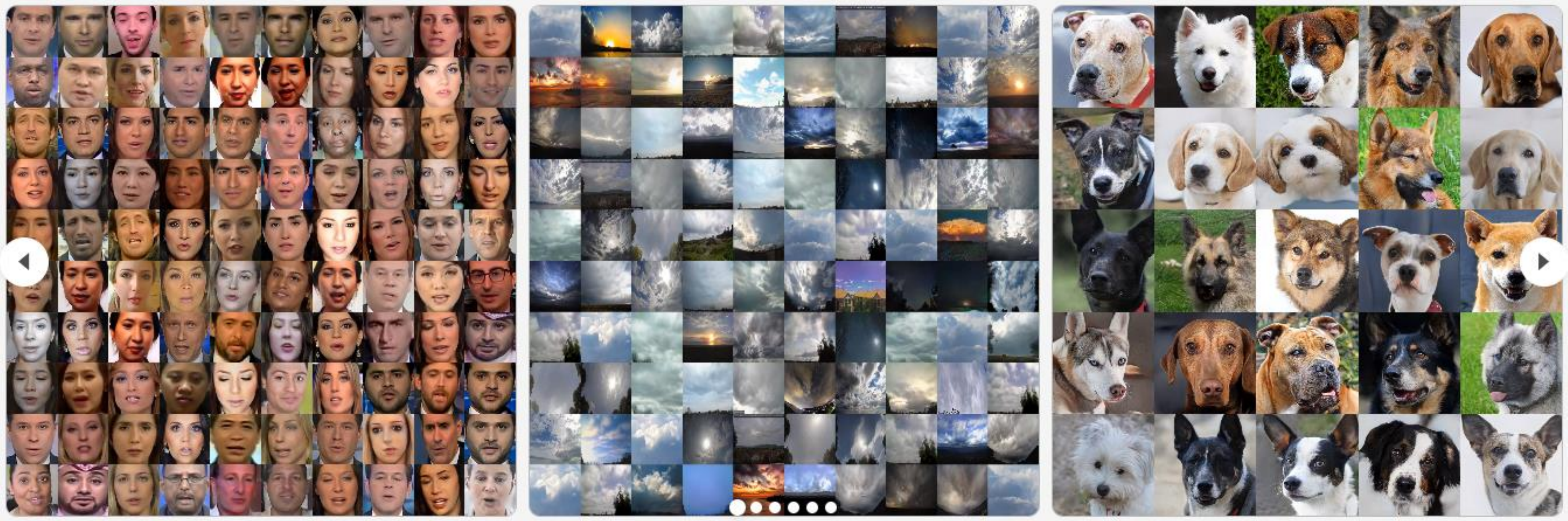
A Good Image Generator Is What You Need for High-Resolution Video Synthesis (ICLR 2021, Spotlight)

Yu Tian[1], Jian Ren[2], Menglei Chai[2], Kyle Olszewski[2], Xi Peng[3], Dimitris N. Metaxas[1], Sergey Tulyakov[2]
[1]Rutgers University, [2]Snap Inc., [3]University of Delaware



link

발표: 정채연

Difficulties in Video Synthesis

1. Low resolution, low quality
2. High training cost
3. Lack of training data

MoCoGAN (64 res)

TGANv2 (256 res)

