# **THÔNG TIN CHUNG CỦA NHÓM**

* Link YouTube video của báo cáo (tối đa 5 phút):   
  *(ví dụ: https://youtu.be/Vygg2AjB4ZM)*
* Link slides *https://github.com/thivinhhuy/CS519.O11/blob/main/Text-to-Image%20Diffusion%20Models%20are%20Zero-Shot%20Video%20Generators.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: Thi Vĩnh Huy * MSSV: 21522165   A person in a black shirt  Description automatically generated | * Lớp: CS519.O11 * Tự đánh giá (điểm tổng kết môn): 8/10 * Số buổi vắng: 1 * Số câu hỏi QT cá nhân: 2 * Link Github: *https://github.com/thivinhhuy/CS519.O11/* |

# **ĐỀ CƯƠNG NGHIÊN CỨU**

|  |
| --- |
| **TÊN ĐỀ TÀI (IN HOA)**  TEXT2VIDEO-ZERO: MÔ HÌNH KHUYẾN TÁN VĂN BẢN THÀNH HÌNH ẢNH TẠO VIDEO KHONG HUẤN LUYỆN |
| **TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)**  TEXT2VIDEO-ZERO: TEXT TO IMAGE DIFFUSION MODEL ARE ZERO-SHOT VIDEO GENERATORS |
| **TÓM TẮT** *(Tối đa 400 từ)*  Trong bài này, chúng tôi giới thiệu một nhiệm vụ mới mang tên tạo video từ văn bản không huấn luyện (zeroshot) và đề xuất một cách tiếp cận tiết kiệm chi phí, không cần bất kỳ huấn luyện hay tối ưu hóa nào. Thay vào đó, chúng tôi tận dụng sức mạnh của các phương pháp tổng hợp văn bản thành hình ảnh hiện có như Stable Diffusion, khiến chúng phù hợp với lĩnh vực video.  Để đạt được mục tiêu này, chúng tôi đã thực hiện một số thay đổi quan trọng. Thứ nhất, chúng tôi làm giàu mã tiềm ẩn của các khung hình được tạo ra với thông tin về chuyển động để đảm bảo bối cảnh chung và nền xuyên suốt thời gian được nhất quán. Thứ hai, chúng tôi tái lập trình sự chú ý tự thân cấp khung hình bằng cách sử dụng một dạng chú ý giữa các khung hình mới, của mỗi khung hình trên khung hình đầu tiên, nhằm giữ lại ngữ cảnh, ngoại hình và bản sắc của đối tượng nền trước. Thí nghiệm của chúng tôi cho thấy những thay đổi này dẫn đến việc tạo video chất lượng cao, nhất quán đáng kể, với chi phí thấp. |
| **GIỚI THIỆU** *(Tối đa 1 trang A4)*   * Trong những năm gần đây, trí tuệ nhân tạo tổng hợp (generative AI) đã thu hút sự quan tâm lớn trong cộng đồng nghiên cứu thị giác máy tính. Sự ra đời của các mô hình khuếch tán đã khiến cho việc tổng hợp văn bản thành hình ảnh, tức là tạo ra hình ảnh chất lượng cao từ mô tả bằng văn bản, trở nên phổ biến và thành công rực rỡ. Tuy nhiên, những nỗ lực gần đây nhằm mở rộng thành công này sang lĩnh vực tạo và chỉnh sửa video bị hạn chế bởi nhu cầu huấn luyện đáng kể với một lượng lớn dữ liệu dán nhãn, điều này có thể tốn kém và không khả thi cho nhiều người dùng. * Để giải quyết vấn đề này, Tune-A-Video đã giới thiệu một cơ chế có thể áp dụng mô hình Stable Diffusion (SD) cho lĩnh vực video, giảm đáng kể công sức huấn luyện xuống chỉ còn tinh chỉnh một video duy nhất. Mặc dù đây là một cách tiếp cận hiệu quả hơn nhiều so với các phương pháp trước đây, nó vẫn yêu cầu quá trình tối ưu hóa và chỉ giới hạn trong các ứng dụng chỉnh sửa video được hướng dẫn bằng văn bản. * Trong bài này, chúng tôi tiến thêm một bước trong nghiên cứu vấn đề mới lạ về tổng hợp văn bản thành video "không huấn luyện" (zeroshot). Điều này liên quan đến việc tạo video từ mô tả bằng văn bản mà không cần bất kỳ tối ưu hóa hoặc tinh chỉnh nào. Phương pháp của chúng tôi bao gồm sửa đổi một mô hình tổng hợp văn bản thành ảnh đã được huấn luyện trước, chẳng hạn như Stable Diffusion, và làm cho nó phong phú hơn với khả năng tạo ra video có tính nhất quán về thời gian. Bằng cách xây dựng dựa trên các mô hình tổng hợp văn bản thành ảnh đã được huấn luyện, phương pháp của chúng tôi tận dụng lợi thế về chất lượng tạo hình ảnh tuyệt vời của chúng và mở rộng khả năng áp dụng sang lĩnh vực video mà không cần huấn luyện thêm. * Để đảm bảo tính nhất quán về thời gian, chúng tôi trình bày hai cải tiến sáng tạo và nhẹ nhàng: thứ nhất, chúng tôi làm phong phú mã tiềm ẩn của các khung hình được tạo ra với thông tin chuyển động để giữ cho bối cảnh tổng thể và nền xuyên suốt thời gian được nhất quán; thứ hai, chúng tôi sử dụng sự chú ý giữa các khung hình, của mỗi khung hình trên khung hình đầu tiên, để duy trì ngữ cảnh, ngoại hình và bản sắc của đối tượng nền trước trong toàn bộ chuỗi. |
