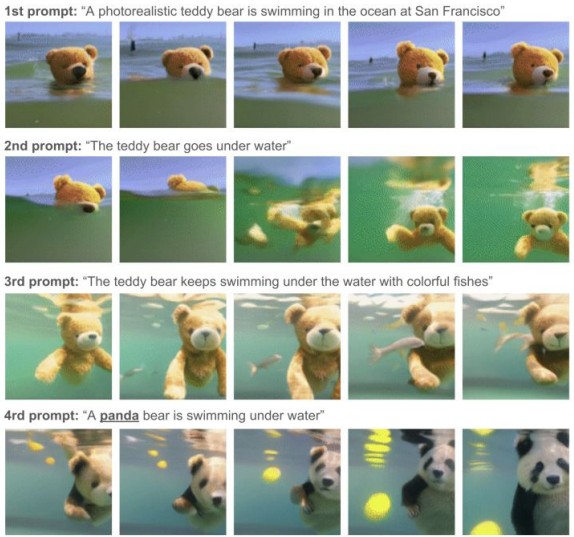
Phenaki: Variable Length Video Generation From Open Domain Textual Descriptions.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, Dumitru Erhan

Presented By: Noah, Daniel

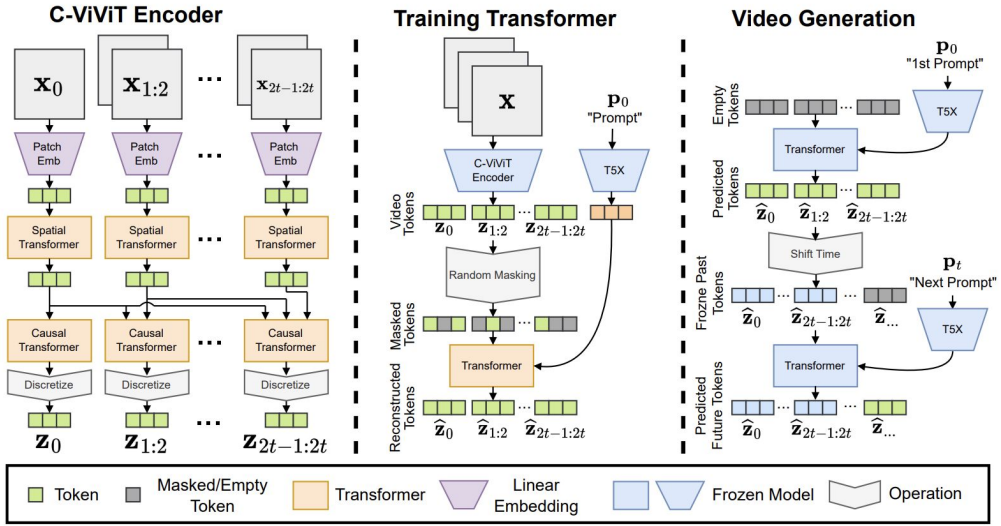
Motivation

- Problems generating arbitrarily long videos due to computational cost
- The lack of high quality text video data sets
- Single prompt is not enough to create a good video

