Phenaki: Variable Length Video Generation From Open Domain Textual Descriptions

Shoubin Yu, Han Lin

Department of Computer Science, University of North Carolina at Chapel Hill

COMP790-170

11/06/2023

1st prompt: "A photorealistic teddy bear is swimming in the ocean at San Francisco"

2nd prompt: "The teddy bear goes under water"

3rd prompt: "The teddy bear keeps swimming under the water with colorful fishes"

4rd prompt: "A **panda** bear is swimming under water"

Introduction

Overview

- How **PHENAKI** address those:
  - learning video representation which **compresses** the video to a small representation of discrete tokens
  - This tokenizer uses **causal attention** in time, allows it to work with variable-length videos
  - joint training on a large corpus of **image-text** pairs as well as a smaller number of video-text examples

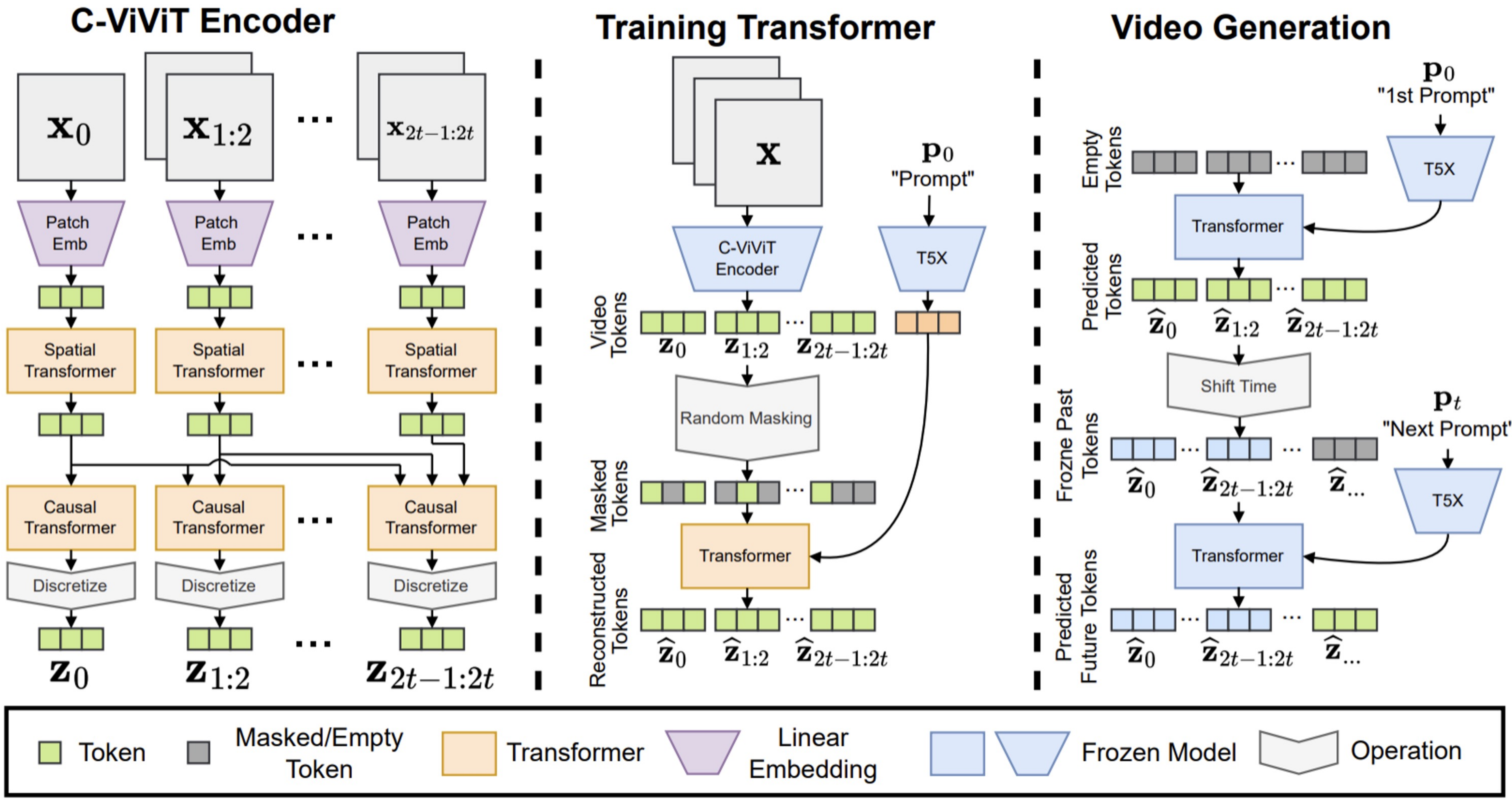
Overview

- What is **PHENAKI**
  - A realistic video synthesis conditional on a sequence of textual prompts

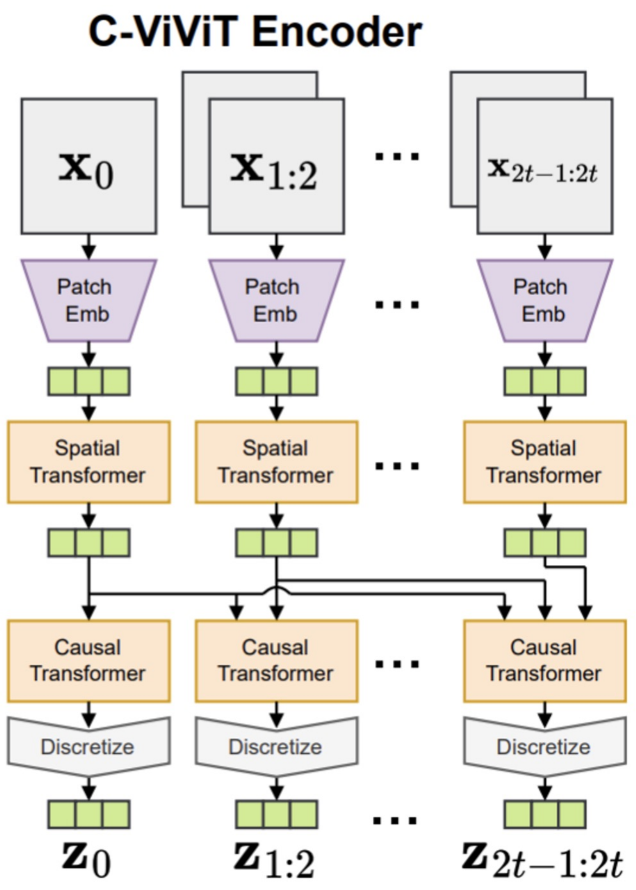
- What are the challenges behind text-2-video generation
  - computational cost
  - lack of high quality data
  - variable length of videos

Model

Model Overview



C-ViViT Encoder



Related Encoder works:

- VQ-GAN: allows for generating videos of arbitrary length, but highly redundant

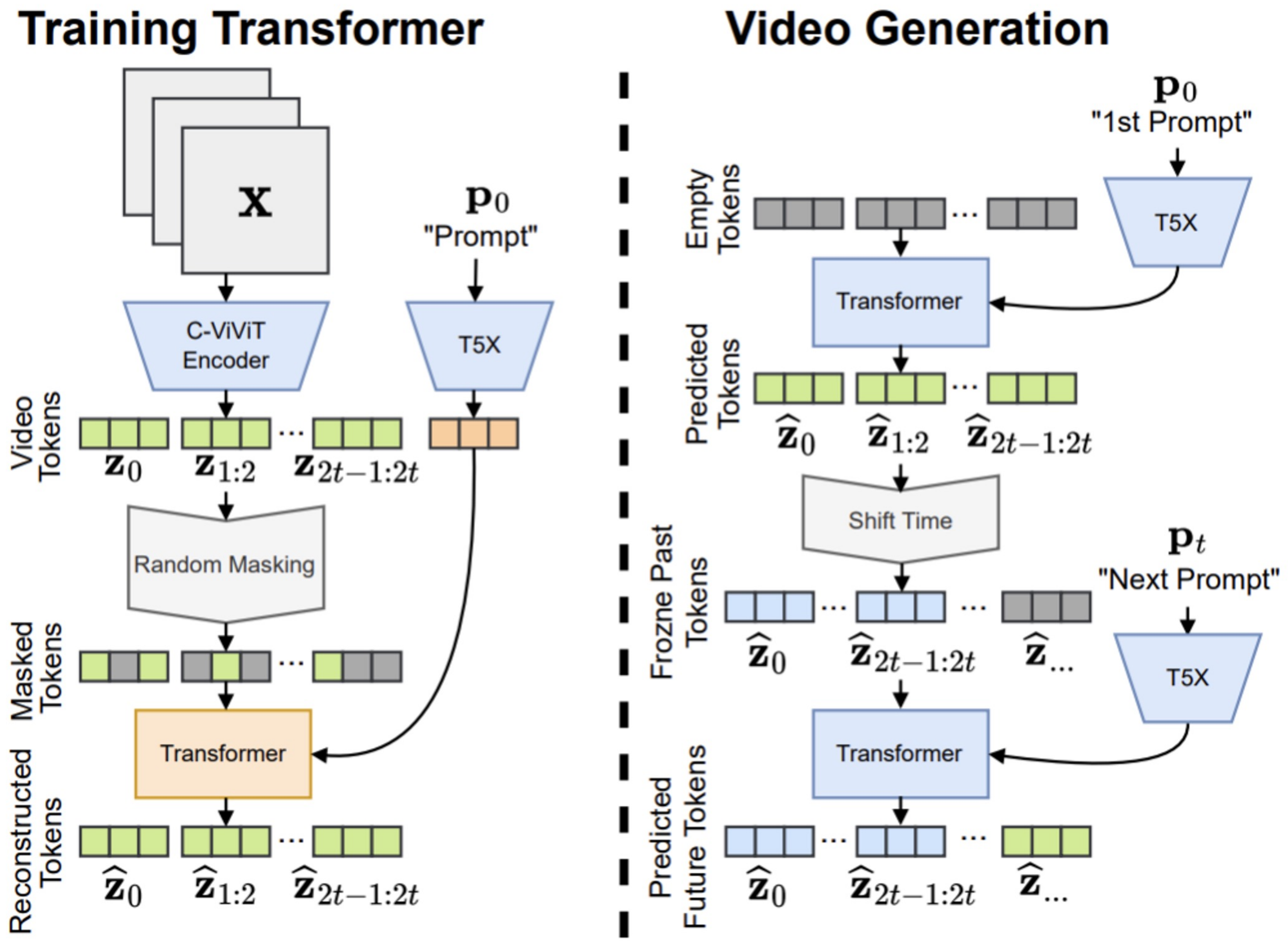
- VideoVQVAE: efficient but does not allow to generate variable length videos

C-ViViT

- generate videos of variable length while keeping the number of video tokens to a minimum

- Discretize: VQVAEs

Transformer



- Masked Visual Token Modeling (MVTM)

$$L_{\text{mask}} = -\sum_{\forall i \in [1,N], m_i=1} \log p(a_i | \mathbf{a}_{\bar{M}}, \mathbf{p}),$$

- Video Generation with Multiple prompts

Experiments

Evaluation Tasks:

- Text conditional video generation
- Text-image conditional video generation
- Story generation from dynamic text inputs
- Video reconstruction
- Video prediction

Evaluation Tasks:

