

Survey of Text to Video Models

Phan Đức Huy

Tháng 9 năm 2024

1 Giới thiệu

Trong khoảng thời gian gần đây, Text to Video đang là chủ đề vô cùng phổ biến nhờ khả năng sinh tạo vô cùng thực tế mà các model SOTA có thể đạt được, phục vụ cho rất nhiều lĩnh vực khác nhau.

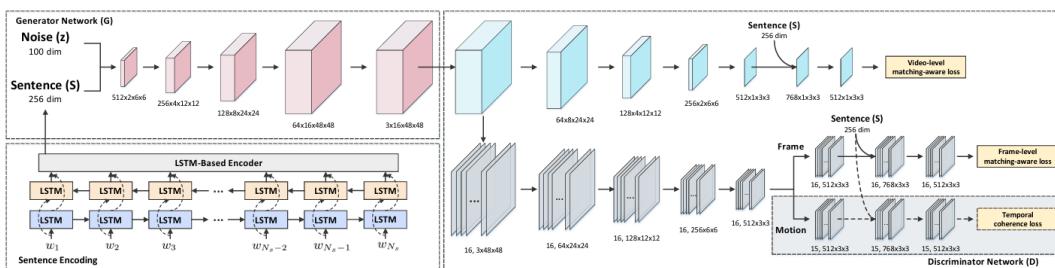
Khởi nguồn từ những năm 2017-2018, làn sóng đầu tiên của các model Text to Video được tạo nên dựa trên cấu trúc GAN (Generative Adversarial Network). Cho đến năm 2021, dựa trên hiệu quả mà cấu trúc Transformers đạt được trong tác vụ sinh tạo hình ảnh hay văn bản, các model Text to Video cũng áp dụng cấu trúc này và đạt được kết quả vượt trội. Đến hiện nay, giống với Text to Image, Diffusion Models đang áp đảo thị trường Text to Video với khả năng sinh tạo vô cùng chân thực, có thể nắm bắt được những chuyển động phức tạp giữa các frame trong một video .

Bài nghiên cứu này sẽ đánh giá các model tiêu biểu với cấu trúc của từng thời kì và chỉ ra khả năng cũng như hạn chế của chúng. Ở phần cuối, một app ứng dụng sẽ được deploy trên Hugging Face Spaces sử dụng các model open-source để xem kết quả của chúng với prompt đưa ra.

2 TGANs-C

2.1 Kiến trúc

Kiến trúc của model TGANs-C được thiết kế với mục đích tạo ra sự liên kết giữa chuyển động trong các frame của video với văn bản được đưa ra. Model được chia ra thành hai mạng Generator và Discriminator, trong quá trình train hai mạng này sẽ được tối ưu. Mạng Generator sẽ nhận văn bản được mã hoá và tạo ra video còn mạng Discriminator sẽ phân biệt xem video được tạo ra là thật hay giả.



Hình 1: Toàn bộ cấu trúc model TGANs-C

2.1.1 Mạng Generator

Văn bản đầu vào sẽ được xử lý thành biểu diễn từ (S) sau khi đi qua mạng LSTM hai chiều và một LTSM-based encoder đã được học sẵn. Mạng Generator sau đó sử dụng biểu diễn từ trên cùng với các biến nhiễu ngẫu nhiên (z) để sinh tạo video hay cụ thể là một chuỗi các frame hình ảnh.

Trước đây, các lớp tích chập 3D đã được sử dụng cho vài tác vụ xử lý video ví dụ như Action Recognition cho thấy chúng có thể nắm bắt được thông tin không-thời gian (spatio-temporal), hay có thể gọi là "hiểu" được video. Dựa vào đó, các tầng tích chập chuyển vị 3D được sử dụng để biến giá trị ở không gian ngầm (kết hợp từ S và z) thành một video. Weight ở các lớp tích chập chuyển vị sẽ được update theo feedback từ mạng Discriminator.

