Make-A-Video: Text-to-Video Generation without Text-Video Data

Meta AI

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Devi Parikh, Sonal Gupta, Yaniv Taigman

Introduction

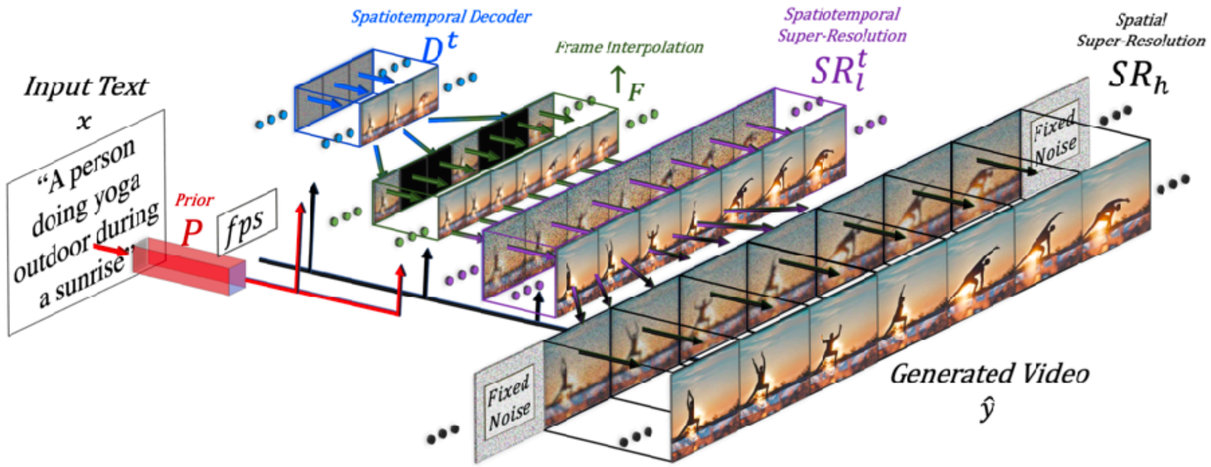
✓**(Alt-Text, Image) 쌍의 데이터는 웹 상에 넘쳐나기 때문에 Text-to-Image(T2I) 분야의 성공을 이룸**