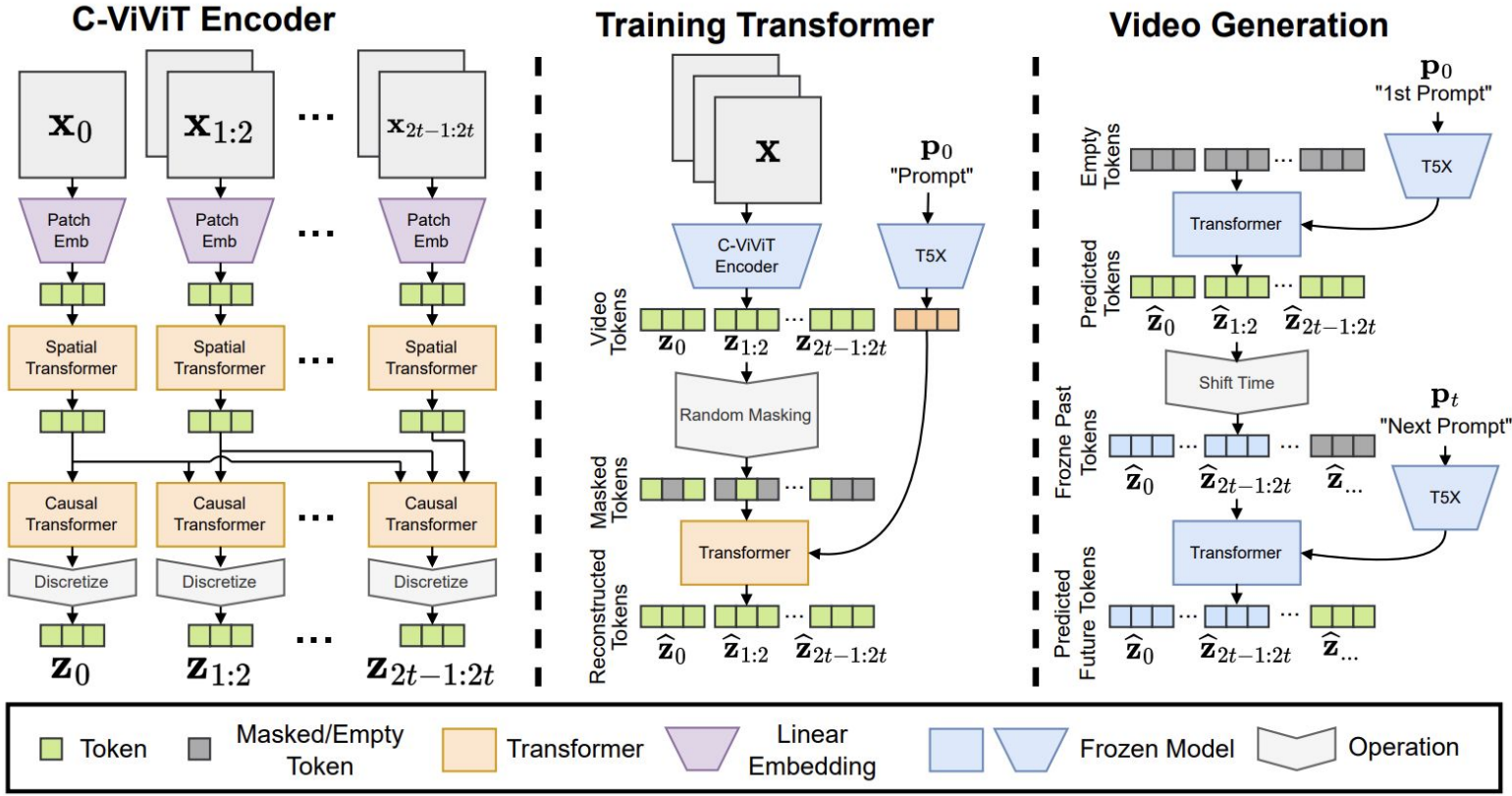


Model Architecture

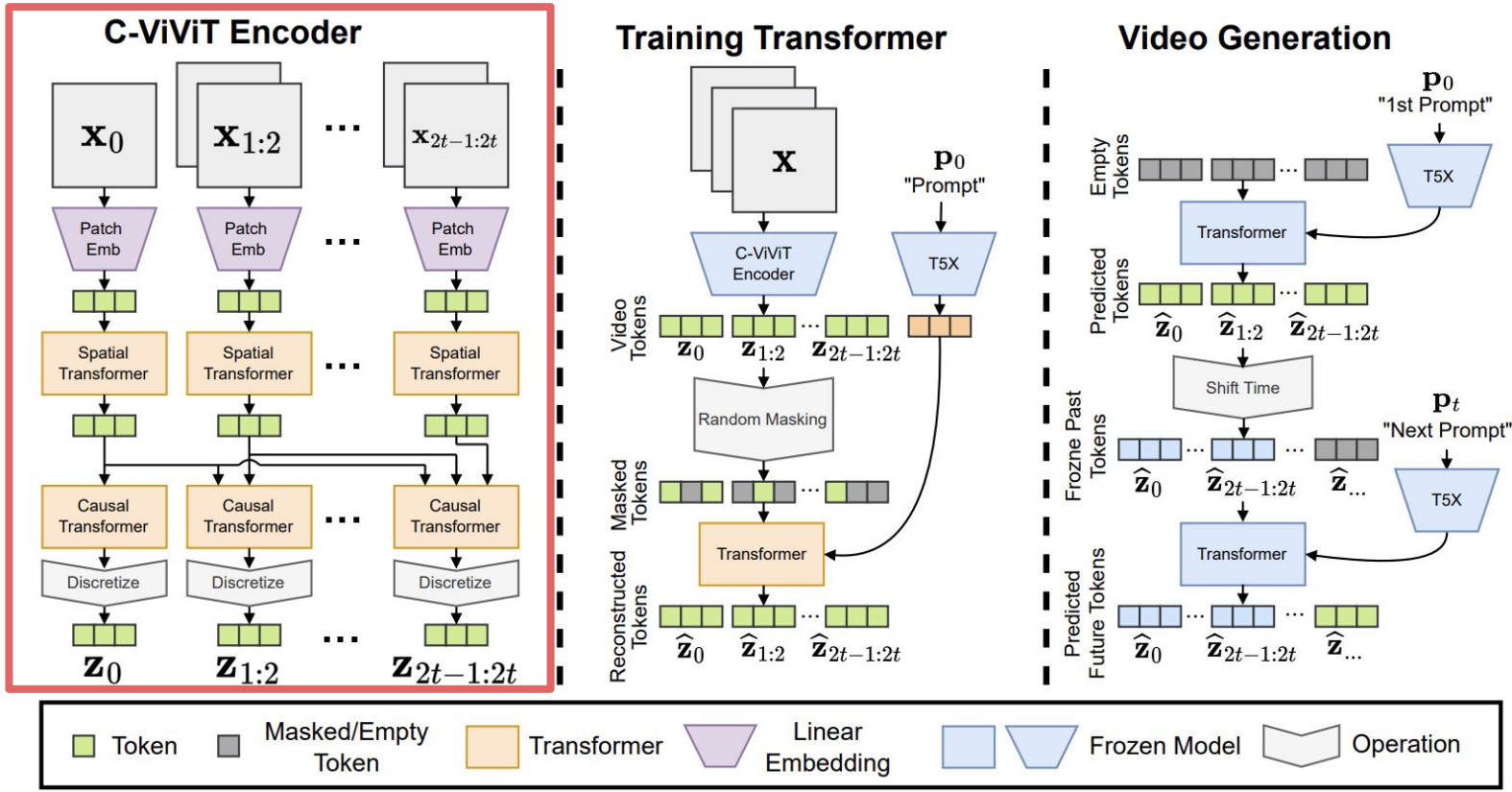
- C-ViViT architecture
  - Encoder architecture
  - Decoder
  - Losses
- Bidirectional Transformers
  - How masking is used
  - Losses
- How arbitrarily long videos are generated
  - Page 7 3.3 details specifics on this
  - Also page 5 Inference and auto-regressive generation of long videos
  - Define auto regressive
- We should give a high level summary of everything once we have gone through in detail