2.1.2 Mạng Discriminator

Để có thể tối ưu hóa khả năng sinh tạo của mạng Generator, mạng Discriminator được thiết kế với 3 tiêu chí đánh giá:

- Qua cả video: Video được tạo ra sẽ đi qua các lớp tích chập 3D và gắn với biểu diễn từ tương ứng để bộ phân biệt đánh giá xem video có phải thật và có tương ứng với văn bản đưa ra hay không.
- Qua từng frame: Các frame nằm trong một video sẽ được đưa qua các lớp tích chập 2D, yêu tố thêm biểu diễn văn bản tương ứng để phân biệt frame thật với tiêu đề chính xác.
- Qua chuyển động giữa các frame: Mạng đọc chuyển động giữa các frame của video bằng cách trừ giá trị của frame sau cho frame trước. Giá trị này được sử dụng để đánh giá xem chuyển động có mượt và mạch lạc không.

Hàm mất mát được tính qua 3 cặp video-text bao gồm cặp giữa video thật với tiêu đề tương xứng $\{v_{real^+}, S\}$, video giả với tiêu đề tương xứng $\{v_{syn^+}, S\}$ và cuối cùng là video thật với tiêu đề không tương xứng $\{v_{real^-}, S\}$. Từng frame trong video cũng được ghép thành các cặp frame-text tương tự và còn được sử dụng để tính khoảng cách Euclid giữa 2 frame liền kề, khoảng cách được tính cũng sẽ được ghép cặp với tiêu đề. Công việc của mạng Discriminator là phân biệt giữa video thật hay giả và đảm bảo video được tạo sinh có nội dung, chuyển động tương xứng với tiêu đề được đưa ra.

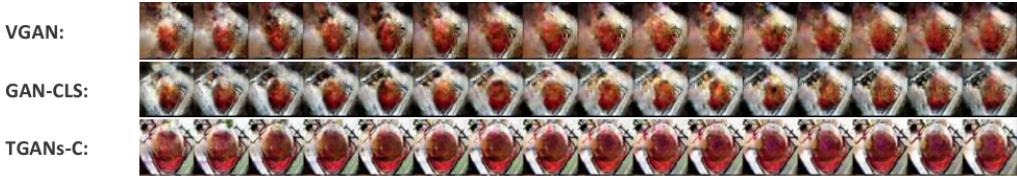
2.2 Kết quả sinh tạo

Model được huấn luyện trên các tập dữ liệu với độ phức tạp tăng dần bao gồm Single-Digit Bouncing MNIST GIFs (SBMG), Two-digit Bouncing MNIST GIF (TBMG) và tập Microsoft Research Video Description Corpus (MSVD) bao gồm các video nấu ăn. Dữ liệu huấn luyện dài 16 frame với độ phân giải 48x48.



Hình 2: "Số 1 sang trái rồi phải và số 9 đi lên rồi xuống"

Như Hình 2, model có thể tạo ra các vật thể đơn giản như các chữ số một cách hoàn chỉnh. Các chuyển động lên xuống hay trái phải cũng được thể hiện rõ ràng, lành mạch. Đối với video phức tạp hơn như



Hình 3: "Đầu bếp đảo nồi súp"

video nấu ăn, các frame được sinh ra từ các model GAN trước đều vô cùng mờ, khó có thể quan sát được bất cứ thứ gì ngoài màu sắc của súp. Với TGANs-C, hình dạng của nồi súp có thể được xác định cùng với phần bàn tay mờ của đầu bếp.

2.3 Thủ thách và hạn chế

Chất lượng: Như đã thấy ở trên, dù có cải thiện rõ rệt so với các model trước, vật thể tay và chuyển động của tay gần như không thể xác định rõ. Điều này cho thấy model vẫn gặp hạn chế trong việc sinh tạo các vật thể phức tạp và nắm rõ chuyển động của chúng với chất lượng cao.