- 반면에 (Text, Video) 쌍의 데이터는 수집하기 힘듦
- Text-to-Video(T2V)를 Scratch부터 학습하기에는 낭비임

✓**비지도 학습은 많은 양의 데이터를 학습하게 해주고, 보다 중요한 Representation 학습을 도와줌**

- 이는 NLP 분야를 발전시키는 데 큰 성공을 이루었고, 지도 학습보다 더 높은 성능을 이룰 수 있음

Introduction



✓**Make-A-Video 모델**

-Spatiotemporally Factorized Diffusion 모델을 통한 T2I 모델의 T2V로의 효율적인 확장

✓**(Text, Video) 쌍의 데이터를 우회하기 위해 (Text, Image) 쌍의 Prior를 활용**

✓**사용자 입력 텍스트 기반의 고화질, 높은 프레임의 비디오 생성을 하는 고화질 전략을 사용**

Method

✓**Base T2I 모델**

-(Text, Image) 쌍의 데이터로 학습

✓**시공간 Conv.와 Att. 레이어**

-신경망의 블록을 시간 차원으로 확장

✓**Frame Interpolation 네트워크**

-시공간 신경망과 높은 프레임 속도 추가

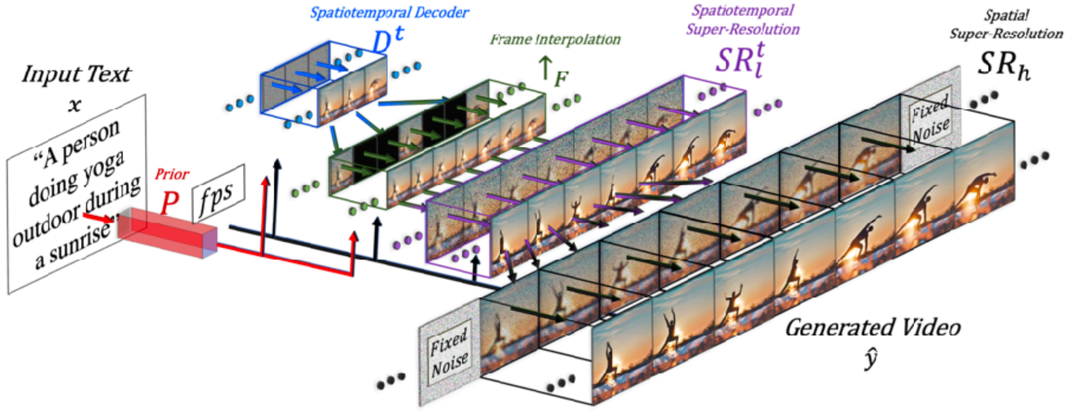


Figure 2: **Make-A-Video high-level architecture.** Given input text x translated by the prior P into an image embedding, and a desired frame rate fps, the decoder D^t generates 16 64×64 frames, which are then interpolated to a higher frame rate by $\uparrow_F$, and increased in resolution to 256×256 by SR_l^t and 768×768 by SR_h, resulting in a high-spatiotemporal-resolution generated video $\hat{y}$.

$$\hat{y}_t = \mathrm{SR}_h \circ \mathrm{SR}_l^t \circ \uparrow_F \circ \mathrm{D}^t \circ \mathrm{P} \circ (\hat{x}, \mathrm{C}_x(x)),$$