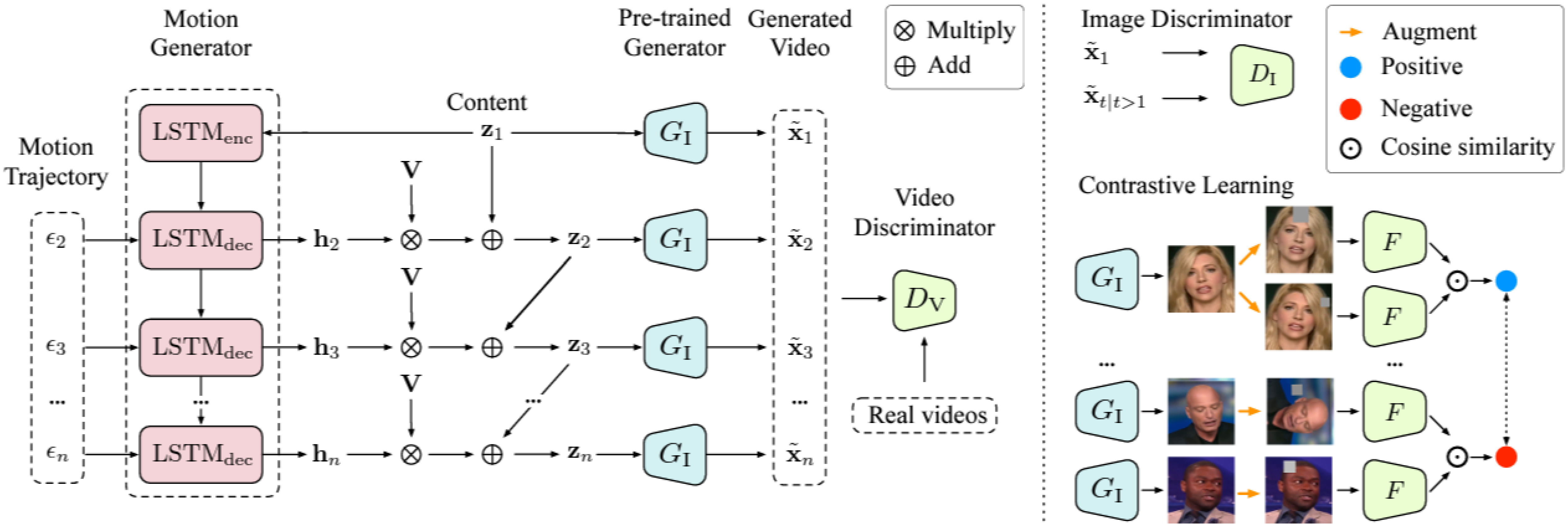
Contributions

1. High quality even in high resolution (~1024x1024) using pre-trained image generator
2. Computationally more efficient (less training time)
3. Cross-domain video synthesis: move images using video dataset from different domain via motion/content disentanglement

StyleGAN2 (1024 res)

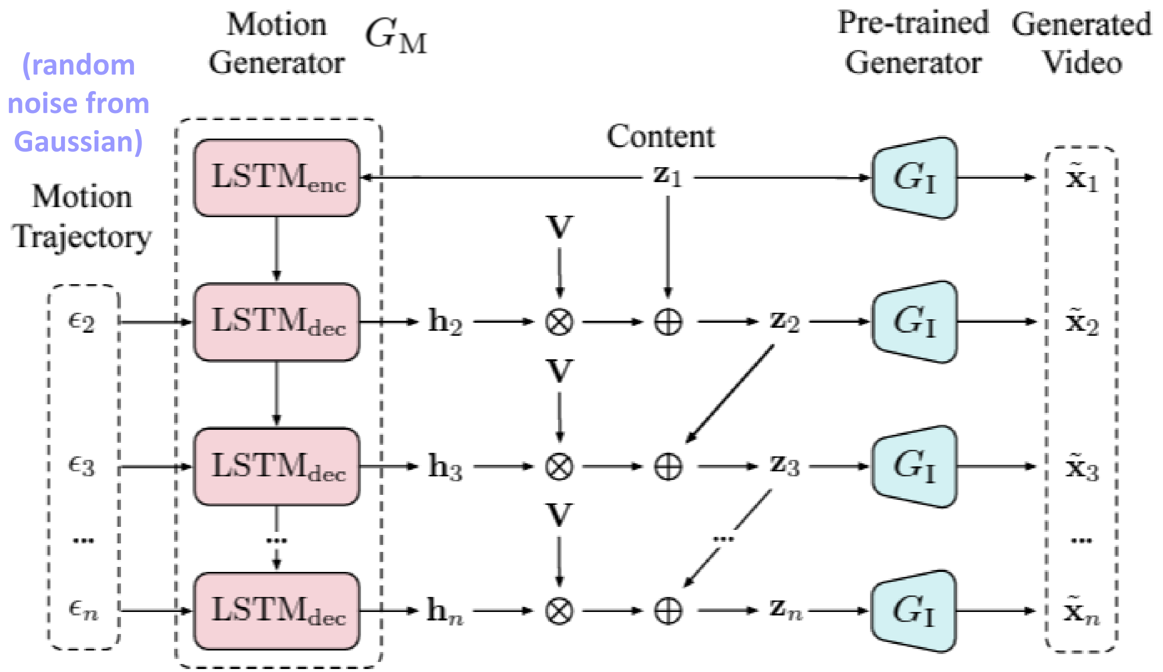
Method

Overview



Method

Motion Generator



$$\mathbf{h}_1, \mathbf{c}_1 = \mathrm{LSTM}_{\mathrm{enc}}(\mathbf{z}_1)$$

$$\mathbf{h}_t, \mathbf{c}_t = \mathrm{LSTM}_{\mathrm{dec}}(\epsilon_t, (\mathbf{h}_{t-1}, \mathbf{c}_{t-1})), \quad t = 2, 3, \cdots, n,$$

$$\mathbf{z}_t = \mathbf{z}_{t-1} + \lambda \cdot \mathbf{h}_t \cdot \mathbf{V}, \quad t = 2, 3, \cdots, n, \quad \mathbf{h}_t \in [-1, 1]$$

