

# Text-to-Video Generation Guide

Hoang-Bach Ngo

Minh-Hung An

Ngày 28 tháng 2 năm 2024



## Phần I: Giới thiệu tổng quan

Trong những năm gần đây, sự phát triển không ngừng của công nghệ đã mở đường cho một bước đột phá đáng kinh ngạc trong lĩnh vực tạo sinh nội dung: khả năng chuyển đổi văn bản thành video, một tiến bộ mới nhất trong chuỗi các thành tựu công nghệ. Điều này không chỉ là một bước tiến trong việc tạo ra nội dung đa phương tiện từ văn bản mô tả mà còn mở ra cánh cửa mới cho nhiều ngành công nghiệp khác nhau. Mô hình Sora, phát triển bởi OpenAI, hay Genie, phát triển bởi Google, là những minh chứng nổi bật cho tiến trình này, đánh dấu một cột mốc quan trọng trong việc mô hình hóa và tạo sinh video dựa trên văn bản, giúp biến những mô tả văn bản thành video sinh động với mạch lạc thời gian và không gian cao.

Lý do phía sau việc nghiên cứu và phát triển text-to-video không chỉ là để mở rộng biên giới của thị giác máy tính mà còn nhằm giải quyết những thách thức liên quan đến việc mô hình hóa video, một tác vụ phức tạp hơn đáng kể so với tạo sinh hình ảnh từ văn bản. Sự phức tạp này đến từ việc cần phải hiểu và tái tạo được sự liên kết thời gian và không gian trong video, đòi hỏi những phương pháp tiếp cận sáng tạo và tiên tiến. Bài viết này nhằm mục đích giới thiệu về cách thức hoạt động của các mô hình tạo sinh video, phân biệt chúng với mô hình tạo sinh hình ảnh, và khám phá các nghiên cứu hàng đầu trong lĩnh vực hấp dẫn nhưng cũng đầy thách thức này.

# Phần II: Các thách thức cho bài toán text-to-video

Trước hết, chúng ta hãy cùng nhau khám phá những thách thức lớn mà công nghệ tạo sinh video phải đối mặt. Những thách thức này chủ yếu liên quan đến việc giữ cho video có sự nhất quán về không gian và liên kết mạch lạc theo thời gian, đồng thời cũng cần phải quan tâm đến yếu tố tài nguyên tính toán và sự khan hiếm của dữ liệu dataset chất lượng cao.

## 1 Sự nhất quán trong không gian và sự liên kết về mặt thời gian

Duy trì sự nhất quán trong không gian và sự liên kết về mặt thời gian giữa các vật thể, nhân vật và bối cảnh là một trong những thử thách lớn nhất trong bài toán tạo sinh video [8], [11]. Sự nhất quán về không gian yêu cầu mô hình phải có sự hiểu biết về không gian ba chiều trong video và theo sát vị trí, phương hướng và sự tương tác giữa các vật thể trong video. Vấn đề này trở nên khó khăn khi câu văn bản mô tả những phân cảnh có tính phức tạp cao, với nhiều vật thể tương tác với nhau, yêu cầu mô hình phải có hiểu biết phức tạp về bối cảnh.

Bên cạnh đó, sự nhất quán về thời gian yêu cầu video được tạo ra không chỉ tuân theo mạch tường thuật được mô tả trong văn bản mà chuyển động và chuyển tiếp giữa các khung hình cũng mượt mà và chân thực. Để đạt được sự mạch lạc về mặt thời gian đòi hỏi mô hình phải hiểu được mối quan hệ nhân quả và trình tự của các sự kiện trong câu text prompt, chuyển chúng thành một chuỗi khung hình mạch lạc về mặt trực quan.

## 2 Độ phức tạp trong tính toán

Nhu cầu duy trì đồng thời sự nhất quán về không gian và tính liên kết về thời gian làm tăng đáng kể yêu cầu tính toán của việc tạo văn bản thành video. Mỗi khung hình trong video được tạo ra đều phải liên kết đến tất cả các khung hình trước đó, đến từng chi tiết để đảm bảo độ chính xác về không gian và quá trình chuyển đổi giữa các khung hình phải được mô hình hóa cẩn thận để duy trì dòng thời gian. Quá trình này yêu cầu tài nguyên tính toán khổng lồ [8] [11], đặc biệt đối với các video có độ phân giải cao trong đó độ trung thực của từng pixel có thể ảnh hưởng đến chất lượng tổng thể và độ chân thực của video.

Hơn nữa, các mô hình tổng hợp được sử dụng để tổng hợp văn bản thành video, chẳng hạn như Mạng GAN, VAE hay mạng Diffusion, đều cần một lượng tính toán khổng lồ. Khi được giao nhiệm vụ tạo video, các mô hình này phải hoạt động trên không gian đầu ra lớn hơn nhiều so với việc tạo hình ảnh, xử lý nhiều khung hình thậm chí chỉ trong vài giây video. Điều này không chỉ đòi hỏi sức mạnh xử lý đáng kể mà còn đặt ra những thách thức về bộ nhớ và lưu trữ, vì mô hình phải duy trì và xử lý lượng lớn dữ liệu để tạo ra đầu ra video cuối cùng.

## 3 Thiếu các datasets có chất lượng cao

Trong việc chuyển đổi văn bản thành video, một trong những khó khăn lớn nhất chính là việc tìm kiếm và sử dụng dữ liệu phù hợp. Có một sự thiếu hụt rõ ràng về các bộ dữ liệu lớn, nơi mô tả văn bản chi tiết đi kèm với video tương ứng [9], [7] [3] khiến việc thu thập dữ liệu trở nên cực kỳ phức tạp. Cần phải có những mô tả rất cụ thể và chính xác về nội dung video, điều này làm tăng thêm độ khó trong việc tạo ra những bộ dữ liệu đa dạng và chất lượng cao.

Để giải quyết vấn đề này, các nhà nghiên cứu đang áp dụng nhiều phương pháp như tạo dữ liệu giả, phương pháp tăng cường dữ liệu, sử dụng crowdsourcing để thu thập mô tả video, và áp dụng kỹ

thuật học chuyển giao từ các nhiệm vụ có liên quan. Mặc dù đã có nhiều cố gắng, việc vượt qua những trở ngại liên quan đến việc tiếp cận dữ liệu vẫn là một bước tiến quan trọng để phát triển công nghệ chuyển đổi văn bản thành video.

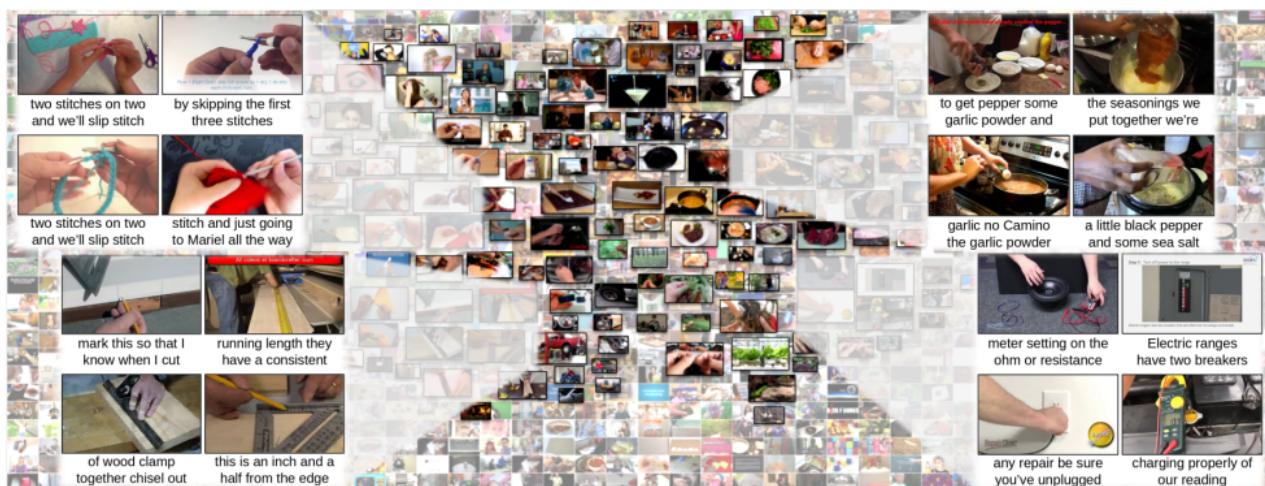
Nhìn chung đó là những thách thức chính cho bài toán tạo sinh video từ văn bản. Ở phần sau, chúng ta hãy cùng đi vào tìm hiểu xem có những phương pháp nào được sử dụng để vượt qua những thách thức trên.

## Phần III: Các Dataset phổ biến

Việc sử dụng các bộ dữ liệu gồm video kèm theo mô tả (caption) để huấn luyện các mô hình tạo sinh video là một phần không thể thiếu trong quá trình phát triển công nghệ này. Những bộ dữ liệu này cung cấp cho mô hình một cơ sở dữ liệu phong phú, giúp mô hình học cách hiểu và tái tạo lại các sự kiện, hành động, và mối liên kết giữa các nhân vật trong video dựa trên mô tả văn bản. Qua đó, mô hình có thể tăng cường khả năng nhận thức về không gian và thời gian, cũng như cách các yếu tố trong video tương tác với nhau. Việc này đòi hỏi bộ dữ liệu phải đa dạng về nội dung, cảnh quay và phong cách kể chuyện, giúp mô hình có khả năng áp dụng vào nhiều tình huống và bối cảnh khác nhau, từ đó mở rộng khả năng ứng dụng của công nghệ tạo sinh video trong thực tế. Sau đây là một số bộ dữ liệu phổ biến, thường xuyên được sử dụng để huấn luyện các mô hình tạo sinh video

### 3.1 HowTo100M

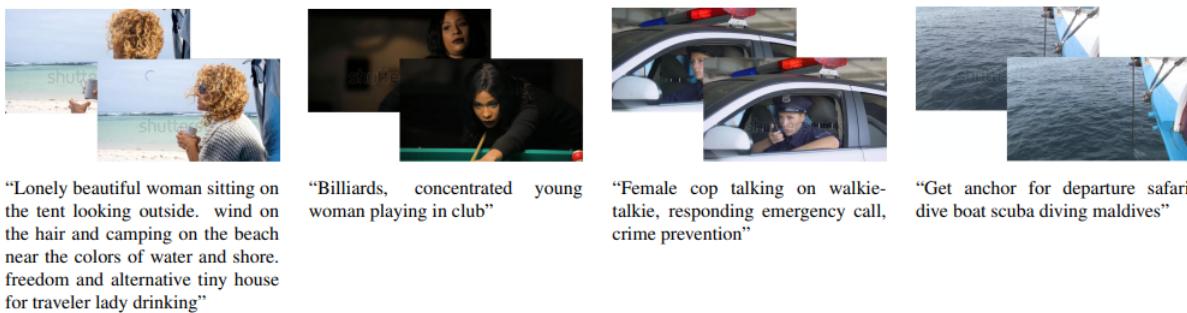
HowTo100M (<https://www.di.ens.fr/willow/research/howto100m>) là một tập dữ liệu quy mô lớn bao gồm 136 triệu đoạn clip video được thu thập từ 1.22 triệu video hướng dẫn có lời bình trên web, mô tả con người thực hiện và mô tả hơn 23 nghìn nhiệm vụ trực quan khác nhau. Tập dữ liệu được tạo ra bằng cách sử dụng WikiHow để xác định 23,611 nhiệm vụ trực quan từ các hoạt động tương tác với thế giới vật lý, loại trừ các nhiệm vụ trừu tượng. Các video hướng dẫn được tìm kiếm trên YouTube dựa trên các tiêu chí như có phụ đề tiếng Anh và thu hút ít nhất 100 lượt xem, đồng thời loại bỏ video quá dài hoặc thiếu nội dung. Tập dữ liệu được tinh lọc để loại bỏ trùng lặp và cải thiện chất lượng, mặc dù vẫn có khả năng chứa bản sao do tải lên nhiều lần, nhưng điều này không ảnh hưởng lớn đến quy mô dự án của tập dữ liệu này.



Hình 1: Ví dụ về các cặp clip-caption được chọn dựa trên sự tương đồng giữa ảnh và mô tả tương ứng.