Model Architecture



Model Architecture

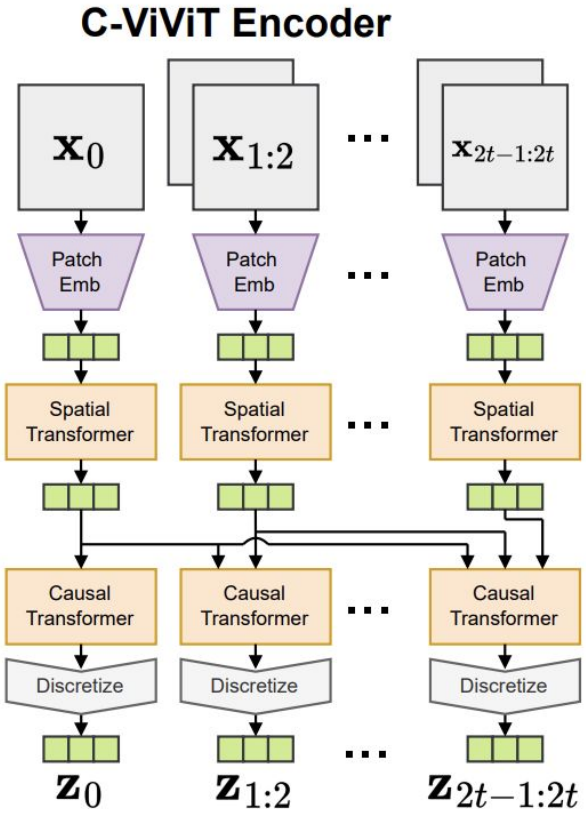


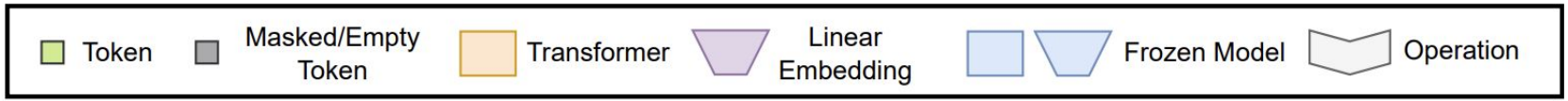
C-ViViT Encoder Architecture

$$t_p \times w_p \times h_p \times c_p$$

$$(t_z+1) \times w_z * h_z \times d_z$$

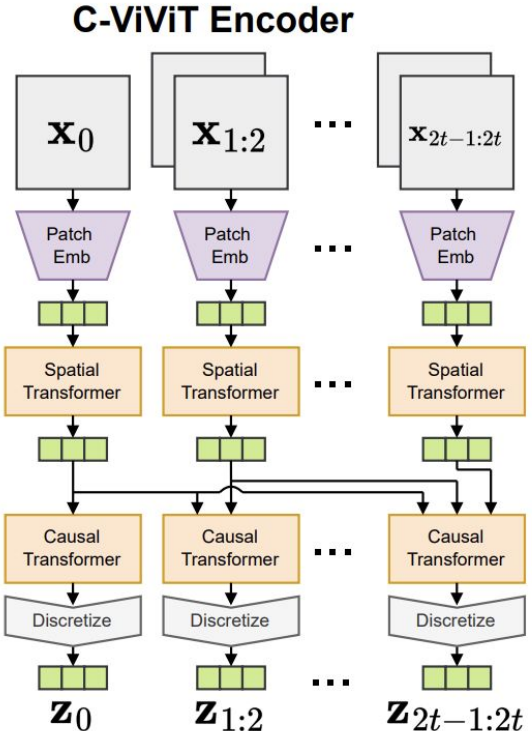
$$t_z \times w_z \times h_z \times d_z$$



