THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút): https://youtu.be/_SwtsKSPrZ4
- Link slides (dạng .pdf đặt trên Github): https://github.com/DragonPow/CS2205.CH181/blob/main/reserved_from_video_to_prompt.Slide.pdf
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

• Họ và Tên: Vũ Ngọc Thạch • MSSV: 230201054 	• Lớp: CS2205.CH181 • Tự đánh giá (điểm tổng kết môn): 8/10 • Số buổi vắng: 0 • Số câu hỏi QT cá nhân: 0 • Link Github: https://github.com/DragonPow/CS2205.CH181

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

TỪ VIDEO SANG PROMPT, KỸ THUẬT CHUYỂN ĐỔI NGƯỢC DÀNH CHO VIỆC XÂY DỰNG CONTEXT TRONG STABLE DIFFUSION

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

FROM VIDEO TO PROMPT, A RESERVED TECHNICAL TO BUILD CONTEXT FOR STABLE DIFFUSION

TÓM TẮT *(Tối đa 400 từ)*

Hiện nay, bằng sự phát triển mạnh mẽ của các công cụ trí tuệ nhân tạo, các công ty về công nghệ hàng đầu đã cho ra mắt hàng loạt các tính năng nổi trội. Gần nhất có các mô hình cho phép nhập vào nội dung (prompt) từ đó sinh ra được những video theo như mong muốn của người dùng, một vài mô hình nổi bật nhất hiện nay trong lĩnh vực này là Sora (thuộc OpenAI) và Stable Video Diffusion. Sức mạnh này của AI đã mở ra rất nhiều tiềm năng mới cho người dùng trong việc tăng trải nghiệm, giảm thiểu chi phí và thời gian tự thiết kế, vẽ lại để có thể có được video như vậy.

Tuy nhiên, đa phần chưa có nhiều bài báo kỹ thuật hỗ trợ viết prompt trong AI, người dùng muốn bắt đầu phải tự học cách viết thế nào cho tốt, thế nên để có thể tạo ra được video như ý mình mong muốn, họ phải tiến hành thực hiện lặp việc viết prompt và xuất kết quả rất nhiều lần cho đến khi đạt được đúng kết quả. Việc lặp lại này sẽ tốn khá nhiều chi phí để có thể xây dựng ra được đúng video tốt.

Chính vì vậy, điều này khiến cho việc mở rộng các hướng nghiên cứu để xem xét việc đảo ngược quá trình cũng được đẩy mạnh theo, nhằm hạn chế, giảm bớt được số lần thử trong 1 mô hình. Ở đây tôi đề xuất một nhiệm vụ mới là dự đoán prompt từ video được tạo bởi mô hình diffusion. Với tập dữ liệu danh sách prompt và video trong bộ dataset VidProM được cung cấp bởi Stable Fiffussion, tôi sẽ đưa ra các phương pháp mang lại hiệu quả tốt nhất cho kỹ thuật chuyển đổi ngược này.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Gần đây, có rất nhiều mô hình AI cho phép chuyển đổi từ văn bản (prompt) sang video được ra đời, trong đó có thể kể tới mô hình Stable Video Diffusion [1]. Các tác vụ sản xuất video giờ đây có thể sử dụng AI để hỗ trợ người dùng thuận tiện hơn. Đối với tập người dùng mới, khi sử dụng mô hình, cần phải học thêm các kiến thức về viết prompt để có được video tốt. Việc này khiến người dùng phải thử tạo rất nhiều lần mới được kết quả như mong muốn, chi phí cho mỗi lần tạo video là rất tốn kém và cần được hạn chế.

Để giảm thiểu số lần thử này, tôi mong muốn cung cấp được 1 mô hình có thể dự đoán được prompt sinh ra từ 1 video với kết quả gần đúng nhất. Nhờ đó người dùng có thể tái sử dụng prompt mà không cần phải chỉnh sửa, thay đổi nhiều.

Sau khi khảo sát tôi nhận thấy rằng vẫn chưa có nhiều nhóm nghiên cứu nào thực hiện tìm hiểu về đề tài này, trong số ít đó có bài báo của Croitoru [1] có đề cập tới việc chuyển đổi ngược từ hình ảnh sang promp, và đạt kết quả khá tốt. Ở phần đề tài này tôi cũng sẽ thực hiện một số phương pháp dựa trên bài báo trên, sử dụng vào việc chuyển đổi ngược từ prompt sang video, nhằm xem hiệu quả, đồng thời bổ sung thêm 1 số bước để cải tiến cho phù hợp với tập dữ liệu video.

Bộ dữ liệu tôi sử dụng chính trong đề tài này là bộ VidProM [2] được phát hành ngày 10/3/2024 bao gồm 6.69 triệu video và thông tin prompt dùng để tạo ra video.

Input: Video được tạo bởi Stable Diffusion

(hình ảnh 1 vài khung hình trong một video tạo bởi Stable Diffusion)

Output: Prompt được dự đoán dùng để sinh ra video trên

Prompt = *“ A horse galloping through van Gogh’s ‘Starry Night’ ”*

MỤC TIÊU

(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)

- Nghiên cứu về các kỹ thuật học máy hiện tại tốt nhất có thể đọc và hiểu được thông tin từ video, đồng thời kết hợp với các mô hình ngôn ngữ để đề xuất được prompt trong phạm vi mô hình Stable Diffusion.
- Xây dựng được mô hình cho phép dự đoán prompt từ các video của Stable Diffusion.
- Đánh giá độ hiệu quả của prompt được dự đoán và so sánh kết quả của video được sinh ra từ prompt dự đoán so với video mẫu.
- Phân tích được một số hạn chế, giới hạn trong video ảnh hưởng tới việc dự đoán prompt.