### 3.2 WebVid

WebVid (<https://maxbain.com/webvid-dataset>) là tập dữ liệu được thu thập từ web, bao gồm 2.5 triệu cặp video-text, vượt trội hơn hẳn so với các tập dữ liệu trước đó. Quá trình thu thập dữ liệu này tuân thủ một phương pháp tương tự như Google Conceptual Captions, với việc nhận thấy một phần đáng kể các hình ảnh từ CC3M thực chất là hình thu nhỏ của video, WebVid khai thác các nguồn video tương tự để tạo ra tập dữ liệu phong phú. Tập dữ liệu video này, mặc dù nhỏ hơn đáng kể so với HowTo100M về thời lượng và số lượng cặp clip-caption, nhưng lại nổi bật với chất lượng caption thủ công cao, tạo ra câu văn rõ ràng và mô tả chính xác nội dung hình ảnh. Bên cạnh đó, tập dữ liệu Google Conceptual Captions cho hình ảnh của WebVid, với 3.3 triệu cặp text-image, mở rộng phạm vi đa dạng phong cách mô tả và hình ảnh, phản ánh một cách chân thực hơn các nguồn thông tin đa dạng từ web.



Hình 2: Ví dụ minh họa về cặp video và văn bản mô tả trong WebVid dataset:

### 3.3 CelebV-Text

CelebV-Text (<https://celebvt-text.github.io>), một tập dữ liệu lớn, đa dạng và chất lượng cao của cặp video về khuôn mặt và văn bản mô tả, nhằm hỗ trợ nghiên cứu về nhiệm vụ tạo video từ văn bản mô tả cho khuôn mặt. CelebV-Text bao gồm 70,000 đoạn clip video về khuôn mặt trong môi trường tự nhiên với nội dung hình ảnh đa dạng, mỗi đoạn được ghép nối với 20 đoạn văn bản mô tả được tạo ra thông qua chiến lược tạo văn bản bán tự động. Các văn bản mô tả được cung cấp có chất lượng cao, mô tả chính xác cả các thuộc tính tĩnh và động. Sự vượt trội của CelebV-Text so với các tập dữ liệu khác được thể hiện qua phân tích thống kê toàn diện về video, văn bản mô tả và mối liên quan text-video.



Hình 3: CelebV-Text bao gồm (a) 70,000 mẫu video và (b) 1,400,000 văn bản mô tả. Mỗi mẫu video được ghi chú với hình dáng tổng quát, hình dáng chi tiết, điều kiện ánh sáng, hành động, cảm xúc, và hướng ánh sáng.

## Phần IV: Các hướng tiếp cận chính

Khi nói đến việc tạo video từ văn bản, có hai phương pháp tiếp cận chính mà các nhà nghiên cứu thường theo đuổi. Một là sử dụng mô hình dựa trên cấu trúc Transformer, vốn nổi tiếng với khả năng hiểu và xử lý ngôn ngữ tự nhiên một cách hiệu quả. Phương pháp thứ hai là áp dụng kỹ thuật diffusion, một cách tiếp cận mới mẻ và đầy hứa hẹn trong việc tạo ra hình ảnh và video từ văn bản.

Gần đây, một số nghiên cứu đã bắt đầu kết hợp những ưu điểm của cả hai mô hình này, tích hợp cấu trúc Transformer vào bên trong mô hình diffusion để tạo ra những kết quả còn ấn tượng hơn. Một ví dụ điển hình cho sự kết hợp này là mô hình Sora của OpenAI, được kỳ vọng sẽ mở ra những khả năng mới trong lĩnh vực tạo sinh video từ văn bản.

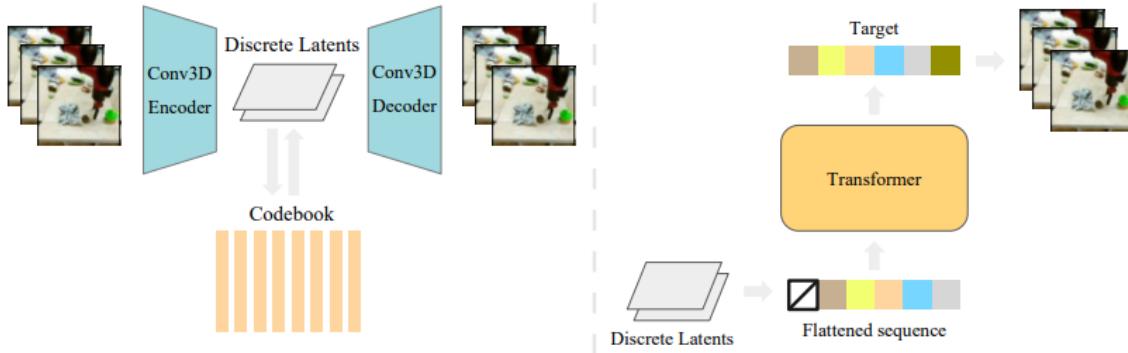
### 4 Transformer-based Pretraining

Transformer, kể từ khi được giới thiệu, đã tạo ra một cuộc cách mạng trong nhiều lĩnh vực của trí tuệ nhân tạo, từ xử lý ngôn ngữ tự nhiên đến nhận dạng hình ảnh. Gần đây, nhiều sự chú ý đã được hướng tới việc áp dụng kiến trúc transformer trong tác vụ tạo sinh video, một lĩnh vực đầy thách thức nhưng cũng rất hứa hẹn. Trong bối cảnh này, transformer được sử dụng để hiểu và dự đoán các mô hình thời gian không gian phức tạp trong dữ liệu video, cho phép chúng tạo ra những đoạn video mới mẻ và độc đáo từ mô tả văn bản. Sau đây là những công trình nghiên cứu nổi bật với hướng tiếp cận này.

#### 4.1 VideoGPT:

Trong nghiên cứu tiên phong về việc áp dụng công nghệ Transformer vào việc tạo video, bài báo khoa học "VideoGPT: Video Generation using VQ-VAE and Transformers" [10] đưa ra một kiến trúc độc đáo nhằm khai thác tiềm năng của các mô hình tạo sinh dựa trên xác suất. Kiến trúc này được thiết kế để mở rộng và tối ưu hóa quá trình tạo video, làm cho nó trở nên mạnh mẽ và linh hoạt hơn. Trong VideoGPT, có một số thành phần chính được đưa ra, mỗi thành phần đều đóng vai trò quan trọng

trong việc tạo ra các video chất lượng cao và đa dạng từ văn bản hoặc các dạng đầu vào khác. Một số thành phần chính của VideoGPT được thể hiện trong Hình 7 bao gồm:

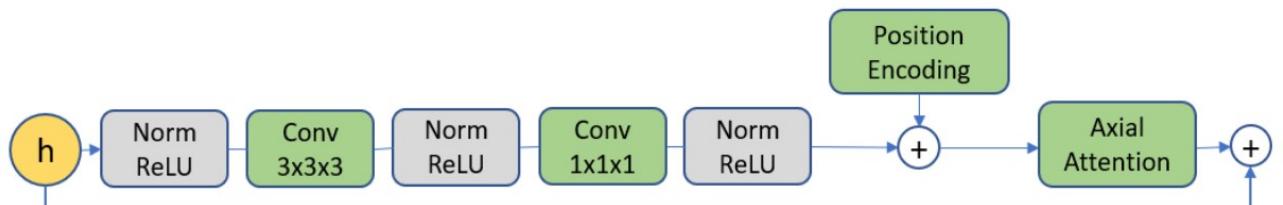


Hình 4: Kiến trúc của VideoGPT bao gồm 2 thành phần: khối VQ-VAE và khối GPT

- **VQ-VAE (Vector Quantized Variational AutoEncoder):** được đề xuất bởi [4], là một dạng mô hình nén các điểm dữ liệu ở không gian nhiều chiều về không gian latent rời rạc và tái tạo lại những điểm dữ liệu đó. Thành phần này có chức năng chính là học những biểu diễn của các video trong không gian latent, cụ thể như sau:

- Đầu tiên một khối encoder  $E(x) \rightarrow h$  mã hóa video  $x$  thành một chuỗi các vector  $h$ . Sau đó chuỗi vector này sẽ được rời rạc hóa bằng cách sử dụng phương pháp nearest neighbor để map về các vector codebook có sẵn.
- Sau đó một khối  $D(e) \rightarrow \hat{x}$  học để tái tạo lại  $x$  từ những vector đã được rời rạc hóa.

Để có thể học mô hình hóa được data dạng video, VideoGPT sử dụng một chuỗi các khối 3D convolutions để lấy mẫu xuống (downsample) thông tin không gian và thời gian, theo sau đó bằng một khối attention residual. Khối attention residual được thiết kế như hình 5:



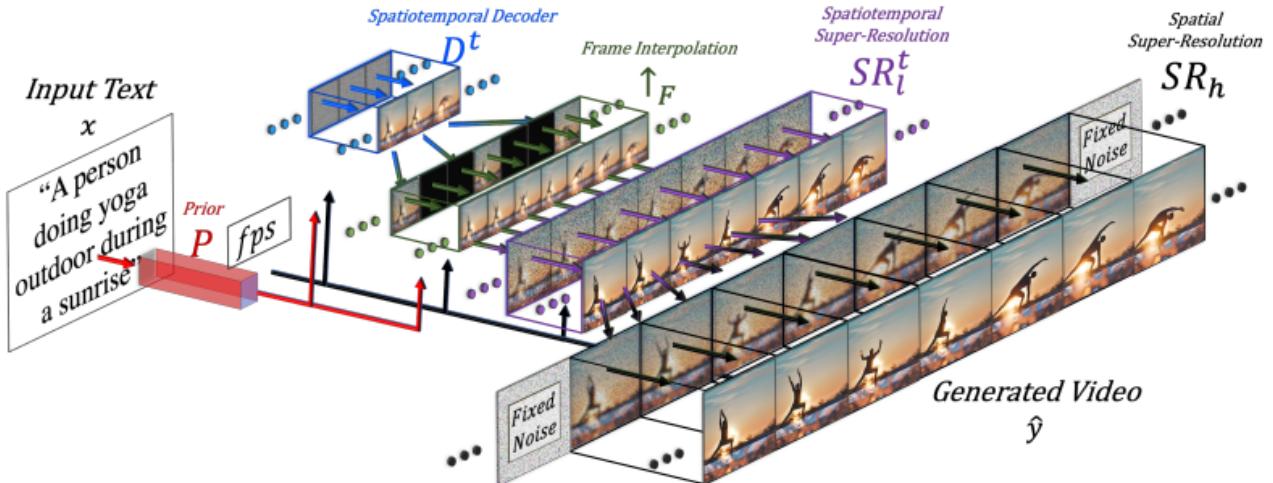
Hình 5: Khối attention residual

- **GPT:** Ở giai đoạn tiếp theo, sau khi đã có những vector rời rạc trong không gian latent ở bước trước, tác giả dùng một kiểu kiến trúc tương tự GPT (Generative pretrained transformer) để mô hình hóa những vector này thành một chuỗi liên tục. Và train bằng cách dự đoán vector của frame kế tiếp.

Tóm lại, videoGPT giới thiệu một kiến trúc đơn giản và hiệu quả cho bài toán tạo sinh video, sử dụng 2 thành phần chính là khối VQ-VAE có chức năng học không gian latent rời rạc của các video và khối kiến trúc tương tự như GPT để mô hình hóa và tạo sinh chuỗi. Sự đơn giản của phương pháp này cùng với tính hiệu quả của nó mang đến một hướng nghiên cứu về các mô hình tạo video.

## 4.2 Make-A-Video

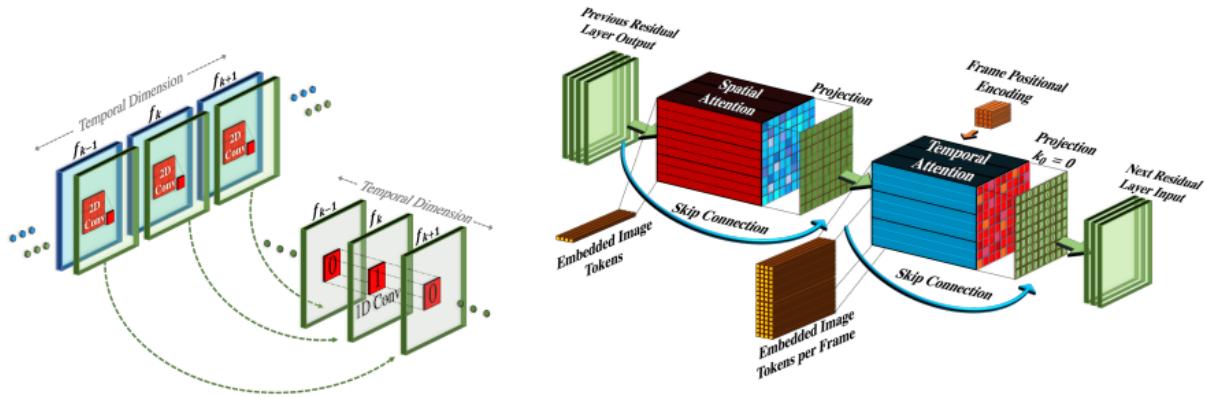
Được giới thiệu trong [7], đây là một phương pháp tạo video từ văn bản mà không cần dữ liệu video đi kèm văn bản mô tả, từ đó giải quyết vấn đề về thiếu hụt data trong tác vụ tạo sinh video. Phương pháp này mở rộng từ mô hình tạo hình ảnh từ văn bản mô tả (Text to Image-T2I) bằng cách thêm vào các mô-đun spatial-temporal, cho phép tạo video có độ phân giải cao và tốc độ khung hình cao từ văn bản đầu vào. Make-A-Video không yêu cầu dữ liệu video có kèm theo văn bản mô tả và có thể tạo ra video với chất lượng và mức độ chi tiết cao, mở ra khả năng áp dụng trong nhiều lĩnh vực khác nhau.



