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

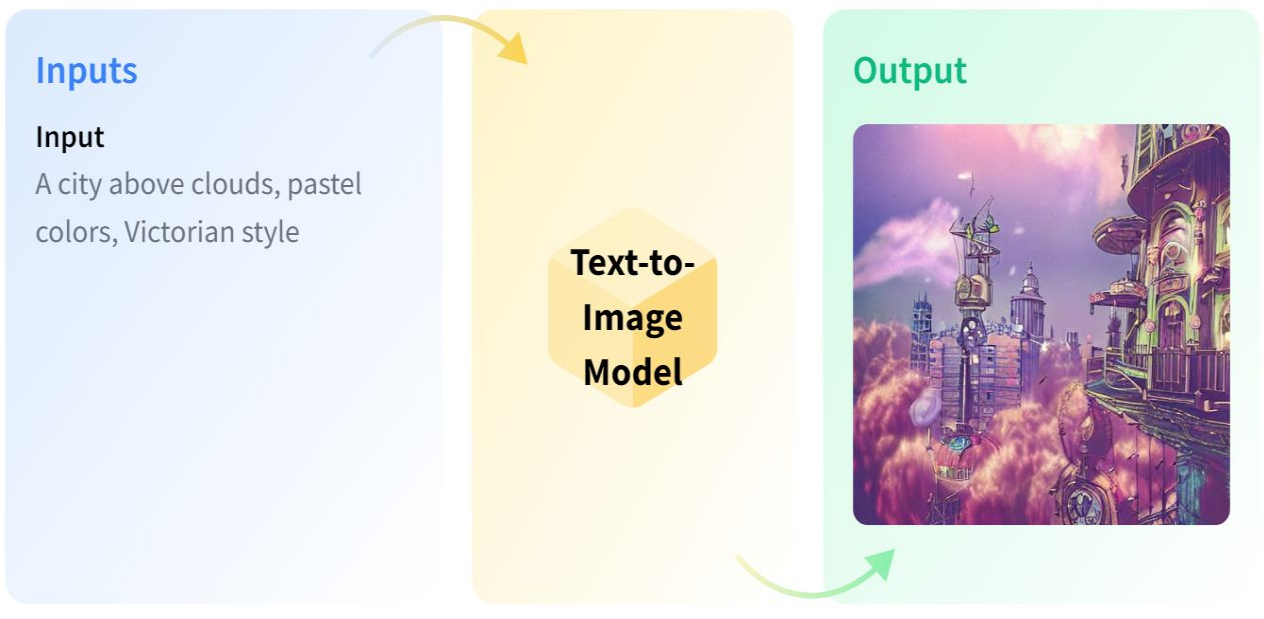
Trần Siêu - 21520097

Tóm tắt

- Lớp: CS519.O11
- Link Github của nhóm: https://github.com/transieu102/CS519.O11
- Link YouTube video:
- Người thực hiện:

Trần Siêu

Giới thiệu



DALL·E

Diffusion-based

Giới thiệu

Two sheep, one eating grass with a tree in front of a mountain; the sky has a cloud

Two cars, one parked on a street with a tree along it, and a window in front of a house and a house with a roof.

- Thiếu sự giám sát về ngữ nghĩa trong quá trình huấn luyện.
- Thông tin được biểu diễn thiếu trực quan.

Mục tiêu

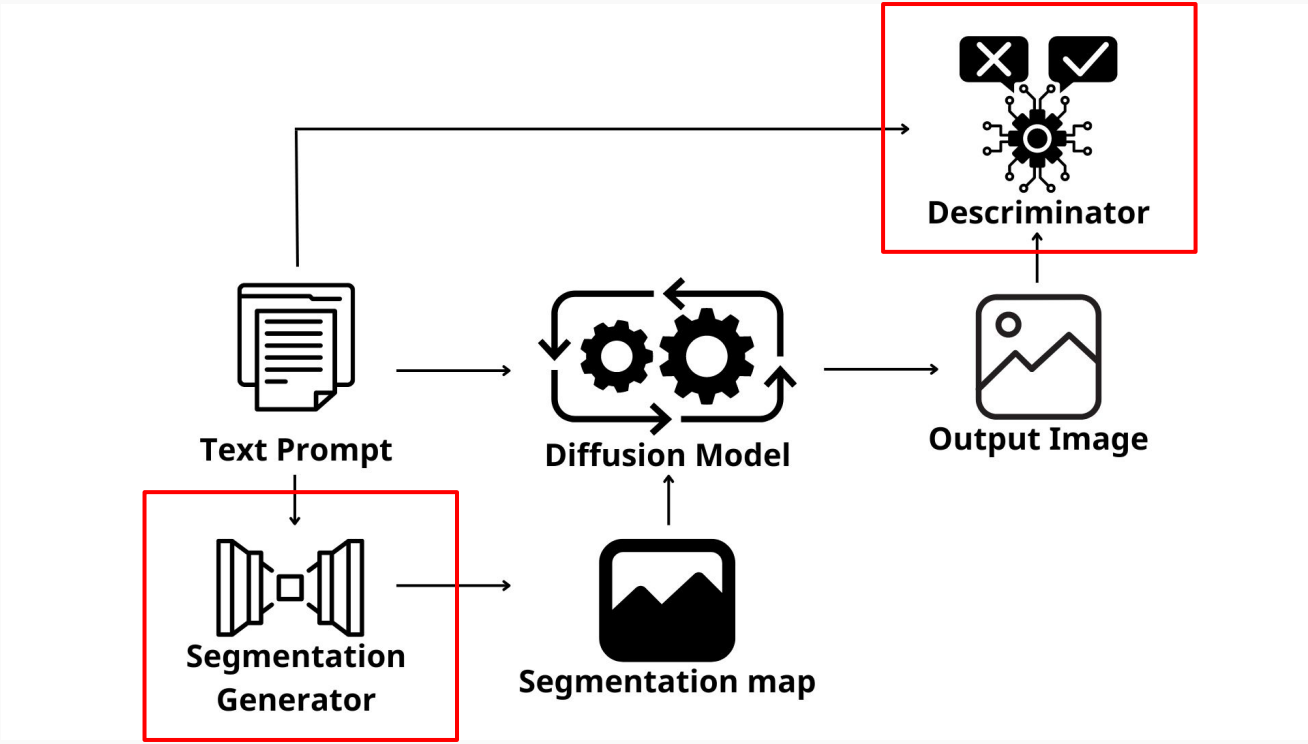
- Nâng cao tính chính xác về mặt ngữ nghĩa của ảnh tạo bởi mô hình Diffusion, cụ thể là Latent Diffusion.

- 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

- **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.

Nội dung



Hình minh họa kiến trúc mô hình.

Phương pháp

- Tìm hiểu về quá trình huấn luyện của mô hình Diffusion-based.
- Tìm hiểu các phương pháp trực quan hóa thông tin không gian từ văn bản.
- Tìm hiểu phương pháp mã hóa hiệu quả segmentation map để đưa vào latent space của mô hình Latent Diffusion.
- Tìm hiểu các phương pháp, độ đo để đánh giá.
- Thu thập, chuẩn bị các bộ dữ liệu phù hợp.
- Thực hiện cài đặt, huấn luyện và đánh giá các phương pháp.

Kết quả dự kiến

- **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.

- **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.**

Tài liệu tham khảo

[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