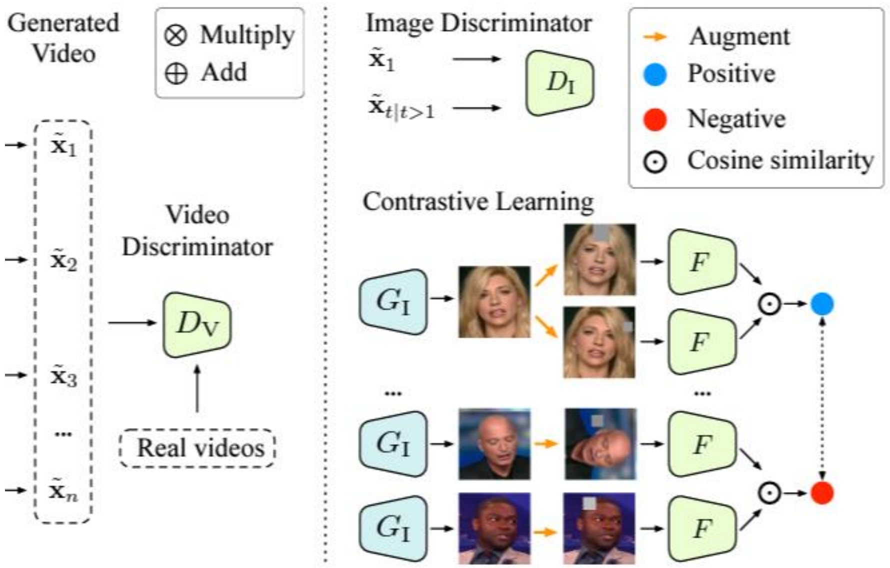
(from PCA)

$$G_{\mathrm{M}}(\mathbf{z}_1) = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n\}$$

$$\tilde{\mathbf{v}} = G_{\mathrm{I}}(G_{\mathrm{M}}(\mathbf{z}_1))$$

Method

Training Losses

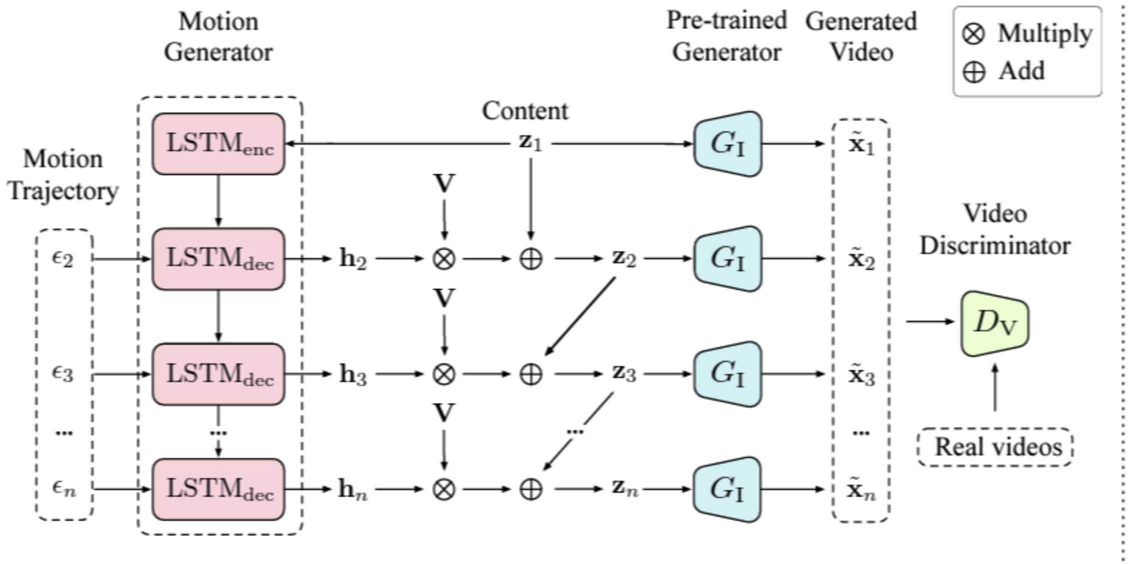


$$\min_{G_M}(\max_{D_V}\mathcal{L}_{D_V}+\max_{D_I}\mathcal{L}_{D_I})+\max_{G_M}(\lambda_m\mathcal{L}_m+\lambda_f\mathcal{L}_f)+\min_{D_I}(\lambda_{contr}\mathcal{L}_{contr})$$

1. Video discriminator loss
2. Image discriminator loss (for quality matching)
3. Contrastive loss & feature matching loss (for content matching)
4. Mutual information loss (for motion diversity)