Hiệu năng: Điểm hạn chế tiếp theo nằm ở lượng tài nguyên yêu cầu lớn của model. Kèm theo với mạng LSTM để xử lý văn bản, model sử dụng 3 hàm loss trên nhiều giá trị khác nhau dẫn đến tài nguyên huấn luyện vô cùng lớn.

Độ đa dạng: Cấu trúc GAN còn mang một vấn đề nghiêm trọng chính là mode collapse khi mạng Generator tìm ra được điểm mù của mạng Discriminator và chỉ sinh ra các kết quả giống nhau dựa trên điểm mù đó, làm mất đi sự đa dạng của kết quả sinh tạo.

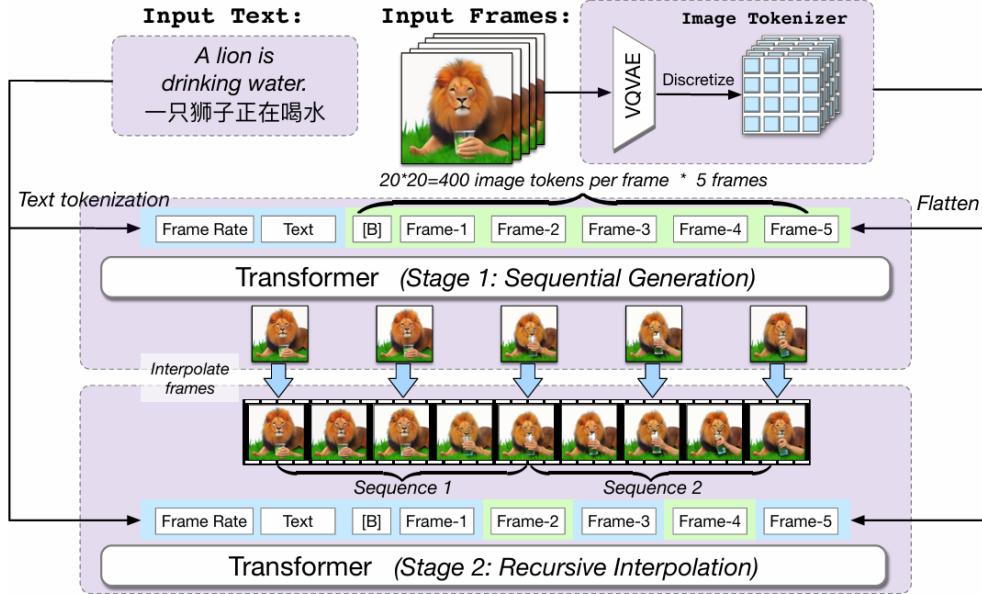
Kết luận: Mặc dù model thành công trong việc tạo ra được video từ text nhập vào, kết quả sinh tạo vẫn còn rất nhiều hạn chế với độ phân giải thấp, thời lượng ngắn và chưa thể nắm bắt được các chuyển động dài và phức tạp. Không chỉ vậy, quá trình huấn luyện tiêu tốn lượng lớn thời gian, tài nguyên và yêu cầu phải tinh chỉnh hợp lý các hyperparameters để phòng tránh các vấn đề như mode collapse, non-convergence. Các hạn chế này khiến cho model không thể đưa ra được video với chất lượng như mong đợi.

3 CogVideo

3.1 Kiến trúc

CogVideo là một model autoregressive tạo ra video bằng cách dự đoán các frame nối tiếp nhau dựa vào các token văn bản trước đó. Nhà phát triển đã tìm ra cách nâng cao chất lượng, hiệu quả của model bằng cách tận dụng các model được train sẵn, cụ thể ở đây là model Text to Image CogView2 do chính họ thiết kế.

Ở giai đoạn một của quá trình sinh tạo, với frame rate được xác định trước và văn bản đầu vào, model có thể dự đoán các image token trong các frame kế tiếp và ghép lại thành một frame hoàn chỉnh. Ở stage 2, model sẽ sử dụng phép nội suy để ước lượng giá trị của frame ở giữa từ 2 frame liền kề được sinh tạo



Hình 4: Cấu trúc của CogVideo

ở giai đoạn 1.

