Generating Videos with Dynamics-aware Implicit Generative Adversarial Networks

Sihyun Yu*,[1], Jihoon Tack*,[1], Sangwoo Mo*,[1],
Hyunsu Kim[2], Junho Kim[2], Jung-Woo Ha[2], Jinwoo Shin [1]

[1]Korea Advanced Institute of Science and Technology (KAIST)
[2]NAVER AI Lab

ICLR 2022

*Equal contribution.

Remarkable Success of Deep Generative Models

Deep generative models have shown remarkable success in various domains:

- Including images [Karras et al., 2021; Nichol et al., 2022], texts [Brown et al., 2021], and audios [Dhariwai et al., 2021]

StyleGAN3-R (ours), FID 3.66

[Karras et al, 2021]

"a hedgehog using a calculator"

"a corgi wearing a red bowtie and a purple party hat"

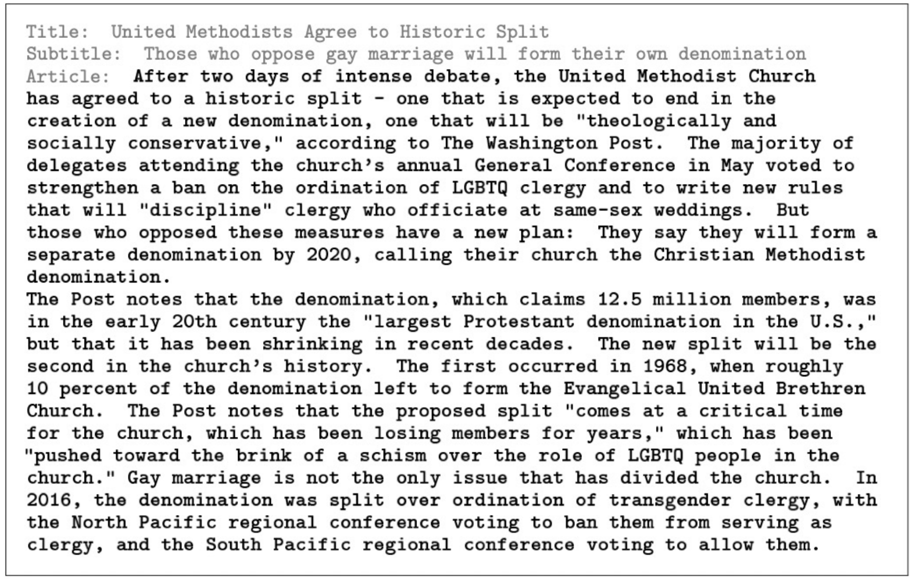
"robots meditating in a vipassana retreat"

"a surrealist dream-like oil painting by salvador dalí of a cat playing checkers"

"a professional photo of a sunset behind the grand canyon"

"a high-quality oil painting of a psychedelic hamster dragon"

[Nichol et al., 2022]



Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.," but that it has been shrinking in recent decades. The new split will be the second in the church's history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

[Brown et al., 2020]

[Karras et al., 2021] Alias-Free Generative Adversarial Networks, NeurIPS 2021.
[Nichol et al., 2022] GLIDE: Towards Photorealistic Image Generation and Editing with Text-guided Diffusion Models, 2022.
[Brown et al., 2020] Language Models are Few-Shot Learners, NeurIPS 2020.
[Dhariwai et al., 2021] Jukebox: A Generative Model for Music, 2020.

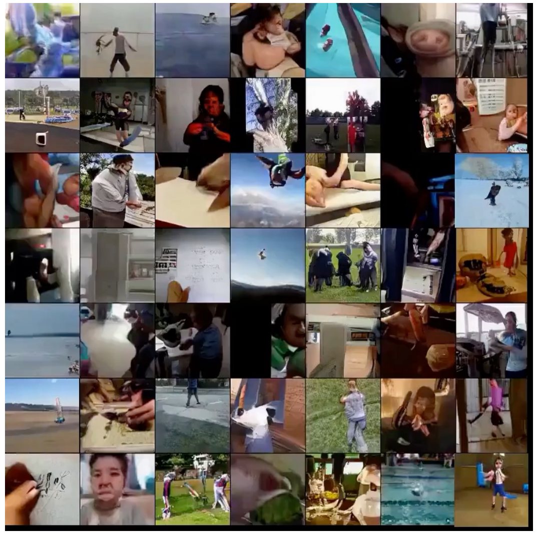
Video Generative Modeling Still Remains as a Challenge

Video generative modeling still remains as a challenge!