Method

Training Losses - Video Discriminator Loss

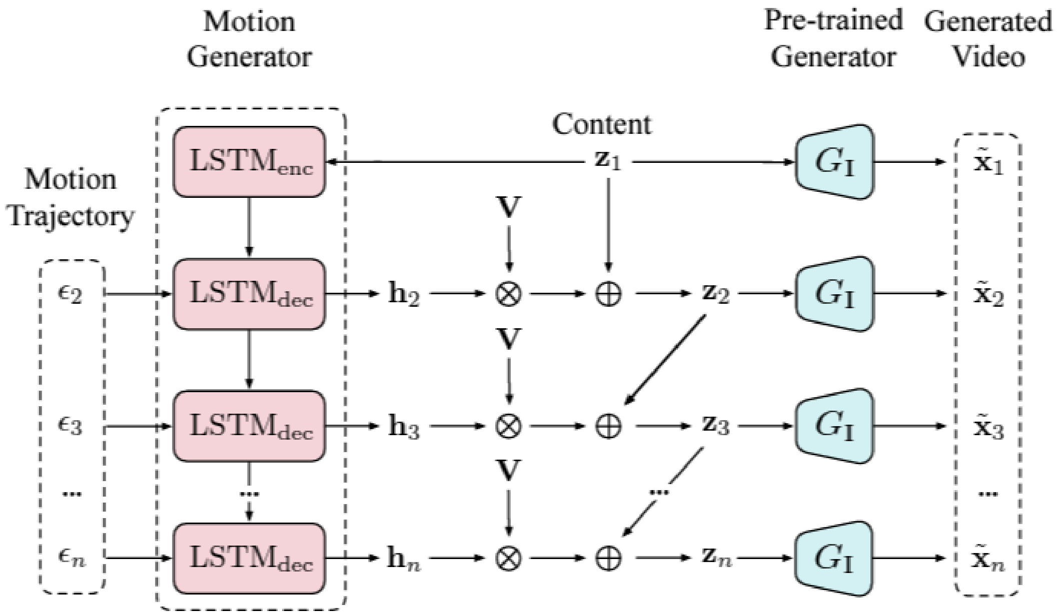


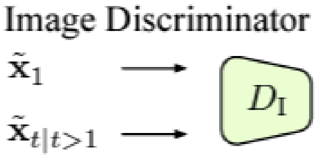
$$\mathcal{L}_{D_\mathrm{V}} = \mathbb{E}_{\mathbf{v}\sim p_v}[\log D_\mathrm{v}(\mathbf{v})] + \mathbb{E}_{\mathbf{z}_1\sim p_z}[\log(1 - D_\mathrm{V}(G_\mathrm{I}(G_\mathrm{M}(\mathbf{z}_1))))]$$

D_v : multi-scale PatchGAN discriminator with 3D Conv

Method

Training Losses - Image Discriminator Loss





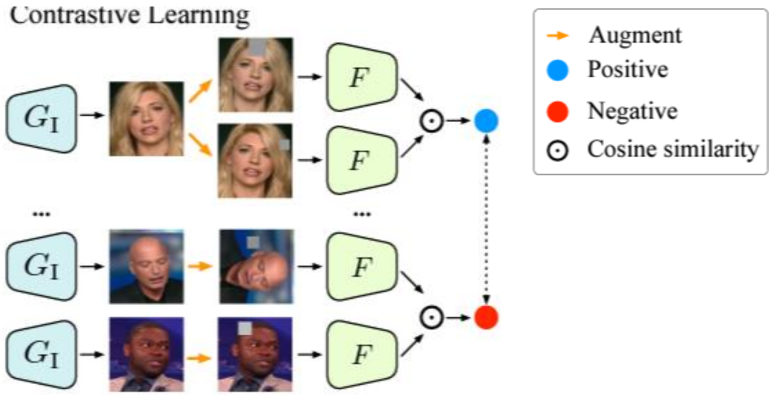
$$\mathcal{L}_{D_\mathrm{I}} = \mathbb{E}_{\mathbf{z}_1 \sim p_z}\left[\log D_\mathrm{I}(G_\mathrm{I}(\mathbf{z}_1))\right] + \mathbb{E}_{\mathbf{z}_1 \sim p_z, \mathbf{z}_t \sim G_\mathrm{M}(\mathbf{z}_1)|t>1}\left[\log(1 - D_\mathrm{I}(G_\mathrm{I}(\mathbf{z}_t)))\right]$$

