Compared to previous video synthesis methods, LEO demonstrates a significantly improved spatio-temporal coherency, even on synthesized videos of length of 512 frames. In addition, LEO shows great potential in two extended tasks, namely infinite-length video synthesis, as well as video editing of a style in a synthesized video, while maintaining the content of the original video.

2 Related Works

Unconditional video generation aims to generate videos by learning the full distribution of training dataset. Most of the previous works (Brooks et al., 2022; Clark et al., 2019; Saito et al., 2017; Tulyakov et al., 2018; Vondrick et al., 2016; Wang et al., 2020; Wang, 2021; Wang et al., 2021) are built upon GANs (Brock et al., 2019; Goodfellow et al., 2014; Karras et al., 2019, 2020; Radford et al., 2015) towards benefiting from the strong performance of the image generator. Approaches (Bhagat et al., 2020; Denton et al., 2017; Li & Mandt, 2018; Xie et al., 2020) based on VAEs (Kingma &

Welling, 2014) were also proposed while only show results on toy datasets. Recently, with the progress of deep generative models (e.g., VQVAE (Van Den Oord et al., 2017), VQGAN (Esser et al., 2021), GPT (Radford et al., 2018) and Denoising Diffusion Models (Ho et al., 2020; Nichol & Dhariwal, 2021; Song et al., 2021)) on both image (Ramesh et al., 2021, 2022) and language synthesis (Radford et al., 2019), as well as the usage of large-scale pre-trained models, video generation also started to be explored with various approaches.

MoCoGANHD (Tian et al., 2021) builds the model on top of a well-trained StyleGAN2 (Karras et al., 2020) by integrating an LSTM in the latent space towards disentangling content and motion. DIGAN (Yu et al., 2022) and StyleGAN-V (Skorokhodov et al., 2022) and MoStGAN-V (Shen et al., 2023), inspired by NeRF (Feichtenhofer et al., 2019), proposed an implicit neural representation approach to model time as a continuous signal aiming for long-term video generation. VideoGPT (Yan et al., 2021) and TATS (Ge et al., 2022) introduced to first train 3D-VQ models to learn discrete spatio-temporal codebooks, which are then be refined temporally by modified transformers (Vaswani et al., 2017). Recently, several works (Ho et al., 2022b; Luo et al., 2023; Yu et al., 2023) have shown promising capacity to model complex video distribution by incorporating spatio-temporal operations in Diffusion Models. While previous approaches have proposed various attempts either in training strategies (Tian et al., 2021; Ge et al., 2022; Yan et al., 2021) or in model architectures (Skorokhodov et al., 2022; Wang et al., 2020, 2021; Yu et al., 2022) to disentangle appearance and motion, due to the lack of strong constrains, it is still difficult to obtain satisfying results.

In contrast to unconditional video generation, conditional video generation seeks to produces high-quality videos, following image-to-image generation pipeline (Chu et al., 2017; Huang et al., 2018; Isola et al., 2017). In this context, additional signals such as semantic maps (Pan et al., 2019; Wang et al., 2018, 2019), human key-points (Chan et al., 2019; Jang et al., 2018; Yang et al., 2018; Walker et al., 2017; Wang et al., 2019, 2021; Zakharov et al., 2019), motion labels (Wang et al., 2020), 3DMM (Yang et al., 2022; Zhao et al., 2018) and optical flow (Li et al., 2018; Ohnishi et al., 2018) have been exploited to guide motion generation. In addition, text description, has been used in large-scale video diffusion models (Blattmann et al., 2023a, b; Chen et al., 2023, 2024; Ho et al., 2022a, b; Ma et al., 2024; Menapace et al., 2024; Singe et al., 2023; Wang et al., 2023; Zhang et al., 2023) for high-quality video generation. Our framework also supports for conditional video generation based on a single image. However, unlike previous approaches, our method follows the image animation pipeline (Siarohin et al., 2019, 2021; Wang et al., 2022) which leverages the dense flow maps for

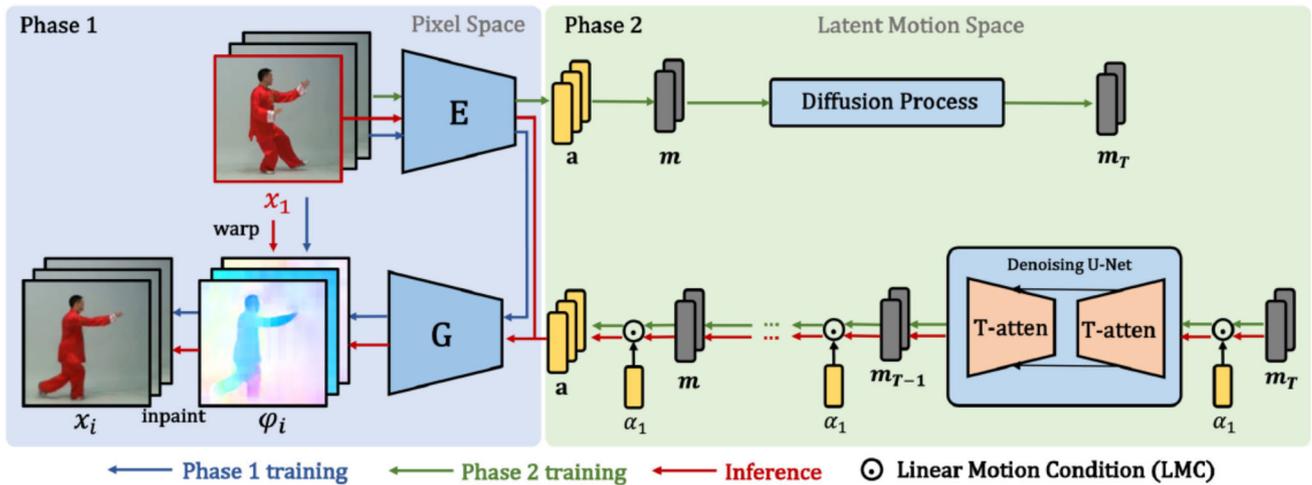


Fig. 3 Overview of LEO. Our framework incorporates two main parts, (i) an image animator, aiming to generate flow maps and synthesize videos in the pixel space, and (ii) Latent Motion Diffusion Model (LMDM), focusing on modeling the motion distribution in a latent motion space. Our framework requires a two-phase training. In the first phase, we train the image animator in a self-supervised manner towards mapping latent codes to corresponding flow maps ϕ_i . Once the image animator is well-trained, motion codes \mathbf{a} are extracted from a frozen encoder and used as inputs of LMDM. In the second phase, LMDMs