Hình 6: Make-A-Video high-level architecture

Make-A-Video bao gồm ba thành phần chính:

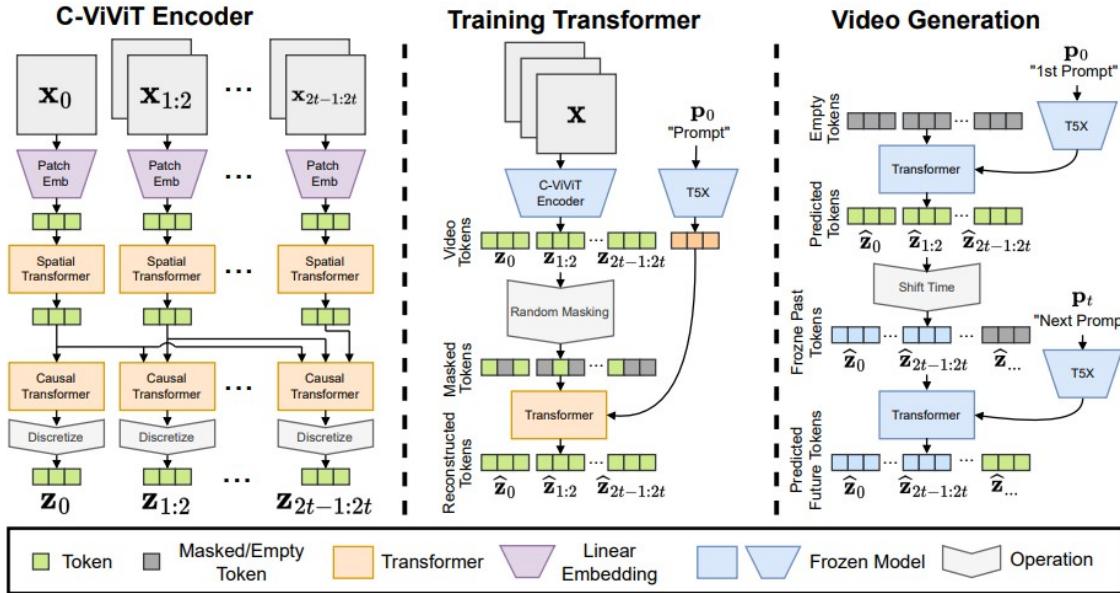
- **Một mô hình cơ sở T2I được huấn luyện trên các cặp text-image:** Sử dụng 3 networks để tạo ra hình ảnh độ phân giải cao từ văn bản. **Prior network P** trong quá trình inference tạo ra các image embedding từ text embedding và các BPE encoded text tokens. **Decoder network D** tạo ảnh RGB với low-resolution ( $64 \times 64$ ) dựa trên các image embedding. **Hai super-resolution networks  $SR_1, SR_h$**  tăng độ phân giải của ảnh lên thành  $256 \times 256$  và  $768 \times 768$ .
- **Spatiotemporal network sử dụng các spatiotemporal convolution và attention layers để mở rộng các blocks.** Điều này đòi hỏi phải điều chỉnh các lớp tích chập và các lớp attention. Ngoài ra, spatiotemporal decoder network tạo ra những khung hình RGB độ phân giải thấp ban đầu. Các khung hình này sau đó được cải thiện bởi interpolation network và các super-resolution networks đảm bảo sự nhất quán giữa các khung hình để tránh flickering artifacts. Super-resolution được sử dụng do hạn chế về bộ nhớ và khả năng tính toán cùng với sự khởi tạo consistent noise giữa các khung hình để duy trì nội dung chi tiết giữa chúng.



Hình 7: Kiến trúc và cơ chế khởi tạo của Pseudo-3D convolutional và attention layers, cho phép chuyển đổi mô hình T2I đã được huấn luyện trước sang chiều không gian thời gian.

- **Pseudo-3D convolutional layers:** Việc sử dụng Pseudo-3D convolutional layers nhằm nâng cao 2D convolutional network cho temporal learning mà không cần đến tính toán của 3D convolutions. Xếp 1D convolution sau 2D convolution layer nhằm thúc đẩy việc chia sẻ thông tin giữa không gian và thời gian đồng thời duy trì sự tách biệt giữa 2D convolution đã huấn luyện trước và 1D convolution mới. Lớp này được định nghĩa theo mặt toán học, với tensor được biểu diễn trong không gian đa chiều nơi mà B, C, F, H, và W lần lượt ký hiệu cho batch, channels, frames, height, width. Nó cũng lưu ý rằng khi khởi tạo, mạng có khả năng tạo ra nhiều hình ảnh chính xác về mặt văn bản nhưng không nhất quán về thời gian do nhiễu ngẫu nhiên.
- **Pseudo-3D attention layers:** với mục đích là chèn thông tin văn bản vào network mà không tốn nhiều tài nguyên tính toán như sử dụng full 3D convolutions. Việc này được thực hiện bằng cách xếp chồng các 1D attention layer lên trên pre-trained spatial attention layer. Những lớp này cho phép network duy trì sự chú ý trong không gian đồng thời kết hợp thông tin thời gian. Cách tiếp cận này giúp khả thi về quản lý tài nguyên bộ nhớ và tính toán, đồng thời giải quyết thách thức tích hợp thời gian vào tạo ảnh và video.
- Spatiotemporal network còn có thêm một thành phần quan trọng cho việc tạo video là **frame interpolation network** nhằm cải thiện tốc độ tạo khung hình. Network có thể tăng số lượng khung hình của video được tạo ra thông qua nội suy khung hình để tạo ra các chuỗi smooth hơn hoặc ngoại suy khung hình trước/sau để kéo dài video. Để quản lý giới hạn về bộ nhớ và tính toán, network tinh chỉnh một spatiotemporal decoder cho nhiệm vụ nội suy khung hình với mask để nâng cao chất lượng video. Quá trình này bao gồm việc thêm các channels bổ sung vào input và sử dụng zero-padding cho các masked frames. Network cũng có khả năng bỏ qua một số khung hình biến đổi và điều kiện fps trong thời gian suy luận, cho phép linh hoạt trong việc nâng cao chất lượng thời gian từ một số lượng khung hình nhất định sang một số lượng lớn hơn.

### 4.3 Phenaki



Hình 8: Kiến trúc chung của mô hình Phenaki

Mô hình Phenaki, được trình bày trong công trình "Phenaki: Variable Length Video Generation From Open Domain Textual Description" [8], mở ra một hướng mới trong việc tạo video từ văn bản. Được thiết kế để tạo ra những đoạn video chân thực từ các mô tả văn bản, Phenaki giải quyết nhiều thách thức đáng kể trong lĩnh vực này. Đó là việc tối ưu hóa hiệu suất tính toán, đổi mới với tình trạng thiếu hụt nguồn dữ liệu văn bản-video chất lượng cao và duy trì sự nhất quán trong không gian cũng như tính liên kết về thời gian cho các video dài. Bằng cách biến đổi các đoạn văn bản thành những câu chuyện video liền mạch, Phenaki không chỉ đặt ra một tiêu chuẩn mới mà còn cho thấy khả năng ứng dụng rộng rãi của nó trong thực tế.

#### 4.3.1 Kiến trúc Encoder-Decoder cho video: C-ViViT

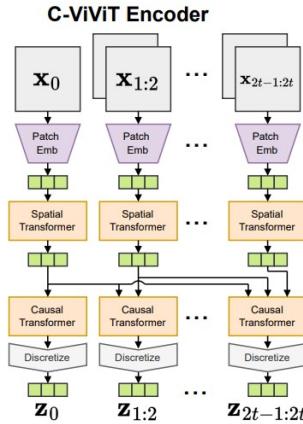
Để có thể làm được những điều trên, Phenaki giới thiệu một kiến trúc encoder-decoder mới, đặt tên là C-ViViT, được mô tả trong Hình 8. Kiến trúc này có 2 khả năng chính.

- Kiến trúc này có thể khai thác sự dư thừa thời gian trong video để cải thiện chất lượng tái tạo mỗi khung hình đồng thời nén số lượng video token từ 40% trở lên.
- Cho phép mã hóa và giải mã các video có độ dài thay đổi dựa trên cấu trúc của nó.

Một trong những thách thức lớn nhất đối với các mô hình tạo sinh video từ văn bản đó chính là quá trình nén video thành các vector biểu diễn. Các công trình trước đó đi theo 2 hướng chính để làm việc này, đó là sử dụng các image encoder theo từng frame một, hoặc là sử dụng một video encoder có số frame cố định. Hướng thứ nhất cho phép tạo sinh video với độ dài tùy thích, tuy nhiên trên thực tế thì việc tạo sinh video với độ dài lớn là rất khó khăn và tốn nhiều tài nguyên. Phương pháp thứ 2 có thể giúp giảm tài nguyên tính toán, tuy nhiên lại không thể tạo sinh video với độ dài tùy thích. Mục tiêu của Phenaki là giải quyết 2 vấn đề này: tạo sinh video với độ dài tùy thích, đồng thời nén số lượng token video về mức tối thiểu. Kiến trúc C-ViViT, là một biến thể của mô hình ViViT có khả năng tạo

sinh video và nén video trong các chiều thời gian và không gian, trong khi vẫn có tính hồi quy theo thời gian.

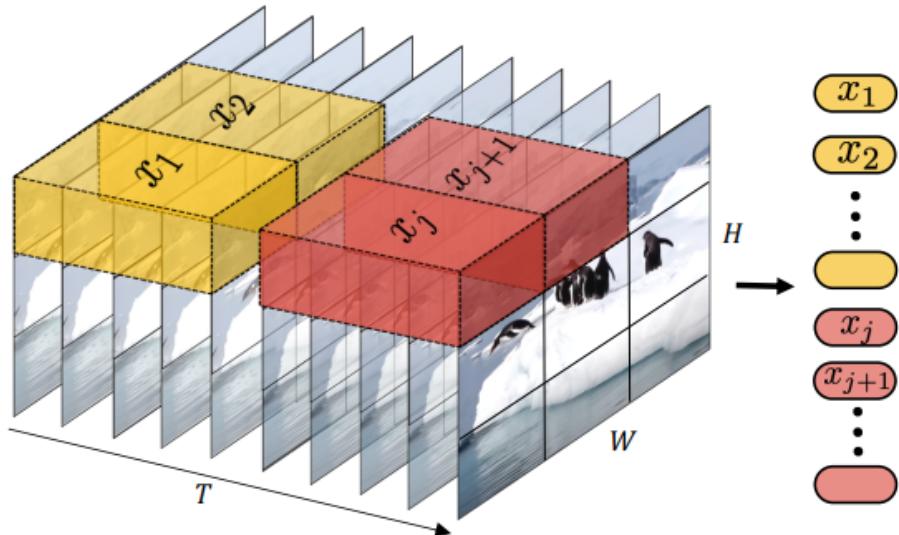
### **Khối Encoder:**



Hình 9: Kiến trúc của khối encoder

Khối encoder chuyển chuỗi video thành các token biểu diễn nhỏ gọn phù hợp để xử lý bởi các lớp transformers của mô hình. Quy trình bắt đầu với một chuỗi video ban đầu ở không gian bốn chiều biểu diễn thời gian, chiều cao, chiều rộng và kênh màu. Chuỗi này sau đó được nén thành biểu diễn token nhỏ hơn phân biệt giữa khung hình đầu tiên và các token video không gian-thời gian tiếp theo, phụ thuộc vào các khung hình trước.

Quá trình nén được thực hiện bằng cách trích xuất các patches không chồng lấn từ khung hình đầu tiên và các khung hình video tiếp theo, sau đó được làm phẳng và chiếu vào không gian chiều thấp hơn. Các chiều không gian và thời gian của các patches này được sắp xếp lại thành định dạng tensor phân biệt rõ ràng giữa chiều không gian và chiều thời gian. Quá trình này được biểu diễn trong hình 8.