for quality matching

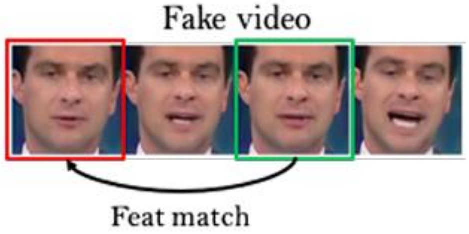
Method

Training Losses - Contrastive Loss & Feature Matching Loss

$$\mathcal{L}_{\text{contr}} = -\sum_{i=1}^{N}\sum_{\alpha=a}^{b} \log \frac{\exp(\text{sim}(F(\tilde{\mathbf{x}}_t^{(ia)}), F(\tilde{\mathbf{x}}_t^{(ib)}))/\tau)}{\sum_{j=1}^{N}\sum_{\beta=a}^{b} \mathbb{1}_{[j\neq i]}(\exp(\text{sim}(F(\tilde{\mathbf{x}}_t^{(i\alpha)}), F(\tilde{\mathbf{x}}_t^{(j\beta)}))/\tau)}$$



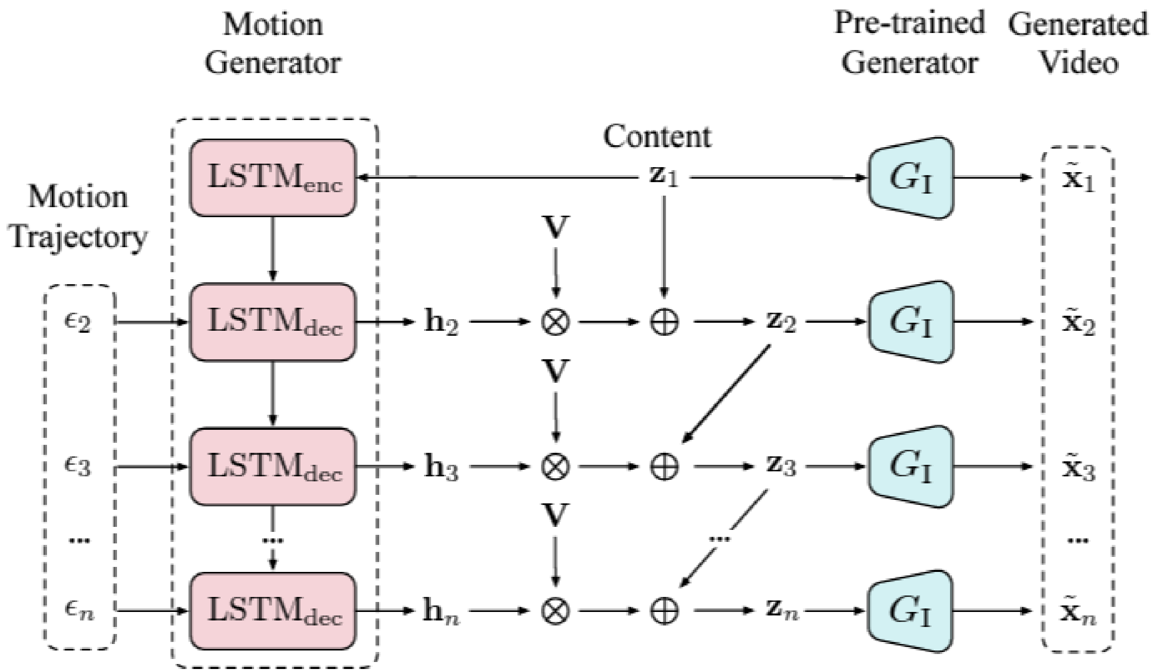
$\mathcal{L}_f$ = sim(F($\tilde{x}_0$), F($\tilde{x}_t$)) (t>0)



for content matching

Method

Training Losses – Mutual Information Loss



for motion diversity

(2-layer MLP)

$$\mathcal{L}_{\mathrm{m}} = \frac{1}{n-1}\sum_{t=2}^{n} \mathrm{sim}(H(\mathbf{h}_t), \epsilon_t)$$

$$\mathrm{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v} / \|\mathbf{u}\| \, \|\mathbf{v}\|$$

(cosine similarity)

Experiments

Video Generation

- UCF-101 with 101 sport categories

Table 1: IS and FVD on UCF-101.

Method	IS (↑)	FVD (↓)
VGAN	$8.31 \pm .09$	-
TGAN	$11.85 \pm .07$	-
MoCoGAN	$12.42 \pm .07$	-
ProgressiveVGAN	$14.56 \pm .05$	-
LDVD-GAN	$22.91 \pm .19$	-
TGANv2	$26.60 \pm .47$	1209 ± 28
DVD-GAN	$27.38 \pm .53$	-
Ours	$\mathbf{33.95 \pm .25}$	$\mathbf{700 \pm 24}$

Experiments

Video Generation

* ACD: Average Content Distance
(diff of average colors between frames)

- Face Forensics

Table 2: FVD, ACD, and Human Preference on FaceForensics.