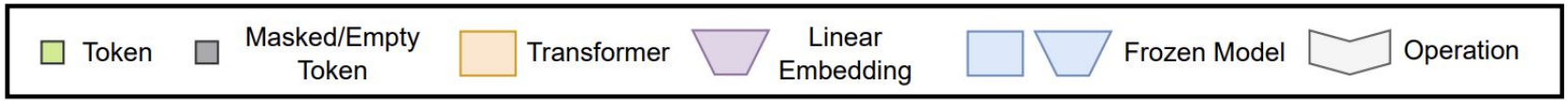


C-ViViT Encoder Architecture

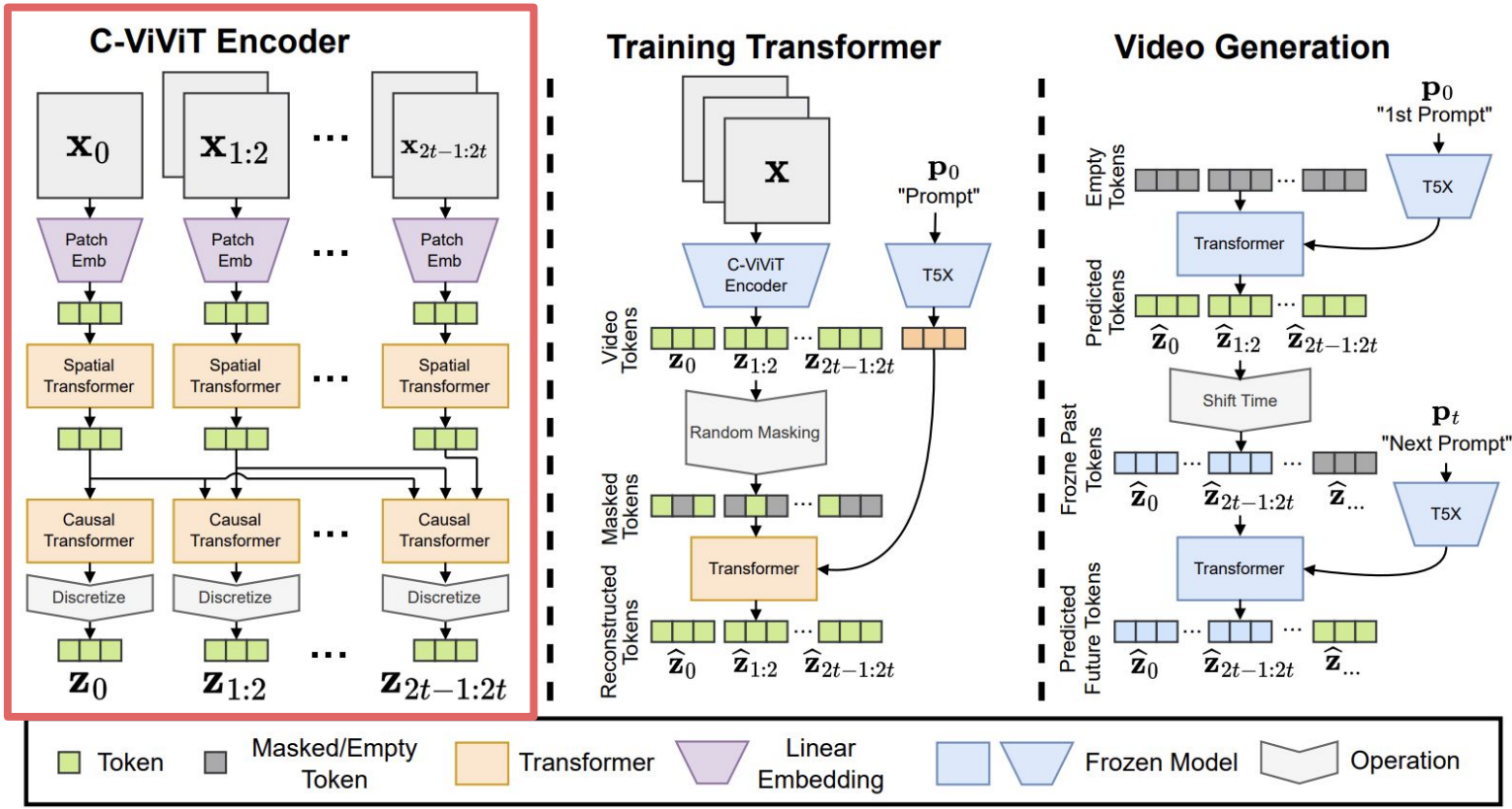
$$L_{VQ} = \|\mathrm{sg}(\mathbf{z}) - \mathbf{e}\|_2^2 + \beta\|\mathbf{z} - sg(\mathbf{e})\|_2^2$$



