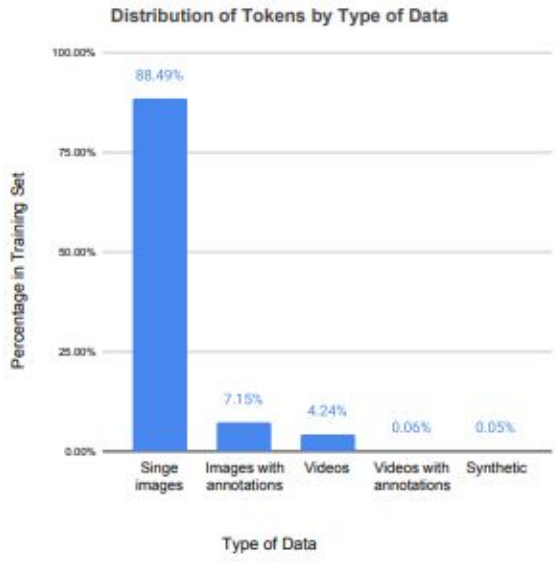
Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning

Motivation

What makes text-to-video (T2V) generation hard?

1. Video datasets are smaller
2. T2V models a significantly more complex distribution given the same information



Motivation

But we can leverage existing text-to-image (T2I) models

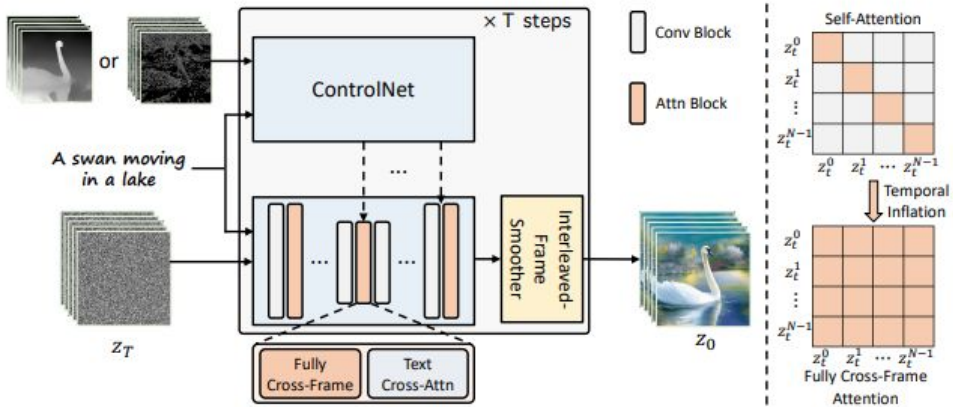
1. Learned priors from image datasets
2. Stronger conditioning signal

Motivation

How might we leverage pretrained T2I models?

1. With minimal/no training?

    -Add motion dynamics to T2I generations



Motivation

How might we leverage pretrained T2I models?

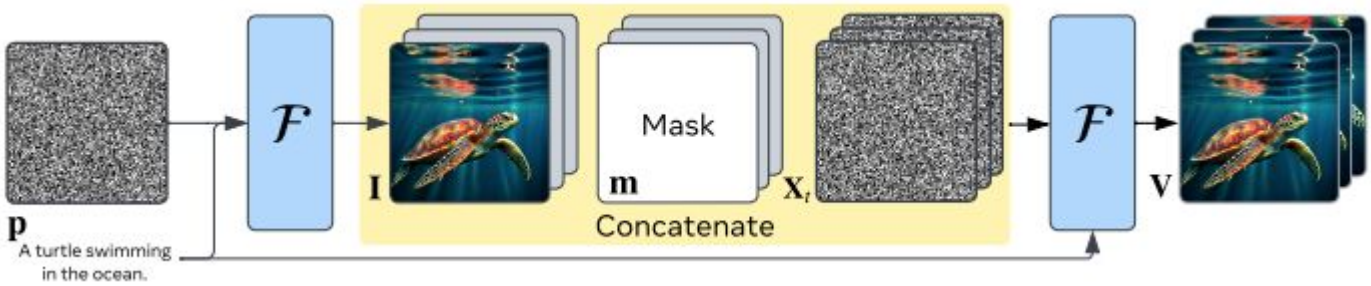
2. With known “tricks” from T2I literature?

