


# THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):  
[https://www.youtube.com/watch?v=VSUE\\_bQfTQU](https://www.youtube.com/watch?v=VSUE_bQfTQU)
- Link slides (dạng .pdf đặt trên Github của nhóm):  
<https://github.com/transieu102/CS519.O11/blob/main/slides.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*

<ul style="list-style-type: none"><li>● Họ và Tên: Trần Siêu</li><li>● MSSV: 21520097</li></ul> 	<ul style="list-style-type: none"><li>● Lớp: CS519.O11</li><li>● Tự đánh giá (điểm tổng kết môn): 9.5/10</li><li>● Số buổi vắng: 1</li><li>● Số câu hỏi QT cá nhân: 11</li><li>● Link Github: <a href="https://github.com/mynameuit/CS519.O11">https://github.com/mynameuit/CS519.O11</a></li></ul>
--	---

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

ĐẢM BẢO NGŨ NGHĨA ẢNH TẠO BỞI MÔ HÌNH DIFFUSION BẰNG PHƯƠNG PHÁP GIÁM SÁT

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

ENSURING SEMANTIC PRECISION IN DIFFUSION-GENERATED IMAGES WITH A SUPERVISED METHODOLOGY

## TÓM TẮT (*Tối đa 400 từ*)

Text-to-image synthesis là một lĩnh vực nghiên cứu có tính ứng dụng cao, và các mô hình Diffusion đang thể hiện khả năng xuất sắc trong lĩnh vực này. Các hình ảnh được tạo ra bởi các mô hình Diffusion đạt chất lượng cao về cả độ phân giải, nội dung và độ chân thực. Để tạo ra những hình ảnh theo mong muốn của người dùng, các mô hình này mã hóa các đề xuất đầu vào (prompt) vào không gian ẩn (latent space) như là các gợi ý. Tuy nhiên, do thiếu giám sát trong quá trình đào tạo, việc đảm bảo độ chính xác ngữ nghĩa giữa các đề xuất và hình ảnh được tạo ra là một thách thức. Trong nghiên cứu này, chúng tôi giới thiệu một phương pháp đào tạo có giám sát về mặt ngữ nghĩa cho các mô hình Diffusion bằng cách sử dụng discriminator. Ngoài ra, chúng tôi tạo ra segmentation map từ prompt đầu vào để trực quan hóa thông tin về không gian, giúp mô hình dễ dàng hơn trong việc sinh ra ảnh đúng với yêu cầu. Điều này không chỉ đảm bảo tính ngữ nghĩa của các hình ảnh được tạo ra mà còn bảo toàn khả năng tạo ra bố cục hòa hợp của mô hình pretrained.

## GIỚI THIỆU (*Tối đa 1 trang A4*)

Text-to-image synthesis đã thu hút sự chú ý đáng kể do khả năng ứng dụng rộng rãi của nó. Người dùng có thể tạo ra những hình ảnh theo mong muốn bằng cách cung cấp cho mô hình các câu prompt. Cụ thể đầu vào và đầu ra của bài toán:

- Input: Câu prompt mô tả hình ảnh mong muốn.
- Output: Hình ảnh phù hợp với câu prompt đầu vào.

Trong bài toán này, các mô hình Diffusion [1, 2] đã nổi lên như những công cụ mạnh mẽ, thể hiện khả năng ấn tượng trong việc tạo ra những hình ảnh chất lượng với độ phân giải cao,

nội dung phong phú và tính chân thực xuất sắc. Các mô hình này cho phép người dùng truyền đạt ý tưởng về hình ảnh mong muốn bằng cách đưa các câu prompt đã được mã hóa vào latent space. Tuy nhiên, do sự liên kết ngữ nghĩa của kết quả đầu ra và prompt đầu vào không được giám sát trong quá trình huấn luyện, nên tính đúng đắn ngữ nghĩa của kết quả không được đảm bảo, đặc biệt là đối với các câu prompt phức tạp có nhiều đối tượng. Một nghiên cứu được công bố gần đây [3] đã thực hiện giám sát quá trình huấn luyện của các mô hình Diffusion-based bằng Discriminator từ đó nâng cao sự đảm bảo về kết quả đầu ra tương ứng với điều kiện đầu vào (layout). Bên cạnh đó, các nghiên cứu [4, 5] đã có thể trích các đặc trưng tương quan giữa ngữ nghĩa của hình ảnh và văn bản. Không chỉ vậy, một nghiên cứu khác [6] đã biến đổi prompt đầu vào thành segmentation maps để cung cấp tốt hơn thông tin về không gian cho mô hình. Những nghiên cứu này không chỉ cho thấy những kết quả khả quan trong việc giám sát mô hình Diffusion mà còn đạt được những kết quả ấn tượng trong biểu diễn mối quan hệ giữa văn bản và hình ảnh.

Trong nghiên cứu này, để đảm bảo được tính chính xác về nội dung của ảnh sinh ra, chúng tôi thực hiện việc giám sát ngữ nghĩa trong quá trình huấn luyện và tăng cường tính biểu đạt của thông tin đầu vào. Bằng cách sử dụng kết hợp Discriminator trong quá trình huấn luyện, chúng tôi hướng mô hình đến mục tiêu giảm thiểu sự khác biệt về đặc trưng ngữ nghĩa - được trích xuất bằng mô hình CLIP [4] - giữa prompt và hình ảnh. Bên cạnh đó, chúng tôi thực hiện việc tạo ra các segmentation map từ prompt để cung cấp thông tin về không gian cho mô hình một cách trực quan hơn. Những điều này không chỉ giúp cho mô hình sinh ra hình ảnh chính xác theo prompt mà còn bảo toàn được khả năng sinh ảnh có tính hài hòa trong bố cục của mô hình pretrained.