are trained to learn the motion distribution by providing the starting motion α_1 as condition. Instead of directly learning the distribution of \mathbf{a} , we adopt a Linear Motion Condition (LMC) mechanism in LMDM towards synthesizing sequences of residuals with respect to x_1 . At the inference stage, given a starting image x_i and corresponding motion code α_i , LMDM firstly generates a motion code sequence, which is then used by the image animator to generate flow maps to synthesize output videos in a warp-and-inpaint manner

motion modeling. We introduce our method in details in the following.

3 Method

Figure 3 illustrates the training of LEO, comprising of two-phases. We firstly train an image animator towards learning high-quality latent motion codes of the datasets. In the second phase, we train the Latent Motion Diffusion Model (LMDM) to learn a motion prior over the latent motion codes. To synthesize a video, the pre-trained image animator takes the motion codes to generate corresponding flow maps, which are used to warp and inpaint starting frame. The warp-and-inpaint operation is conducted in two modules inside image animator. The warping module firstly produces flow fields based on motion codes to warp starting frame, then the inpainting module learns to fill in the holes in the warped starting frame and refine the entire image. Each video sequence is produced frame by frame.

We formulate a video sequence $v = \{x_i\}_{i=1}^L$, $x_i \sim \mathcal{X} \in \mathbb{R}^{3 \times H \times W}$ as $v = \{\mathcal{T}(x_1, G(\alpha_i))\}_{i=2}^L$, $\alpha_i \sim \mathcal{A} \in \mathbb{R}^{1 \times N}$, where x_i denotes the i^{th} frame, α_i denotes a latent motion code at timestep i , G represents the generator in the image animator aiming to generate a flow map ϕ_i from α_i .

3.1 Learning Latent Motion Codes

Towards learning a frame-wise latent motion code, we adopt the state-of-the-art image animator LIA (Wang et al., 2022) as

it enables to encode input images into corresponding motion codes. LIA consists of two modules, an encoder E and a generator G . During training, given a source image x_s and a driving image x_d , E encodes x_s, x_d into a motion code $\alpha = E(x_s, x_d)$, and G generates a flow field $\phi = G(\alpha)$ from the code. LIA is trained in a self-supervised manner with the objective to reconstruct the driving image.

Training LIA in such a self-supervised manner brings two notable benefits for our framework, (i) it enables LIA to achieve high-quality perceptual results, and (ii) as a motion code is strictly equivalent to flow maps, there are guarantees that α is only motion-related without any appearance interference.

3.2 Learning a Motion Prior

Once LIA is well-trained on a target dataset, for any given video $v = \{x_i\}_{i=1}^L$, we are able to obtain a motion sequence $\mathbf{a} = \{\alpha_i\}_{i=1}^L$ with the frozen E . In the second phase of our training, we propose to learn a motion prior by temporal Diffusion Models.

Unlike image synthesis, data in our second phase is a set of sequences. We firstly apply a temporal Diffusion Model for modeling the temporal correlation of \mathbf{a} . The general architecture of this model is a 1D U-Net adopted from Ho et al. (2020). To train this model, we follow the standard training strategy with a simple mean-squared loss,

$$L_{\text{LMDM}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_{\theta}(\mathbf{a}_t, t)\|_2^2 \right], \quad (1)$$

where ϵ denotes the unscaled noise, t is the time step, \mathbf{a}_t is the latent noised motion code to time t . During inference, a random Gaussian noise \mathbf{a}_T is iteratively denoised to $\mathbf{a}_0 = \{\alpha_i\}_{i=1}^L$, and the final video sequence is obtained through the generator.

At the same time in our experiments, we found that learning motion sequences in a complete unconditional manner brings to the fore limitations, namely (i) the generated codes are not consistent enough for producing smooth videos, as well as (ii) the generated motion codes can only be used to produce fixed length videos. Hence, towards addressing those issues, we propose a conditional Latent Motion Diffusion Model (cLMDDM) which aims for high-quality and long-term human videos.

One major characteristic of LIA has to do with the linear motion space. Any motion code α_i in \mathbf{a} can be re-formulated as

$$\alpha_i = \alpha_1 + m_i, i \geq 2, \quad (2)$$

where α_1 denotes the motion code at the first timestep and m_i denotes the motion difference between timestep 1 and i , so that we can re-formulate \mathbf{a} as

$$\mathbf{a} = \alpha_1 + \mathbf{m}, \quad (3)$$

where $\mathbf{m} = \{m_i\}_{i=2}^L$ denotes the motion difference sequence. Therefore, Eqs. 2 and 3 indicate that a motion sequence can