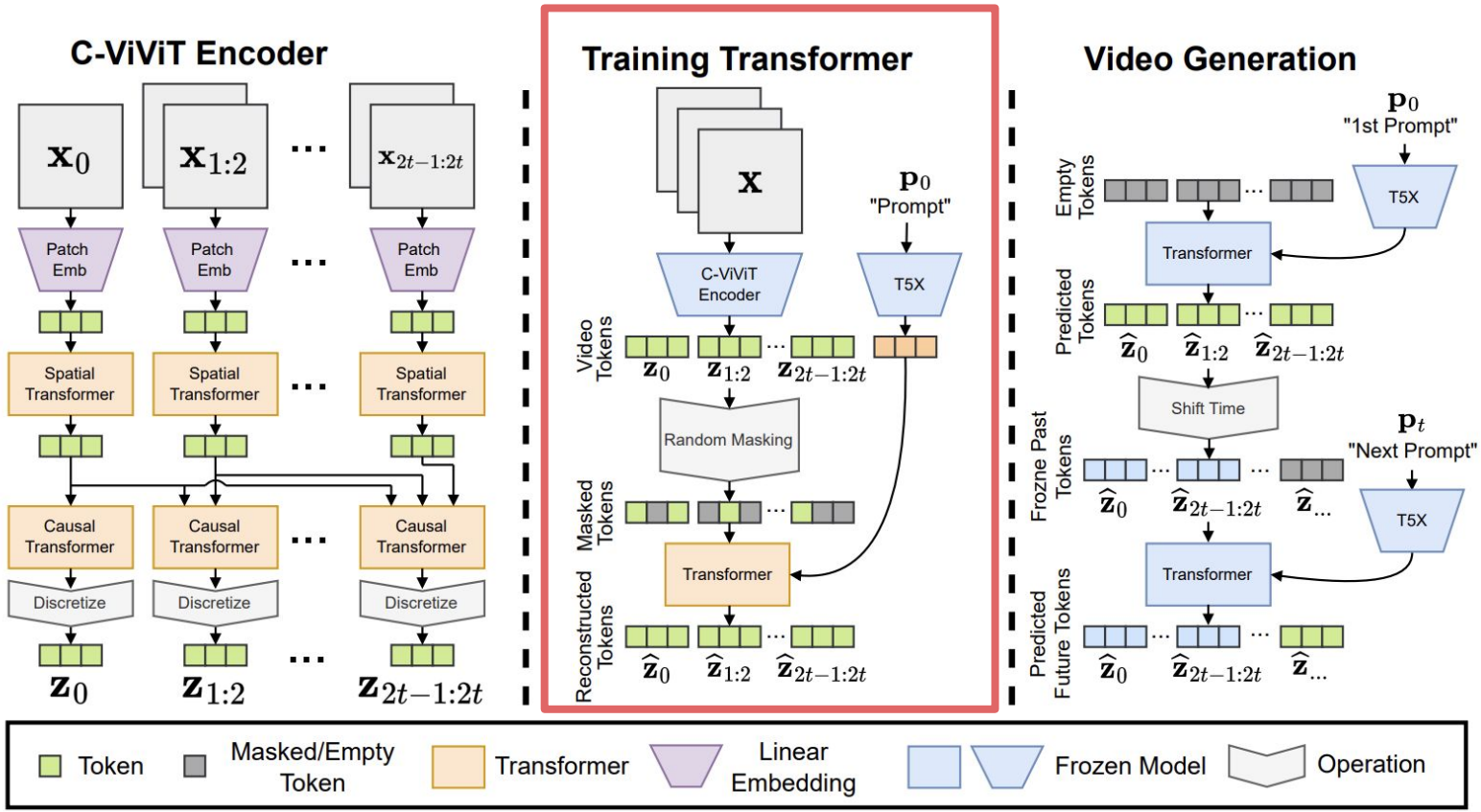
$$L = L_{VQ} + 0.1 \times L_{Adv} + 0.1 \times L_{IP} + 1.0 \times L_{VP} + 1.0 \times L_2$$