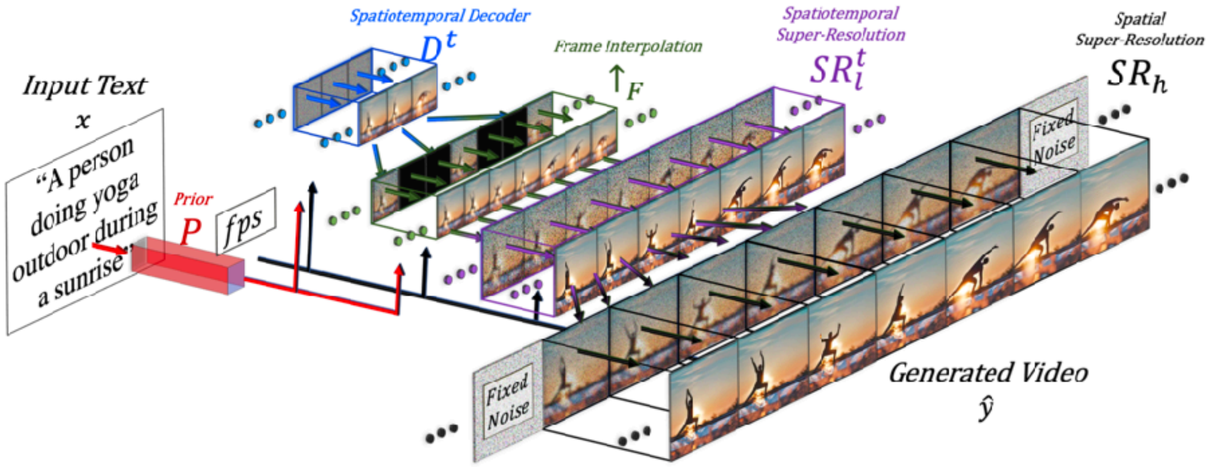
Method - T2I

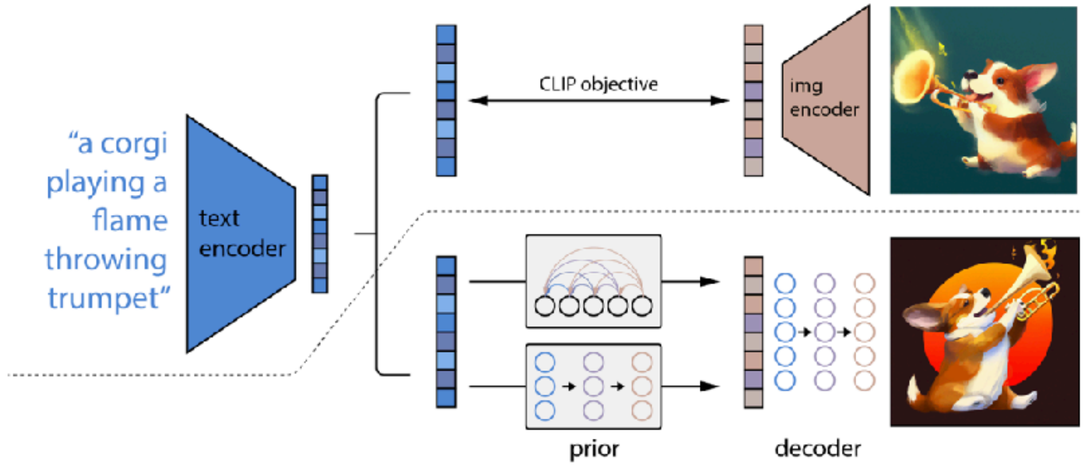
✓**Prior Network (P)**

-텍스트 임베딩(x_e)과 BPE 텍스트 토큰($\hat{x}$)에 대한 이미지 임베딩(y_e) 생성

✓**Decoder Network (D)**

-이미지 임베딩(y_e)을 조건으로 하여 64x64 RGB 이미지($\hat{y}_l$) 생성





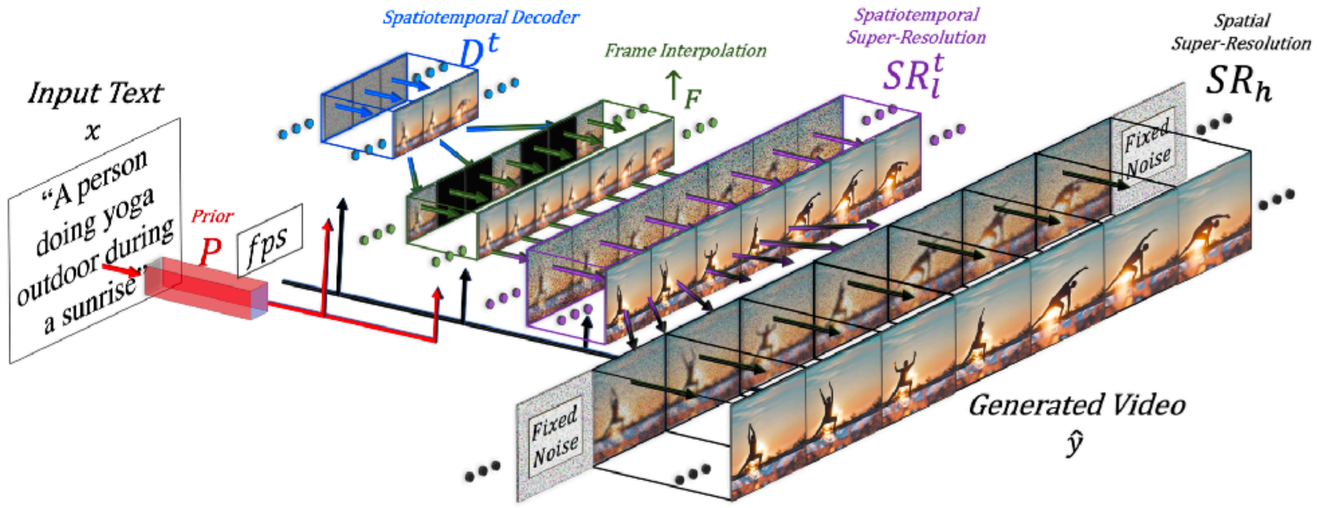
Method - T2I

✓**Super-Resolution Network (SR_l, SR_h)**

-64x64 RGB 이미지($\hat{y}_l$)를 256, 768 크기로 올려 최종 이미지($\hat{y}$) 생성

-Flickering을 보완하기 위해 프레임 간 Hallucinating(환영) 정보 포함

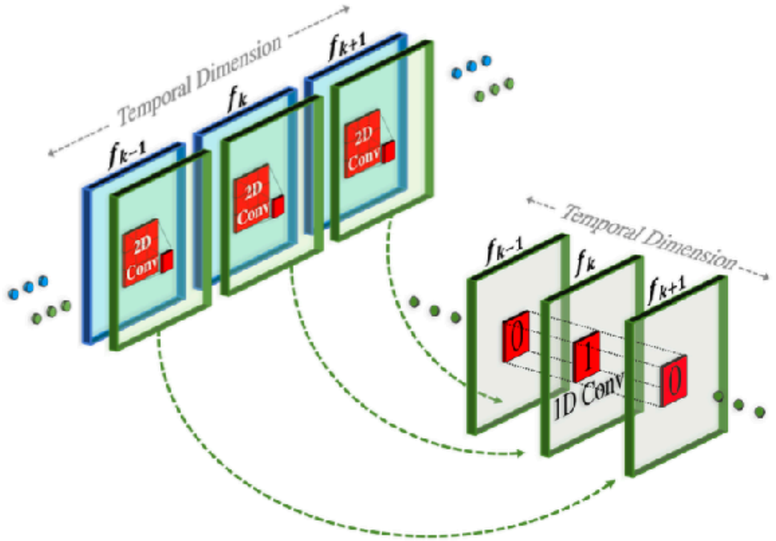
-Temporal Dimension을 다루기에는 메모리/컴퓨팅 제약으로 공간 정보만 다룸