Figure 6: The cascaded sampling pipeline starting from a text prompt input to generating a 5.3-second, 1280×768 video at 24fps. “SSR” and “TSR” denote spatial and temporal super-resolution respectively, and videos are labeled as frames×width×height. In practice, the text embeddings are injected into all models, not just the base model.

Introduction

Core idea: **Factorized** video generation representation

1) Generate first frame given a text prompt
2) Generate T frames given text prompt + first frame condition



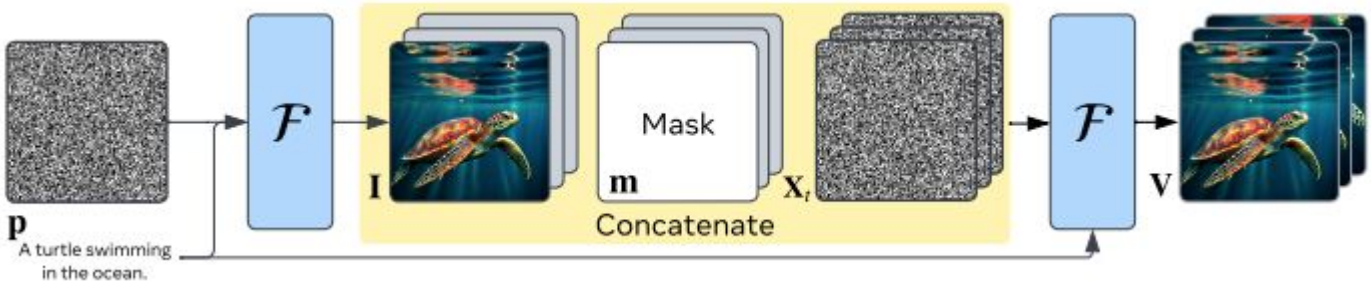
Emu Video - Inputs

Take video (T x 3 x H x W) and transform to latents (T x C x H x W)

Everything is done framewise

-Encoder is applied per-frame

-Latents are noised iid in the temporal dimension



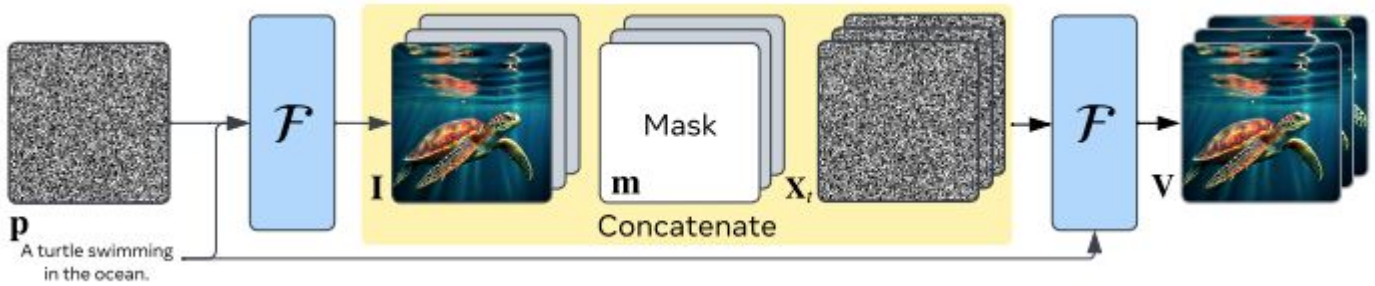
Emu Video - Inputs

Generate initial frame **I** with T2I model

Represent **I** is a single frame “video”

Zero-pad out to video latents shape (T x C x H x W)

Concatenate with temporal mask **m** (T x 1 x H x W)



Emu Video - Architecture

Core architecture is similar to Make-A-Video:

1. Initialize with pretrained T2I (**frozen**)
2. Add learnable temporal parameters
    i) 1D temporal conv after every spatial conv
    ii) 1D temporal attn after every spatial attn

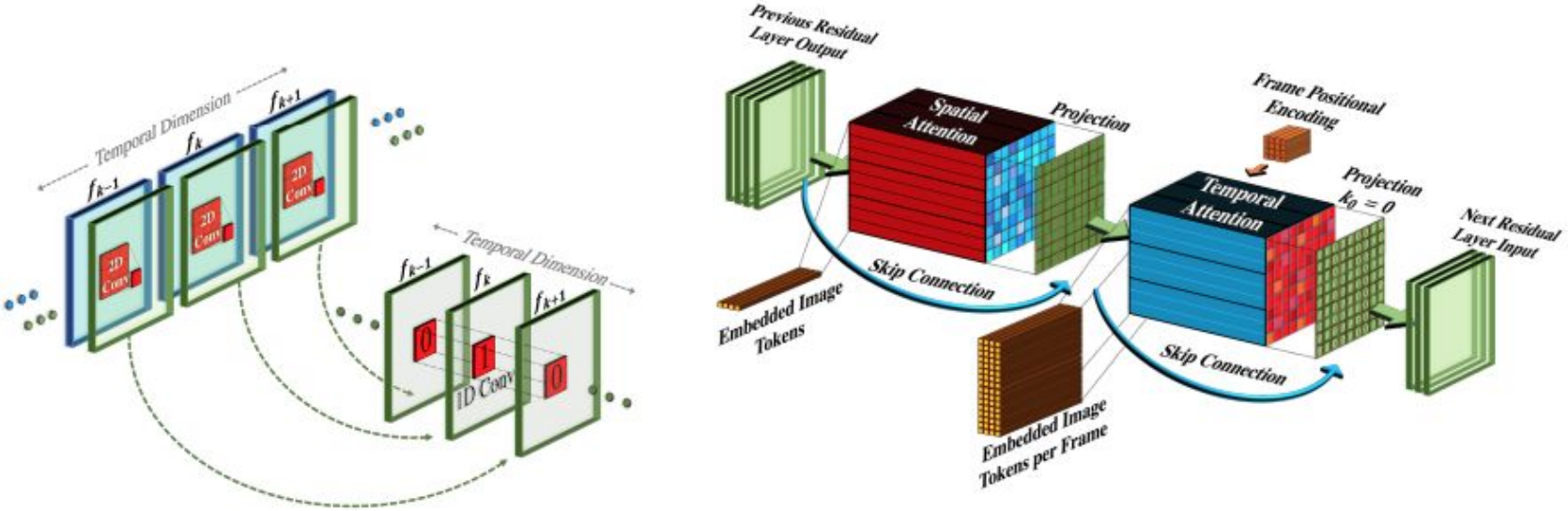
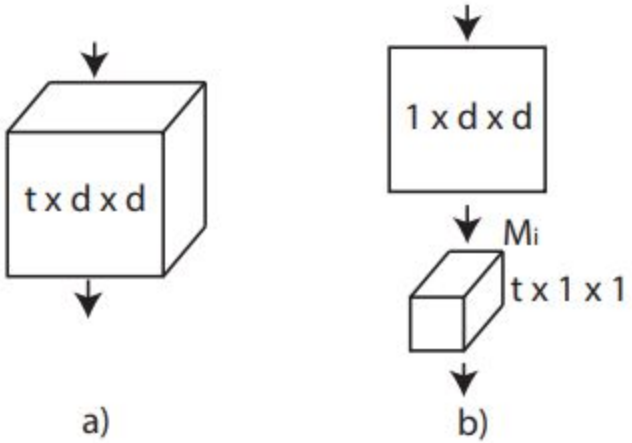


Figure 3: **The architecture and initialization scheme of the Pseudo-3D convolutional and attention layers, enabling the seamless transition of a pre-trained Text-to-Image model to the temporal dimension. (left)** Each spatial 2D conv layer is followed by a temporal 1D conv layer. The temporal conv layer is initialized with an identity function. **(right)** Temporal attention layers are applied following the spatial attention layers by initializing the temporal projection to zero, resulting in an identity function of the temporal attention blocks.