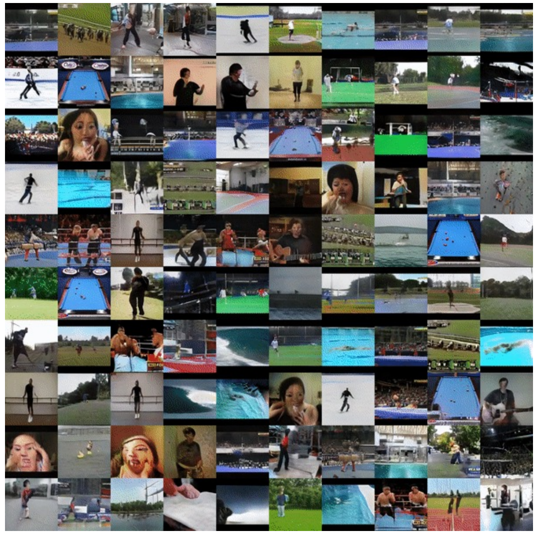
- Videos: spatiotemporally complex signals → Most approaches interpret videos as a 3D grid of RGB values
- ...limits the scalability of videos → up to 512 TPUs are required to train DVD-GAN [Clark et al., 2019]

Question: Can we interpret video signals differently **for efficient, scalable video synthesis?**

- Videos are **"continuous" signals**, where frames are **highly correlated with temporal dynamics**



[Clark et al., 2019]



[Tian et al., 2021]

[Clark et al., 2019] Adversarial Video Generation on Complex Datasets, 2019
[Tian et al., 2021] A Good Image Generator is What You Need for High-Resolution Video Synthesis, ICLR 2021

Video Generative Modeling Still Remains as a Challenge

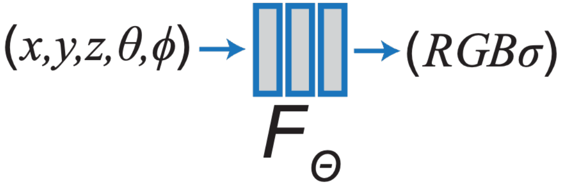
🤔 **How can we interpret video signals differently for efficient, scalable video synthesis?**

💡 **Idea**: We interpret videos as continuous signals and leverage implicit neural representations!

- Videos are **continuous signals**, where frames are **highly correlated with temporal dynamics**

Implicit neural representations (INRs): New paradigm of representing complex continuous signals

- Represent a signal into a neural network from input coordinates to corresponding signal values [Sitzmann et al., 2020]
- e.g.) Video = a neural network $f_\theta : (x, y, t) \rightarrow (r, g, b)$
- Have shown its effectiveness at representing complex signals such as 3D scenes [Mindelhall et al., 2020]



[Mindelhall et al., 2020]

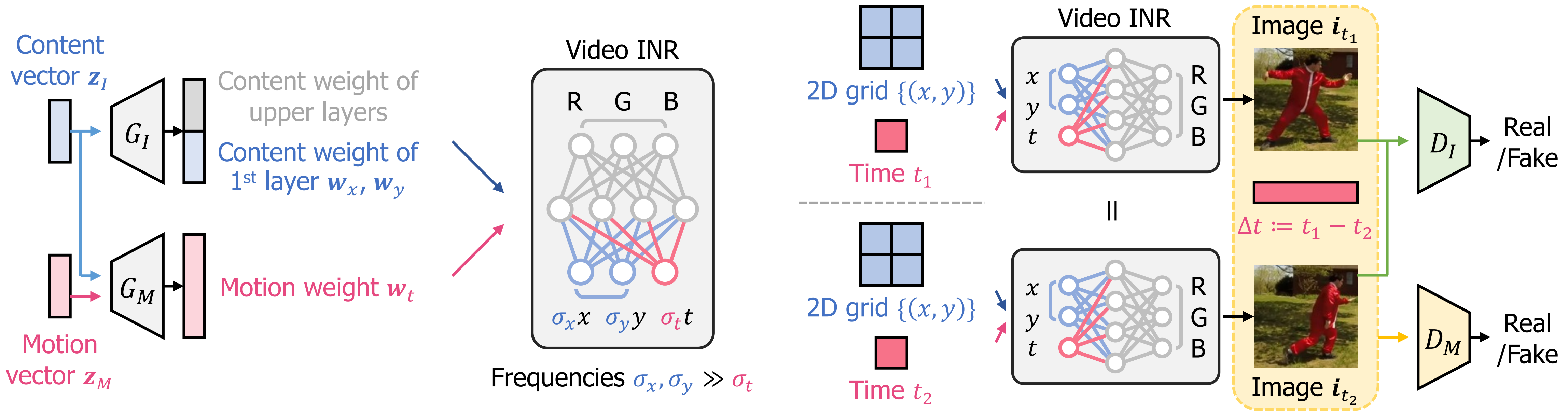
[Sitzmann et al., 2020] Implicit Neural Representations with Periodic Activation Functions, NeurIPS 2020.
[Mindelhall et al., 2020] NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020.

Implicit Neural Representations for Video Generation?

🤔 How can we interpret video signals differently for efficient, scalable video synthesis?

