THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
  Presentation
- Link slides (dạng .pdf đặt trên Github của nhóm):
  Slides
- *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*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Phạm Thị Bích Nga
- MSSV: 240101018

- Lớp: CS2205.CH183
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 6
- Link Github: Link

ĐỀ CƯƠNG NGHIÊN CỨU

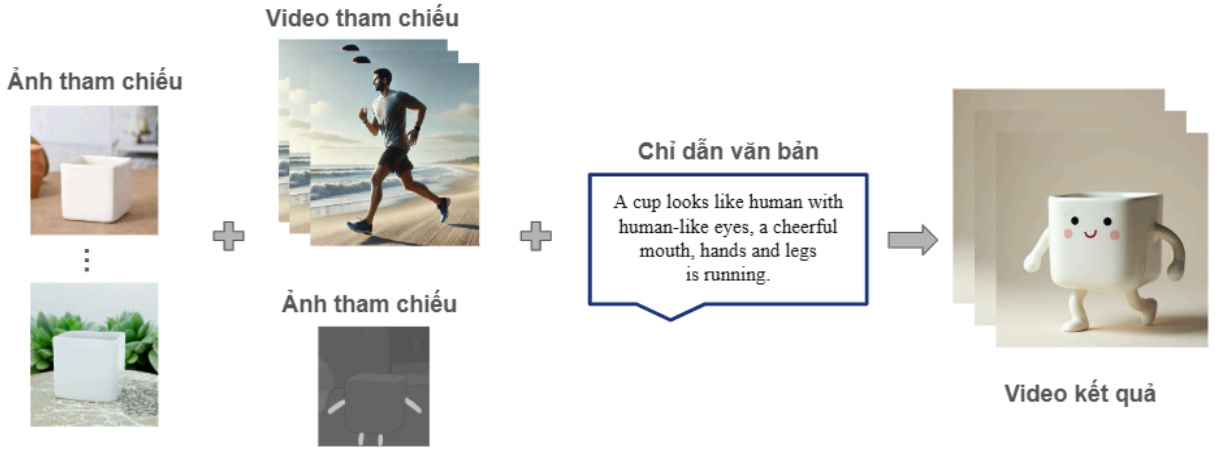
TÊN ĐỀ TÀI (IN HOA)

TẠO SINH VIDEO VẬT THỂ THỰC HIỆN HÀNH ĐỘNG CỦA CON NGƯỜI DỰA TRÊN CHỈ DẪN VĂN BẢN VỚI ĐA RÀNG BUỘC

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

PROMPT-BASED VIDEO GENERATION OF OBJECTS PERFORMING HUMAN ACTIONS WITH MULTIPLE CONSTRAINTS

TÓM TẮT *(Tối đa 400 từ)*



Hình 1. Đầu vào và đầu ra của bài toán sinh dữ liệu video.

Sự phát triển mạnh mẽ của các mô hình sinh ảnh và video trong những năm gần đây đã mở ra nhiều ứng dụng trong phim ảnh, quảng cáo và hoạt hình. Một hướng nghiên cứu tiềm năng trong lĩnh vực này là **tạo sinh video nhân hóa vật thể**, trong đó các vật thể không phải con người được mô phỏng với đặc trưng và hành động giống con người (ví dụ: ly nước "chạy bộ", con chim "đạp xe"...).

Bài toán tạo sinh video nhân hóa vật thể gặp nhiều thách thức do các mô hình sinh ảnh và video hiện tại chưa thể diễn giải tốt các chỉ dẫn mang tính sáng tạo, như mô phỏng hành vi con người trên vật thể không phải người. Hai hướng tiếp cận chính gồm: (1) Fine-tuning mô hình sinh ảnh để tạo video [1-2] và (2) Sử dụng mô hình sinh video trực tiếp từ văn bản [3], tuy nhiên cả hai đều gặp hạn chế trong việc mô tả

hình dáng và chuyển động nhân hóa do vấn đề của bộ mã hóa văn bản không hiểu được chỉ dẫn của người dùng.