Hình 10: Quá trình trích xuất các patches không chồng lấn sau đó làm phẳng và giảm chiều.

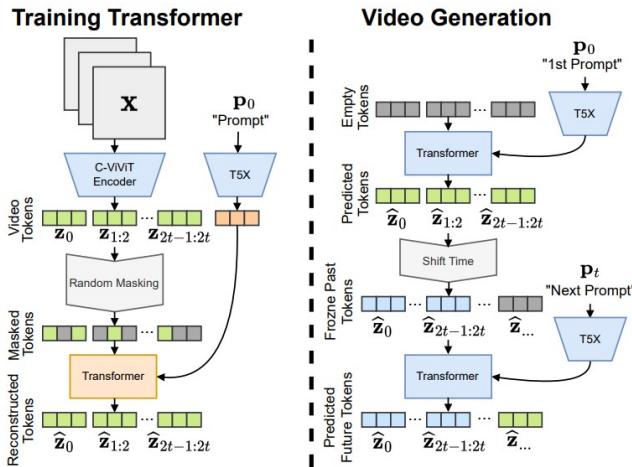
Bộ mã hóa áp dụng nhiều lớp transformer trên các chiều không gian sử dụng Spatial Transformer với attention toàn cục. Khối transformer này có nhiệm vụ học thông tin về không gian. Tiếp theo là

các lớp transformer bổ sung trên chiều thời gian sử dụng causal attention. Các lớp này có nhiệm vụ học thông tin về thời gian. Điều này đảm bảo rằng mỗi token không gian chỉ tương tác với các token không gian từ các khung hình trước, cho phép khung hình đầu tiên được mã hóa độc lập. Thiết kế này tạo điều kiện cho việc kết hợp các mô hình text-to-image trong mô hình video và cho phép quá trình tạo video được điều kiện hóa (conditioned) bởi một tập hợp khung hình ban đầu.

Các vector đầu ra sau khi đi qua các lớp transformer sẽ được lượng tử hóa về không gian latent rời rạc. Tương tự như trong kiến trúc VQ-VAE.

**Khối Decoder:** Khối decoder của C-ViT đơn giản là một khối encoder được lật ngược. Đầu tiên token được chuyển đổi thành các embedding. Sau đó là temporal transformer, tiếp theo là spatial transformer. Đầu ra sau đó được áp dụng một phép chiếu tuyến tính đơn để ánh xạ các token trở lại không gian pixel.

#### 4.3.2 Tạo sinh video từ văn bản với bidirectional transformers



Hình 11: Quá trình training bidirectional transformer

Việc chuyển đổi từ văn bản sang video có thể được định nghĩa như một tác vụ sequence-to-sequence, thường được giải quyết bằng cách sử dụng các mô hình autoregressive transformer dự đoán tuần tự các token video từ nhúng văn bản. Tuy nhiên, cách tiếp cận này trở nên không hiệu quả cho các video dài do thời gian lấy mẫu tăng tuyến tính.

Để cải thiện hiệu quả, tác giả sử dụng mô hình bidirectional transformer có khả năng dự đoán đồng thời nhiều token video, giảm đáng kể thời gian lấy mẫu bắt kể độ dài của chuỗi video. Trong quá trình huấn luyện, một phần các token video được che kín và các token này được dự đoán bằng các embedding văn bản và các token không bị che kín, với mô hình học thông qua hàm loss cross-entropy.

Phương pháp này, lấy cảm hứng từ các kỹ thuật như được sử dụng trong MaskGIT, giảm đáng kể số bước lấy mẫu cần thiết (thông thường từ 12 đến 48 bước), có khả năng cải thiện chất lượng video được tạo ra trong khi đảm bảo quá trình xử lý nhanh hơn.

**Quá trình inference** Trong quá trình inference, trước tiên tất cả các video token được đánh dấu là token đặc biệt [MASK]. Sau đó, tại mỗi bước inference, tất cả các token video bị che được dự đoán cùng một lúc dựa trên các embedding văn bản và các token video không bị che (đã dự đoán). Sau đó, một tỷ lệ  $\beta_i$  của các token được dự đoán tại bước lấy mẫu  $i$  được giữ lại, và các token còn lại sẽ được che lại và dự đoán lại ở bước tiếp theo.

### 4.3.3 Tổng kết

Phenaki được giới thiệu là một mô hình có khả năng tạo ra video có độ dài biến thiên dựa trên chuỗi prompt văn bản từ các chủ đề mở. Nó sử dụng C-ViViT làm bộ mã hóa video, một mô hình mới cung cấp khả năng nén không gian-thời gian hiệu quả trong khi vẫn duy trì tính tự hồi quy theo thời gian. Phenaki cho thấy kết quả hứa hẹn trong việc dự đoán video và có thể tạo ra video dài từ lời nhắc văn bản, với sự linh hoạt để bắt đầu từ một khung hình nhất định. Phenaki có thể tạo ra các câu chuyện video dài, có mạch lạc từ nhiều prompt văn bản, minh họa tiềm năng của nó như một công cụ sáng tạo cho việc kể chuyện bằng video.

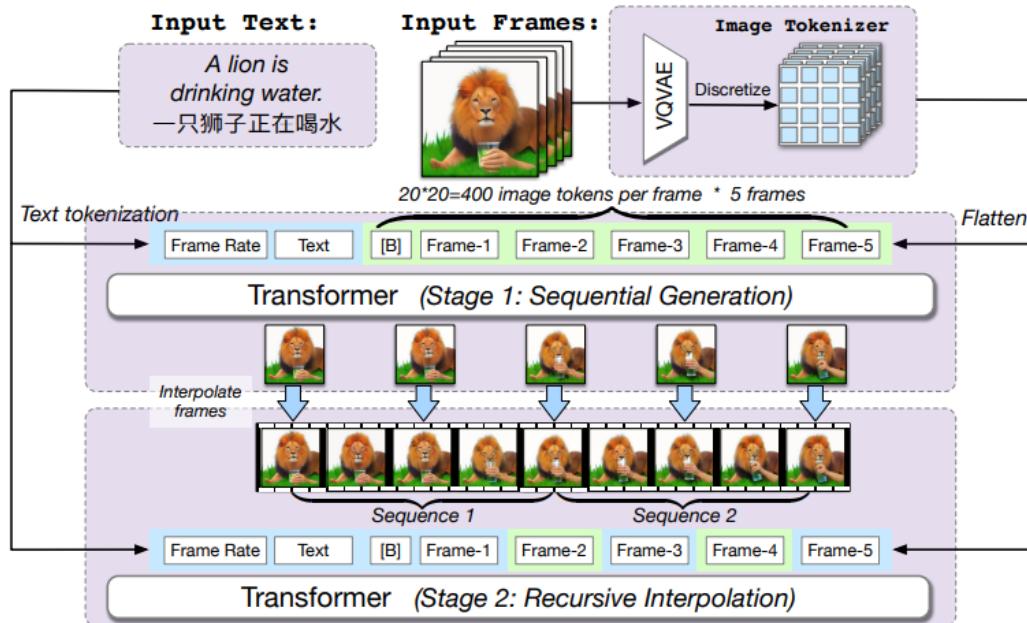
## 4.4 CogVideo

CogVideo [3] là một mô hình tạo video từ văn bản mô tả sử dụng kiến trúc Transformer với 9 tỷ tham số, được huấn luyện dựa trên mô hình sinh ảnh từ văn bản mô tả CogView2. CogVideo sử dụng chiến lược huấn luyện phân cấp nhiều tốc độ khung hình để cải thiện việc căn chỉnh giữa văn bản mô tả và video, cũng như tinh chỉnh khả năng kiểm soát độ chính xác trong quá trình tạo video. Trong paper CogVideo đề cập đến việc mở rộng và áp dụng cơ chế Swin attention trong tạo video tự động, nhằm tăng cường tốc độ huấn luyện và suy luận.

CogVideo sử dụng Multi-frame-rate Hierarchical Training để huấn luyện mô hình. Ý tưởng chính là thêm một frame-rate token vào văn bản mô tả và lấy mẫu các khung hình ở frame-rate này để tạo thành một chuỗi huấn luyện cố định. Động lực dựa trên hai phần:

- Tách video dài thành các đoạn ở frame-rate cố định thường dẫn đến sự không khớp nghĩa. CogVideo vẫn sử dụng toàn bộ văn bản nhưng đoạn clip bị cắt có thể chỉ chứa hành động không hoàn chỉnh.
- Các khung hình liền kề thường rất giống nhau. Một sự thay đổi lớn so với khung hình trước có thể gây ra sai số lớn. Điều này khiến các mô hình ít có xu hướng khám phá mối liên hệ dài hạn vì đơn giản sao chép khung hình trước đó giống như một lối tắt.

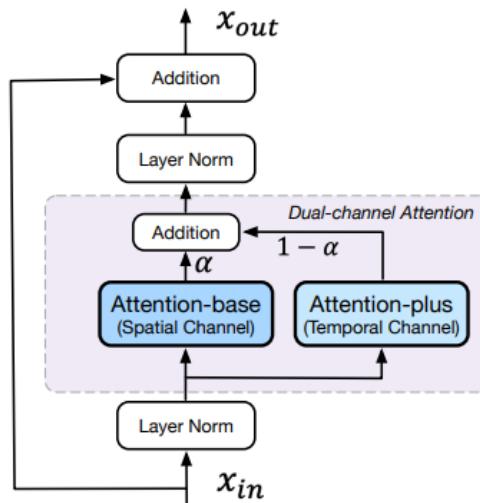
Trong quá trình huấn luyện, nhằm đạt được sự khớp chính xác giữa văn bản mô tả và hình ảnh. CogVideo thiết lập trước một dãy các frame-rate và chọn ra frame-rate thấp nhất có thể cho từng cặp text-video, đảm bảo có thể lấy mẫu được ít nhất là 5 khung hình. Dù phương pháp này cải thiện sự phù hợp giữa văn bản mô tả và video, nhưng video tạo ra ở frame-rate thấp có thể không liền mạch. Do đó, một frame interpolation model được tạo ra để thêm các khung hình chuyển tiếp vào video, giúp cho quá trình sinh video trở nên mượt mà hơn. Nhờ vào sự linh hoạt của CogLM, hai mô hình này có thể sử dụng chung một cấu trúc và quy trình huấn luyện, chỉ khác biệt ở điểm sử dụng các attention masks.



Hình 12: Multi-frame-rate hierarchical generation framework in CogVideo

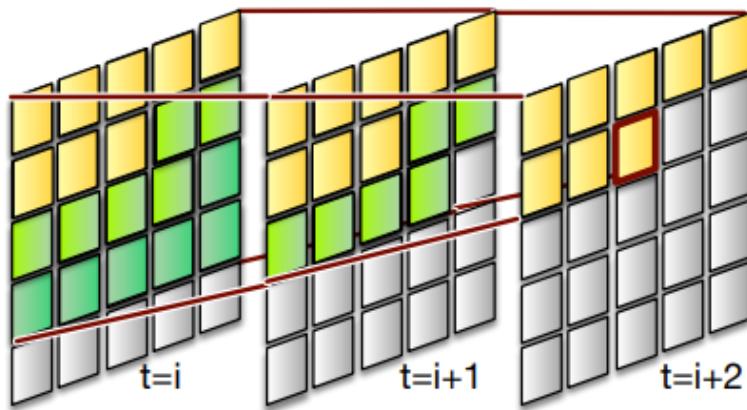
Quá trình Multi-frame-rate hierarchical generation của CogVideo là một quá trình đệ quy gồm hai giai đoạn. Đầu tiên là sinh tuần tự khung hình chủ chốt dựa trên frame-rate thấp và văn bản mô tả; thứ hai là nội suy đệ quy các khung hình dựa trên văn bản mô tả, frame-rate và các khung hình đã biết, với mục tiêu tạo ra video có nhiều khung hình hơn thông qua việc chia các khung hình đã sinh ra và nội suy thêm khung hình giữa chúng.

Cũng trong bài báo CogVideo tác giả đề xuất việc sử dụng các mô hình tạo ảnh đã được huấn luyện trước để hỗ trợ việc tạo video từ văn bản mô tả, thay vì dựa trên việc thu thập dữ liệu ảnh và video chất lượng cao, quá trình này thường tốn kém và mất thời gian. Cụ thể áp dụng Dual-channel Attention bằng cách thêm spatial-temporal attention channel vào mỗi lớp chuyển đổi của mô hình CogView2 đã được huấn luyện trước, giữ nguyên các tham số cũ và chỉ huấn luyện các tham số mới của lớp attention. Điều này giúp giữ lấy kiến thức về mối quan hệ text-image từ CogView2 mà không làm hỏng trọng số đã học được khi chuyển sang tạo video, vốn đòi hỏi sự chú ý đến cả không gian và thời gian.