💡 **DIGAN**: We propose an **implicit GAN which generates weights of video INRs**

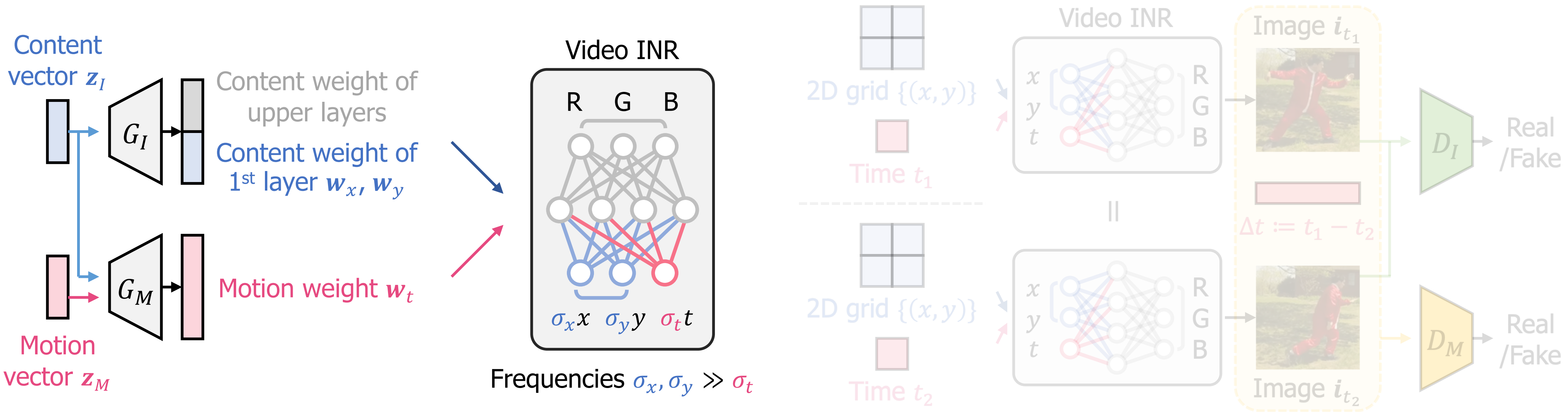
- Extends recent implicit GANs on image synthesis which generates weights of image INRs [Skorokhodov et al., 2020]
- **Idea**: We more focus on "temporal aspects" of videos to design implicit GANs for video synthesis


Content vector $\boldsymbol{z}_I$
G_I
Content weight of upper layers
Content weight of 1st layer $\boldsymbol{w}_x, \boldsymbol{w}_y$
Motion vector $\boldsymbol{z}_M$
G_M
Motion weight $\boldsymbol{w}_t$
Video INR
R G B
$\sigma_x x$ $\sigma_y y$ $\sigma_t t$
Frequencies $\sigma_x, \sigma_y \gg \sigma_t$
2D grid $\{(x, y)\}$
Time t_1
2D grid $\{(x, y)\}$
Time t_2
Video INR
x y t
R G B
$\parallel$
Image $\boldsymbol{i}_{t_1}$
$\Delta t := t_1 - t_2$
Image $\boldsymbol{i}_{t_2}$
D_I
Real /Fake
D_M
Real /Fake

[Skorokhodov et al., 2021] Adversarial Generation of Continuous Images, CVPR 2021

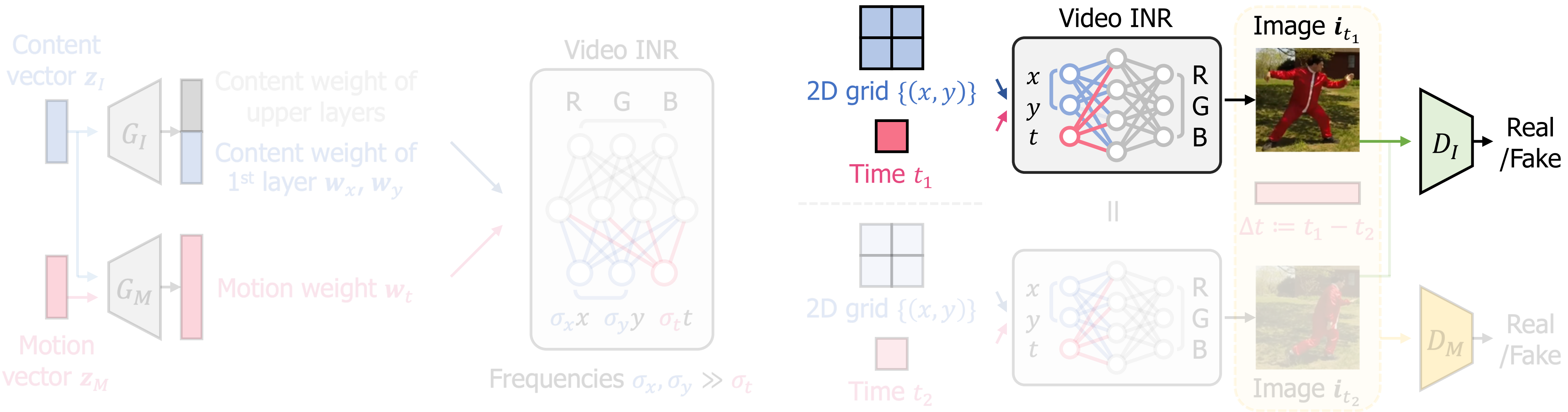
DIGAN: Dynamics-aware Implicit GAN

1. **Generators synthesize the weights of video INRs from a content vector and a motion vector**
   - z_I determines overall spatial contents (or style) of generated videos
   - z_M determines the temporal motion of generated videos
   - Smaller time-frequency σ_t than space-frequencies σ_x, σ_y
   - Since frames **"change relatively slowly over time"** compared to spatial variations