Nghiên cứu đề xuất phương pháp **kết hợp đa ràng buộc**, trong đó tích hợp thông tin ngữ nghĩa như **mặt nạ phân đoạn ngữ nghĩa (semantic segmentation mask)** để giúp mô hình hiểu và định vị các bộ phận như tay, chân, từ đó cải thiện chất lượng mô phỏng hình dáng và chuyển động của vật thể nhân hóa trong video.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Sự phát triển nhanh chóng của các mô hình tạo sinh ảnh và video trong những năm gần đây đã mở ra nhiều ứng dụng đột phá trong các lĩnh vực như phim ảnh, quảng cáo, hoạt hình. Các mô hình như **Diffusion** đã chứng tỏ khả năng vượt trội trong việc tạo ra hình ảnh có độ phân giải cao và thẩm mỹ từ chỉ dẫn văn bản. Không chỉ dừng lại ở việc tạo sinh ảnh tĩnh, các mô hình này đã và đang được mở rộng để giải quyết bài toán **tạo sinh video**, cho phép tạo ra các chuỗi hình ảnh liên tục với chuyển động mượt mà. Đặc biệt, một hướng nghiên cứu đầy tiềm năng là **tạo sinh video nhân hóa vật thể**, trong đó các vật thể vô tri được mô phỏng với đặc trưng và hành động giống con người.

Tuy nhiên, các mô hình tạo sinh video hiện tại chủ yếu dựa trên **chỉ dẫn văn bản**, trong khi khả năng diễn giải các chỉ dẫn mang tính sáng tạo như "nhân hóa vật thể" vẫn còn hạn chế. Điều này dẫn đến việc video đầu ra chưa đảm bảo được tính tự nhiên về hình dáng và chuyển động của vật thể. Để giải quyết vấn đề này, nghiên cứu đề xuất **một phương pháp kết hợp đa ràng buộc**, trong đó các ràng buộc như ảnh vật thể, video hành động, chỉ dẫn văn bản và ảnh mặt nạ ngữ nghĩa (semantic segmentation mask). Việc sử dụng ảnh mặt nạ giúp mô hình xác định vị trí và ngữ nghĩa của các bộ phận tương ứng với tay, chân,... từ đó cải thiện chất lượng mô phỏng hình dáng và chuyển động của vật thể nhân hóa trong video. Như minh hoạ ở Hình 1, đầu vào và đầu ra của hệ thống bao gồm:

- Đầu vào: (1) chỉ dẫn văn bản về nội dung video, (2) các ràng buộc trực quan như ảnh vật thể, video hành động, ảnh mặt nạ ngữ nghĩa.
- Đầu ra: Video được sinh ra dựa trên chỉ dẫn và các ràng buộc đầu vào.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

1. **Tìm hiểu và phân tích** các mô hình tạo sinh ảnh/video hiện có, đánh giá ưu điểm và hạn chế trong việc nhân hóa vật thể, đặc biệt là khả năng tích hợp đa ràng buộc để cải thiện chất lượng hình ảnh và chuyển động.
2. **Đề xuất phương pháp tạo sinh video nhân hóa vật thể** dựa trên mô hình tạo sinh, kết hợp đa ràng buộc gồm chỉ dẫn văn bản, ảnh vật thể, video hành động và ảnh mặt nạ để cải thiện độ chân thực của hình dáng và chuyển động.
3. **Thực nghiệm và đánh giá** phương pháp đề xuất dựa trên các tiêu chí: chất lượng hình ảnh, độ mượt của chuyển động, mức độ tuân thủ ràng buộc đầu vào, và so sánh với các phương pháp tạo sinh hiện có.

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung 1: Tìm hiểu và phân tích ưu/nhược điểm của các mô hình sinh ảnh và video kết hợp đa ràng buộc

Mục tiêu:

- Hiểu rõ kiến trúc và nguyên lý hoạt động của các mô hình cơ sở như Diffusion Models, cũng như các mô hình sinh ảnh và video dựa trên nó.
- Đánh giá ưu và nhược điểm của các phương pháp tiếp cận khác nhau trong việc tạo sinh video với đa ràng buộc.
- Phân tích cơ chế và hiệu quả của các mô hình học đặc trưng vật thể, cùng khả năng tích hợp các điều kiện ràng buộc khác nhau như chỉ dẫn văn bản, ảnh mặt nạ ngữ nghĩa, …

Phương pháp:

1. **Nghiên cứu lý thuyết và khảo sát tài liệu:**
    - Tiến hành nghiên cứu sâu về mô hình Diffusion Models, bao gồm kiến

trúc, cơ chế hoạt động và ứng dụng trong việc tạo sinh ảnh và video.