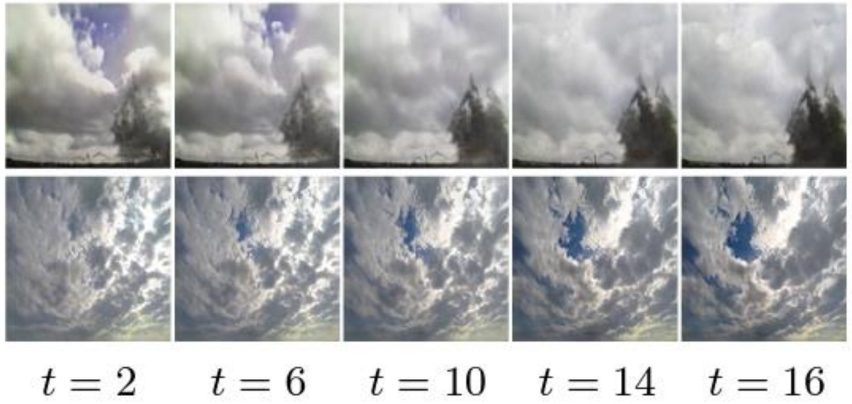
Method	FVD (↓)	ACD (↓)
GT	9.02	0.2935
TGANv2	58.03	0.4914
Ours	**53.26**	**0.3300**

Method	Human Preference (%)
Ours / TGANv2	**73.6** / 26.4

Experiments

Video Generation

- Sky Time-Lapse



Method	FVD (↓)	PSNR (↑)	SSIM (↑)
Up-B	-	25.367	0.781
MDGAN	840.95	13.840	0.581
DTVNet	451.14	21.953	0.531
Ours	**77.77**	**22.286**	**0.688**