NỘI DUNG VÀ PHƯƠNG PHÁP

(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)

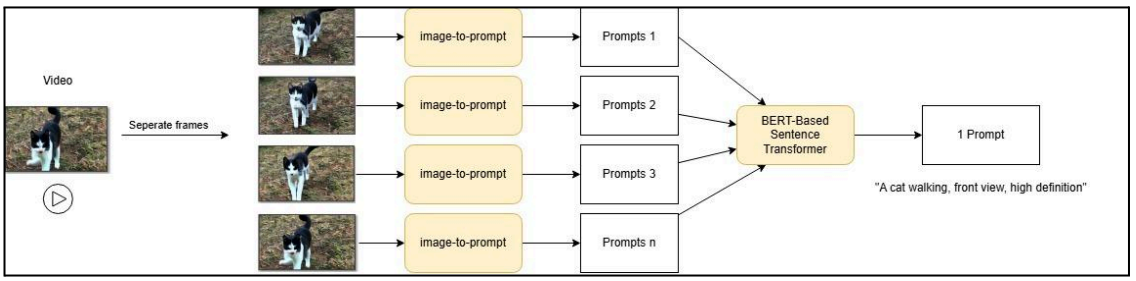
1. Chuẩn bị tập dữ liệu:

- Thu tập tập dữ liệu video VidProM. Phân tích đánh giá các video có trong tập dữ liệu.
- Thu thập thêm 1 tập dữ liệu video từ nguồn khác để kiểm tra hiệu năng của prompt được dự đoán có tốt hay không.

2. Thiết kế thuật toán:

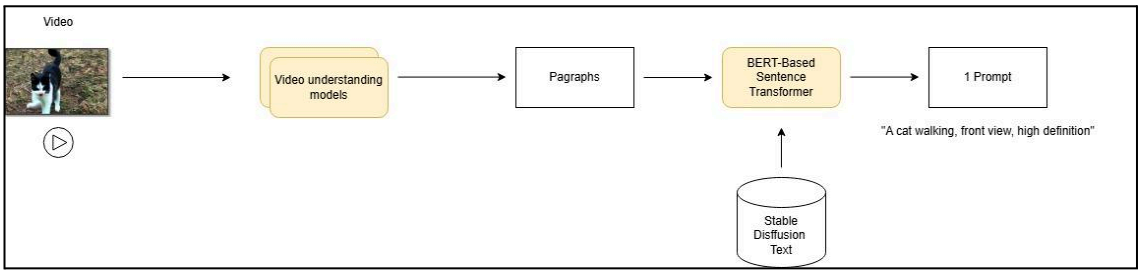
Nguyên cứu và tiếp cận bài toán theo 2 hướng:

a. Cắt lát video và sử dụng thuật toán image-to-prompt [1] để trích xuất prompt theo từng frame hình, tổng hợp lại và đề xuất prompt chung cho toàn bộ mô hình.



b. Sử dụng các kỹ thuật trích xuất thông tin video hiện có, dựa trên kỹ thuật Video

Swin Transformer [3], phân tích và đưa ra bộ text mô tả về video, sau đó dùng các kỹ thuật tạo prompt hiện có để chuyển đổi prompt.



3. Ghi nhận kết quả và thống kê:

Đối với từng hướng tiếp cận ở trên cần đánh giá rõ ràng các tiêu chí:

- So sánh prompt được dự đoán và prompt thực tế: có thể thông qua các tiêu chí BLEU, ROUGE, Word2Vec, BERT embeddings,...
- Sử dụng mô hình CLIP để đánh giá sự tương đồng giữa video được sinh ra từ prompt dự đoán và video gốc. Mục tiêu ở đây để xem thích ứng của mô hình đối với tập video không được sinh ra từ Stable Diffusion là bao nhiêu.
- Liên tục thực hiện việc thống kê và thay đổi, tuỳ chỉnh lại các tham số để đạt được một kết quả tốt nhất.
- Phân tích sự các yếu tố của video ảnh hưởng tới việc dự đoán prompt.

4. Xây dựng ứng dụng từ mô hình đạt được:

- Từ mô hình đạt được ở trên, xây dựng một ứng dụng/trang web cho phép người dùng upload video lên và nhận kết quả prompt trả về, có mục đánh giá xem liệu prompt có đạt đúng yêu cầu mà người dùng mong muốn hay không.

KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

- Xây dựng được một mô hình cho phép chuyển đổi từ video sang prompt với hiệu suất ổn định, ít nhất trên 60%, thời gian dự đoán ngắn dưới 1 phút.
- Xây dựng được 1 ứng dụng sử dụng mô hình được đào tạo ở trên, nhằm đánh giá chất lượng của mô hình trong thực tế với tập người dùng thật. Bảo đảm rằng prompt được sinh ra có thể xây dựng được video gần giống nhất so với người dùng mong muốn, không cần chỉnh sửa quá nhiều về mặt hình thức.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, Tim Salimans:
Imagen Video: High Definition Video Generation with Diffusion Models. CoRR abs/2210.02303 (2022)

[2]. Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Mubarak Shah: Reverse Stable Diffusion: What prompt was used to generate this image? CoRR abs/2308.01472 (2023)

[3]. Wenhao Wang, Yi Yang:
VidProM: A Million-scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models. CoRR abs/2403.06098 (2024)

[4]. Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, Han Hu: Video Swin Transformer. CoRR abs/2106.13230 (2021)