Kiến trúc của CogVideo trong bài báo thiếu nhiều phần giải thích nên cần nghiên cứu thêm source code để nắm bắt rõ hơn.

3.2 Kết quả sinh tạo

Cogvideo có thể sinh ra các video dài 4 giây với 8 fps (frame per second) ở độ phân giải 480x480. Khi nhập vào văn bản, chúng sẽ được mã hoá và sử dụng để tạo ra frame đầu tiên của video nhờ model Text to Image đã được train sẵn. Model ở giai đoạn một sau đó sẽ sinh ra các frame tuân tự tuỳ theo frame rate được chọn. Sau khi đi qua cả 2 giai đoạn, video được tạo ra có thể đạt được chất lượng về cả vật thể lẫn chuyển động.

Bài công bố CogVideo cũng được đánh giá cao trên mạng xã hội nhờ kết quả sinh tạo đa dạng trên nhiều loại prompt khác nhau. Các model Text to Video trước đó vẫn gặp khó khăn trong việc xử lý độ phân giải lớn, phần lớn chỉ đạt được khoảng 128x128 nên đạt được thành tựu 480x480 với 8fps là thành quả lớn.

3.3 Thủ thách và hạn chế

Chất lượng: Do các frame được chia thành các token hình ảnh riêng biệt, vật thể được sinh tạo ở các frame có một số đặc điểm khác nhau ví dụ như kết cấu các bộ phận trên mặt người hay màu da. Sự không đồng nhất này gây ảnh hưởng lớn đến độ chân thực của video.

Hiệu năng: Điểm hạn chế tiếp theo mà CogVideo gặp phải nằm ở thời gian sinh tạo. Ngoài việc các frame phía sau vẫn cần phải đợi kết quả từ các frame phía trước, giai đoạn 2 của CogVideo cũng phải đợi kết quả sinh tạo từ giai đoạn 1, dẫn đến thời gian sản xuất 1 video có thể lên đến hơn 10 phút. CogVideo

Hình 5: A lion typing on a computer (GIF chạy trên Adobe Player và Okular)

cũng vô cùng lớn với hơn 9 triệu tham số do sử dụng 2 model khác nhau ở các giai đoạn cùng với 1 model Text to Image.

Giới hạn về ngôn ngữ: Các nhà phát triển từ Trung Quốc cũng không hỗ trợ prompt bằng tiếng Anh, những prompt đơn giản như " Subject + Verb + Object " vẫn đưa ra kết quả đúng khi được dịch sang tiếng Trung nhưng các prompt đặc thù, chi tiết hơn có thể gây khó khăn cho người sử dụng.

Kết luận: CogVideo thành công trong việc vận dụng các model được train sẵn để áp dụng cho lĩnh vực Text to Video với kết quả vượt trội hơn hẳn so với các sản phẩm lúc bấy giờ, cho thấy sức mạnh của Transfomers đối với các model AI đa phương thức. Tuy vậy, các nhược điểm của model autoregressive như thời gian sinh tạo chậm, thiếu đồng nhất giữa các frame khiến chúng chưa thể được đón nhận rõ ràng bởi đại chúng.