Experiments

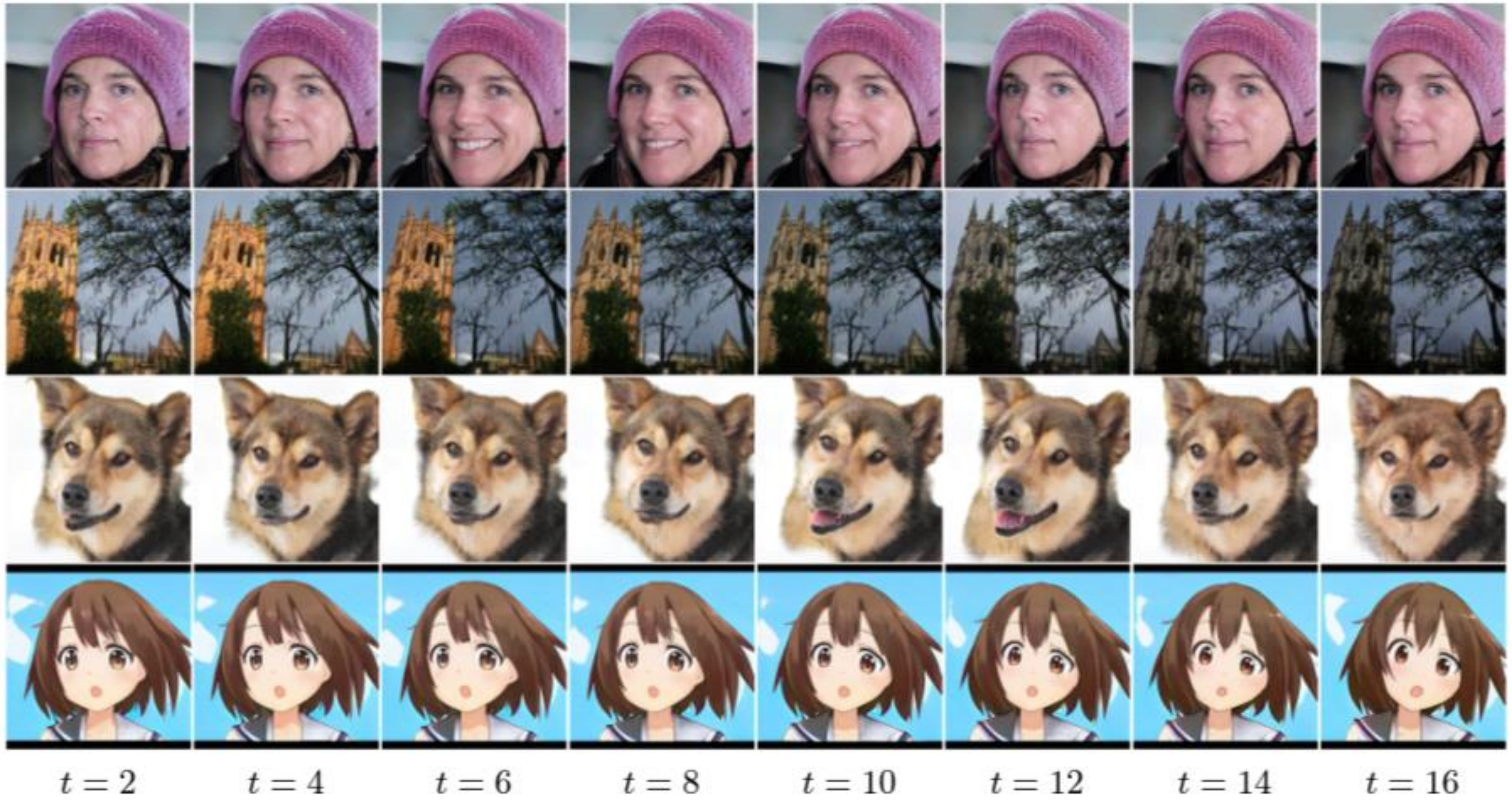
Cross-Domain Video Generation

(FFHQ, VoxCeleb)

(LSUN-Church, TLVDB)

(AFHQ-Dog, VoxCeleb)

(AnimeFaces, VoxCeleb)



Experiments

Cross-Domain Video Generation

(FFHQ, VoxCeleb)

(LSUN-Church, TLVDB)

(AFHQ-Dog, VoxCeleb)

(AnimeFaces, VoxCeleb)

Experiments

Content/Motion Disentanglement

Figure 5: The first and second row (also the third and fourth row) share the same initial content code but with different motion codes.

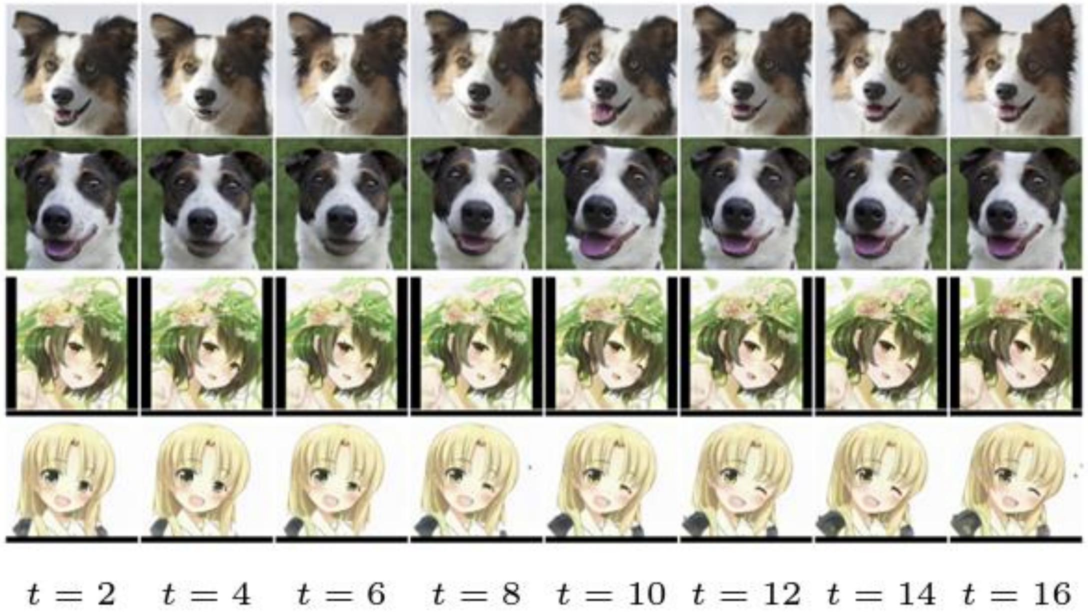


Figure 6: The first and second row (also the third and fourth row) share the same motion code but with different content codes.

Experiments

Ablation Study

Table 4: Ablation study on UCF-101.

Method	IS (↑)	FVD (↓)
w/o Eqn. 2	28.20	790.87
w/o D_{I}	33.22	796.67
w/o D_{V}	33.84	867.43
Full-128	32.36	838.09
Full-256	**33.95**	**700.00**

Table 5: Ablation study on (FFHQ, VoxCeleb).