t x d x d
a)
1 x d x d
Mi
t x 1 x 1
b)

Emu Video - Architecture

Use another (pretrained) T2I to increase FPS (i.e. upsample T frames to T_p frames)

1. Interleave T frames with zero frames to get T_p x C x H x W input
2. Concatenate with mask **m'** indicating which frames are input frames T

Add temporal parameters like the core model, but only train temporal parameters

Underlying assumption is that temporal superresolution only requires temporal understanding

Emu Video - Training

Diffusion schedule is super important!

Typical schedulers induce a distribution shift at inference

Solution: manually set terminal diffusion step to zero-SNR at training

Emu Video - Training

Data: 34M video-text pairs

Model: 2.7B frozen spatial parameters (Emu), and 1.7B trainable temporal parameters

Multi-level optimization

1. Train for 80K iterations on simpler task: 256px 8fps 1s videos
2. Train for 15K iterations on end task: 512px 4ps 2s videos

Train interpolation model to consume 8 frame inputs and output 37 frames

Emu Video - Inference

1. Run model without temporal layers to get image condition
2. Run model with image condition + text prompt to get T frames
3. Increase frames with interpolation model

Emu Video - Evaluation

1. Quality
2. Faithfulness (to prompt)

Crucially, evaluators needs to JUstify their choICE (JUICE)

Quality: pixel sharpness, motion smoothness, recognizable objects/scenes

Faithfulness: spatial + temporal text alignment

Results

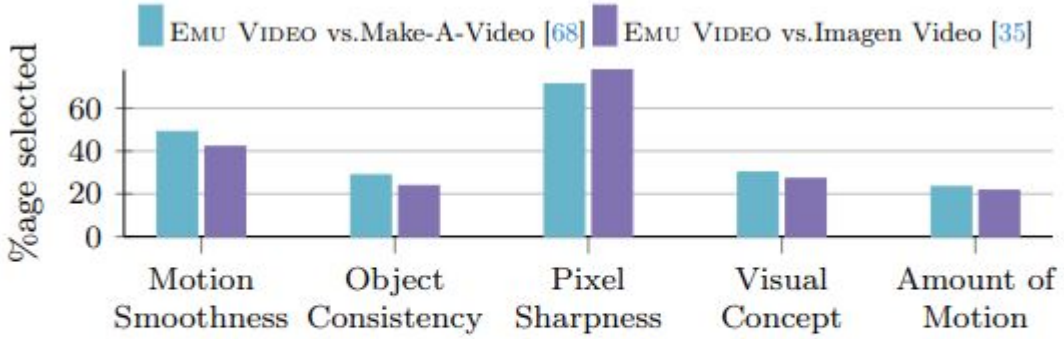


Fig. 6: Percentage of each reason selected for samples where EMU VIDEO wins against Make-A-Video [68] or Imagen Video [35] on Quality. Human raters pick EMU VIDEO primarily due to their pixel sharpness and motion smoothness, with an overall preference of 96.8% and 81.8% to each baseline, respectively.

Impact of Design Choices

a) Factorized (conditioning on first frame) vs. direct T2V
b) Zero-SNR schedule vs. linear schedule
c) Multi-stage training vs. direct 512px training
d) Finetuning on pre-identified “higher quality” videos
e) Freezing spatial parameters vs. not

Method	Q	F
Factorized	70.5	63.3

(a)

Method	Q	F
Zero SNR	96.8	88.3

(b)

Method	Q	F
Multi-stage	81.8	84.1

(c)

Method	Q	F
HQ finetuned	65.1	79.6

(d)

Method	Q	F
Frozen spatial	55.0	58.1

(e)

Final Discussion

1. All at once vs. autoregressive prediction?

2. Are cascades an engineering trick or a grounded necessity?

3. Will factorized representations like this stand the test of time?