Hình 13: Dual-channel attention block

Để giảm thiểu áp lực về thời gian và bộ nhớ trong quá trình huấn luyện và suy luận, CogVideo áp dụng Swin Attention, mở rộng nó cho các tình huống auto-regressive và temporal scenario bằng cách sử dụng auto-regressive attention mask trong shifted windows. Phát hiện thú vị là Swin Attention cho phép tạo ra song song ở các vùng xa của các khung hình khác nhau, tăng tốc độ cho auto-regressive, với sự phụ thuộc vào việc tạo token dựa trên auto-regressive mask và shifted windows, giới hạn sự chú ý chỉ trong phạm vi kích thước cửa sổ.

Hình 14: Autoregressive swin attention (window size  $2 \times 2$ )

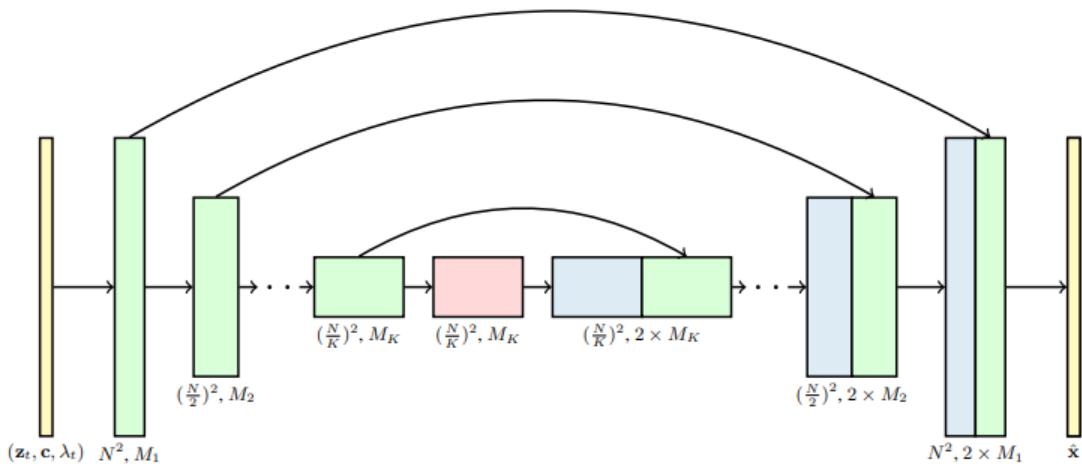
## 5 Diffusion-based models

Một hướng tiếp cận khác cho tác vụ tạo sinh video từ văn bản là hướng áp dụng mô hình diffusion. Hướng tiếp này là một trong những phát triển đáng chú ý gần đây, được thúc đẩy bởi thành công mà các mô hình diffusion đã đạt được trong việc sinh ra hình ảnh chất lượng cao. Các nghiên cứu tiên tiến trong lĩnh vực này đang khám phá cách thức áp dụng các nguyên tắc diffusion để tạo ra video không chỉ có chất lượng hình ảnh sắc nét mà còn đảm bảo tính liên tục và mượt mà về mặt chuyển động. Dưới đây là một số công trình nổi bật đã đưa ra cách tiếp cận sử dụng mô hình diffusion cho việc sinh video, mở ra những khả năng mới cho việc tạo ra nội dung video phong phú và đa dạng từ văn bản mô tả.

## 5.1 Video Diffusion Models

Dây là một trong những công trình nghiên cứu đầu tiên tiếp cận theo hướng sử dụng mô hình diffusion để tạo sinh video. Video Diffusion Models [2] là sự mở rộng tự nhiên của kiến trúc image diffusion và cho phép đào tạo chung từ dữ liệu ảnh và video. Để tạo ra video dài và độ phân giải cao hơn, các tác giả đã giới thiệu một kỹ thuật lấy mẫu điều kiện mới cho việc mở rộng video theo không gian và thời gian.

Trong công trình nghiên cứu về mô hình hình ảnh, U-Net là kiến trúc tiêu chuẩn cho image diffusion, bao gồm quá trình giảm và tăng mẫu không gian với các skip connections. U-Net được xây dựng từ các khối 2D convolutional và mỗi khối được sau bởi một khối spatial attention. Trong bài báo Video Diffusion Models tác giả đề xuất mở rộng kiến trúc này cho dữ liệu video với U-Net 3D, phân tách theo không gian và thời gian, và thêm khối temporal attention, cho phép đào tạo chung trên video và hình ảnh, cải thiện chất lượng mẫu.

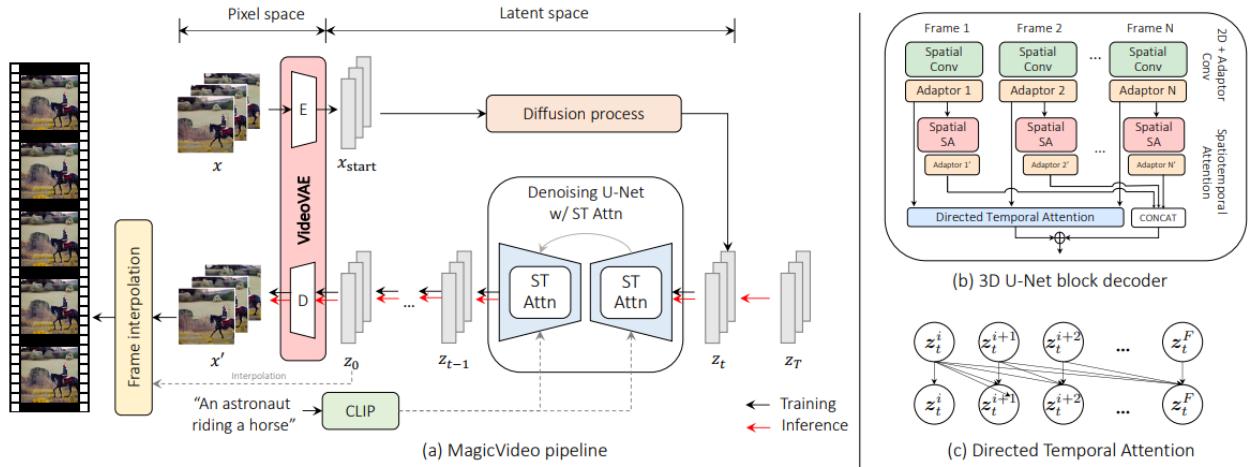


Hình 15: 3D U-Net architecture trong Video Diffusion Models

Thông thường việc giải quyết thách thức tính toán khi mô hình hóa video, thường gồm hàng trăm đến hàng nghìn khung hình, bằng cách đào tạo model trên một tập hợp con khung hình và sau đó mở rộng mẫu để tạo video dài hơn. Hai phương pháp lấy mẫu điều kiện từ diffusion model gồm: sử dụng phương pháp thay thế - không hiệu quả cho mô hình video do thiếu sự liên kết giữa các khung hình được sinh ra, và phương pháp được đề xuất trong paper gọi là reconstruction-guided sampling. Phương pháp này cải thiện chất lượng mẫu bằng cách điều chỉnh denoising model với gradient term based dựa trên sự tái tạo của model, đặc biệt khi kết hợp với Langevin diffusion samplers. Nó cũng mở rộng sang nội suy không gian cho super-resolution, thể hiện tính linh hoạt của phương pháp trong việc tạo video độ phân giải cao từ đầu vào độ phân giải thấp.

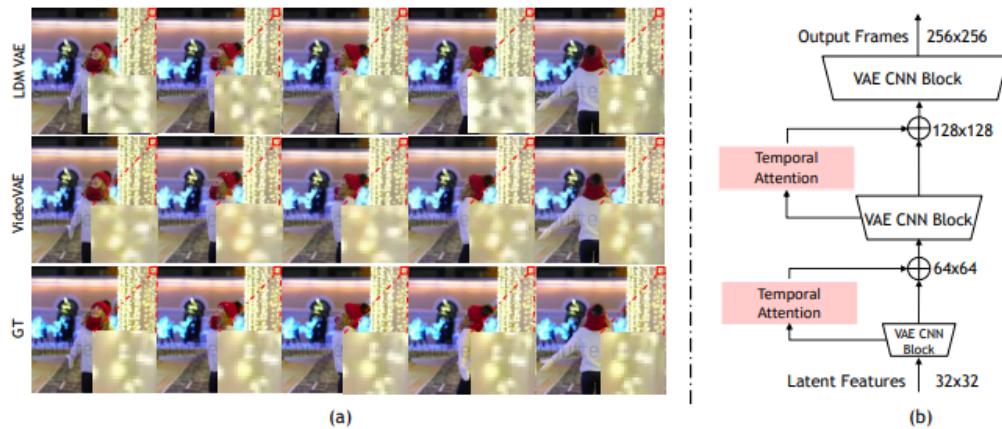
## 5.2 MagicVideo

MagicVideo [11] có thể tạo ra các đoạn video mượt mà phù hợp với văn bản mô tả đã cho. Nhờ vào kiến trúc 3D U-Net và hiệu quả cùng với việc mô hình hóa phân phối video trong không gian low-dimensional. MagicVideo có thể tổng hợp video  $256 \times 256$  spatial resolution trên một card GPU, giảm đến 64 lần lượng tính toán so với Video Diffusion Models (VDM) về FLOPs. MagicVideo còn giới thiệu hai thiết kế mới để thích nghi bộ lọc U-Net được đào tạo trên dữ liệu hình ảnh sang dữ liệu video, và chứng minh khả năng tạo ra các đoạn video chất lượng cao với nội dung thực tế hoặc tưởng tượng.



Hình 16: (a) data flow cho cả quá trình training và inference: trong quá trình training, timestep t sẽ được chọn mẫu ngẫu nhiên từ  $[0, T]$  và các khung hình video đầu vào bị làm nhiễu qua quá trình lan truyền, U-Net được sử dụng để học cách tái tạo các khung hình video. Gaussian noise được chọn mẫu ngẫu nhiên trong suy luận, và denoising process được lặp lại T lần. Latent vector z sau đó được đưa vào bộ giải mã VAE và chuyển đổi sang không gian RGB. (b) là cấu trúc của spatiotemporal attention (ST-Attn). (c) là directed attention được sử dụng trong ST-Attn

Trong quá trình huấn luyện model tạo video từ văn bản mô tả, MagicVideo tiếp cận bằng cách lấy mẫu một phần nhỏ các khung hình liên tiếp và xác định rõ frame-rate mới dựa trên độ dài mẫu. Để cải thiện chất lượng, model được huấn luyện trước mà không cần ghép cặp văn bản-video sử dụng dữ liệu video chất lượng cao, sau đó được tinh chỉnh với mục tiêu huấn luyện cụ thể, áp dụng loss function cho từng khung hình và sử dụng các embeddings để tinh chỉnh mô hình, cho phép tạo video liền mạch và chất lượng cao từ văn bản mô tả.



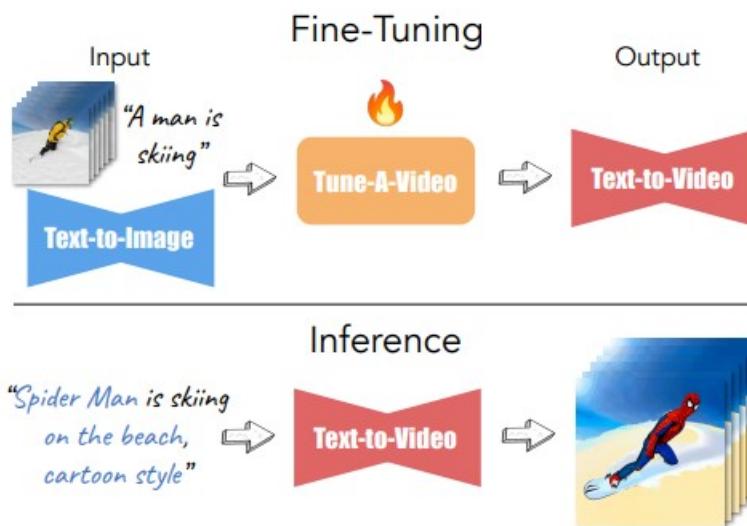
Hình 17: VideoVAE decoder

Trong quá trình tổng hợp hình ảnh RGB từ các decoding latent features qua VAE decoder đã được đào tạo trước, quá trình tái tạo từng khung hình video dẫn đến hiện tượng pixel dithering, làm giảm chất lượng thẩm mỹ hình ảnh. Các đặc trưng không gian với kích thước lớn hơn sẽ giảm bớt dithering nhưng cũng làm tăng chi phí tính toán. Để cải thiện chất lượng hình ảnh mà không tăng tính toán, tác

giả đã giữ kích thước thấp cho latent features và thêm vào decoder hai khối temporal directed attention layers, tạo nên VideoVAE decoder giúp hiệu quả trong việc giảm thiểu dithering.