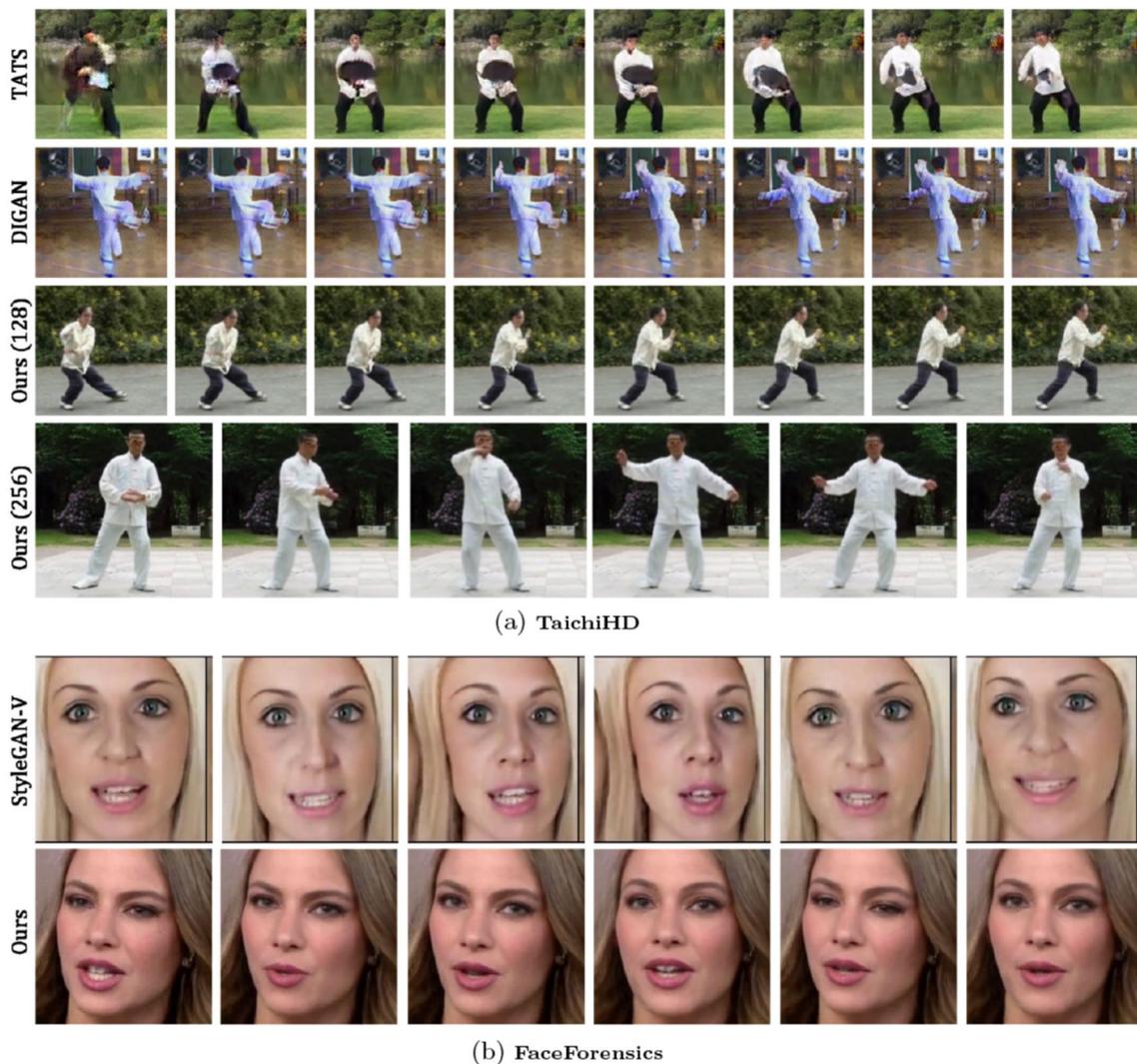


Fig. 4 Qualitative Comparison. We qualitatively compare LEO with DIGAN, TATS, StyleGAN-V on short video generation. The results indicate that on both **a** TaichiHD (128 and 256 resolutions) and **b** FaceForensics datasets, our proposed method achieves the best visual quality

and is able to capture the human structure well. Other approaches either modify the facial structure (e.g., StyleGAN-V) or fail to generate a complete human body (e.g., TATS and DIGAN)

be represented by α_1 and \mathbf{m} . Based on this, we propose a Linear Motion Condition (LMC) mechanism in cLMDM to condition the generative process with α_1 . During training, at each time step, we only add noise onto \mathbf{m}_t instead of the entire \mathbf{a} and leave α_1 intact. The objective function of cLMDM is

$$L_{\text{cLMDM}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_{\theta}(\mathbf{m}_t, \alpha_1, t)\|_2^2 \right], \quad (4)$$

where α_1 denotes the condition signal and \mathbf{m}_t stands for the noised \mathbf{m} to time t . α_1 is first added on \mathbf{m}_t and then concatenated along temporal dimension. LMC will be applied at each time step until we reach \mathbf{m}_0 . The final motion sequence is obtained as $\mathbf{a}_0 = [\alpha_1, \mathbf{m}_0]$. We find that following this, a related generated motion sequence is more stable and contains fewer artifacts, as α_1 serves as a strong signal to constrain the generated \mathbf{m} to follow the initial motion.

While the results from cLMDM outperforms previous models, the groundtruth α_1 is necessitated during both, training and inference stage. Towards unconditional generation, we train an additional simple DM to fit the distribution $p(\alpha_i)$ in a frame-wise manner. We refer to the cLMDM and such simple DM jointly as Latent Motion Diffusion Model (LMDM). By this way, LMDM are able to work in both conditional and unconditional motion generation.

Towards generating videos of arbitrary length, we propose an autoregressive approach based on proposed LMDM. By taking the last motion code from the previous generated sequence as the α_1 in the current sequence, with a randomly sampled noise, LMDM are able to generate an infinite-length motion sequence. By combining such sequence in pre-trained LIA with a starting image, LEO can synthesize photo-realistic and long-term videos.

3.3 Learning Starting Frames

In our framework, a starting image x_1 is required to synthesize a video. As image space is modeled independently, here we propose two options to obtain x_1 .

Option 1: existing images The first option is to directly take the images either from a real distribution or from an image generation network. In this context, our model is a conditional video generation model, which learns to predict future motion from x_1 . Starting motion α_1 is obtained through $\alpha_1 = E(x_1)$.

Option 2: conditional Diffusion Models The second option is to learn a conditional DDPM (Ho et al., 2020) (cDDPM) with α_1 as a condition to synthesize x_1 . By combining LEO with LMDM as well as cDDPM, we are able to conduct unconditional video synthesis.

