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

Vũ Ánh Minh - 250101041

Tóm tắt

- Lớp: CS2205.CH201
- Link Github của nhóm: https://github.com/VuAnhMinh/CS2205.CH201
- Link YouTube video: https://youtu.be/jx5BcXFZd9o
- Thành viên:
  - Vũ Ánh Minh:

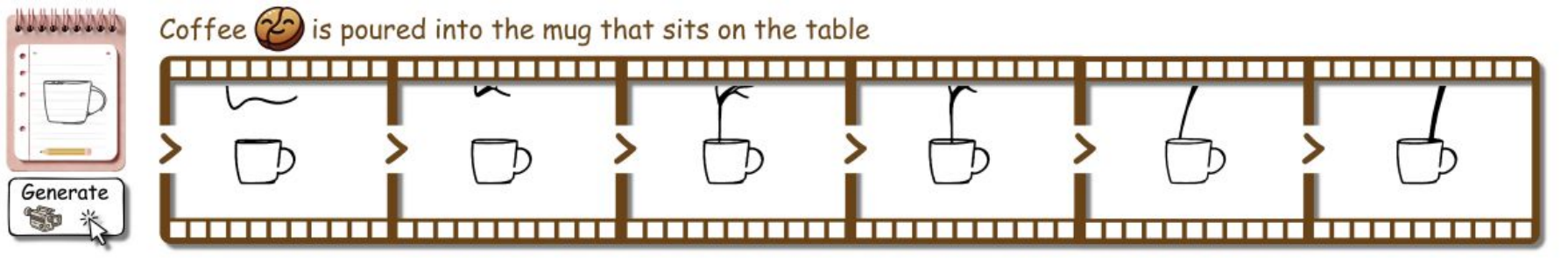
Giới thiệu

Vấn đề nghiên cứu:

- 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 như:
  - Tốn nhiều công sức vẽ tay
  - Chưa có mô hình T2V có thể sáng tạo ra các phác thảo vẽ tay mà đa số chỉ sử dụng lại vector có sẵn.

Giới thiệu

- Hệ thống **FlipSketch** tận dụng mô hình khuếch tán video từ văn bản 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 hoạt cảnh phác thảo ngắn khoảng 10 khung hình, trong đó đối tượng chuyển động khớp với mô tả đầu vào



Mục tiêu

- **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.
- **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.

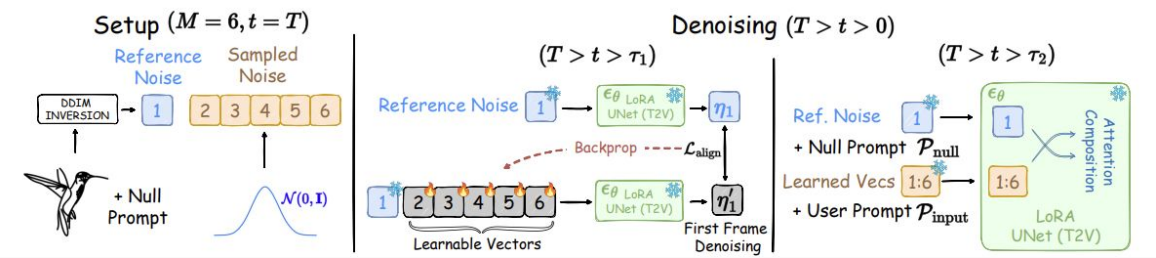
Nội dung và Phương pháp

- **Bước 1**: **Thu thập và chuẩn bị dữ liệu huấn luyện**: 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.
- **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**, triển khai kỹ thuật **LoRA** mạng 3D UNet của mô hình khuếch tán video.

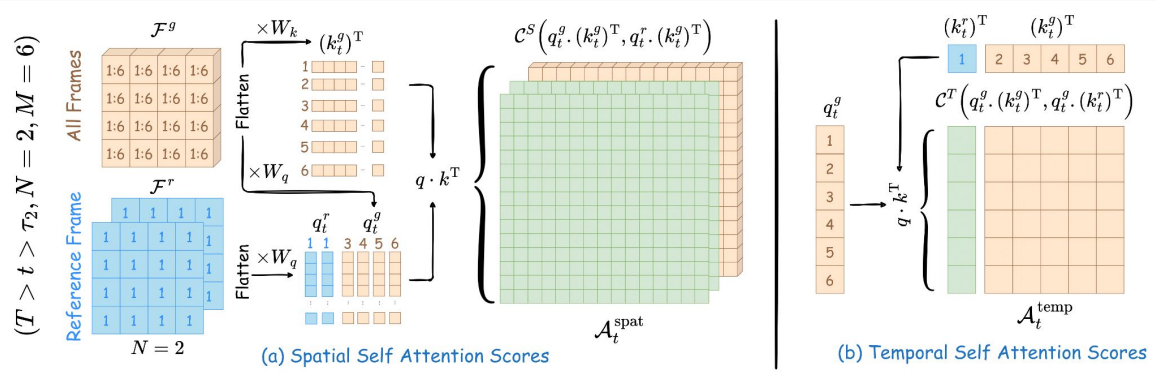
Nội dung và Phương pháp

- **Bước 3:** Trích xuất nhiễu tham chiếu (DDIM Inversion): mã hóa bản vẽ phác thảo đầu vào vào không gian tiềm ẩn (latent space) để **DDIM** tìm ra nhiễu tham chiếu (reference noise).
- **Bước 4:** Duy trì tính nhất quán và tạo chuyển động:
  - Dùng (i) Sắp xếp khung hình lặp lại để tối ưu hóa nhiễu của các khung hình tiếp theo
  - Dùng (ii) Cơ chế chú ý kép cho phép đối tượng chuyển động linh hoạt mà nét vẽ không bị thay đổi phong cách hoặc bị mờ nhòe

Nội dung và Phương pháp



Hình 2: Các bước sinh ra hình ảnh kết hợp Iterative Frame Alignment và Dual-Attention Composition



Hình 3: Trong bước 4 (Attention Composition), so sánh bản vẽ với nhiễu được tạo ra tại bước 3

Nội dung và Phương pháp

- **Bước 5:** Thử nghiệm và Đánh giá
  - Thử nghiệm trên đa dạng các chủ đề (động vật, đồ vật, ký tự)
  - So sánh với các baseline như Live-Sketch hay các mô hình I2V truyền thống
  - Đo lường bằng chỉ số CLIP, thực hiện khảo sát người dùng.

Kết quả dự kiến

- **Xây dựng mô hình FlipSketch:** có khả năng tạo video hoặc gif từ ảnh phác thảo gốc và văn bản hướng dẫn
- **Đảm bảo đầu ra có tính nhất quán và chuyển động:** 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**

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.