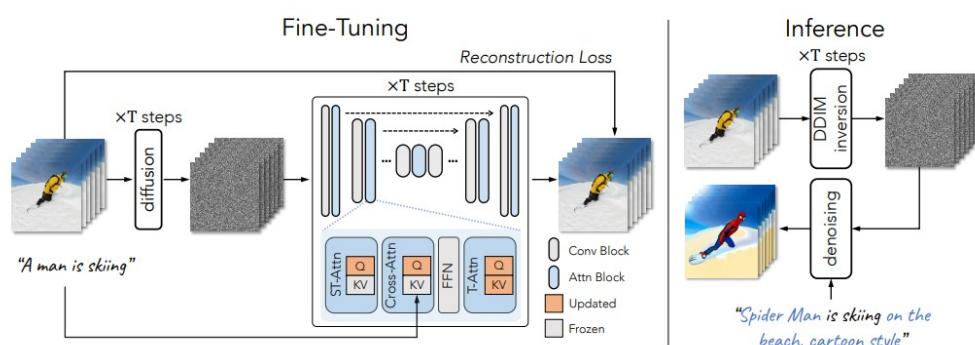
### 5.3 Tune-A-Video

"Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation" [9] là paper giới thiệu một phương pháp mới để tạo ra video từ các prompt văn bản bằng cách tận dụng khả năng của các mô hình diffusion text-to-image (T2I) đã được pretrained. Nghiên cứu này giải quyết vấn đề thiếu hụt các cặp văn bản-video chất lượng cao cho việc đào tạo bằng cách đề xuất phương pháp one-shot video-tuning, từ đó không cần phải phụ thuộc vào các cặp video-văn bản caption.



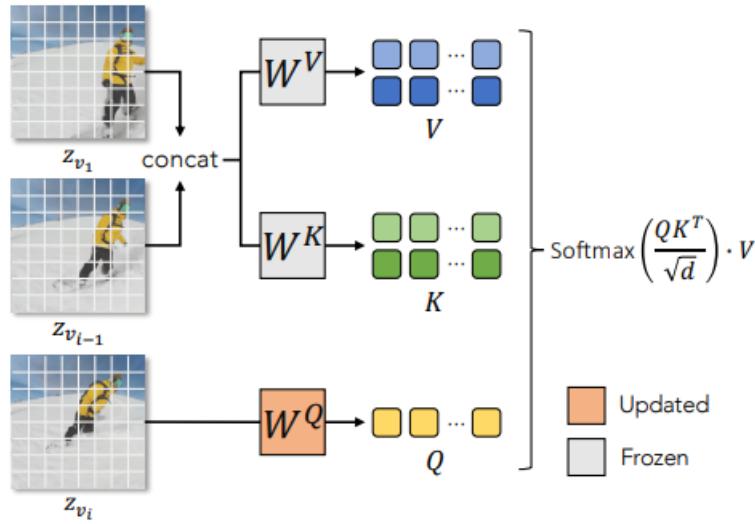
Hình 18: Cái nhìn tổng quát về paper Tune-A-Video

Cụ thể với một video cho trước, nhiệm vụ của model là fine-tune cụ thể trên video đó, để từ đó có thể edit video đó dựa theo yêu cầu từ text prompt của user. Điều này giúp giảm chi phí tính toán, khi thay vì phải train trên một tập dataset về video lớn, với chi phí về gán nhãn và về phần cứng khổng lồ, ta có thể tận dụng các mô hình Text-to-image có sẵn để tune duy nhất cho một video, và tương tác trên video đó. Phương pháp này được gọi là zero-shot video-tuning.



Hình 19: Pipeline của paper Tune-a-video

Paper này sử dụng Latent Diffusion Model (LDM, [6]) đã được pretrained như một mô hình text-to-image. Trong đó LDM sử dụng kiến trúc U-Net, sử dụng nhiều khối Convolution 2D và khối transformer, với từng khối transformer sử dụng các layer self-attention, cross-attention và feed-forward. Để sử dụng các khối này cho tác vụ video, nghiên cứu này có một số chỉnh sửa cho khối U-Net. Đầu tiên, khối Convolution 2D được thay thế bằng khối Convolution 3D giả, với các filter 3x3 được thay thế bằng các layer 1x3x3. Về khối attention, một khối temporal self-attention (T-Attn) được thêm vào ở cuối cùng ở mỗi khối transformer để mô tả thông tin về thời gian. Để tăng độ mạch lạc về thời gian cho video, paper này sử dụng một khối spatial-temporal attention thay cho khối self-attention. Trong đó, để giảm tài nguyên cần cho tính toán, attention chỉ được tính giữa frame hiện tại, frame liền trước nó và frame đầu tiên, trong đó, thông tin về feature giữa frame đầu tiên và frame liền trước được concat lại với nhau. Quá trình này được mô tả trong hình 20.



Hình 20: Minh họa Spatial-Temporal Attention

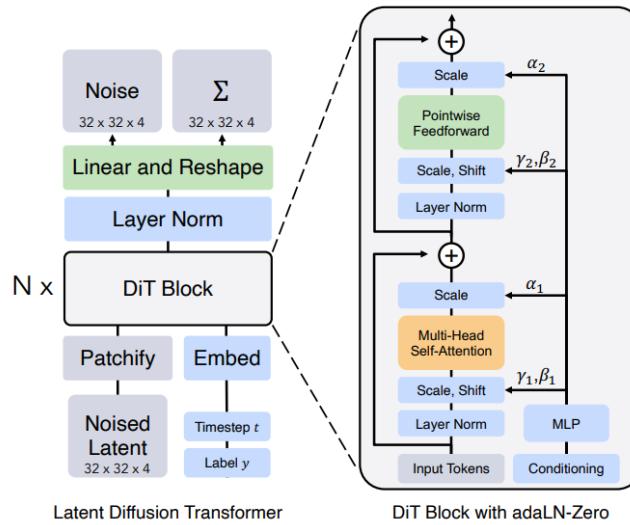
Tổng kết lại, bài báo giới thiệu một task mới có tên one-shot video-tuning cho việc tạo video từ văn bản (T2V), mà trong đó việc huấn luyện mô hình tạo sinh T2V chỉ sử dụng một cặp văn bản-video duy nhất với sự hỗ trợ của các mô hình chuyển đổi từ văn bản sang hình ảnh (T2I) đã được huấn luyện từ trước. Công cụ được giới thiệu, Tune-A-Video, hỗ trợ việc tạo và chỉnh sửa video dựa trên văn bản thông qua một chiến lược tuning đặc biệt và sự đảo ngược cấu trúc, đảm bảo tính nhất quán về thời gian trong các video được tạo ra. Hiệu quả của phương pháp này được chứng minh thông qua các thử nghiệm rộng rãi trên nhiều ứng dụng, cho thấy khả năng ẩn tượng của nó.

## 6 Diffusion Models with Transformer

Đây là một phương pháp kết hợp những điểm mạnh của cả hai cách làm trước đây. Bằng cách dùng mô hình diffusion cùng với kiến trúc transformer - thay vì dùng kiến trúc UNet thông thường - cách tiếp cận này mở ra khả năng thiết kế mô hình linh hoạt hơn và tận dụng khả năng nhận diện hình ảnh ưu việt của transformer. Điều này giúp tạo ra những hình ảnh, video chất lượng cao, với hiểu biết sâu sắc về nội dung và bối cảnh hơn.

Nghiên cứu "Scalable Diffusion Models with Transformers" [5] là một công trình tiên phong khám phá một nhánh mới của mô hình diffusion sử dụng kiến trúc transformer. Khác với các mô hình diffusion truyền thống thường sử dụng cốt lõi U-Net, nghiên cứu này giới thiệu việc sử dụng transformer hoạt động trên các patches trong không gian latent của hình ảnh. Điểm chính của nghiên cứu là về khả năng

mở rộng của các Diffusion Transformers (DiTs) thông qua phân tích độ phức tạp của quá trình lan truyền xuôi, được đo bằng Gflops. Nghiên cứu đã tìm thấy một xu hướng nhất quán, khi mô hình DiTs có Gflops cao hơn, đạt được thông qua việc tăng độ sâu/rộng của transformer hoặc số lượng token đầu vào, thể hiện điểm Frechet Inception Distance (FID) thấp hơn, cho thấy hiệu suất tốt hơn. Đáng chú ý, mô hình lớn nhất, DiT-XL/2, đã vượt qua các mô hình khuếch tán trước đó về hiệu suất trên các benchmark phổ biến.



Hình 21: Caption

Một trong những đột phá chính của nghiên cứu này là thay thế kiến trúc U-Net thường được sử dụng trong các mô hình diffusion thành kiến trúc diffusion transformer. Hình 21 cho ta thấy cái nhìn tổng quan về thiết kế của khối DiT. Đầu vào của khối DiT là biểu diễn trong không gian latent  $z$  của một tấm ảnh (với một tấm ảnh kích thước  $256 \times 256 \times 3$ ,  $z$  có kích thước  $32 \times 32 \times 4$ ). Sau đó  $z$  sẽ được chia thành từng patches ảnh nhỏ, sau đó trải phẳng để tạo thành một chuỗi  $T$  các token hình ảnh, với mỗi token có chiều là  $d$ .

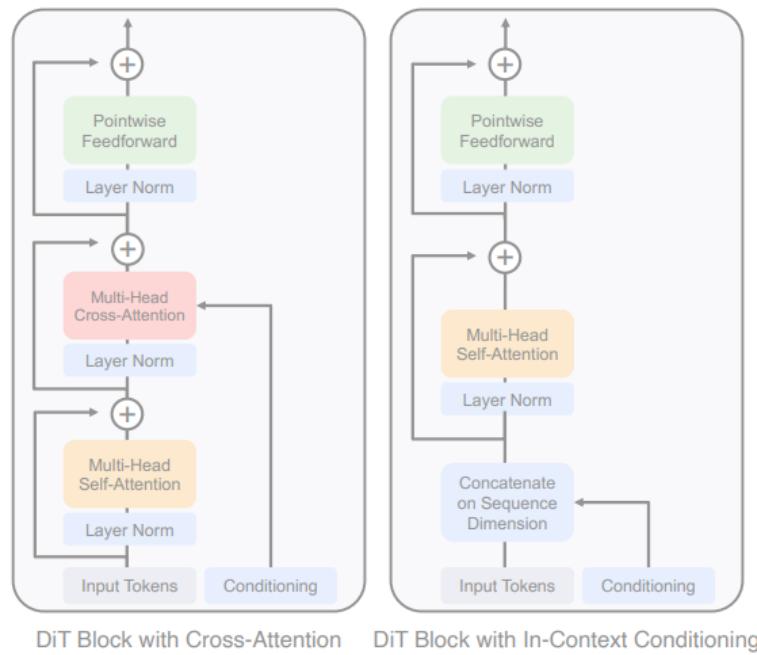
## 6.1 Diffusion Transformer

Bây giờ chúng ta sẽ tìm hiểu tổng quan về Diffusion Transformer được đề cập trong bài báo.

**Diffusion formulation** - Diffusion model thêm dần nhiều vào dữ liệu và sau đó học cách đảo ngược quá trình này. Quá trình đào tạo dựa vào việc giảm thiểu sai số giữa nhiều dự đoán và nhiều thực tế.

**Classifier-free guidance** - Phương pháp này cải thiện khả năng tạo mẫu của mô hình bằng cách điều chỉnh đầu ra dựa trên thông tin bổ sung như nhãn lớp để tạo ra ảnh chất lượng cao hơn.

**Latent diffusion models** - Đây là một phương pháp hiệu quả về mặt tính toán, nén ảnh vào biểu diễn không gian nhỏ hơn trước khi tạo ảnh, giúp tiết kiệm tài nguyên tính toán.



Hình 22: DiT Block.