Model Architecture



Model Architecture



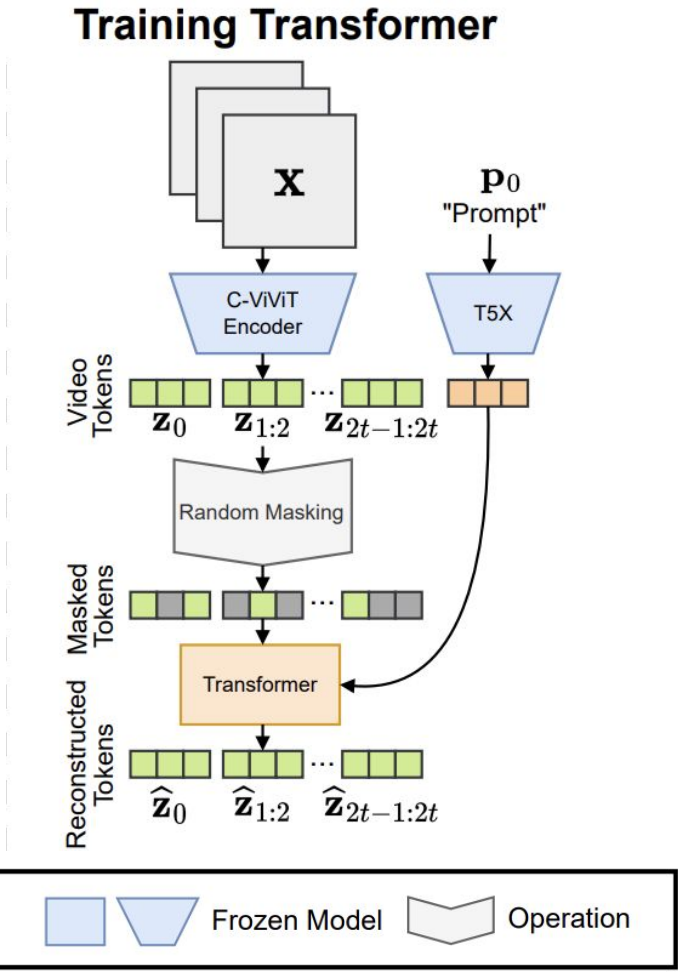
MaskGiT Architecture

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

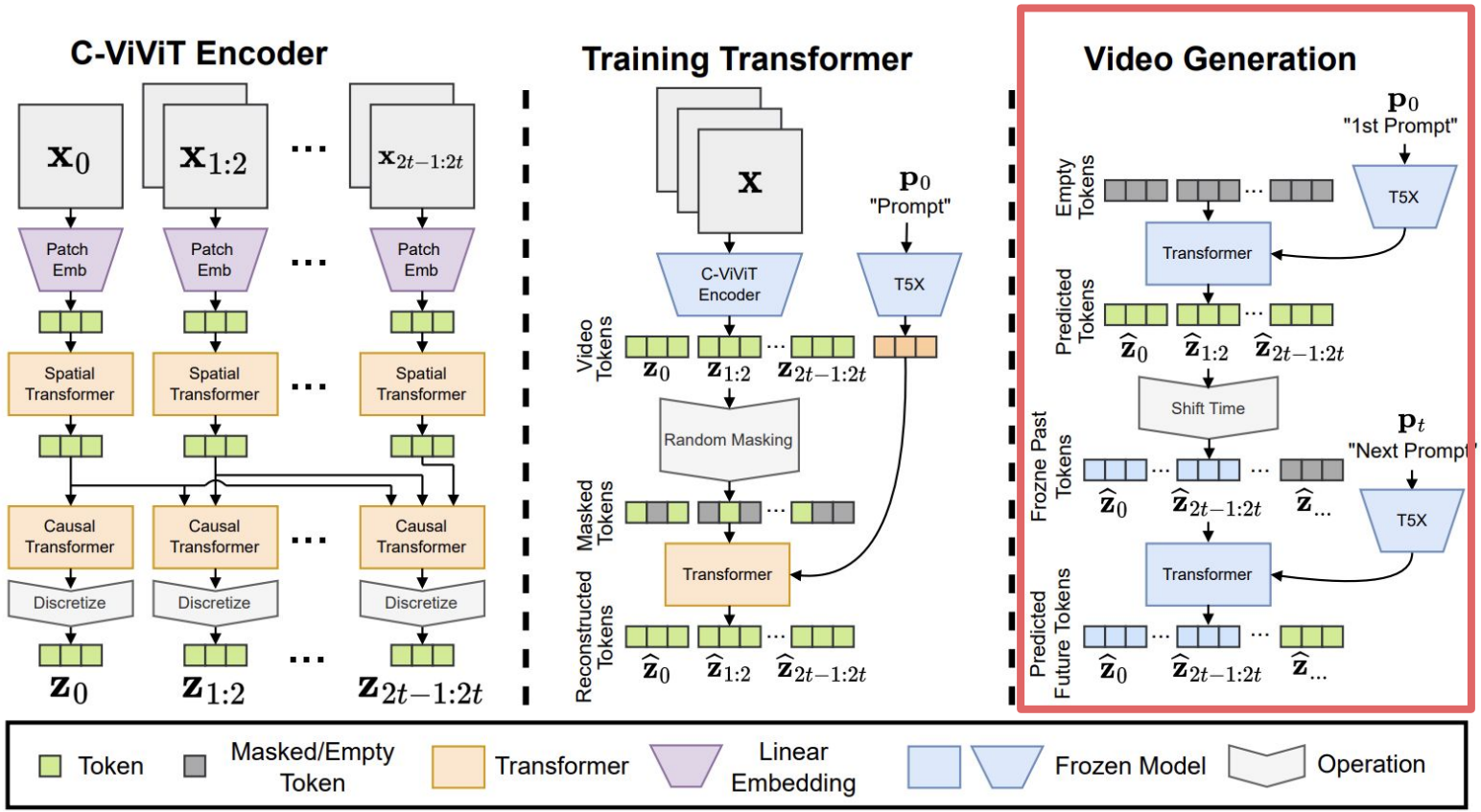
$\mathbf{a}_{\bar{M}}$ represents the masked version of $\mathbf{a}$