DIGAN: Dynamics-aware Implicit GAN

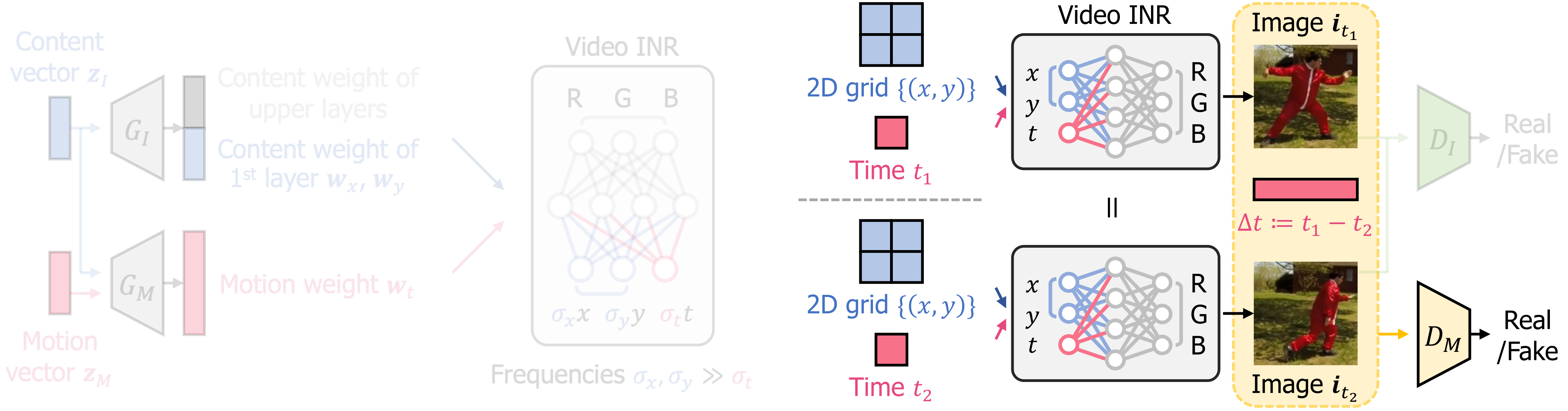
2. **DIGAN utilizes two discriminators: an image discriminator D_I and a motion discriminator D_M**
   - D_I determines real/fake of the generated frame (image) at arbitrary time $t \in [0, 1]$
   - Use a similar architecture to prior image GANs like StyleGAN2 [Karras et al., 2020]



[Karras et al., 2020] Anaylzing and Improving the Improve Quality of StyleGAN, CVPR 2020

DIGAN: Dynamics-aware Implicit GAN

2. DIGAN utilizes two discriminators: an image discriminator D_I and a motion discriminator D_M
   - D_M determines real/fake of the triplet $(\mathbf{i}_{t_1}, \mathbf{i}_{t_2}, \Delta t)$ consists of a pair of images and their time difference
   - D_M is not a 3D convolutional discriminator → same architecture as D_I (input channel 3 → 7)
   - We generate video INR → efficiently produce arbitrary time frames (unlike prior autoregressive video models)

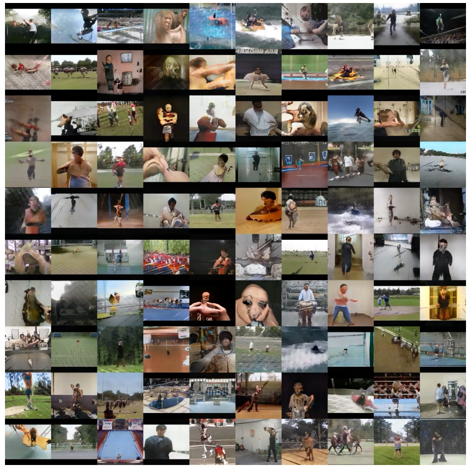


Experiments: DIGAN Beats Other Video Synthesis Methods

DIGAN significantly outperforms prior video synthesis methods (both quantitatively/qualitatively)

- **e.g.) UCF-101:** Shows 30.7% improvement on UCF-101 (measured with FVD)

