THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo: *https://youtu.be/jx5BcXFZd9o*
- Link slides: https://github.com/VuAnhMinh/CS2205.CH201/blob/554e9eaa9fed4966137779d36c5d38eaad23e7ed/Minh%20V%C5%A9%20%C3%81nh%20-%20CS2205.SEP2025.DeCuong.FinalReport.Template.Slide.pdf

- Họ và Tên: Vũ Ánh Minh
- MSSV: 250101041

- Lớp: CS2205.CH201
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 3
- Số câu hỏi QT của cả nhóm: 0
- Link Github: https://github.com/VuAnhMinh/CS2205.CH201

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI

TẠO HOẠT CẢNH TỪ BẢN VẼ PHÁC THẢO DỰA TRÊN HƯỚNG DẪN BẰNG VĂN BẢN.

TÊN ĐỀ TÀI TIẾNG ANH

FLIPPING STATIC DRAWINGS TO TEXT-GUIDED SKETCH ANIMATIONS.

TÓM TẮT

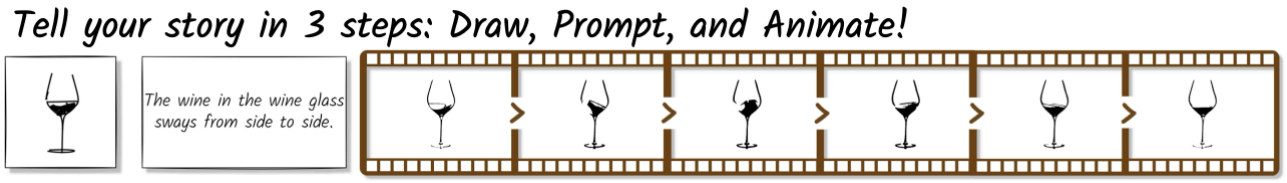
Việc sử dụng hoạt cảnh từ các bản vẽ phác thảo (sketch) là một phương tiện thu hút thị giác mạnh mẽ nhờ vào tính đơn giản và khả năng thúc đẩy trí tưởng tượng, có thể ví như một video lật từng trang vẽ phác thảo của một cuốn sổ. Các phương pháp truyền thống hiện tại đòi hỏi người sáng tác phải vẽ từng khung hình chính và trung gian, hoặc ở thiết bị điện tử thì bị bó buộc trong đồ họa vector, và phải chỉ dẫn hoạt cảnh cho chúng. Với sức mạnh của các mô hình khuếch tán video (video diffusion model) thông qua hướng dẫn văn bản ngày nay, việc tận dụng mô hình khuếch tán video để tạo ra hoạt cảnh phác thảo chỉ với một khung hình gốc là hoàn toàn khả thi. Để hạn chế sự sáng tạo của mô hình khuếch tán video, nghiên cứu cần phải tìm ra một mô hình dễ kiểm soát và áp dụng thuật toán hạn chế các tham số của mô hình để mô hình chỉ sinh ra hình ảnh phác thảo chứ không phải hình ảnh giống đời thực. Trong nghiên cứu, hai ứng viên được chọn ra là mô hình khuếch tán video **ModelScope** và phương pháp tinh chỉnh tham số **LoRA** (Low-Rank Adaptation). Mô hình sẽ tận dụng cơ chế Chú ý kép (Dual-Attention Composition) để vừa áp đặt các chi tiết phác thảo được sinh mới trung thành với bản gốc, cũng như vừa áp đặt các khung hình phải trung thành với bố cục thời gian để cho ra khung hình mượt mà. Mô hình kết quả sẽ được đặt tên là **FlipSketch**, hướng đến khả năng bảo toàn đặc điểm của bản vẽ gốc, mở ra hướng đi mới cho việc sản xuất nội dung hoạt hình sáng tạo một cách tự động và hiệu quả.

GIỚI THIỆU

Hoạt cảnh phác thảo (sketch animation) từ lâu đã là một phương tiện truyền tải thông điệp thị giác mạnh mẽ. Tuy nhiên, các phương pháp hiện tại vẫn đối mặt với những thách thức lớn. Kỹ thuật truyền thống yêu cầu người vẽ phải vẽ thủ công từng khung hình chính (key frames), còn các bản vẽ điện tử - đa số là đồ họa vector - lại bị giới hạn vì chưa thể sáng tạo các điểm ảnh mới, mà vẫn giữ các vector ảnh cũ và cố gắng làm chúng chuyển động.