**DiT block design** - Với DiT block tác giả đã có một số những thay đổi nhỏ so với ViT block chuẩn thông thường:

- In-context conditioning: DiT thêm các vector embedding của timestep và class labels như là hai token bổ sung trong chuỗi đầu vào, xử lý chúng không khác gì so với các token ảnh. Điều này tương tự như token cls trong ViTs, và nó cho phép DiT sử dụng các khối ViT chuẩn mà không cần chỉnh sửa.
- Cross-attention block: tại đây nối các embedding của timestep và class labels thành một chuỗi có độ dài hai, tách biệt từ chuỗi token ảnh. Khối transformer được chỉnh sửa để bao gồm thêm một lớp multi-head cross-attention theo sau multi-head self-attention.
- Adaptive layer norm (adaLN) block: thay thế các layer norm thông thường trong các khối transformer bằng adaptive layer norm - adaLN. Tại đây, thay vì học trực tiếp các tham số dimensionwise scale và shift parameters thì DiT hồi quy chúng từ tổng của các vector embedding của timestep và class labels.
- adaLN-Zero block: Trong nghiên cứu về ResNets, việc khởi tạo các residual block để hoạt động như identity function đã được chứng minh là có ích. Mô hình Diffusion U-Net áp dụng phương pháp tương tự với lớp tích chập cuối của mỗi khối. DiT áp dụng chiến lược này trong khối adaLN của mô hình DiT, đồng thời hồi quy thêm các tham số tỷ lệ để tối ưu hóa hiệu suất trước khi thực hiện kết nối phần dư.

## Phần V: Mô hình Sora

Sora là một mô hình AI có khả năng tạo ra các video lên đến một phút, với chất lượng hình ảnh cao và nội dung bám sát theo yêu cầu của người dùng. Mô hình hiện đang được mở cho các nhóm kiểm định

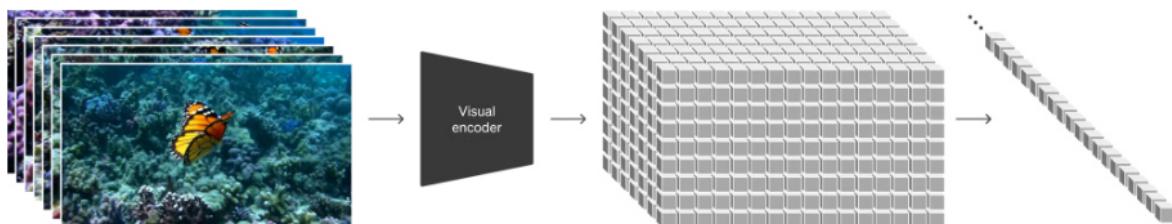
để đánh giá về các rủi ro tiềm ẩn và cũng được cung cấp cho các nghệ sĩ, nhà thiết kế và nhà làm phim để thu thập phản hồi nhằm cải thiện mô hình cho nhu cầu sáng tạo. Sora có thể tạo ra các cảnh phức tạp với nhiều nhân vật, các loại chuyển động cụ thể, và chi tiết chính xác về đối tượng và phông nền. Mặc dù có khả năng hiểu biết sâu sắc về ngôn ngữ và tạo ra các nhân vật biểu cảm, mô hình vẫn còn hạn chế như khó khăn trong mô phỏng vật lý của cảnh phức tạp và nhầm lẫn các chi tiết không gian trong yêu cầu.

**Safety** - Trước khi Sora được tích hợp vào các sản phẩm của OpenAI, mô hình sẽ được áp dụng các biện pháp an toàn quan trọng, bao gồm hợp tác với các chuyên gia từ các lĩnh vực như thông tin sai lệch và nội dung độc hại để kiểm thử mô hình. Hiện OpenAI cũng đang phát triển công cụ để phát hiện nội dung lừa đảo và dự định tích hợp metadata C2PA. Sử dụng các phương pháp an toàn từ sản phẩm sử dụng DALL-E 3 và tiếp tục phát triển kỹ thuật mới nhằm đảm bảo vào việc kiểm soát nội dung vi phạm chính sách sử dụng trước khi hiển thị cho người dùng.

## 7 Research techniques

Ở phần này chúng ta sẽ tìm hiểu tổng quan về mô hình Sora. Trong khi nghiên cứu trước đây tập trung vào video ngắn hoặc dữ liệu hình ảnh hạn chế, Sora phá vỡ giới hạn này bằng cách tạo ra video và hình ảnh đa dạng, kéo dài đến một phút với độ nét cao.

**Xử lý dữ liệu hình ảnh** - Áp dụng phương pháp từ các mô hình ngôn ngữ lớn trước đó là dùng dữ liệu quy mô internet để phát triển khả năng tổng quát sau đó chuyển đổi qua dữ liệu ảnh với các image patches thay cho các tokens văn bản. Cách tiếp cận này, đã được chứng minh là hiệu quả trong việc đại diện cho dữ liệu hình ảnh giúp Sora trở nên mạnh mẽ và có thể mở rộng để huấn luyện trên video và hình ảnh đa dạng. Quá trình biến đổi video thành các patches được thực hiện bằng cách nén chúng vào lower-dimensional latent space sau đó phân tách chúng vào spacetime patches.



**Mạng nén thông tin video** Một trong những thành phần rất quan trọng của Sora, nhưng lại không được đề cập quá kĩ trong báo cáo kĩ thuật của OpenAI đó chính là mô hình nén video. Mô hình này có chức năng chính là nén video gốc thành những vector biểu diễn trên không gian latent, và những vector biểu diễn này đã nén cả về chiều không gian lẫn chiều thời gian. Một mạng decoder cũng được train để ánh xạ những biểu diễn này về lại không gian pixel. Có một số suy đoán về kiến trúc của thành phần này, chẳng hạn một mô hình VAE, hay một dạng tương tự như VQGAN như các mô hình ở phần III có đề cập. Tuy nhiên ta không thể nào biết chính xác kiến trúc và cách training của thành phần này.

**Spacetime latent patches** - Với một video đầu vào đã được nén, Sora trích xuất một chuỗi các spacetime patches hoạt động như các token transformer. Cơ chế này cũng áp dụng cho ảnh vì ảnh chỉ là video với một khung hình duy nhất. Biểu diễn dựa trên patches của chúng cho phép Sora huấn luyện trên video và hình ảnh với độ phân giải, thời lượng và tỷ lệ khung hình đa dạng. Tại thời điểm suy luận,

chúng ta có thể kiểm soát kích thước của video được sinh ra bằng cách sắp xếp các patches được khởi tạo ngẫu nhiên trong một lưới có kích thước phù hợp.

**Mở rộng quy mô mạng transformer cho việc tạo sinh video** - Như chúng ta đã biết diffusion model là một loại mô hình sinh mô phỏng quá trình xoá mờ và tái tạo dữ liệu, thường được sử dụng trong việc sinh ra hình ảnh, video, hoặc âm thanh. Bằng cách bắt đầu từ dữ liệu nhiễu và dần dần loại bỏ nhiễu đó qua nhiều bước lặp, mô hình học cách tái tạo dữ liệu gốc từ dữ liệu đã bị làm nhiễu, từ từ sinh ra sản phẩm cuối cùng có chất lượng cao. Quá trình này thường được hỗ trợ bởi thông tin điều kiện, như văn bản mô tả, để hướng dẫn quá trình tái tạo dữ liệu theo mong muốn.

Trong bối cảnh của Sora, diffusion model được áp dụng để dự đoán và tái tạo các clean patches từ các noisy patches, với sự hỗ trợ của thông tin điều kiện như văn bản mô tả. Đặc biệt, Sora kết hợp cấu trúc của diffusion transformer (DiT), tận dụng khả năng mở rộng mạnh mẽ của bộ biến đổi qua nhiều lĩnh vực như mô hình hóa ngôn ngữ và sinh hình ảnh. Qua quá trình huấn luyện, chất lượng của dữ liệu sinh ra bởi Sora tiếp tục cải thiện, chứng minh khả năng mở rộng hiệu quả của diffusion model dưới dạng bộ biến đổi cho việc mô hình hóa video.



## 8 Một số khả năng của Sora

**Khả năng thông hiểu ngôn ngữ:** Đầu tiên đó chính là khả năng hiểu ngôn ngữ của Sora, từ đó có thể tạo ra video từ các prompt văn bản. Sora được train trên dữ liệu video-text với caption cho các video được thu thập bằng phương pháp re-captioning được giới thiệu trong paper DALLE 3. Tương tự như DALLE 3, Sora cũng sử dụng các mô hình GPT để biến những prompt của người dùng thành những câu captions dài và chi tiết.

**Prompting từ hình ảnh và video** Ngoài khả năng tạo sinh video từ văn bản, ta còn có thể prompt sora bằng ảnh và video. Khả năng này cho phép Sora có thể thực hiện những tác vụ về edit hình ảnh và video, chẳng hạn như tạo những video loop, tạo ảnh động từ ảnh tĩnh, kéo dài video tiến hoặc lùi theo thời gian, v.v.

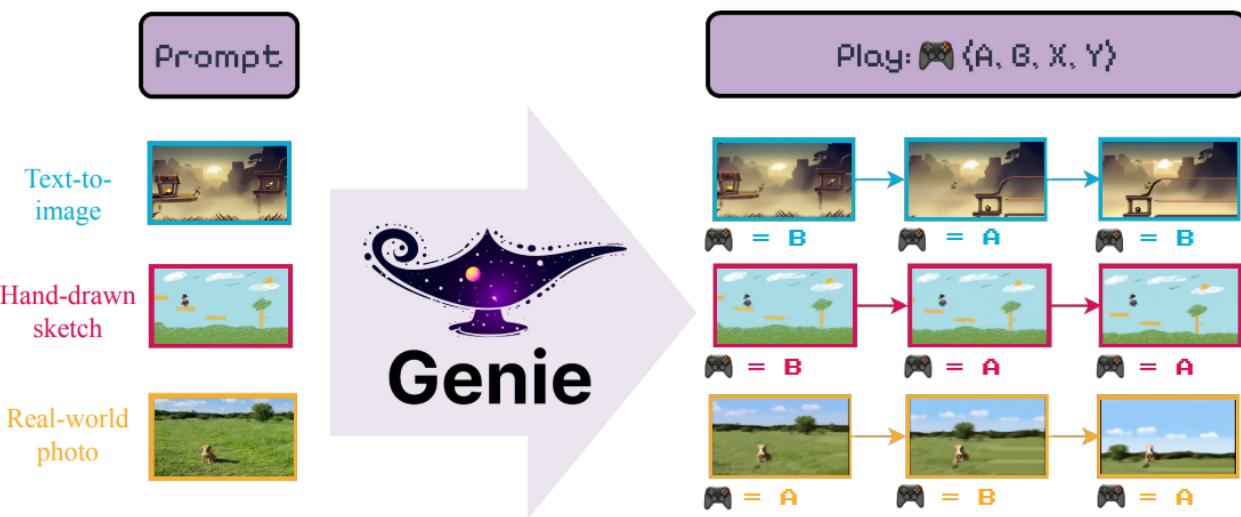
**Khả năng mô phỏng thế giới** Và cuối cùng, một khả năng của Sora khiến giới khoa học thảo luận và tranh cãi khá nhiều đó chính là khả năng mô phỏng thế giới của Sora. Khi được huấn luyện với lượng dữ liệu khổng lồ, Sora có thể có được một số khả năng thú vị, cho phép mô hình có thể mô phỏng một số khía cạnh của con người, động vật và môi trường từ thế giới vật chất. Sora hoàn toàn không được train để làm những tác vụ này, đây hoàn toàn là thành quả tuyệt vời được sinh ra từ dữ liệu lớn.

Những khả năng này cho thấy rằng việc mở rộng kích thước của các mô hình video là một con đường đầy hứa hẹn, hướng tới sự phát triển của các mô hình có khả năng mô phỏng thế giới vật chất và tương tác giữa các vật thể, con người và động vật trong đó.

Tóm lại, SORA đã chứng minh mình là một bước tiến đáng kể trong lĩnh vực công nghệ, mở ra một tương lai đầy hứa hẹn với khả năng tạo sinh video từ văn bản. Mặc dù còn tồn tại một số hạn chế,

nhưng tiềm năng để phát triển thành một hệ thống mô phỏng thực tế ảo, nơi vật thể và con người có thể tương tác một cách tự nhiên, là điều không thể phủ nhận. SORA không chỉ mở ra cánh cửa cho những cải tiến công nghệ tiếp theo mà còn hứa hẹn sẽ mang lại những ứng dụng sáng tạo và cách mạng trong các ngành như metaverse, điện ảnh, quảng cáo và giáo dục, đánh dấu một bước ngoặt mới trong cách chúng ta tương tác và tạo ra nội dung số.