UCF-101



Sky

TaiChi

Kinetics-food

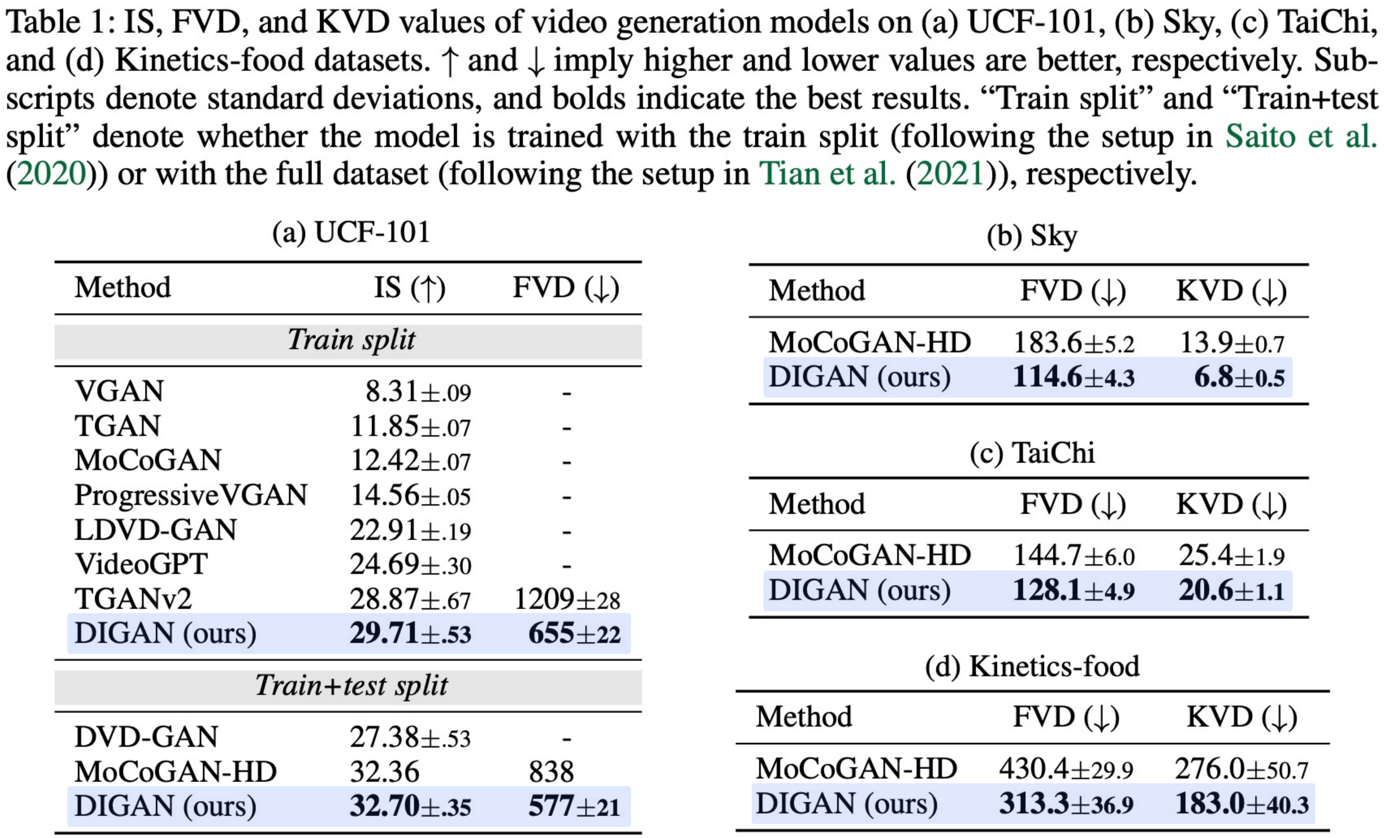


Table 1: IS, FVD, and KVD values of video generation models on (a) UCF-101, (b) Sky, (c) TaiChi, and (d) Kinetics-food datasets. ↑ and ↓ imply higher and lower values are better, respectively. Subscripts denote standard deviations, and bolds indicate the best results. "Train split" and "Train+test split" denote whether the model is trained with the train split (following the setup in Saito et al. (2020)) or with the full dataset (following the setup in Tian et al. (2021)), respectively.

(a) UCF-101

Method	IS (↑)	FVD (↓)
Train split		
VGAN	8.31±.09	-
TGAN	11.85±.07	-
MoCoGAN	12.42±.07	-
ProgressiveVGAN	14.56±.05	-
LDVD-GAN	22.91±.19	-
VideoGPT	24.69±.30	-
TGANv2	28.87±.67	1209±28
DIGAN (ours)	**29.71**±.53	**655**±22
Train+test split		
DVD-GAN	27.38±.53	-
MoCoGAN-HD	32.36	838
DIGAN (ours)	**32.70**±.35	**577**±21

(b) Sky

Method	FVD (↓)	KVD (↓)
MoCoGAN-HD	183.6±5.2	13.9±0.7
DIGAN (ours)	**114.6**±4.3	**6.8**±0.5

(c) TaiChi

Method	FVD (↓)	KVD (↓)
MoCoGAN-HD	144.7±6.0	25.4±1.9
DIGAN (ours)	**128.1**±4.9	**20.6**±1.1