Nghiên cứu này lấy ý tưởng từ các trang giấy vẽ tay liền mạch và việc lật từng trang giấy một sẽ tạo nên 1 câu chuyện chuyển động, do vậy hệ thống được đặt tên là **FlipSketch**. Hệ thống **FlipSketch** tận dụng mô hình khuếch tán video từ văn bản (T2V diffusion models) để tạo ra hoạt cảnh sống động chỉ từ một khung hình tĩnh và mô tả văn bản đầu vào.

- **Input**: Một bản vẽ tay phác thảo, một câu mô tả chuyển động.
- **Output**: Một đoạn video hoặc gif, mô tả lại hoạt cảnh phác thảo ngắn khoảng 10 khung hình, trong đó nhân vật/đối tượng chuyển động khớp với mô tả nhưng vẫn giữ nguyên hình dạng và phong cách của bản vẽ đầu vào. Ví dụ:



MỤC TIÊU

Đề tài hướng đến ba mục tiêu chính sau:

- **Mục tiêu 1:** Nghiên cứu và triển khai hệ thống **FlipSketch** nhằm tự động hóa quy trình tạo hoạt cảnh từ bản vẽ phác thảo tĩnh bằng cách tận dụng mô hình T2V có sẵn, áp dụng các kỹ thuật tối ưu hoá tham số đầu vào để mô hình khuếch tán video chỉ tạo ra bản vẽ phác thảo.
- **Mục tiêu 2:** Tối ưu hóa tính nhất quán của sản phẩm đầu ra bằng các kỹ thuật xử lý video.

- **Mục tiêu 3:** Đánh giá hiệu quả của mô hình thông qua các chỉ số định lượng như CLIP và chỉ số khảo sát từ người dùng. Đồng thời, thực nghiệm khả năng mở rộng của hệ thống trong các ứng dụng thực tế.

NỘI DUNG VÀ PHƯƠNG PHÁP

Hệ thống **FlipSketch** sẽ được xây dựng dựa trên sự kết hợp giữa tri thức chuyển động của mô hình khuếch tán video từ văn bản **ModelScope**[2] và khả năng kiểm soát phong cách phác thảo từ kỹ thuật **LoRA**[3]. Quy trình thực hiện dự kiến bao gồm các bước:

Bước 1: Thu thập và chuẩn bị dữ liệu huấn luyện (Data Synthesis): Vì dữ liệu video phác thảo tay thực tế rất khan hiếm cũng như rất khó để tự đạt được số lượng lớn dữ liệu, đề tài sẽ thực hiện tạo tập dữ liệu tổng hợp (synthetic dataset).

- **Phương pháp:** Sử dụng các công cụ tạo hoạt cảnh vector hiện có để sinh ra các cặp dữ liệu bao gồm: (i) mô tả văn bản và (ii) các chuỗi khung hình hoạt cảnh phác thảo tương ứng [4].
- **Đầu ra:** Tập dữ liệu gồm các cặp văn bản - video phác thảo dạng điểm ảnh (raster), đóng vai trò là dữ liệu đầu vào cho quá trình tinh chỉnh mô hình.

Bước 2: Chỉnh sửa các tham số của mô hình bằng kỹ thuật LoRA (Fine-tuning): Thay vì huấn luyện lại toàn bộ mô hình **ModelScope** rồi bắt nó phù hợp với phác thảo vẽ tay, kỹ thuật **LoRA (Low-Rank Adaptation)** sẽ được triển khai trên mạng **3D UNet**[5] của mô hình khuếch tán video.

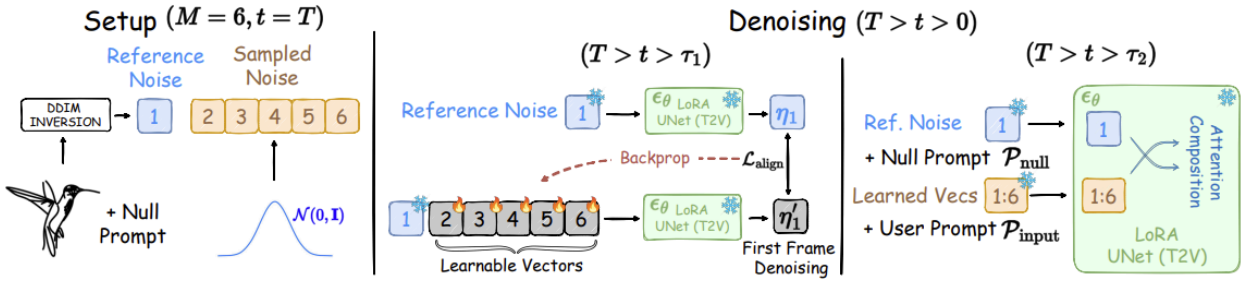
- **Cách thức:** Các ma trận hạng thấp (rank = 4) sẽ được chèn vào các lớp chú ý và tích chập của mô hình.
- **Mục tiêu:** Quá trình huấn luyện này sẽ ép mô hình chuyển từ việc tạo ra video ảnh thực sang việc tạo ra các khung hình có phong cách nét vẽ phác thảo đen trắng, tối giản mà vẫn giữ được logic chuyển động vật lý.