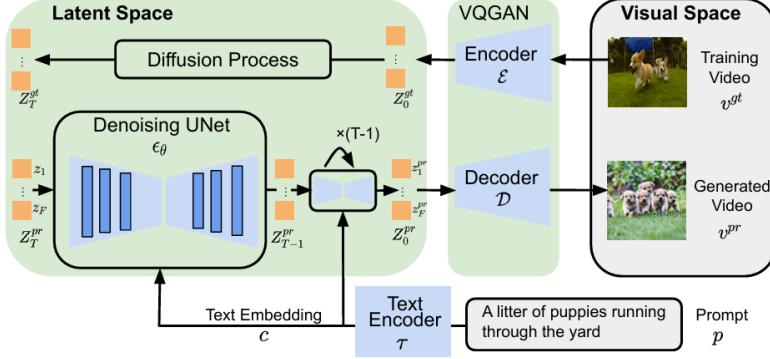
4 ModelScopeT2V

4.1 Kiến trúc

Kiến trúc của ModelScopeT2V được thiết kế dựa trên Stable Diffusion với một số điều chỉnh để có thể thực hiện tác vụ sinh tạo video. ModelScopeT2V có 3 thành phần chính bao gồm VQGAN, U-NET và Text Encoder như hình dưới đây.

Như hình trên, các video huấn luyện sẽ được mã hoá thành biểu diễn ở không gian ẩn để thực hiện quá trình diffusion và denoise sau đó được giải mã thành video mà mắt người có thể hiểu được thông qua VQGAN. Dữ liệu đầu vào $v^{gt} = [f_1, \dots, f_F]$ với F frame sau khi đi qua Encoder \mathcal{E} sẽ cho $Z_0^{gt} = [\mathcal{E}(f_1), \dots, \mathcal{E}(f_F)]$ trong đó $v^{gt} \in \mathbb{R}^{F \times H \times W \times 3}$ là một video RGB, $Z_0^{gt} \in \mathbb{R}^{F \times \frac{H}{8} \times \frac{W}{8} \times 4}$ là biến ở không gian ẩn. Đây là bước quan trọng để giảm tài nguyên và chi phí tính toán của model mà vẫn đảm bảo chất lượng độ phân giải của video đầu ra.

Để có thể đảm bảo độ tương xứng giữa nội dung trực quan và nội dung văn bản, ModelScopeT2V



Hình 6: Cấu trúc của ModelScopeT2V

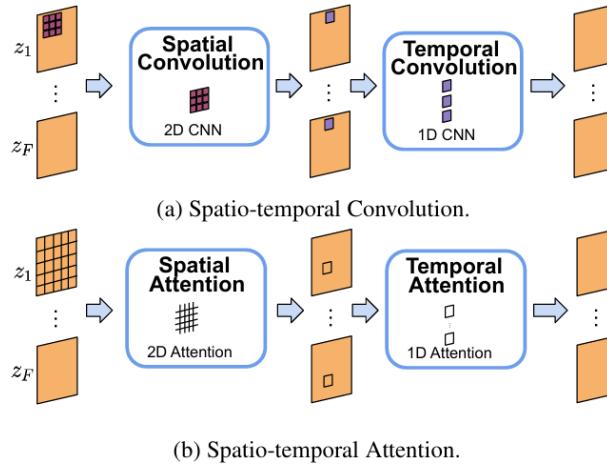
sử dụng text encoder CLIP ViT-H/14 được train sẵn để mã hoá prompt văn bản thành Text Embedding c . Text Embedding c sẽ được đưa vào module Cross-Attention nằm trong mạng U-Net, giúp các textual features có thể ảnh hưởng tới visual features.

Phần quan trọng nhất của model chính là mạng Denoising U-Net ϵ_θ . Mục tiêu của U-Net là dự đoán noise được thêm vào ở mỗi timestep để khôi phục frame ban đầu. Với mỗi timestep $\hat{t} \in [1, 2, \dots, T]$, noise dự đoán được tính với công thức:

$$\epsilon_{\hat{t}}^{pr} = \epsilon_\theta(Z_{\hat{t}}, c, \hat{t})$$

trong đó c là text embedding, $Z_{\hat{t}}$ là biến ở không gian ẩn tại timestep thứ \hat{t} . Nhiệm vụ huấn luyện của model là minimize sự khác biệt giữa $\epsilon_{\hat{t}}^{pr}$ và $\epsilon_{\hat{t}}^{gt}$.