(d) Kinetics-food

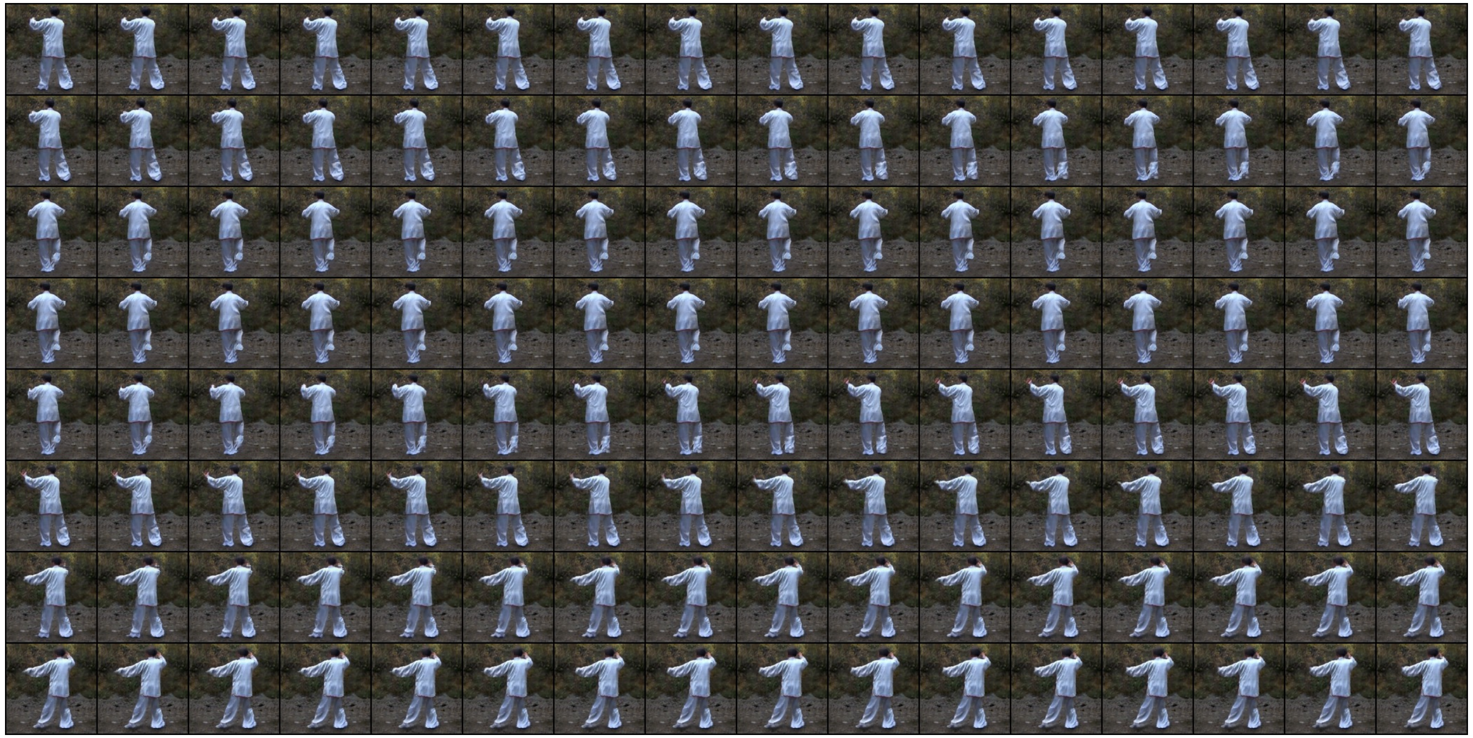
Method	FVD (↓)	KVD (↓)
MoCoGAN-HD	430.4±29.9	276.0±50.7
DIGAN (ours)	**313.3**±36.9	**183.0**±40.3

Experiments: DIGAN Can Generate Long Videos

DIGAN can generate long videos (up to 256 frames) with reasonable visual quality

- It is 5.3x longer than prior state-of-the-art method [Clark et al., 2019]
- Training time and resources are not demanding (4.4 days with 4 NVIDIA V100 GPUs)

→ Follow the arrow direction, and move to the next line at the end

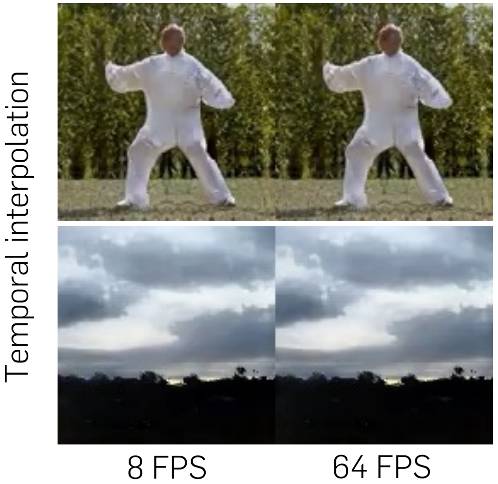


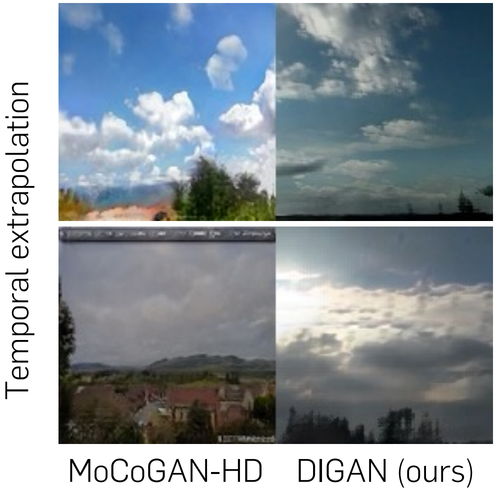
[Clark et al., 2019] Adversarial Video Generation on Complex Datasets

Experiments: DIGAN Can Inter-/Extra-polate Generated Videos

Intriguingly, DIGAN achieves successful "spatiotemporal" inter-/extra-polation of generated videos