4 Experiments

In this section, we firstly briefly describe our experimental setup, introducing datasets, evaluation metrics and implementation details. Secondly, we qualitatively demonstrate generated results on both, short and long video synthesis. Then we show quantitative evaluation w.r.t. video quality, comparing LEO with SoTA. Next, we conduct an ablation study to prove the effectiveness of proposed conditional mechanism LMC. Finally, we provide additional analysis of our framework, exhibiting motion and appearance disentanglement, video editing and infinite-length video generation.

Datasets As we focus on human video synthesis, evaluation results are reported on three human-related datasets, TaichiHD (Siarohin et al., 2019), FaceForensics (Rössler et al., 2018) and CelebV-HQ (Zhu et al., 2022). We use both 128×128 and 256×256 resolution TaichiHD datasets, and only 256×256 resolution FaceForensics and CelebV-HQ datasets.

- *TaichiHD* (Siarohin et al., 2019) comprises 3100 video sequences downloaded from YouTube. In train and test splits, it contains 2815 and 285 videos, respectively. We conducted all our experiments on the train split and used both 128×128 and 256×256 resolutions in our experiments.
- *FaceForensics* (Rössler et al., 2018) includes 1000 video sequences downloaded from YouTube. Following the preprocessing of previous methods (Saito et al., 2020; Skorokhodov et al., 2022), face areas are cropped based on the provided per-frame meshes. We resized all videos to 256×256 resolution.
- *CelebV-HQ* (Zhu et al., 2022) comprises 35666 high-quality talking head videos of 3 to 20s each. In total, it represents 15653 celebrities. We resized the original videos to 256×256 resolution, in order to train our models.

Evaluation metric For quantitative evaluation, we apply the commonly used metrics FVD and KVD, in order to compare with other approaches on video quality and apply Average Content Distance (ACD) towards evaluating the identity consistency of faces and bodies in the generated videos. In addition, we conduct a user study with 20 users towards comparing with objective quantitative evaluation.



Fig. 5 Comparison on long-term video generation. We compare with TATS by generating 512-frame videos. Videos from TATS start crashing around 50 frames while our model is able to continue producing high-quality frames with diverse motion

- *Frechet video distance (FVD) and Kernel Video Distance (KVD)* We use I3D (Carreira & Zisserman, 2017) trained on Kinetics-400 as feature extractor to compute FVD and KVD. However, we find FVD is a very sensitive metric, which can be affected by many factors such as frame-rate, single image quality, video length and implementation, which also mentioned in Skorokhodov et al. (2022). Therefore, towards making a fair comparison, on the TaichiHD dataset, we adopt the implementation from DIGAN (Yu et al., 2022). As for FaceForensics and CelebV-HQ, we chose to follow the implementation of StyleGAN-V (Skorokhodov et al., 2022).
- *Average Content Distance (ACD)* ACD measures the content consistency in generated videos. To evaluate results from FaceForensics and TaichiHD, we extract features from each generated frame and proceed to extract a per-frame feature vector in a video. The ACD was then computed using the average pairwise L2 distance of the per-frame feature vectors. We follow the implementation in Tian et al. (2021) to compute ACD for FaceForensics. As for TaichiHD, we employ the pre-trained person-reID model (Zheng et al., 2018) to extract person identity features.
- *User study* We asked 20 human raters to evaluate generated video quality, as well as video coherency. In each user study, we show paired videos and ask the raters, to rate 'which clip is more realistic/which clip is more coherent'. Each video-pair contains one generated video from our method, whereas the second video is either real or generated from other methods.

Implementation details Our framework requires two-phase training. In the first phase, we follow the standard protocol to train LIA (Wang et al., 2022) to encode input images into low-dimensional latent motion codes, and map such codes to flow maps, which are used for reconstruction via warp-and-inpaint. Therefore, once trained, LIA naturally provides a space of motion codes that are strictly constrained to only containing motion-related information. In the second phase, we only train LMDM on the extracted motion codes from Encoder. We note that the LMDM is a 1D U-Net adopted

from Nichol and Dhariwal (2021), we set the input size as 64×20 , where 64 is the length of the sequence and 20 is the dimension of the motion code. We use 1000 diffusion steps and a learning rate of $1e-4$. As the training of LMDM is conducted in the latent space of LIA, the entire training is very efficient and only requires one single GPU.

4.1 Qualitative Evaluation

We qualitatively compare LEO with SoTA by visualizing the generated results. We firstly compare our method with DIGAN, TATS and StyleGAN-V on the FaceForensics and TaichiHD datasets for short video generation. As shown in Figs. 1 and 4, the visual quality of our generated results outperforms other approaches w.r.t both, appearance and motion. For both resolutions on TaichiHD datasets, our method is able to generate complete human structures, whereas both, DIGAN and TATS fail, especially for arms and legs. When compared with StyleGAN-V on FaceForensics dataset, we identify that while LEO preserves well facial structures, StyleGAN-V modifies such attributes when synthesizing large motion.

Secondly, we compare with TATS for long-term video generation. Specifically, 512 frames are produced for the resolution 128×128 pertained to the TaichiHD dataset. As shown in Fig. 5, the subject in the videos from TATS starts crashing around 50 frames and the entire video sequence starts to fade. On the other hand, in our results, the subject continues to perform diverse actions whilst well preserving the human structure. We note that our model is only trained using a 64-frame sequence.

4.2 Quantitative Evaluation

In this section, we compare our framework with five state-of-the-art for both, conditional and unconditional short video generation, as well as unconditional long-term video generation.