| **MỤC TIÊU**  *(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)*   * Giới thiệu một nhiệm vụ mới: tạo video từ văn bản không huấn luyện (zero-shot text-to-video). * Đề xuất một cách tiếp cận tiết kiệm chi phí (không cần huấn luyện hay tối ưu) để giải quyết nhiệm vụ này. * Đánh giá hiệu suất của cách tiếp cận được đề xuất trên một số bộ dữ liệu video. |
| **NỘI DUNG VÀ PHƯƠNG PHÁP**  *(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)*  **Nội dung:**   * Tận dụng sức mạnh của các phương pháp tổng hợp văn bản thành hình ảnh hiện có (ví dụ: Stable Diffusion) để phù hợp với lĩnh vực video. * Làm giàu mã tiềm ẩn của các khung hình được tạo ra với động lực chuyển động để duy trì sự thống nhất của bối cảnh chung và nền theo thời gian. * Lập trình lại sự chú ý tự thân cấp khung hình bằng cách sử dụng một sự chú ý giữa các khung hình mới của mỗi khung hình trên khung hình đầu tiên, để duy trì ngữ cảnh, ngoại hình và bản sắc của đối tượng nền trước.   **Phương pháp:**    Các bước tạo video từ văn bản:   1. Thu thập dữ liệu video và chú thích văn bản:  * Thu thập một bộ dữ liệu lớn gồm các video và chú thích văn bản tương ứng. * Chú thích văn bản có thể bao gồm mô tả về nội dung của video, lời thoại của các nhân vật, hoặc bất kỳ thông tin nào khác có thể giúp mô hình học cách tạo video từ văn bản.  1. Đào tạo mô hình khuếch tán hình ảnh tới văn bản:  * Đào tạo một mô hình khuếch tán hình ảnh tới văn bản trên bộ dữ liệu đã thu thập được. * Mô hình này học cách tạo ra các latent codes cho các hình ảnh từ các mô tả văn bản.  1. Tạo latent codes cho video:  * Sử dụng mô hình khuếch tán hình ảnh tới văn bản để tạo ra một chuỗi các latent codes cho video. * Các latent codes này đại diện cho nội dung của từng khung hình trong video.  1. Mô hình hóa motion dynamics in latent codes:  * Sử dụng một mô hình hồi quy hoặc mô hình động để mô hình hóa motion dynamics in latent codes. * Mô hình này học cách dự đoán các latent codes cho một khung hình tiếp theo trong video, dựa trên các latent codes cho các khung hình trước đó.  1. Tạo video từ latent codes:  * Sử dụng các latent codes đã được mô hình hóa để tạo ra một chuỗi các hình ảnh. * Các hình ảnh này có thể được tạo ra bằng cách sử dụng một mô hình tổng hợp hình ảnh hoặc một mô hình khuếch tán hình ảnh ngược.  1. Liên kết các hình ảnh lại với nhau để tạo thành video:  * Liên kết các hình ảnh đã tạo ra lại với nhau để tạo thành một video. * Video này có thể có độ dài bất kỳ và có thể bao gồm bất kỳ loại nội dung nào, miễn là nội dung đó có thể được mô tả bằng văn bản.   Các phương pháp cải thiện chất lượng video:  Instruct-Pix2Pix:   * Sử dụng mô hình inpainting để điền vào các vùng bị thiếu trong hình ảnh được tạo ra bởi mô hình tổng hợp hình ảnh. * Giúp tạo ra các video chất lượng cao hơn, có độ phân giải cao và chuyển động mượt mà hơn. * Reprogramming cross frame attention: * Thay đổi cơ chế chú ý của mô hình tổng hợp hình ảnh để nó có thể tập trung vào các phần quan trọng của video. * Giúp tạo ra các video mượt mà hơn và có chất lượng cao hơn.   Background smoothing:   * Làm mịn nền của video để giảm nhiễu và mờ. * Giúp tạo ra các video có nền mượt mà hơn, ít bị nhiễu và mờ hơn. |
| **KẾT QUẢ MONG ĐỢI**  *(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)*   * Tạo ra video chất lượng cao và nhất quán từ văn bản. * Video được tạo ra phải có nội dung phù hợp với văn bản. * Video phải có thể được tạo ra một cách nhanh chóng và dễ dàng. * Video phải có thể được tạo ra với chi phí thấp. |
| **TÀI LIỆU THAM KHẢO** *(Định dạng DBLP)*  [1]. KHACHATRYAN, Levon, et al. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. 2023.  [2]. Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303, 2022. |
| [3]. Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792, 2022.  [4]. Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:2212.11565, 2022.​  [5]. Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792, 2022.​  [6]. Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:2302.01329, 2023.​ |