- **Recall:** We interpret videos as spatiotemporally continuous signals





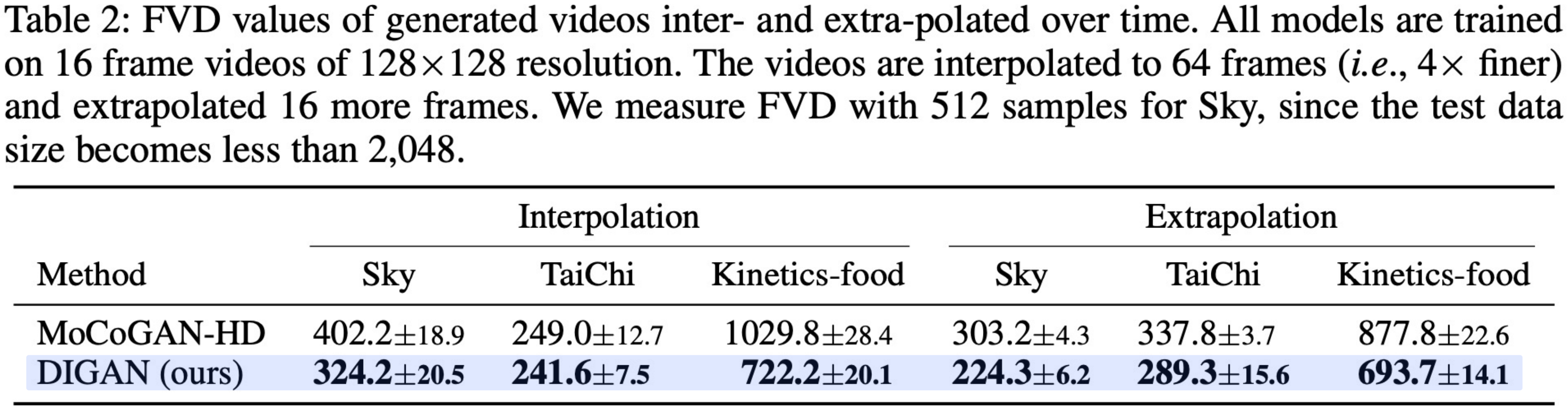
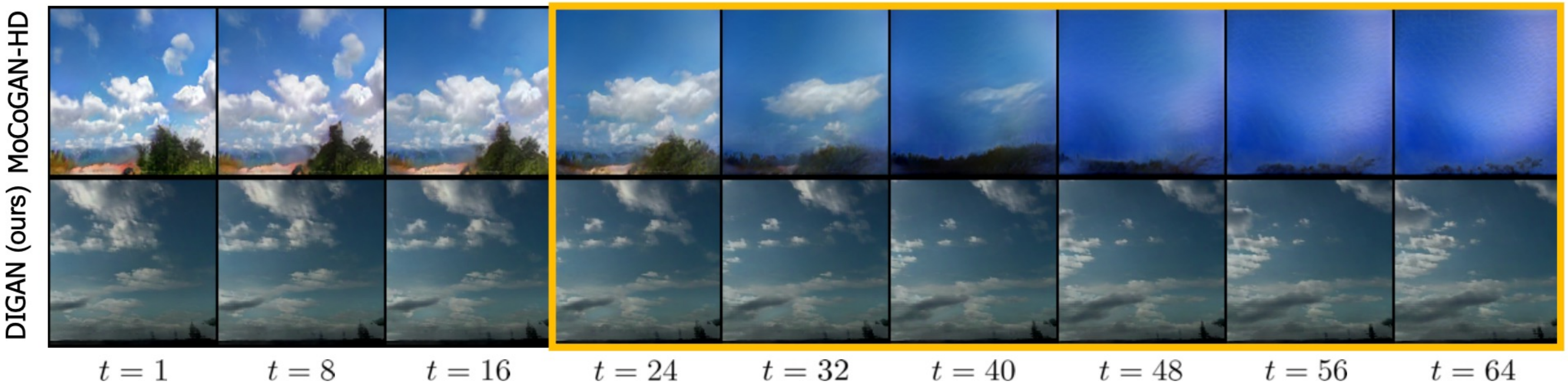


Table 2: FVD values of generated videos inter- and extra-polated over time. All models are trained on 16 frame videos of 128×128 resolution. The videos are interpolated to 64 frames (*i.e.*, 4× finer) and extrapolated 16 more frames. We measure FVD with 512 samples for Sky, since the test data size becomes less than 2,048.

	Interpolation			Extrapolation		
Method	Sky	TaiChi	Kinetics-food	Sky	TaiChi	Kinetics-food
MoCoGAN-HD	402.2±18.9	249.0±12.7	1029.8±28.4	303.2±4.3	337.8±3.7	877.8±22.6
DIGAN (ours)	**324.2±20.5**	**241.6±7.5**	**722.2±20.1**	**224.3±6.2**	**289.3±15.6**	**693.7±14.1**



Experiments: DIGAN Can Inter-/Extra-polate Generated Videos

Intriguingly, DIGAN achieves successful "spatiotemporal" inter-/extra-polation of generated videos