Method - Spatiotemporal Layers

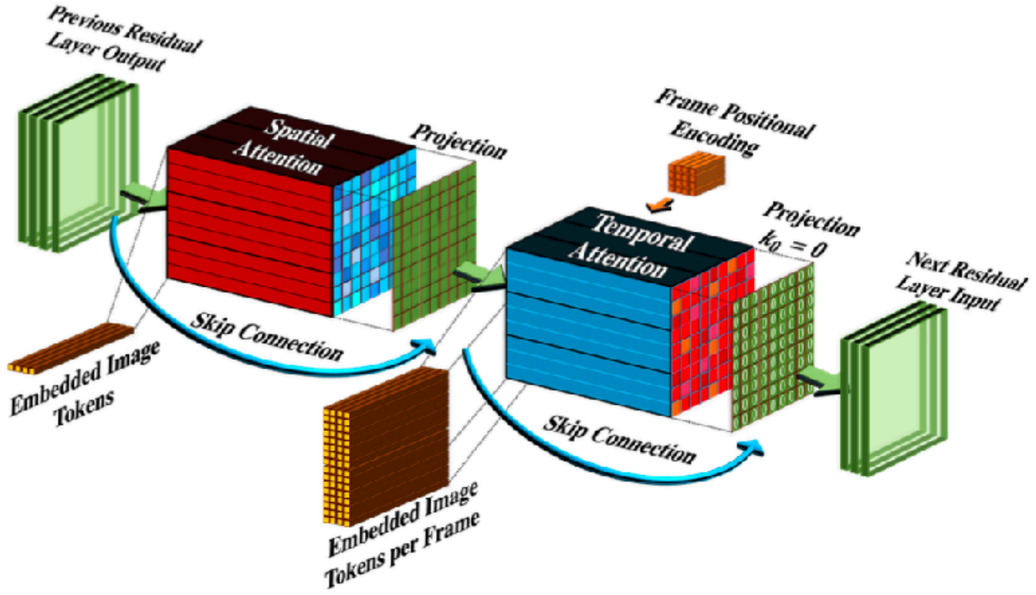
✓Pseudo-3D Convolutional Layers

$$Conv_{P3D}(h) := Conv_{1D}(Conv_{2D}(h) \circ T) \circ T,$$



✓Pseudo-3D Attention Layers

$$ATTN_{P3D}(h) = unflatten(ATTN_{1D}(ATTN_{2D}(flatten(h)) \circ T) \circ T).$$



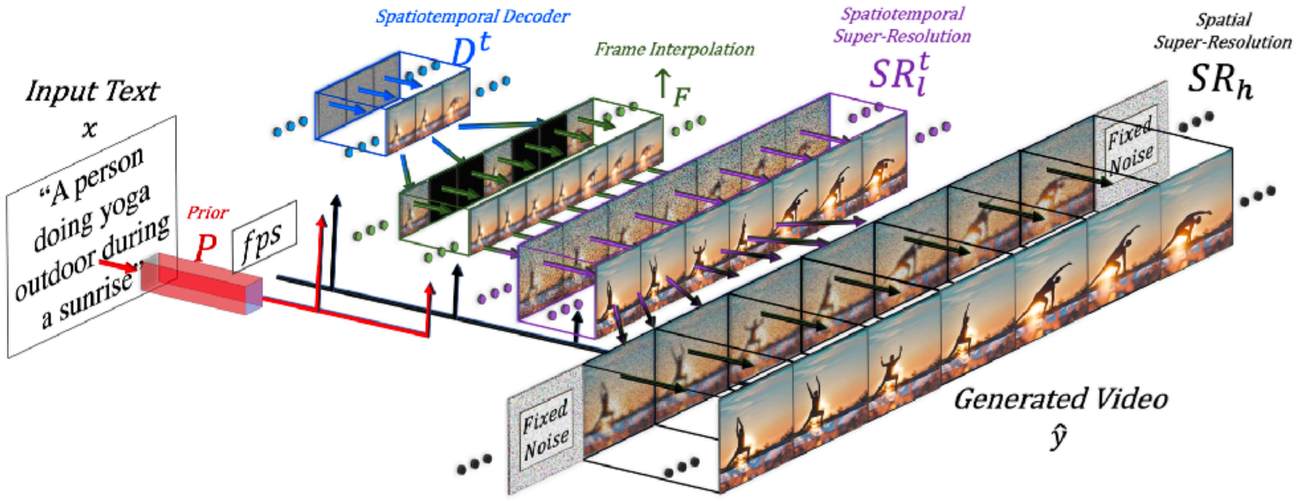
Method - Spatiotemporal Layers

✓**Frame Rate Conditioning (*fps*)**

-T2I의 추가 Conditioning Parameter로 생성된 동영상의 초당 프레임 수를 나타냄

-학습에 사용 가능한 동영상의 제한된 볼륨 처리가 가능한 Augmentation 역할

-추론 과정에서 생성된 동영상에 대한 추가 제어 제공



Method - Frame Inter. Network

✓Interpolation으로 생성된 동영상을 부드럽게 가능

✓전후 프레임 간의 Extrapolation으로 동영상의 길이를 증가할 수 있음

✓제한된 환경(e.g. 메모리, 컴퓨팅)에서 프레임률을 올리기 위해

-동영상 Upsampling이 가능하도록 마스킹된 입력 프레임을 Zero-Padding

-시공간 Decoder를 Finetuning 함

✓Finetuning 과정에서 U-Net의 입력에 4개의 채널을 추가함

-마스킹된 입력 RGB 동영상을 위한 3개의 채널 + 마스킹 여부를 나타내는 1개의 채널

Method - Training

✓**Make-A-Video의 모든 모델들은 독립적으로 학습되어짐**

✓**Prior 모델만 유일하게 Text를 입력으로 받음**

-(Text, Image) 쌍의 데이터만을 활용

✓**Decoder, Prior, SR 모델의 초반 학습**

-처음에 정렬된 텍스트 없이 이미지로 학습이 이루어짐

✓**이미지에 대한 학습 이후**

-새로운 Temporal 레이어를 추가

-Unlabeled 동영상 데이터에 대해서 Finetuning

Experiments