- Khảo sát các phương pháp tiếp cận khác nhau trong việc tạo sinh video với đa ràng buộc, như:
    - Tinh chỉnh (fine-tuning) các mô hình sinh ảnh để tạo video [1-2]
    - Sử dụng các mô hình sinh video trực tiếp từ chỉ dẫn văn bản [3]

2. **Phân tích ưu và nhược điểm của các phương pháp:** dựa trên các tiêu chí như chất lượng hình ảnh/video, khả năng tích hợp và xử lý các ràng buộc đầu vào, độ phức tạp và tài nguyên tính toán.
3. **Phân tích các mô hình học đặc trưng và tích hợp ràng buộc:**
    - Nghiên cứu các mô hình học đặc trưng như Textual Inversion [4] và DreamBooth [5], đánh giá khả năng học đặc trưng từ các ảnh tham chiếu và tích hợp vào quá trình tạo sinh video.
    - Phân tích các mô hình tích hợp các điều kiện ràng buộc để đánh giá khả năng cung cấp và tổng hợp thông tin.

Nội dung 2: Đề xuất và phát triển phương pháp tạo sinh video nhân hóa vật thể

Mục tiêu: Đề xuất một phương pháp mới để tạo sinh video nhân hóa vật thể kết hợp đa ràng buộc nhằm tích hợp thông tin hữu ích cải thiện khả năng sinh ảnh và chuyển động từ các ràng buộc.

Phương pháp:

1. Xây dựng kiến trúc mô hình:
    - Thiết kế mô hình tạo sinh video có khả năng xử lý và kết hợp các ràng buộc đầu vào như chỉ dẫn văn bản, Ảnh vật thể cần nhân hóa, Video hành động mẫu, Ảnh mặt nạ ngữ nghĩa (xác định các bộ phận cơ thể như tay, chân,...)
2. Phát triển cơ chế tích hợp đa ràng buộc: Xây dựng cơ chế tích hợp các ràng buộc đầu vào vào quá trình tạo sinh video, đảm bảo:
    - Mô hình có thể hiểu và kết hợp thông tin từ các nguồn khác nhau.
    - Đảm bảo sự nhất quán và tự nhiên trong hình dáng và chuyển động của vật thể nhân hóa trong video kết quả.

3. Xây dựng mã nguồn và chuẩn bị dữ liệu thử nghiệm:
    - Phát triển mã nguồn cho mô hình đề xuất, đảm bảo tính linh hoạt và khả năng mở rộng.
    - Chuẩn bị bộ dữ liệu thử nghiệm bao gồm: Các ảnh vật thể cần nhân hóa, Video hành động mẫu, chỉ dẫn văn bản và ảnh mặt nạ ngữ nghĩa tương ứng.

Nội dung 3: Thực nghiệm và đánh giá phương pháp đề xuất

Mục tiêu: Đánh giá hiệu quả của phương pháp đề xuất dựa trên các tiêu chí: chất lượng hình ảnh, độ mượt mà của chuyển động, mức độ tuân thủ các ràng buộc đầu vào và so sánh với các phương pháp hiện có.

Phương pháp:

1. **Thực nghiệm** trên bộ dữ liệu đa dạng bao gồm các vật thể và hành động khác nhau, để đánh giá khả năng tổng quát hóa của mô hình.
2. **Đánh giá kết quả:** sử dụng các chỉ số đánh giá khách quan như PSNR, SIM,…để tổng hợp số liệu.
3. **So sánh với các phương pháp hiện có** để đánh giá ưu và nhược điểm của phương pháp đề xuất.

KẾT QUẢ MONG ĐỢI

- **Phân tích và tổng hợp đánh giá** về ưu/nhược điểm các mô hình tạo sinh ảnh/video tích hợp đa ràng buộc hiện có.
- **Phát triển mô hình tạo sinh video nhân hóa vật thể**, kết hợp đa ràng buộc để cải thiện hình dáng và đảm bảo chuyển động mượt mà, tự nhiên và tuân thủ ràng buộc đầu vào.
- **Kết quả đánh giá và phân tích mô hình được đề xuất**
- **Xây dựng bộ dữ liệu thử nghiệm và mã nguồn**, hỗ trợ nghiên cứu tiếp theo.
- **Công bố kết quả nghiên cứu** dưới dạng bài báo hoặc báo cáo khoa học.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei

Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, Mike Zheng Shou: Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. ICCV 2023: 7589-7599
[2]. Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, Yaniv Taigman: Make-A-Video: Text-to-Video Generation without Text-Video Data. ICLR 2023
[3]. Spencer Sterling. Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023
[4]. Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, Daniel Cohen-Or: An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR 2023
[5]. Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, Kfir Aberman: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR 2023: 22500-22510