## Phần VI: Mô hình Genie

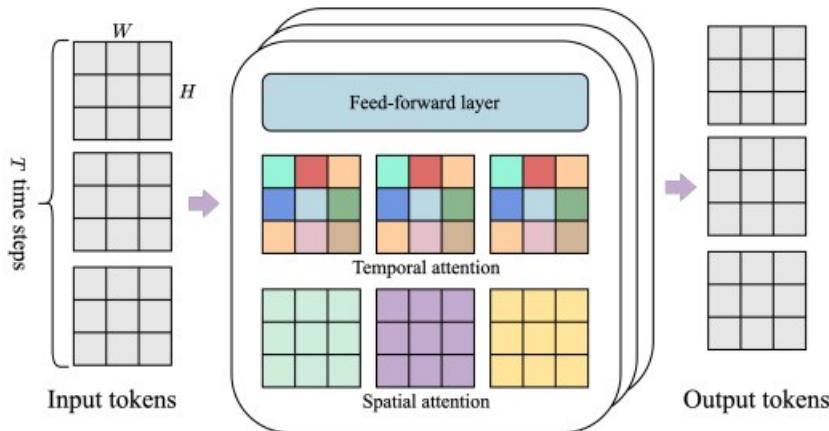


Genie - là một generative interactive environment, đánh dấu bước tiến lớn trong việc tạo dựng thế giới ảo thông qua việc học không giám sát từ kho dữ liệu video trên internet không được gắn nhãn. Điểm nổi bật của Genie là khả năng của nó trong việc phản hồi các yêu cầu để sinh ra các thế giới ảo đa dạng, từ văn bản mô tả, hình ảnh tổng hợp, ảnh chụp đến bản vẽ, mở ra không gian sáng tạo không giới hạn cho người dùng. Với 11 tỷ tham số, mô hình này không chỉ là một cơ sở dữ liệu thế giới mà còn là một công cụ mạnh mẽ, bao gồm spatiotemporal video encoder, autoregressive dynamics model và scalable latent action model.

Đặc biệt, Genie cho phép người dùng tương tác với các môi trường sinh ra, từng khung hình một, mà không cần đến nhãn hành động sự thật cơ bản hay các yêu cầu đặc thù của lĩnh vực thường thấy trong các nghiên cứu về mô hình thế giới. Điều này không chỉ làm giảm bớt gánh nặng về dữ liệu mà còn tạo điều kiện cho việc áp dụng rộng rãi. Hơn nữa, latent action space mà Genie học được mở ra khả năng đào tạo các đại lý để mô phỏng hành vi từ video chưa từng thấy, tiến một bước dài hướng tới mục tiêu phát triển các generalist agents cho tương lai.

Genie bao gồm ba thành phần chính: latent action model dự đoán hành động giữa các cặp khung hình, video tokenizer chuyển đổi khung hình thành token rời rạc, và dynamics model dự đoán khung hình tiếp theo dựa trên latent action và khung hình trước. Quá trình đào tạo mô hình diễn ra qua hai giai đoạn, bắt đầu với video tokenizer, sau đó đồng thời đào tạo latent action model và dynamics model, cho phép tạo ra dòng video liên tục và chân thực từ các pixel, mở ra khả năng tạo ra các trải nghiệm ảo đa dạng và phong phú. Sau đây, chúng ta hãy đi vào tìm hiểu từng thành phần chính của Genie.

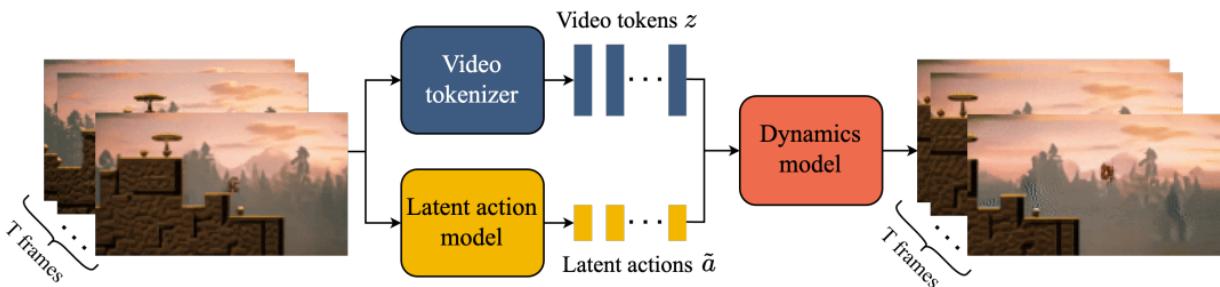
**ST-Transformer** Mô hình Genie tích hợp nhiều yếu tố từ kiến trúc Vision Transformer, nổi bật với đặc thù là độ phức tạp tính toán tăng theo cấp số nhân. Điều này đặt ra thách thức không nhỏ trong việc xử lý video, do đó, nghiên cứu này đã chọn giải pháp sử dụng khối ST-Transformer. Phương



Hình 23: Kiến trúc khối ST-Transformer

pháp này giúp cân đối hiệu quả giữa khả năng xử lý của mô hình và các giới hạn về khả năng tính toán. Kiến trúc của khối này được minh họa trong Hình 23.

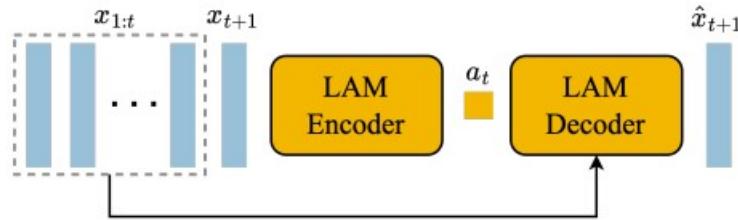
Khác biệt so với các khối transformer thông thường, nơi mỗi phần tử dữ liệu (token) quan tâm đến mọi phần tử khác, ST-transformer lại được thiết kế với các khối không gian-thời gian đa dạng. Các lớp chú ý không gian trong đó tập trung vào việc xem xét các phần tử dữ liệu trong cùng một bức ảnh, trong khi các lớp chú ý theo thời gian lại nhắm đến việc kết nối với dữ liệu từ các khung hình trước đó. Nhờ cách tiếp cận này, độ phức tạp tính toán của ST-transformer chỉ tăng theo chiều dài của video một cách tuyến tính thay vì theo cấp số nhân, làm cho việc xử lý video trở nên hiệu quả hơn nhiều.



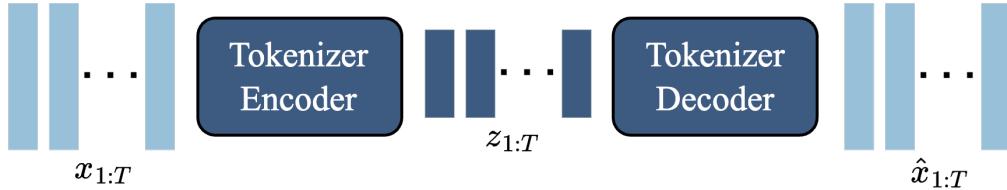
Hình 24: Các thành phần của mô hình Genie

**Latent Action Model (LAM)** Trong mô hình Genie, quá trình tạo ra mỗi khung hình mới phụ thuộc vào hành động của người dùng được ghi nhận từ các khung hình trước đó. Thách thức ở đây là thông tin về những hành động này thường khá hiếm và khó có được từ những video trên internet. Để giải quyết vấn đề này, mô hình sử dụng khối LAM, giúp máy học được các hành động ẩn mà không cần dữ liệu được gán nhãn cụ thể.

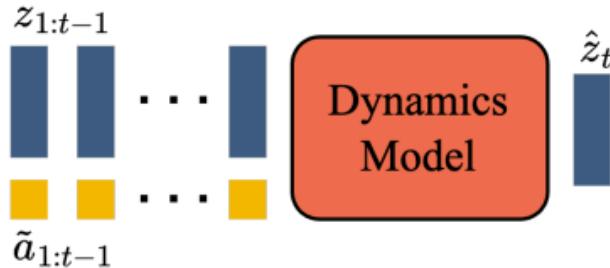
Khối LAM hoạt động bằng cách xem xét một loạt các khung hình trước và sau đó dự đoán chuỗi hành động ẩn có thể dẫn đến khung hình tiếp theo. Một bộ giải mã sau đó sử dụng thông tin này cùng với chuỗi khung hình để dự đoán khung hình kế tiếp. Quá trình này được huấn luyện dựa trên kỹ thuật VQ-VAE, giúp hạn chế số lượng hành động có thể xuất hiện và trong trường hợp này là 8 hành động. Do bộ giải mã chỉ nhìn vào các khung hình trước đó, mô hình cần phải nắm bắt được những biến đổi quan trọng nhất giữa các khung hình đã qua và khung hình tiếp theo để có thể dự đoán chính xác.



**Video Tokenizer** Khối này có nhiệm vụ nén các video thành những token rời rạc, với mục đích để giảm chiều và tăng chất lượng tạo sinh video. Một lần nữa kiến trúc VQ-VAE lại được sử dụng cho thành phần này. Khác với những công trình nghiên cứu khác chỉ tập trung vào việc nén trong chiều không gian, nghiên cứu này sử dụng khối ST-transformer trong cả encoder và decoder để có thể đảm bảo sự liên kết trong chiều thời gian.



**Khối Dynamics Model** Đây là một khối decoder transformer tương tự như trong nghiên cứu MaskGIT [1]. Ở mỗi bước, khối này nhận đầu vào là một chuỗi video đã được tokenize và thông tin về hành động tiềm ẩn được dự đoán bởi khối LAM. Mục tiêu huấn luyện của nó là dự đoán token tiếp theo trong chuỗi video. Trong quá trình train, một số token video sẽ được che đi ngẫu nhiên theo phân phối Bernoulli.



Genie được nhấn mạnh là một mô hình đột phá, mang đến khả năng tạo sinh và tương tác với các môi trường ảo một cách linh hoạt. Đặc biệt, dù chỉ dựa trên dữ liệu video, Genie vẫn có thể tạo ra các môi trường đa dạng và kiểm soát chúng một cách hiệu quả. Sự đa năng và tiềm năng của mô hình này mở ra khả năng ứng dụng rộng rãi trong các lĩnh vực như video game và mô phỏng thực tế ảo, hứa hẹn làm thay đổi cách chúng ta tạo sinh và tương tác với thế giới số.

## Phần VII: Tổng kết

Genie và Sora đều là những mô hình tạo sinh video tiên tiến, được huấn luyện trên tập data khổng lồ. Tiềm năng của 2 mô hình này không chỉ dừng lại trong việc tạo sinh video, mà còn là cách mạng trong

nhiều lĩnh vực khác nhau như điện ảnh, làm game, thực tế ảo, v.v. Hơn hết, cả 2 mô hình đều cho thấy việc huấn luyện các mô hình deep learning trên tập data lớn về video có thể giúp các mô hình trí tuệ nhân tạo học được cách vận hành của thế giới vật chất, mở ra hướng mới cho việc phát triển AI có khả năng tương tác với thế giới và tiềm cận trí tuệ con người.

## References

- [1] Huiwen Chang et al. “MaskGIT: Masked Generative Image Transformer”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022.
- [2] Jonathan Ho et al. “Video diffusion models”. In: *arXiv:2204.03458* (2022).
- [3] Zheng W. Liu X. Tang J. Hong W. Ding M. “CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers”. In: *arXiv preprint arXiv:2205.15868* (2022).
- [4] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. “Neural discrete representation learning”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS’17*. Long Beach, California, USA: Curran Associates Inc., 2017, 6309–6318. ISBN: 9781510860964.
- [5] William Peebles and Saining Xie. “Scalable Diffusion Models with Transformers”. In: *arXiv preprint arXiv:2212.09748* (2022).
- [6] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv: [2112.10752 \[cs.CV\]](https://arxiv.org/abs/2112.10752).
- [7] Thomas Hayes Xi Yin Jie An Songyang Zhang Qiyuan Hu Singer Adam Polyak. “Make-A-Video: Text-to-Video Generation without Text-Video Data”. In: *arXiv preprint arXiv:2209.14792* (2022).
- [8] Ruben Villegas et al. “Phenaki: Variable Length Video Generation From Open Domain Textual Description”. In: *ArXiv abs/2210.02399* (2022).
- [9] Jay Zhangjie Wu et al. “Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 7623–7633.
- [10] Wilson Yan et al. *VideoGPT: Video Generation using VQ-VAE and Transformers*. 2021. arXiv: [2104.10157 \[cs.CV\]](https://arxiv.org/abs/2104.10157).
- [11] Wang W. Yan H. Lv W. Zhu Y. Feng J. Zhou D. “MagicVideo: Efficient Video Generation With Latent Diffusion Models”. In: *arXiv preprint arXiv:2211.11018* (2022).

- *Hết* -