Unconditional short video generation In this context, as described in Sec. 3.3, Option 2, the x_1 is randomly generated by a pre-trained cDDPM. We compare with SoTA by

Table 1 Evaluation for unconditional and conditional short video generation. LEO systematically outperforms other approaches on conditional video generation, and achieves better or competitive results on unconditional generation w.r.t. FVD, KVD and ACD

TaichiHD128			TaichiHD256		FaceForensics		CelebV-HQ	
Method	FVD ₁₆	KVD ₁₆	ACD ₁₆	FVD ₁₆	KVD ₁₆	FVD ₁₆	ACD ₁₆	FVD ₁₆
MoCoGAN-HD	144.7 ± 6.0	25.4 ± 1.9	–	–	–	111.8	0.33	212.4
DIGAN	128.1 ± 4.9	20.6 ± 1.1	2.17	156.7 ± 6.2	–	62.5	–	72.9
TATS	136.5 ± 1.2*	22.2 ± 1.0*	2.28	–	–	–	–	–
StyleGAN-V	–	–	–	–	–	47.4	0.36	69.1
MoStGAN-V	–	–	–	–	–	39.7	0.38	132.1
Ours (uncond)	100.4 ± 3.1	11.4 ± 3.2	1.83	122.7 ± 1.1	20.49 ± 0.9	52.3	0.28	–
Ours (cond)	57.6 ± 2.0	4.0 ± 1.5	1.22	94.8 ± 4.2	13.47 ± 2.3	35.9	0.27	40.2

*Results are reproduced based on official code and released checkpoints

Table 2 Evaluation for unconditional long-term video generation. LEO outperforms other methods on long-term (128 frames) video generation w.r.t. FVD, KVD and ACD

TaichiHD128			FaceForensics		
Method	FVD ₁₂₈	KVD ₁₂₈	ACD ₁₂₈	FVD ₁₂₈	ACD ₁₂₈
DIGAN	–	–	–	1824.7	–
TATS	1194.58 ± 1.1	462.03 ± 8.2	2.85	–	–
StyleGAN-V	–	–	–	89.34	0.49
Ours	155.54 ± 2.6	48.82 ± 5.9	2.06	96.28	0.34

Table 3 User study. We conduct user studies pertaining to the datasets TaichiHD and FaceForensics w.r.t. video quality (up) as well as coherency (down)

Method	TaichiHD (%)	FaceForensics (%)
Ours/TATS	93.00 /7.00	–
Ours/StyleGAN-V	–	91.33 /8.67
Ours/TATS	98.60 /1.40	–
Ours/StyleGAN-V	–	93.20 /6.80

Table 4 Ablation study of proposed LMC. Models with LMC achieved the lowest FVD on both datasets

	TaichiHD	FaceForensics
w/o LMC	118.6	60.03
with LMC	100.4	52.32

generating 16 frames. To compare with DIGAN on high-resolution generation, we also generate videos of 256×256 resolution. Related FVDs and KVDs are reported in Table 1. LEO systematically outperforms other methods w.r.t. video quality, obtaining lower or competitive FVD and KVD on all datasets. On high-resolution generation, our results remain better than DIGAN.

However, by comparing the results between StyleGAN-V and ours, we find FVD is not able to represent the quality of generated videos veritably. We observe that StyleGAN-V is not able to preserve facial structures, whereas LEO is

able to do so, see Fig. 4. We additionally compute ACD, in order to further analyze the identity consistency in 16-frame videos. Table 1 reflects on the fact that our method achieves significantly better results compared to other approaches. In addition, we conduct user study w.r.t. video quality and coherency of generated videos among different methods. Results in Table 3 showcase that as nearly all users rated for our generated results to be superior than other approaches. Hence, we conclude that a metric, replacing FVD is in urgent need in the context of video generation.

Unconditional long video generation We evaluate our approach for long-term video generation w.r.t. FVD and ACD. In this context, we compare LEO with StyleGAN-V on the FaceForensics dataset, and both DIGAN and TATS on the TaichiHD. We report results based on 128-frame generation in Table 2, which clearly shows that our method outperforms others in such context. We hypothesize that consistent and stable motion codes produced by our LMDM are key to producing high-quality long-term videos.

Conditional short video generation As described in Sec. 3.3, Option 1, our framework additionally caters for conditional video generation by taking an existing image to hallucinate the following motion. Specifically, we randomly select 2048 images from both, TaichiHD and FaceForensics datasets as x_1 and compute corresponding α_1 as input of LMDM. As depicted in Table 1, results conditioned on the real images achieve the lowest FVD, KVD and ACD values, suggesting that the quality of a starting image is pertinent for output video quality, which further signifies that in the set-

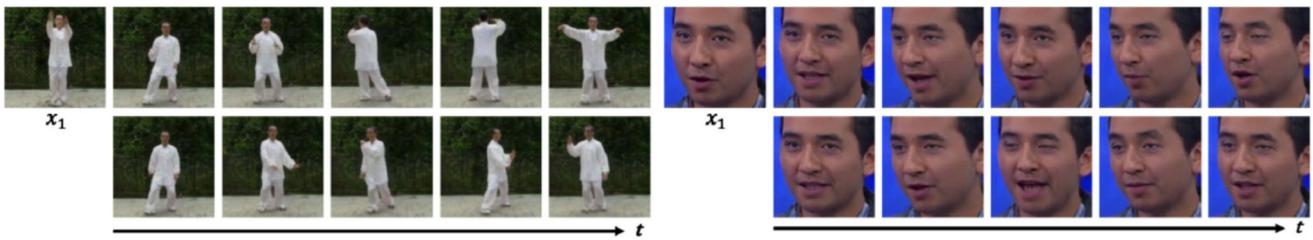


Fig. 6 Disentanglement of motion and appearance. The first and second row share the same appearance, with different motion codes. Results display that our model is able to produce diverse motion from the same content



Fig. 7 Video editing. We show video editing results by combining LEO with off-the-shelf image editing model ControlNet. We are able to edit the appearance of the entire video sequence through only editing the starting image