$\mathbf{p}$ is the text condition embedding

N is the number of video tokens

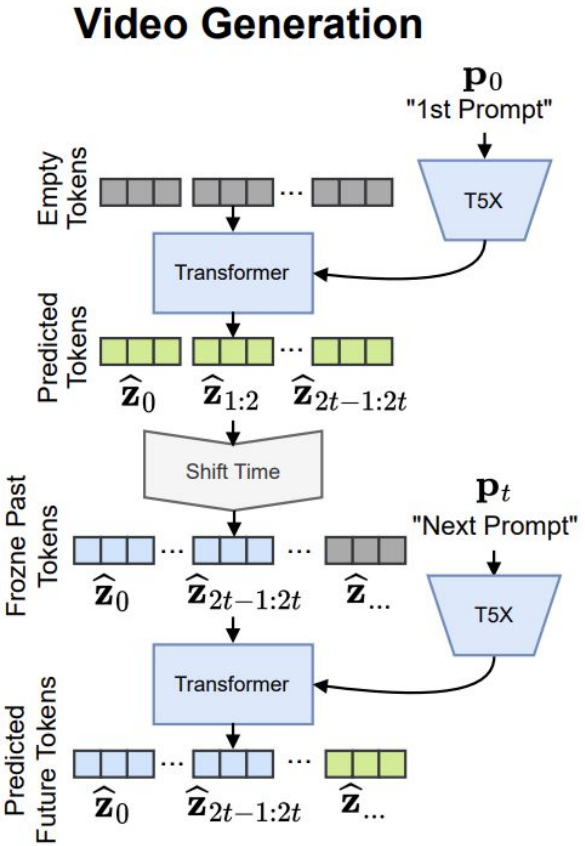


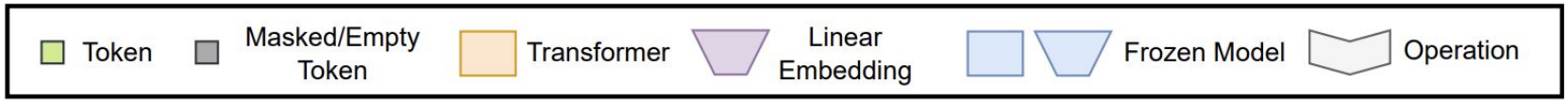
Model Architecture



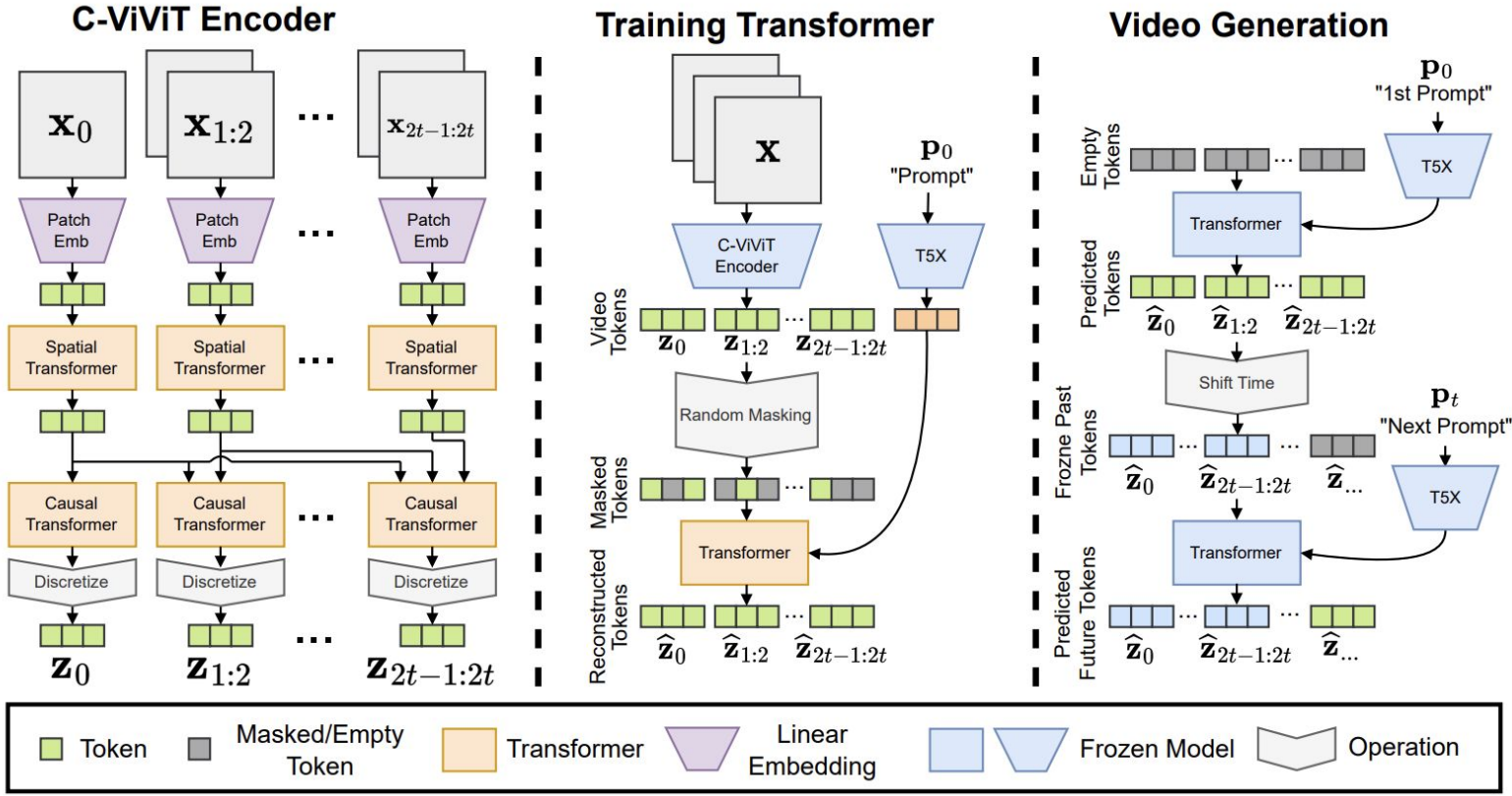
Video Generation Architecture

- Initialize with C-ViViT tokens





Model Architecture



Results

https://phenaki.github.io/

First person view of riding a motorcycle through a busy street.
First person view of riding a motorcycle through a busy road in the woods.
First person view of very slowly riding a motorcycle in the woods.
First person view braking in a motorcycle in the woods.
Running through the woods.
First person view of running through the woods towards a beautiful house.
First person view of running towards a large house.
Running through houses between the cats.
The backyard becomes empty.
An elephant walks into the backyard.
The backyard becomes empty.
A robot walks into the backyard.
A robot dances tango.
First person view of running between houses with robots.
First person view of running between houses; in the horizon, a lighthouse.
First person view of flying on the sea over the ships.
Zoom towards the ship.
Zoom out quickly to show the coastal city.
Zoom out quickly from the coastal city.

Experiments

- Trained a 1.8B parameter Phenaki Model
- 15M text-video pairs at 8 FPS mixed with 50M text-images and 400M pairs of LAION-400M
- No established benchmark for evaluating text to video methods
- Story based conditional video generation is a new task

Text Conditional Video Generation

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	3.84

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

Text Conditional Video Generation

Prompt: "HD Video: A really cute panda washing dishes with yellow gloves in the garden"

Prompt: "A happy panda wearing red boxing gloves and blue shorts standing in front of brandenburg gate with fireworks in the background"

Prompt: "Pencil drawing: A Panda listening to music with headphones knitting a sweater while sitting on the couch"

Prompt: "An astronaut riding a horse on mars with a sunset in the background"

Prompt: "An astronaut diving at a coral reef with many fishes."

Prompt: "A cartoon of an astronaut high fiving a brown bear."

Text Conditional Video Generation

HD Video: A really cute panda washing dishes with yellow globes in the garden.

An astronaut riding a horse on mars with a sunset in the background.

A happy panda wearing red boxing gloves and blue shorts standing in front of brandenburg gate with fireworks in the background.

An astronaut diving at a coral reef with many fishes.