## MỤC TIÊU

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

- Nâng cao tính chính xác về mặt ngữ nghĩa của ảnh tạo bởi mô hình Diffusion, cụ thể là Latent Diffusion [1].
- Đảm bảo sự cân bằng giữa tính đúng đắn về ngữ nghĩa và sự hài hòa về bố cục của ảnh tạo bởi mô hình Latent Diffusion.

## NỘI DUNG VÀ PHƯƠNG PHÁP

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

### Nội dung:

- Giám sát về mặt ngữ nghĩa trong quá trình huấn luyện của mô hình Diffusion.
- Sinh ra segmentation map từ prompt để cung cấp thông tin không gian trực quan hơn cho mô hình.
- Kết hợp thông tin từ segmentation map sinh ra vào latent space của mô hình dưới dạng điều kiện và thực hiện giám sát trong huấn luyện để đảm bảo tính hài hòa trong bố cục đầu ra.

### Phương pháp:

- Tìm hiểu về quá trình huấn luyện của mô hình Diffusion-based như Latent Diffusion. Tìm hiểu về phương pháp kết hợp Discriminator trong quá trình huấn luyện của mô hình Diffusion như ALDM [3]. Tìm hiểu phương pháp so sánh khoảng cách ngữ nghĩa của hình ảnh và văn bản dựa trên các mô hình vision-language như CLIP.
- Tìm hiểu các phương pháp trực quan hóa thông tin không gian từ văn bản như [6]. Tìm hiểu các phương pháp tạo segmentation map đạt hiệu quả cao như dựa trên Unet, Diffusion. Đề xuất phương pháp tạo segmentation map bằng prompt dựa trên những phương pháp tìm hiểu được.
- Tìm hiểu phương pháp mã hóa hiệu quả segmentation map để đưa vào latent space của mô hình Latent Diffusion như ControlNet [2]. Áp dụng phương pháp tìm hiểu được để kết hợp thông tin từ segmentation map sinh ra vào latent space của mô hình dưới dạng điều kiện.
- Tìm hiểu các phương pháp, độ đo để đánh giá sự khác biệt giữa nội dung hình ảnh và văn bản như TIFA [7], CLIP Score, đánh giá sự hài hòa, chính xác về bố cục như FID [8], mIoU [9, 10].
- Thu thập, chuẩn bị các bộ dữ liệu phù hợp cho bài toán như COCO, Visual Genome, ADE20K,...
- Thực hiện cài đặt, huấn luyện và đánh giá các phương pháp về các phương diện đảm bảo đúng đắn về ngữ nghĩa và tính chính xác, hài hòa trong bố cục bằng các độ đo tìm hiểu được.

## KẾT QUẢ MONG ĐỢI

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

- Phương pháp huấn luyện có giám sát các mô hình Diffusion-based như Latent Diffusion, giúp nâng cao sự chính xác giữa nội dung prompt và hình ảnh được đánh giá bởi các độ đo như TIFA, CLIP Score, đồng thời bảo toàn được khả năng tạo ảnh có tính hài hòa trong bố cục của các mô hình pretrained được đánh giá bởi các độ đo như FID, mIoU.
- Mô hình cho phép tạo ra segmentation map từ prompt có độ chính xác cao, đánh giá bởi độ đo như mIoU. Giúp bổ sung thông tin vào latent space của mô hình Diffusion, từ đó nâng cao sự chính xác trong quá trình sinh ảnh.
- Báo cáo về kỹ thuật, cách cài đặt và kết quả thực nghiệm của các phương pháp.

## TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

- [1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer: High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022: 10674-10685
- [2] Lvmin Zhang, Maneesh Agrawala: Adding Conditional Control to Text-to-Image Diffusion Models. CoRR abs/2302.05543 (2023)
- [3] Li, Yumeng and Keuper, Margret and Zhang, Dan and Khoreva, Anna: Adversarial Supervision Makes Layout-to-Image Diffusion Models Thrive. arXiv preprint arXiv:2401.08815
- [4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever: Learning Transferable Visual Models From Natural Language Supervision. ICML 2021: 8748-8763
- [5] Junnan Li, Dongxu Li, Caiming Xiong, Steven C. H. Hoi: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022: 12888-12900
- [6] Justin Johnson, Agrim Gupta, Li Fei-Fei: Image Generation From Scene Graphs. CVPR 2018: 1219-1228
- [7] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna,

Noah A. Smith: TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering. CoRR abs/2303.11897 (2023)

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Sepp Hochreiter: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. NIPS 2017: 6626-6637

[9] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva: OASIS: Only Adversarial Supervision for Semantic Image Synthesis. Int. J. Comput. Vis. 130(12): 2903-2923 (2022)

[10] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, Wenjun Zhang: Freestyle Layout-to-Image Synthesis. CVPR 2023: 14256-14266