Bước 3: Trích xuất nhiễu tham chiếu (DDIM Inversion)[7]**:** Để bản vẽ đầu vào của người dùng trở thành "gốc" cho cả đoạn phim, kỹ thuật nghịch đảo **DDIM** sẽ được áp dụng.

- **Thực hiện:** Bản vẽ phác thảo đầu vào sẽ được mã hóa vào không gian tiềm ẩn (latent space). Sau đó, thuật toán nghịch đảo **DDIM** sẽ tìm ra một vector "nhiễu tham chiếu" (reference noise) mà khi khử nhiễu, nó sẽ tái tạo lại chính xác bản vẽ ban đầu.

Bước 4: Duy trì tính nhất quán và tạo chuyển động[6]**:** Trong quá trình sinh video thực tế, hai kỹ thuật can thiệp sẽ được triển khai đồng thời:

1. **Sắp xếp khung hình lặp lại (Iterative Frame Alignment):** Tại các bước khử nhiễu đầu tiên, hệ thống sẽ thực hiện tối ưu hóa nhiễu của các khung hình tiếp theo để chúng luôn bám sát đặc điểm của khung hình tham chiếu đầu tiên.
2. **Cơ chế chú ý kép (Dual-Attention Composition):** Các của khung hình đầu tiên sẽ được truyền dẫn sang các khung hình sau thông qua cơ chế Cross-attention. Điều này cho phép đối tượng chuyển động linh hoạt mà nét vẽ không bị thay đổi phong cách hoặc bị mờ nhòe.



Hình 2: Toàn bộ bước sinh hình ảnh kết hợp giữa DDIM, và Dual-Attention composition

Bước 5: Thử nghiệm và Đánh giá: Hệ thống sẽ được thử nghiệm trên đa dạng các chủ đề (động vật, đồ vật, ký tự) và so sánh với các baseline như Live-Sketch hay các mô hình I2V truyền thống.

- **Định lượng:** Đo lường bằng chỉ số CLIP[8] để đánh giá độ tương đồng văn bản và tính nhất quán của phác thảo.
- **Định tính:** Thực hiện khảo sát người dùng (User Study) để đánh giá tính thẩm mỹ và mức độ sinh động của hoạt cảnh.

KẾT QUẢ MONG ĐỢI

- **Xây dựng mô hình FlipSketch:** Hoàn thiện mô hình khuếch tán video dựa trên **ModelScope** đã được tích hợp kỹ thuật **LoRA**. Mô hình có khả năng tạo video hoặc gif từ ảnh phác thảo gốc và văn bản hướng dẫn, với 10 khung hình tiếp theo theo phong cách phác thảo.
- **Đảm bảo đầu ra có tính nhất quán và chuyển động:** Triển khai thành công kỹ thuật nghịch đảo DDIM cùng cơ chế chú ý kép (Dual-Attention Composition) và căn chỉnh khung hình lặp lại. Dự kiến kết quả thực nghiệm sẽ đạt chỉ số **CLIP** cao (độ tương đồng văn bản - video), đảm bảo các đối tượng chuyển động của hình vẽ linh hoạt mà không bị biến dạng nét vẽ hay thay đổi phong cách so với bản phác thảo gốc.
- **Triển khai ứng dụng Demo:** Xây dựng một ứng dụng web hoàn chỉnh, cho phép người dùng thực hiện:
    - Tải lên một bản vẽ phác thảo tĩnh (Input Sketch) và nhập câu lệnh mô tả hành động.
    - Hệ thống xử lý và hiển thị đầu ra là một đoạn video hoạt cảnh phác thảo mượt mà, bám sát nội dung văn bản.

TÀI LIỆU THAM KHẢO

[1].Hmrishav Bandyopadhyay, Yi-Zhe Song: FlipSketch: Flipping Static Drawings to Text-Guided Sketch Animations. CVPR 2025

[2] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571, 2023.

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan AllenZhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In ICLR, 2021.

[4] Jun Xing, Li-Yi Wei, Takaaki Shiratori, and Koji Yatani. Autocomplete hand-drawn animations. ACM TOG, 2015.

[5] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinario Passos, Longbo Huang, Jian Li, and Hang ´ Zhao. Lcm-lora: A universal stable-diffusion acceleration module. arXiv preprint arXiv:2311.05556, 2023

[6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In ICCV, 2023.

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.

[8] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In ACMM, 2022.