Method	w/o $\mathcal{L}_{\mathrm{contr}}$	w/o $\mathcal{L}_{\mathrm{m}}$	Full
ACD (↓)	0.5328	0.5158	**0.4353**

Method	Human Preference (%)
Full *vs* w/o $\mathcal{L}_{\mathrm{contr}}$	**68.3** / 31.7
Full *vs* w/o $\mathcal{L}_{\mathrm{m}}$	**64.4** / 35.6

Eqn. 2 : $\mathbf{z}_t = \mathbf{z}_{t-1} + \lambda \cdot \mathbf{h}_t \cdot \mathbf{V}, \quad t = 2, 3, \cdots, n$

w/o Eqn. 2 : $\mathbf{z}_t = \mathbf{h}_t$

Experiments

Ablation Study

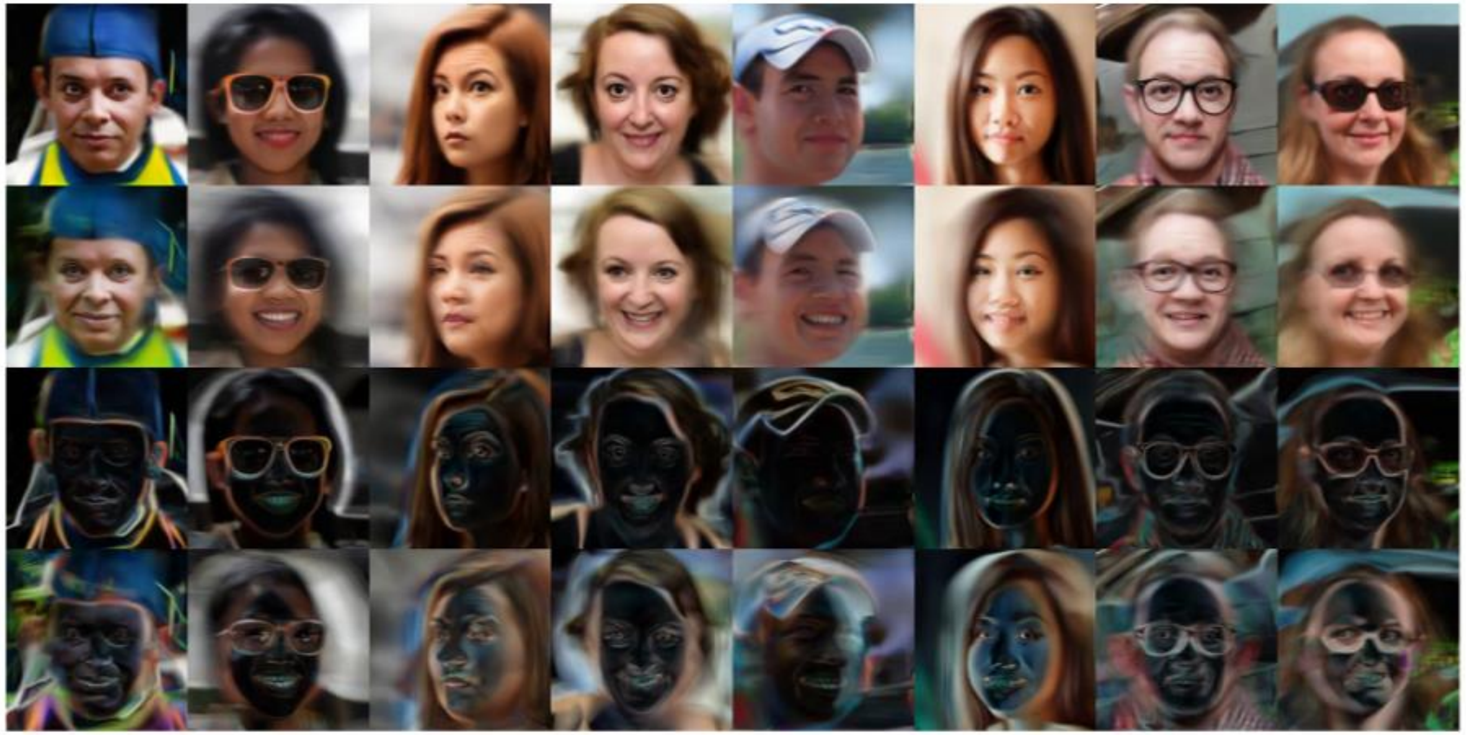


Figure 23: **Row 1 and 3**: The last frame of the mean-video and per-pixel std of *w/o* $\mathcal{L}_{\mathrm{m}}$ model. **Row 2 and 4**: The last frame of the mean-video and per-pixel std of the *Full* model. The *Full* model has a more blurry mean-video and higher per-pixel std, which indicates more diverse motion.

w/o $\mathcal{L}_{\mathrm{m}}$ vs. Full

Experiments

Long Sequence Generation

More steps for LSTM decoder

Motion Interpolation