Table 1: T2V generation evaluation on MSR-VTT. Zero-Shot means no training is conducted on MSR-VTT. Samples/Input means how many samples are generated (and then ranked) for each input.

Method	Zero-Shot	Samples/Input	FID (↓)	CLIPSIM (↑)
GODIVA (Wu et al., 2021a)	No	30	–	0.2402
NÜWA (Wu et al., 2021b)	No	–	47.68	0.2439
CogVideo (Hong et al., 2022) (Chinese)	Yes	1	24.78	0.2614
CogVideo (Hong et al., 2022) (English)	Yes	1	23.59	0.2631
Make-A-Video (ours)	Yes	1	**13.17**	**0.3049**

Table 3: Human evaluation results compared to CogVideo (Hong et al., 2022) on DrawBench and our test set, and to VDM (Ho et al., 2022) on the 28 examples from their website. The numbers show the percentage of raters that prefer the results of our Make-A-Video model.

Comparison	Benchmark	Quality	Faithfulness
Make-A-Video (ours) vs. VDM	VDM prompts (28)	84.38	78.13
Make-A-Video (ours) vs. CogVideo (Chinese)	DrawBench (200)	76.88	73.37
Make-A-Video (ours) vs. CogVideo (English)	DrawBench (200)	74.48	68.75
Make-A-Video (ours) vs. CogVideo (Chinese)	Our Eval. Set (300)	73.44	75.74
Make-A-Video (ours) vs. CogVideo (English)	Our Eval. Set (300)	77.15	71.19

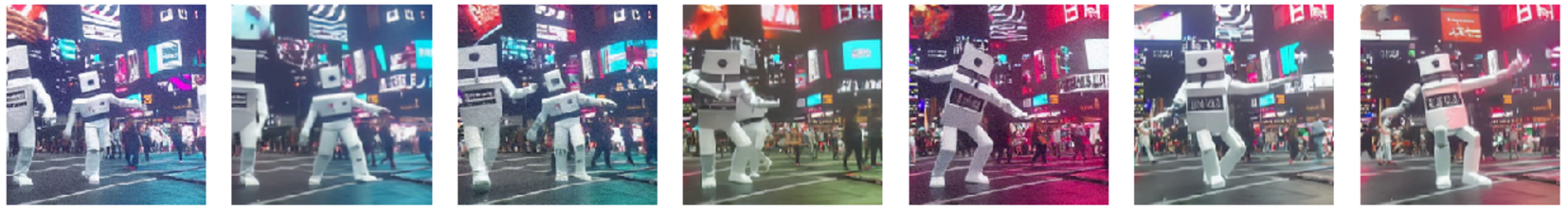
Table 2: Video generation evaluation on UCF-101 for both zero-shot and fine-tuning settings.