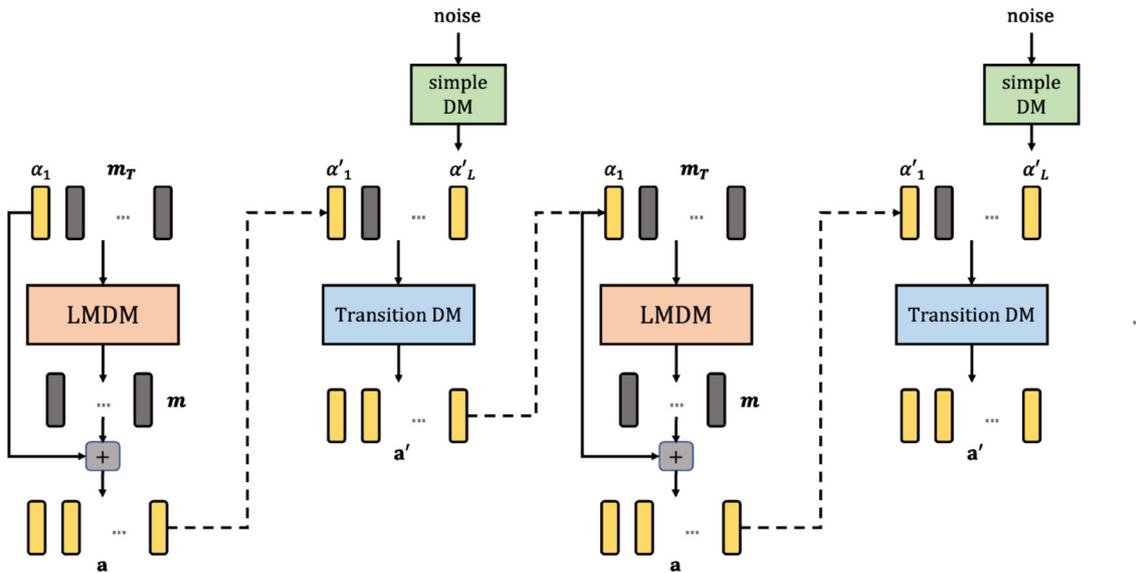


Fig. 8 Infinite-length video generation for TaichiHD

ting of unconditional generation, training a better cDDPM will be instrumental for improving results.

5 Ablation Study

In this section, we place emphasis on analyzing the effectiveness of proposed Linear Motion Condition (LMC) in LMDM. We train two models, with and without LMC on both TaichiHD and FaceForensics datasets. As shown in Table 4,

using LMC significantly improves the generated video quality, which proves that our proposed LMC is an effective mechanism for involving α_1 in LMDM.

6 Additional Analysis

Motion and appearance disentanglement We proceed to combine the same x_1 with different m , aiming to reveal whether m is only motion-related. Figure 6 illustrates that

Table 5 User study of repeated motion. We show the occurrence of repeated motion with and without the usage of Transition DM

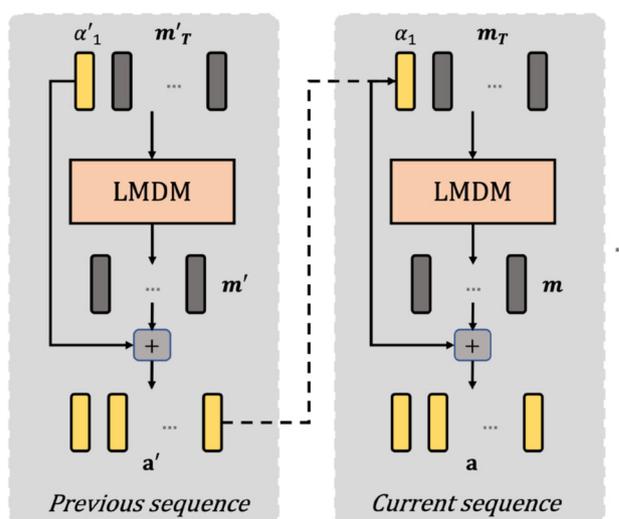
	Occurrence (%)
w/o Transition DM	0.45
with Transition DM	0.02

different \mathbf{m} enables the same subject to perform different motion—which proves that our proposed LMDM is indeed learning a motion space, and appearance and motion are clearly disentangled. This experiment additionally indicates that our model does not overfit on the training dataset, as different noise sequences are able to produce diverse motion sequences.

Video Editing As appearance is modeled in x_1 , we here explore the task of video editing by modifying the semantics in the starting image. Compared to previous approaches, where image-to-image translation is required, our framework simply needs an edit of the semantics in a one-shot manner. Associated results are depicted in Figs. 1 and 7. We apply the open-source approach ControlNet (Zhang & Agrawala, 2023) on the starting frame by entering various different prompts. Given that the motion space is fully disentangled from the appearance space, our videos maintain the original temporal consistency, uniquely altering the appearance.

Infinite-length video generation In addition to presented settings, our framework is able to generate infinite-length videos. To generate long-term FaceForensics, as shown in Fig. 9, we provide the last generated code from the previous sequence as the starting code of the current sequence. The entire long-term video is generated in an autoregressive manner. Surprisingly, we find that such a simple approach is sufficient to produce more than 1000 frames. We note that for TaichiHD dataset, due to limited motion patterns, this setting yields repeated motion. Towards addressing this limitation, as shown in Fig. 8, we design an additional Transition Diffusion Model (Transition DM) aimed at generating transition motion between the last code from original generated sequence and a new motion code generated from the simple DM. Doing so, the Transition DM enforces the network to exit the original motion pattern and transit to new pattern. To evaluate the effectiveness of the proposed method, we generate long videos with and without Transition DM and request human raters to watch respective videos and answer the question ‘Does the clip contain repeated motion?’. Results are reported in Table 5, which shows the effectiveness of Transition DM to prevent repeated motion.

Compared to current diffusion-based methods, our approach is very efficient in long video generation. We autoregressively run LMDM for the generation of long motion code sequences in latent space. Given that LMDM is a small-scale network, which only focuses on generating 1-D motion code, even for

**Fig. 9** Infinite-length video generation for FaceForensics

very long sequence generation, it only requires few seconds during inference stage. In addition, the image animator itself is a one-step inference model which enables our proposed method to be significantly efficient.

7 Limitations

We list several limitations in current framework and proposed potential solutions for future work.