Để có thể nắm bắt được các spatial features và temporal features của video, ModelScopeT2V giới thiệu khối spatio-temporal bao gồm 4 thành phần:



Hình 7: Khối spatio-temporal sử dụng trong U-Net

Ở 5(a), ta có thể thấy cả 2 mạng tích chấp spatial và temporal. Mạng spatial CNN sẽ trích xuất đặc điểm của từng frame $\frac{H}{8} \times \frac{W}{8}$ qua một kernel 3×3 . Trong khi đó, mạng temporal CNN sẽ dùng một kernel 1D với size là 3 để trích xuất feature từ F frame, với F là số frame của một video.

Ở hình 5(b), lớp spatial attention sẽ làm việc với các feature ở spatial dimension với size $\frac{HW}{64}$ còn

temporal attention thi hành ở temporal dimension với size F . Các lớp spatial attention được chia làm 2 loại. Một bên là module cross-attention nhằm đổi xứng textual feature từ text embedding c với visual feature, phần còn lại là module self-attention trên mỗi visual feature.

4.2 Kết quả sinh tạo

ModelScopeT2V có thể sinh ra vật thể, cảnh vật với chuyển động thực tế. Video sinh tạo còn có thể chuyển sang nhiều loại phong cách khác nhau như hoạt hình, truyện tranh, viễn tây ... Số lượng token tối đa cho prompt là 77 token và CLIP cho phép model hiểu được các prompt đa dạng, sáng tạo.

Hình 8: A dog drinks beer (GIF chạy trên Adobe Player và Okular)

Thời gian sinh tạo có cải tiến rõ rệt so với các model ở phần trước nhờ kiến trúc diffusion. Không chỉ vậy, ta hoàn toàn có thể điều khiển số bước của mạng. Số bước lớn giúp tăng chất lượng của video nhưng làm chậm đi thời gian sinh tạo .

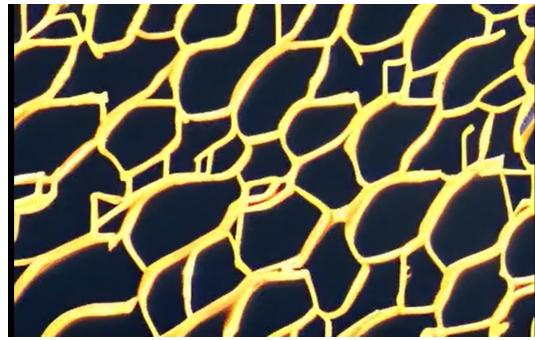
Chuyển động của sinh vật sống khi nhiều bộ phận chuyển động cùng một thời điểm không được lưu loát, hay bị biến dạng giữa các frame. Đối với các vật thể liền mạch như các phương tiện giao thông như xe hơi, máy bay thì hiện tượng này không xảy ra.

4.3 Thủ thách và hạn chế

Chất lượng: Video sinh tạo có thể đạt được độ phân giải cao với số lượng frame lớn nhưng yêu cầu phần cứng rất lớn. Các chuyển động phức tạp với nhiều động lệnh cũng không được tự nhiên như thực tế.

Hiệu năng: ModelScopeT2V tiêu tốn rất nhiều VRAM. Vấn đề này vẫn khó có thể giải quyết do kích cỡ của dữ liệu video là vô cùng lớn. Cộng thêm việc model diffusion không đưa ra kết quả ngay lập tức mà lặp lại các bước denoise nhiều lần, tiêu tốn lượng lớn VRAM trong mỗi vòng lặp.