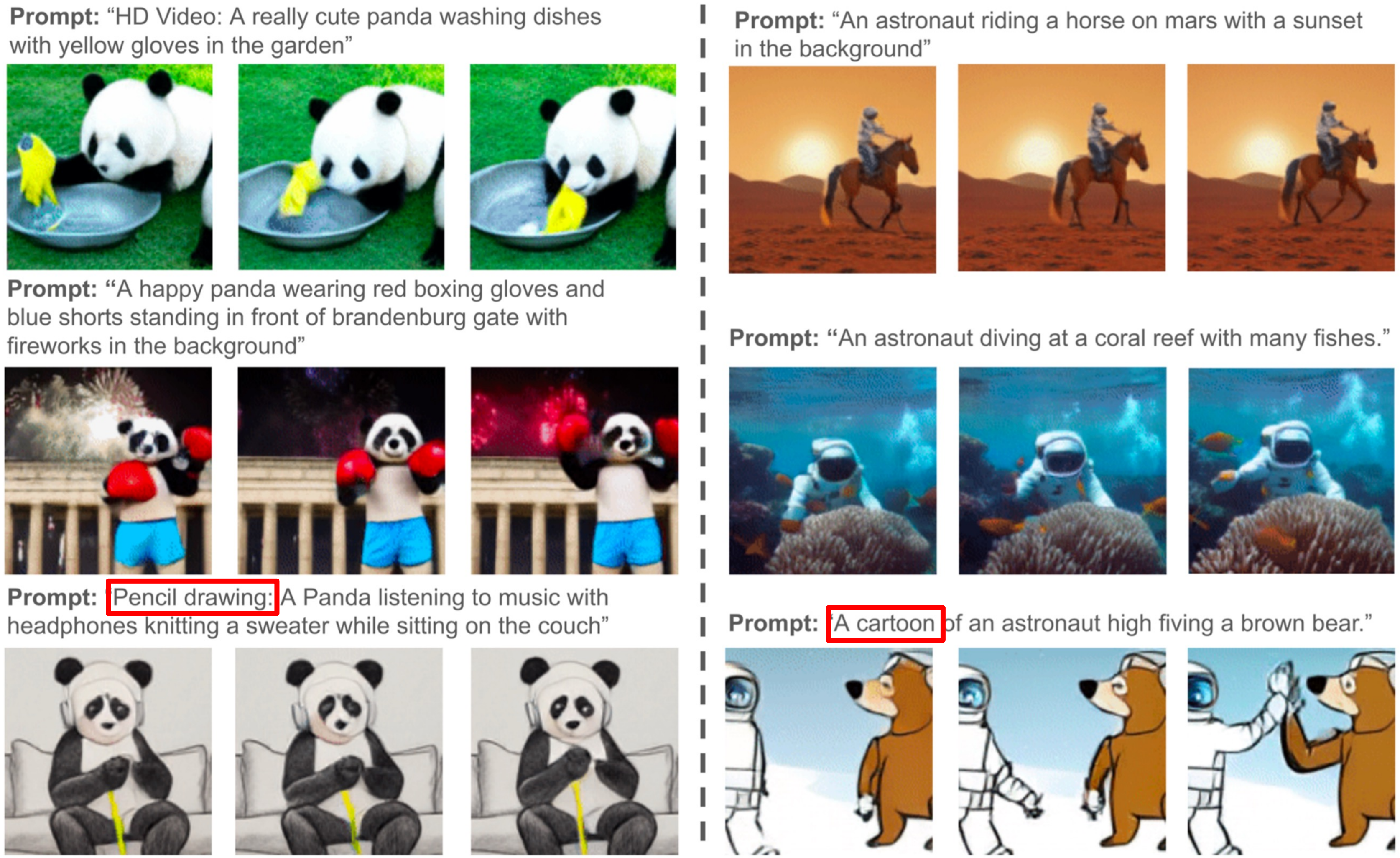
- Text conditional video generation
- Text-image conditional video generation
- Story generation from dynamic text inputs
- Video reconstruction
- Video prediction

Text Conditional Video Generation

- Training dataset:
  - ~15M text-video pairs at 8 FPS
  - ~450M text-image pairs (mostly from LAION-400M dataset)
  - During training, mix the video and image data with ratio 4:1

Text Conditional Video Generation

- Qualitative evaluation:



Text Conditional Video Generation

- Quantitative evaluation:
  - Phenaki: evaluated in zero-shot setting
  - Other baselines: fine-tuned on Kinetics-400 dataset

Table 1. Text to video comparisons on Kinetics-400 [22].