- **Recall:** We interpret videos as spatiotemporally continuous signals

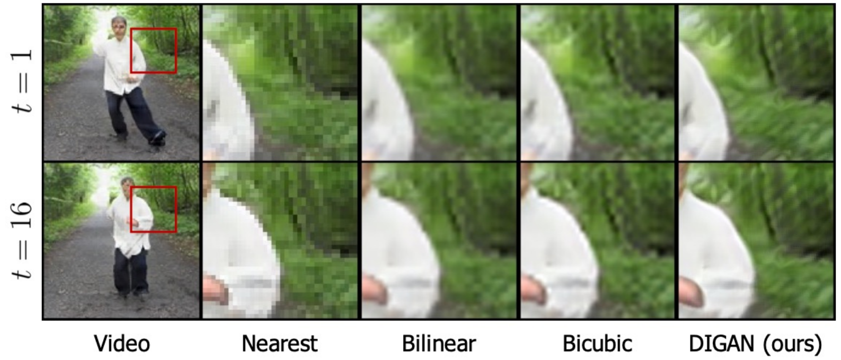


Figure 7: Videos upsampled from 128×128 to 512×512 resolution (4× larger) on TaiChi dataset.

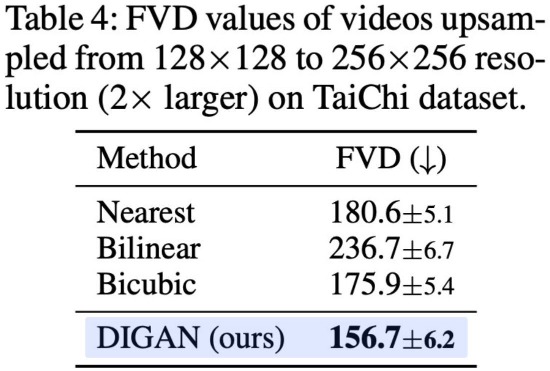
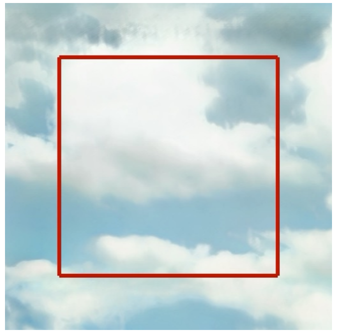


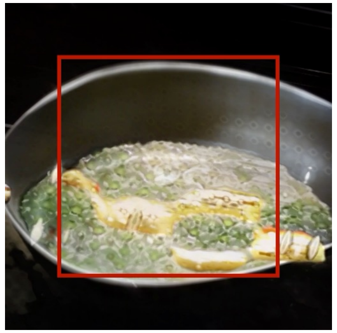
Table 4: FVD values of videos upsampled from 128×128 to 256×256 resolution (2× larger) on TaiChi dataset.

Method	FVD (↓)
Nearest	180.6±5.1
Bilinear	236.7±6.7
Bicubic	175.9±5.4
DIGAN (ours)	**156.7**±6.2

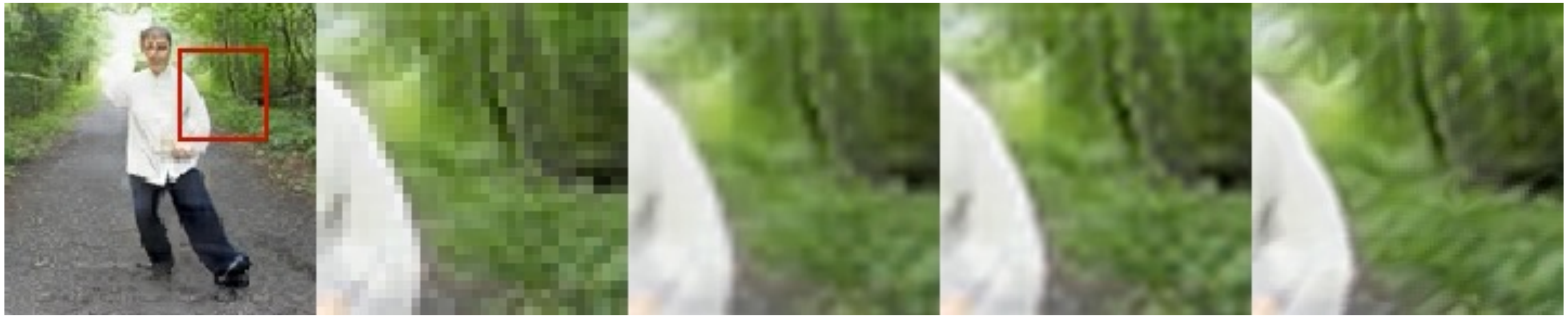
UCF-101



Sky



Kinetics-Food



Summary

Summary: We make video generation scalable leveraging implicit neural representations

We propose DIGAN = Dynamics-aware Implicit Generative Adversarial Networks

1. Achieves state-of-the-art performance on various video generation benchmarks
2. Can generate long videos of high-resolution frames without demanding recourses
3. Have lots of intriguing properties, such as spatiotemporal inter-/extra-polation
4. Non-autoregressive generation of videos is possible

More details can be found:

- Paper: https://arxiv.org/abs/2202.10571
- Code: https://github.com/sihyun-yu/digan
- Project page: https://sihyun-yu.github.io/digan

Please drop by our poster session for more details!