Thời lượng video: Hầu hết dữ liệu huấn luyện của model là các clip ngắn kèm theo tiêu đề, điều này dẫn đến chất lượng video sụt giảm rõ rệt khi gia tăng số frame của video. Với F càng lớn, lớp temporal attention được sử dụng trong U-Net càng gặp khó khăn trong việc nắm bắt các feature ở temporal



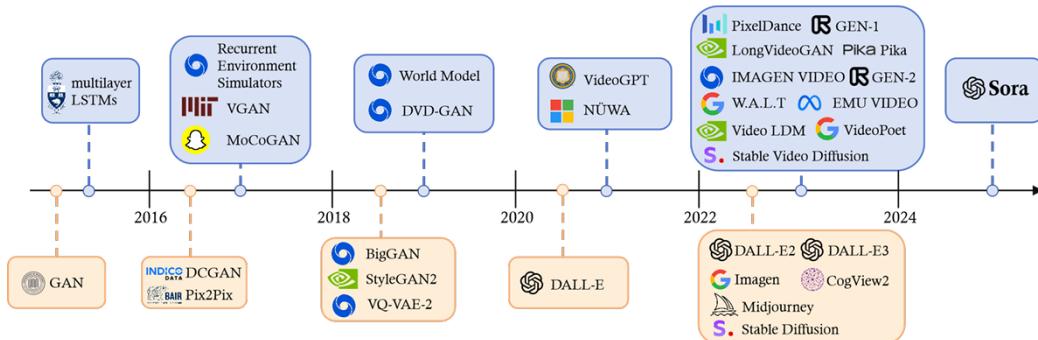
Hình 9: Prompt "A dog drinks beer" khi yêu cầu sinh tạo 64 frame

dimension. Hơn nữa, kết quả sinh tạo có thể gặp lỗi nghiêm trọng khi F nhập vào quá lớn.

Kết luận: ModelScopeT2V cho thấy sức mạnh của diffusion trong tác vụ Text to Video với vài tuỳ chỉnh từ kiến trúc từ Stable Diffusion gốc. Dù là một model open source nhưng model này có thể đạt được kết quả tương đương với các SOTA vào thời điểm giữa năm 2023. Tuy vậy, ModelScopeT2V vẫn gặp khó khăn trong việc sinh tạo video dài và đối xứng một số hành động phức tạp với văn bản.

5 Xu hướng gần đây

Chỉ trong năm 2024, số lượng model Text to Video được công bố là vô cùng nhiều. Hàng loạt các dịch vụ cung cấp model Text to Video trả phí như Fliki, Runway AI Gen-2, Pika. Như hình dưới, các công ty công nghệ đang chạy đua xu thế để tạo ra model Text to Video của riêng mình. Video sinh tạo chất lượng sẽ đem lại rất nhiều giá trị cho các công ty khi có thể được sử dụng để sản xuất quảng cáo, sáng tạo nội dung giải trí,...