Method	FID Image ↓	FID Video ↓
T2V [25]	82.13	14.65
SC [5]	33.51	7.34
TFGAN [5]	31.76	7.19
NUWA	28.46	7.05
Phenaki [0-Shot]	37.74	$\boxed{3.84}$

Text Conditional Video Generation

- Joined text-to-image and text-to-video training:
  - Video-only training ➡️ significantly better FVD
  - Training with more image data ➡️ significantly better FID, and better text-video and text-image alignment (CLIP score)

Table 2. Text to video and text to image results highlighting the importance of image datasets in video models. Text-to-image evaluation is done on ∼40K images of LAION-400M [41].

Data Split	Text to Video			Text to Image	
Vid% / Img%	CLIP ↑	FID ↓	FVD ↓	CLIP ↑	FID ↓
100% / 0%	0.298	19.2	168.9	0.240	53.9
80% / 20%	0.303	21.4	198.4	0.289	29.4
50% / 50%	0.302	21.4	239.7	0.287	30.5

Evaluation Tasks:

- Text conditional video generation
- Text-image conditional video generation
- Story generation from dynamic text inputs
- Video reconstruction
- Video prediction

Text-Image Conditional Video Generation

- **Animate** existing images given a text prompt

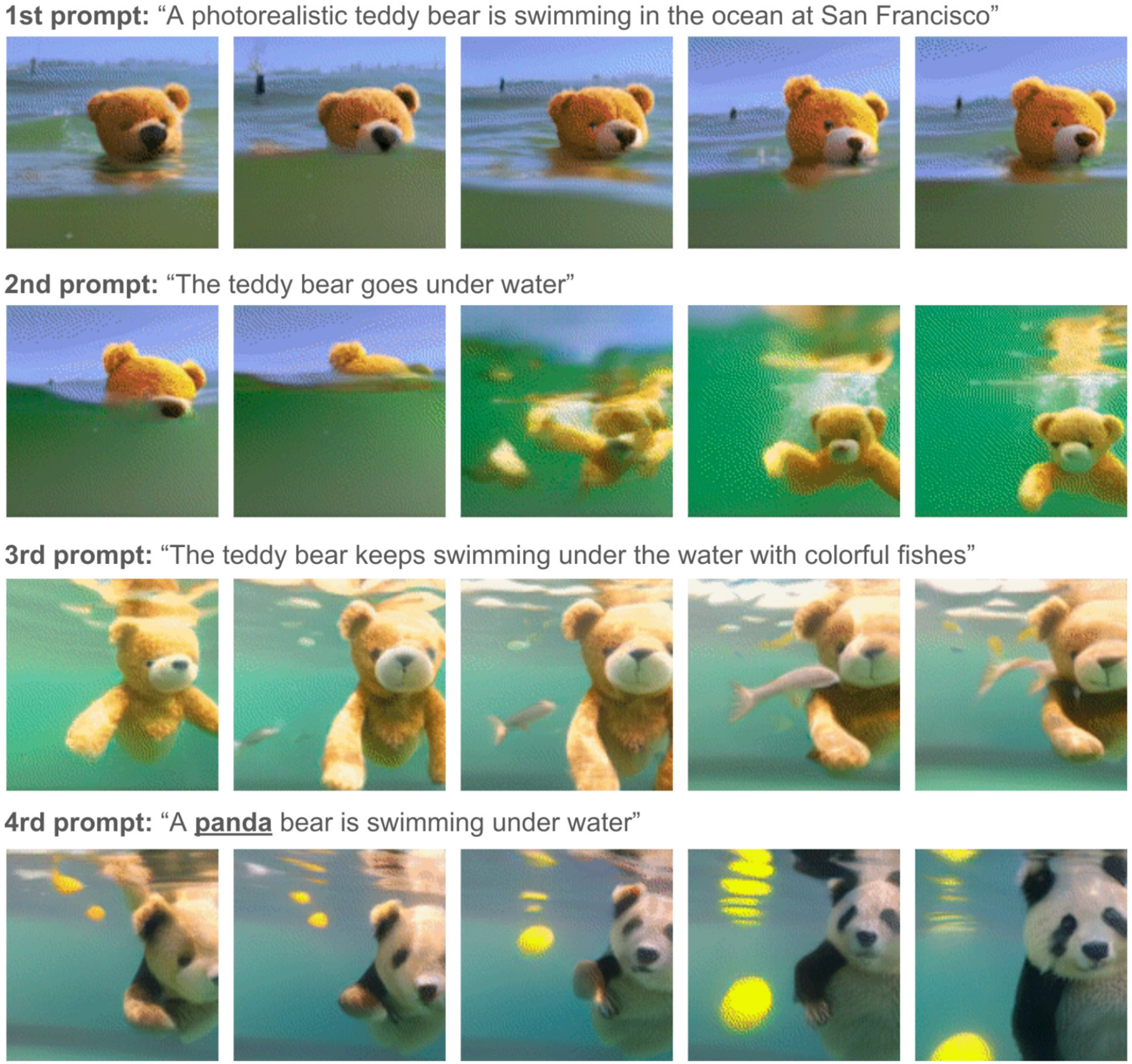
Evaluation Tasks:

- Text conditional video generation
- Text-image conditional video generation
- Story generation from dynamic text inputs
- Video reconstruction
- Video prediction

Story Generation from Dynamic Text Inputs

- Phanaki can generate long videos since it is auto-regressive in time

Steps:
- Generate a video with the first prompt
- Extend it in time by conditioning a new prompt and on the last 5 previously generated frames



Evaluation Tasks:

- Text conditional video generation
- Text-image conditional video generation
- Story generation from dynamic text inputs
- Video reconstruction
- Video prediction

Video Encoding and Reconstruction

- Dataset: Moments-in-Time (MiT), ~802K training, ~33K validation, ~67K text videos at 25 FPS
- Baselines: per-frame image based encoder-decoders (e.g., ViT, VQ-GAN)

Results:
- Per-frame image based method (VQ-GAN and ViT) achieves slightly better FID
- C-ViViT achieves significantly better FVD
- C-ViViT compresses the video input fewer tokens per video compared with image based baselines

Table 3. Video reconstruction results on Moments-in-Time. The number of tokens is computed for 10 frames with the exception of C-ViViT which is for 11, due to the isolated initial frame.

Method	FID ↓	FVD ↓	Number of Tokens ↓
Conv VQ-GAN [12]	7.5	306.1	2560
Conv VQ-GAN + Video loss	13.7	346.5	2560
ViT VQ-GAN [58]	3.4	166.6	2560
ViT VQ-GAN + Video loss	3.8	173.1	2560
C-ViViT VQ-GAN (Ours)	4.5	65.78	1536

Video Encoding and Reconstruction

GT ViT C-ViViT

Evaluation Tasks:

- Text conditional video generation
- Text-image conditional video generation
- Story generation from dynamic text inputs
- Video quantization
- Video prediction

Video Prediction

- Dataset:
  - BAIR Robot Pushing benchmark: predict 15 frames conditioned on a given single frame
  - Kinetics-600: predict 11 frames given 5 frames
- Results:
  - Phenaki is not specifically designed for video prediction
  - Competitive with benchmarks with SOTA video prediction methods

Table 4. Video prediction on Kinetics-600 [7]. While Phenaki is not designed for video prediction it achieves comparable results with SOTA video prediction models.

Method	FVD ↓
Video Transformer [51]	170.0 ± 5.00
CogVideo [18]	109.2
DVD-GAN-FP [9]	69.1 ± 0.78
Video VQ-VAE [49]	64.3 ± 2.04
CCVS [28]	55.0 ± 1.00
TrIVD-GAN-FP [27]	25.7 ± 0.66
Transframer [31]	25.4
RaMViD [19]	16.5
Video Diffusion [17]	16.2 ± 0.34
Phenaki (Ours)	36.4 ± 0.19

Table 5. Video prediction on BAIR [11]

Method	FVD ↓
DVD-GAN [9]	109.8
VideoGPT [55]	103.3
TrIVD-GAN [27]	103.3
Transframer [31]	100.0
HARP [57]	99.3
CCVS [28]	99.0
Video Transformer [51]	94.0
FitVid [3]	93.6
MCVD [47]	89.5
NUWA [54]	86.9
RaMViD [19]	84.2
Phenaki (Ours)	97.0

Thanks for your attention!