- **Geometry ambiguity and temporal coherency** Since we use a 2D generator to predict 2D flow maps, LEO is not able to handle human body occlusion very well especially in Taichi dataset. One solution would be to incorporate the architecture of NeRF or Tri-plane into our generator to support 3D-aware generation. We think in this way, the issues of geometry ambiguity and human body occlusion could be addressed.
- **Generalizability** Since the pre-trained image animator focuses on talking head and human bodies, our proposed framework currently performs better on human-centric videos. However, to analyze the generalizability of LEO, we conducted a small-scale experiment on UCF101 and report quantitative evaluation in Table 6. The results show that under current model design, LEO achieves competitive results with previous GAN-based methods but still has large performance gap compared with large-scale video diffusion models.

We believe our framework is pushing the boundaries of video generation, as it solves a challenge, which constitutes generation of long human-centric videos. While this is a first step, the proposed method has the potential to generalize onto additional settings such as text-to-video

Table 6 Quantitative evaluation on UCF101 w.r.t. FVD

Methods	FVD ₁₆
MoCoGAN-HD	1729.6
DIGAN	1630.2
StyleGAN-V	1431.0
Make-A-Video	367.23
Video LDM	550.61
LaVie	540.30
Ours	1356.2

generation. However, achieving such goals requires scaling up and re-designing (a) the original LIA, as well as (b) LMDM, and (c) training the entire system on larger-scale well-curated video datasets, which requires extremely expensive computational resources. We will explore such research directions in our future work.

- **Architecture** Current architect of LEO still relies on convolutional networks in both image animator and latent motion diffusion models. Advanced techniques such as transformers have not been explored yet. Future work would be involving novel architecture design and training LEO on larger-scale dataset to explore the limits of current approach.

8 Conclusions

In this paper, we introduced LEO, a novel framework incorporating a Latent Image Animator (LIA), as well as a Latent Motion Diffusion Model (LMDM), placing emphasis on spatio-temporal coherency in human video synthesis. By jointly exploiting LIA and LMDM in a two-phase training strategy, we endow LEO with the ability to disentangle appearance and motion. We quantitatively and qualitatively evaluated proposed method on both, human body and talking head datasets and demonstrated that our approach is able to successfully produce photo-realistic, long human videos. In addition, we showcased that the effective disentanglement of appearance and motion in LEO allows for two additional tasks, namely infinite-length human video synthesis by autoregressively applying LMDM, as well as content-preserving video editing (employing an off-the-shelf image editor (e.g., ControlNet)). We postulate that LEO opens a new door in design of generative models for video synthesis and plan to extend our method onto more general videos and applications.

Data Availability The datasets used during and analyzed during the current study are available in the following public domain resources:

- **FaceForensics** (Rössler et al., 2018) <https://github.com/ondyari/FaceForensics>
- **CelebV-HQ** (Zhu et al., 2022) <https://celebv-hq.github.io>

- **TaichiHD** (Siarohin et al., 2019) <https://github.com/AliaksandrSiarohin/first-order-model>

The models and source data generated during and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethic Statement In this work, we aim to synthesize high-quality human-centric videos by combining a pretrained image animator with a proposed latent motion diffusion model. Our approach can be used for digital human, online education, and data synthesis for other computer vision tasks, etc. We note that our framework mainly focuses on learning how to model motion distribution in a pretrained image animator rather than directly model appearance. Therefore, our framework is not biased towards any specific gender, race, region, or social class. It works equally well irrespective of the difference in subjects.

References

- Bar-Tal, O., Ofri-Amar, D., Fridman, R., Kasten, Y., & Dekel, T. (2022). Text2live: Text-driven layered image and video editing. In *ECCV*.
- Bergman, A., Kellnhofer, P., Yifan, W., Chan, E., Lindell, D., & Wetzstein, G. (2022). Generative neural articulated radiance fields. *NeurIPS*, 35, 19900–19916.
- Bhagat, S., Uppal, S., Yin, Z., & Lim, N. (2020). Disentangling multiple features in video sequences using gaussian processes in variational autoencoders. In *ECCV*.
- Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., & Jampani, V. (2023a). Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint [arXiv:2311.15127](https://arxiv.org/abs/2311.15127)
- Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., & Kreis, K. (2023b). Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*.
- Brock, A., Donahue, J., & Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *ICLR*.
- Brooks, T., Hellsten, J., Aittala, M., Wang, T.-C., Aila, T., Lehtinen, J., Liu, M.-Y., Efros, A. A., & Karras, T. (2022). Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35, 31769–31781.
- Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.
- Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2019). Everybody dance now. In *ICCV*.
- Chen, X., Wang, Y., Zhang, L., Zhuang, S., Ma, X., Yu, J., Wang, Y., Lin, D., Qiao, Y., & Liu, Z. (2023). Seine: Short-to-long video diffusion model for generative transition and prediction. In *ICLR*.
- Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., & Shan, Y. (2024). Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*.
- Chu, C., Zhmoginov, A., & Sandler, M. (2017). CycleGAN: a master of steganography. arXiv preprint [arXiv:1712.02950](https://arxiv.org/abs/1712.02950)
- Clark, A., Donahue, J., & Simonyan, K. (2019). Adversarial video generation on complex datasets. arXiv preprint [arXiv:1907.06571](https://arxiv.org/abs/1907.06571)
- Denton, E. L., & Birodkar, V. (2017). Unsupervised learning of disentangled representations from video. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *NeurIPS*.
- Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *CVPR*.
- Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *CVPR*.

- Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.-B., & Parikh, D. (2022). Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *NIPS*.
- Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., & Salimans, T. (2022a). Imagen video: High definition video generation with diffusion models. arXiv preprint [arXiv:2210.02303](https://arxiv.org/abs/2210.02303)
- Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022b). Video diffusion models. arXiv preprint [arXiv:2204.03458](https://arxiv.org/abs/2204.03458)
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *NeurIPS*, 33, 6840–6851.
- Huang, X., Liu, M.-Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *ECCV*.
- Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*.
- Jang, Y., Kim, G., & Song, Y. (2018). Video Prediction with Appearance and Motion Conditions. In *ICML*.
- Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *CVPR*.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In *CVPR*.
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *ICLR*.
- Li, Y., & Mandt, S. (2018). Disentangled sequential autoencoder. *ICML*.
- Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M.-H. (2018). Flow-grounded spatial-temporal video prediction from still images. In *ECCV*.
- Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., & Tan, T. (2023). Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*.
- Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.-F., Chen, C., & Qiao, Y. (2024). Latte: Latent diffusion transformer for video generation. arXiv preprint [arXiv:2401.03048](https://arxiv.org/abs/2401.03048)
- Menapace, W., Siarohin, A., Skorokhodov, I., Deyneka, E., Chen, T.-S., Kag, A., Fang, Y., Stoliar, A., Ricci, E., Ren, J., & Tulyakov, S. (2024). Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *CVPR*.
- Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *ICML*.
- Ohnishi, K., Yamamoto, S., Ushiku, Y., & Harada, T. (2018). Hierarchical video generation from orthogonal information: Optical flow and texture. In *AAAI*.
- Pan, J., Wang, C., Jia, X., Shao, J., Sheng, L., Yan, J., & Wang, X. (2019). Video generation from single semantic label map. arXiv preprint [arXiv:1903.04480](https://arxiv.org/abs/1903.04480)
- Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint [arXiv:1511.06434](https://arxiv.org/abs/1511.06434)
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint [arXiv:2204.06125](https://arxiv.org/abs/2204.06125)
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In *ICML*.
- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint [arXiv:1803.09179](https://arxiv.org/abs/1803.09179)
- Saito, M., Matsumoto, E., & Saito, S. (2017). Temporal generative adversarial nets with singular value clipping. In *ICCV*.
- Saito, M., Saito, S., Koyama, M., & Kobayashi, S. (2020). Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *IJCV*.
- Shen, X., Li, X., & Elhoseiny, M. (2023). Mostgan-v: Video generation with temporal motion styles. In *CVPR*.
- Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. In *NeurIPS*.
- Siarohin, A., Woodford, O., Ren, J., Chai, M., & Tulyakov, S. (2021). Motion representations for articulated animation. In *CVPR*.
- Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., & Taigman, Y. (2023). Make-a-video: Text-to-video generation without text-video data. In *ICLR*.
- Skorokhodov, I., Tulyakov, S., & Elhoseiny, M. (2022). Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*.
- Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. In *ICLR*.
- Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D. N., & Tulyakov, S. (2021). A good image generator is what you need for high-resolution video synthesis. In *ICLR*.
- Tulyakov, S., Liu, M.-Y., Yang, X., & Kautz, J. (2018). MoCoGAN: Decomposing motion and content for video generation. In *CVPR*.
- Van Den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. *NeurIPS*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*.
- Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., & Erhan, D. (2023). Phenaki: Variable length video generation from open domain textual descriptions. In *ICLR*.
- Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Generating videos with scene dynamics. In *NIPS*.
- Walker, J., Marino, K., Gupta, A., & Hebert, M. (2017). The pose knows: Video forecasting by generating pose futures. In *ICCV*.
- Wang, Y. (2021). *Learning to Generate Human Videos*. Theses: Inria - Sophia Antipolis; Université Cote d'Azur.
- Wang, Y., Bilinski, P., Bremond, F., & Dantcheva, A. (2020). G3AN: Disentangling appearance and motion for video generation. In *CVPR*.
- Wang, Y., Bilinski, P., Bremond, F., & Dantcheva, A. (2020). Imaginator: Conditional spatio-temporal gan for video generation. In *WACV*.
- Wang, Y., Bremond, F., & Dantcheva, A. (2021). Inmodegan: Interpretable motion decomposition generative adversarial network for video generation. arXiv preprint [arXiv:2101.03049](https://arxiv.org/abs/2101.03049)
- Wang, T. Y., Ceylan, D., Singh, K. K., & Mitra, N. J. (2021). Dance in the wild: Monocular human animation with neural dynamic appearance synthesis. In *3DV*.
- Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., & Guo, Y. (2023). Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint [arXiv:2309.15103](https://arxiv.org/abs/2309.15103)
- Wang, T.-C., Liu, M.-Y., Tao, A., Liu, G., Kautz, J., & Catanzaro, B. (2019). Few-shot video-to-video synthesis. In *NeurIPS*.
- Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., & Catanzaro, B. (2018). Video-to-video synthesis. In *NeurIPS*.
- Wang, Y., Yang, D., Bremond, F., & Dantcheva, A. (2022). Latent image animator: Learning to animate images via latent space navigation. In *ICLR*.

- Xie, J., Gao, R., Zheng, Z., Zhu, S.-C., & Wu, Y. N. (2020). Motion-based generator model: Unsupervised disentanglement of appearance, trackable and intrackable motions in dynamic patterns. In *AAAI*.
- Yan, W., Zhang, Y., Abbeel, P., & Srinivas, A. (2021). Videogpt: Video generation using vq-vae and transformers. arXiv preprint [arXiv:2104.10157](https://arxiv.org/abs/2104.10157)
- Yang, Z., Li, S., Wu, W., & Dai, B. (2022). 3dhumangan: Towards photo-realistic 3d-aware human image generation. arXiv preprint.
- Yang, C., Wang, Z., Zhu, X., Huang, C., Shi, J., & Lin, D. (2018). Pose guided human video generation. In *ECCV*.
- Yu, S., Sohn, K., Kim, S., & Shin, J. (2023). Video probabilistic diffusion models in projected latent space. In *CVPR*.
- Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.-W., & Shin, J. (2022). Generating videos with dynamics-aware implicit generative adversarial networks. In *ICLR*.
- Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *ICCV*.
- Zhang, L., & Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models.
- Zhang, D. J., Wu, J. Z., Liu, J.-W., Zhao, R., Ran, L., Gu, Y., Gao, D., & Shou, M. Z. (2023). Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv preprint [arXiv:2309.15818](https://arxiv.org/abs/2309.15818)
- Zhao, L., Peng, X., Tian, Y., Kapadia, M., & Metaxas, D. (2018). Learning to forecast and refine residual motion for image-to-video generation. In *ECCV*.
- Zheng, Z., Zheng, L., & Yang, Y. (2018). A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *14*(1), 1–20.
- Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., & Loy, C. C. (2022). CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.