Method	Pretrain	Class	Resolution	IS (↑)	FVD (↓)
Zero-Shot Setting					
CogVideo (Chinese)	No	Yes	480 × 480	23.55	751.34
CogVideo (English)	No	Yes	480 × 480	25.27	701.59
Make-A-Video (ours)	No	Yes	256 × 256	**33.00**	**367.23**
Finetuning Setting					
TGANv2(Saito et al., 2020)	No	No	128 × 128	26.60 ± 0.47	-
DIGAN(Yu et al., 2022b)	No	No		32.70 ± 0.35	577 ± 22
MoCoGAN-HD(Tian et al., 2021)	No	No	256 × 256	33.95 ± 0.25	700 ± 24
CogVideo (Hong et al., 2022)	Yes	Yes	160 × 160	50.46	626
VDM (Ho et al., 2022)	No	No	64 × 64	57.80 ± 1.3	-
TATS-base(Ge et al., 2022)	No	Yes	128 × 128	79.28 ± 0.38	278 ± 11
Make-A-Video (ours)	Yes	Yes	256 × 256	**82.55**	**81.25**

Experiments

(a) A dog wearing a superhero outfit with red cape flying through the sky.

(b) There is a table by a window with sunlight streaming through illuminating a pile of books.



(c) Robot dancing in times square.

(d) Unicorns running along a beach, highly detailed.

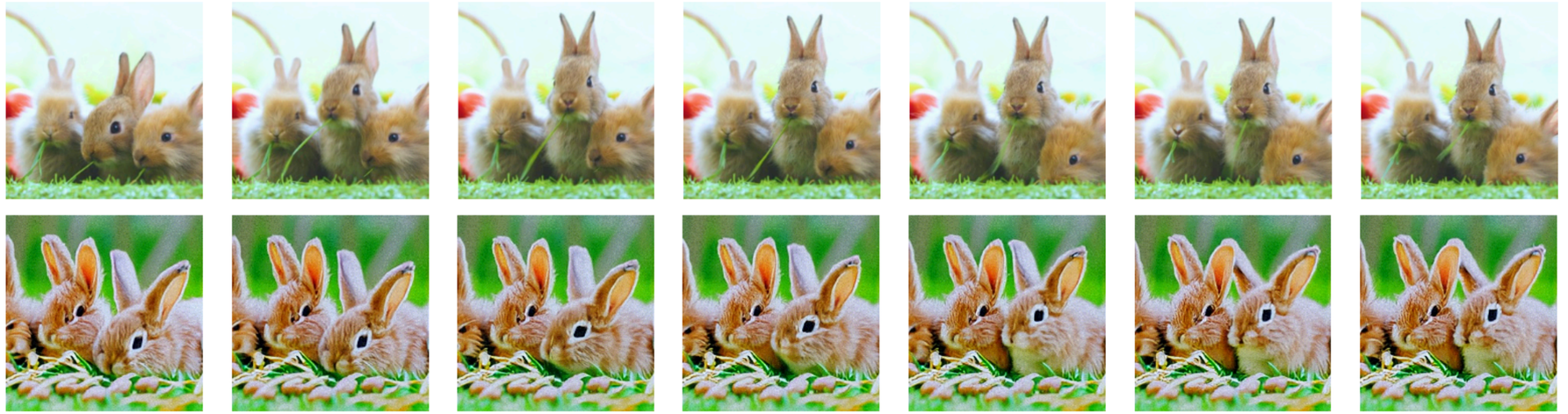
Experiments

(a) **T2V Generation**: comparison between VDM (top), CogVideo (mid), and Ours (bottom) for input “Busy freeway at night”.

Experiments

(b) **Image Animation**: leftmost shows the input image, and we animated it to be a video.

(c) **Image Interpolation**: given two images (leftmost and rightmost), we interpolate frames. Comparing FILM (left) and Ours (right).



(d) **Video Variation**: we can generate a new video (bottom) as a variant to the original video (top).