Hình 10: Lịch sử các các model sinh tạo hình ảnh/video

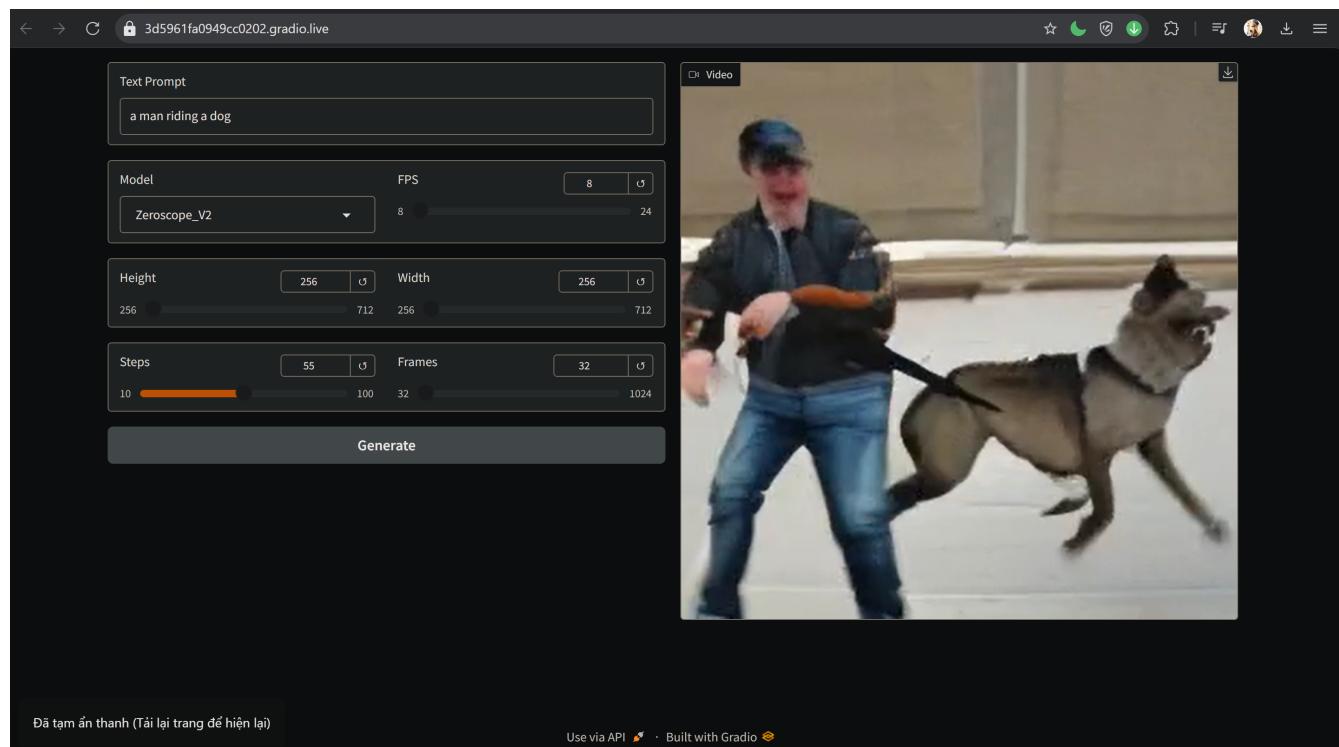
Dù chưa được công bố chính thức, vào tháng 2 năm 2024 OpenAI đã đưa ra những cái nhìn đầu tiên về Sora, model được cộng đồng AI dự đoán sẽ là đột phá mới cho lĩnh vực Text to Video. Qua những sản phẩm giới thiệu mà OpenAI cung cấp, video sinh tạo bởi Sora có thể đạt được độ phân giải 1920x1080 và

thời lượng có thể lên tới hơn 1 phút. Độ chân thực của các vật thể và chuyển động của chúng cũng hoàn toàn vượt trội so với các model tiền nhiệm.

Text to Video là một tác vụ đã được nghiên cứu từ những ngày đầu của GenAI. Từ TGANs-C, ta có thể thấy được cách thức mà các nhà nghiên cứu sử dụng để đổi xứng những chuyển động trong temporal dimension với dữ liệu văn bản. Sự đột phá của cơ chế attention trong các năm tới đã mở ra một chương mới trong tác vụ sinh tạo từ văn bản khi cho phép AI dự đoán được token hình ảnh từ token văn bản. Và trong 2 năm trở lại đây, sự phát triển của Text to Video nhờ diffusion đã vượt xa so với tưởng tượng của cộng đồng AI nói riêng và toàn thế giới nói chung. Có thể sau khi Sora được công bố, tiềm năng của các model Text to Video còn có thể vượt xa hơn hẳn so với hiện tại.

6 App ứng dụng

Gradio là thư viện được sử dụng phổ biến để tạo ra web UI cho các model AI. Phần Input bao gồm Textbox để nhập prompt văn bản, Dropdown để lựa chọn model sử dụng và các slider để điều chỉnh các thông số như số frame, fps, chiều dài chiều rộng video hay số bước denoise. Sau khi bấm nút Generate các biến sẽ được đưa vào model được load từ thư viện Diffuser để sinh ra video ở khung output bên phải.



Hình 11: Interface của app