Pencil drawing: A Panda listening to music with headphones knitting a sweater while sitting on the couch.

A cartoon of an astronaut high fiving a brown bear.

Data Split on Text Conditional Image Generation

	80% Video 20% Image	100% Video
Prompt: Pencil drawing: A Panda listening to music with headphones knitting a sweater while sitting on the couch.		
Prompt: Water color style: A Panda listening to music with headphones knitting a sweater while sitting on the couch.		

Text Image Conditional Video Generation

Text Image Conditional Video Generation

Camera zooms quickly into the eye of the cat

A white cat touches the camera with the paw

A white cat Yawns Loudly

Visual Storytelling by Dynamic Text Inputs

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”

Visual Storytelling by Dynamic Text Inputs

A photorealistic teddy bear is swimming in the ocean at San Francisco.
The teddy bear goes under water.
The teddy bear keeps swimming under the water with colorful fishes.
A panda bear is swimming under water

A teddy bear diving in the ocean
A teddy bear emerges from the water
A teddy bear walks on the beach Camera zooms out to the teddy bear in the campfire by the beach

Side view of an astronaut is walking through a puddle on mars
The astronaut is dancing on mars
The astronaut walks his dog on mars
The astronaut and his dog watch fireworks

Video Reconstruction Performance

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 Reconstruction Performance

Image Conditional Video Generation

- BAIR Robot Pushing Benchmark
  - Task is to generate 15 frames conditioned on a single frame
- Kinetics-600
  - Task is to predict 11 frames given 5 frames

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

Conclusion

https://phenaki.github.io/

- Story based Conditional Image Generation
  - Allows for longer and more complex narratives
- Able to generate coherent and diverse videos
- C-ViViT video encoder
  - Able to encode variable length videos
- Able to animate images
- Ethics Statement

Experiments

- Qualitative
  - Seems very unofficial (no establi
- Quantitative
  - Zero-shot comparison to NUWA without finetuning (Table 1)
- Mention how they do joint txt-img txt-vid training to supplement the fact that there is less high quality txt-vid data
  - Explore difference training splits of the two data types (ablation study)
- Evaluate quality of video encodings and reconstruction
  - Moments in time (results in appendix B1)

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	3.84

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

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

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