# TỪ VIDEO SANG PROMPT KỸ THUẬT CHUYỂN ĐỔI NGƯỢC DÀNH CHO VIỆC XÂY DỰNG CONTEXT TRONG STABLE DIFFUSION

Vũ Ngọc Thạch - 230201054

# Tóm tắt

- Lớp: CS2205.CH181
- Link Github: *https://github.com/DragonPow/CS2205.CH181*
- Link YouTube video: https://youtu.be/_SwtsKSPrZ4

Vũ Ngọc Thạch

# Giới thiệu

- Hiện nay có rất nhiều mô hình AI cho phép chuyển đổi từ prompt sang video [1], giúp người dùng thuận tiện hơn
- Khó khăn của các mô hình này:
  - Người dùng phải tự học cách sử dụng prompt
  - Người dùng phải lặp lại việc sinh video rất nhiều lần

=> Cần có một mô hình cho phép dự đoán prompt từ video sẵn có, giúp nâng cao năng suất trong việc sử dụng AI

# Mục tiêu

- **Nghiên cứu về các kỹ thuật** học máy hiện tại tốt nhất có thể đọc thông tin từ video.
- **Xây dựng được mô hình** cho phép dự đoán prompt từ các video của Stable Diffusion.
- **Đánh giá** độ hiệu quả mô hình và so sánh video được sinh ra từ prompt dự đoán so với video mẫu.
- Phân tích được một số hạn chế, giới hạn trong video ảnh hưởng tới việc dự đoán prompt.

# Nội dung và Phương pháp

- Chuẩn bị tập dữ liệu:
  - Thu tập tập dữ liệu video VidProM [3]
  - Thu thập thêm 1 tập dữ liệu video từ nguồn khác dùng kiểm tra hiệu năng của mô hình.

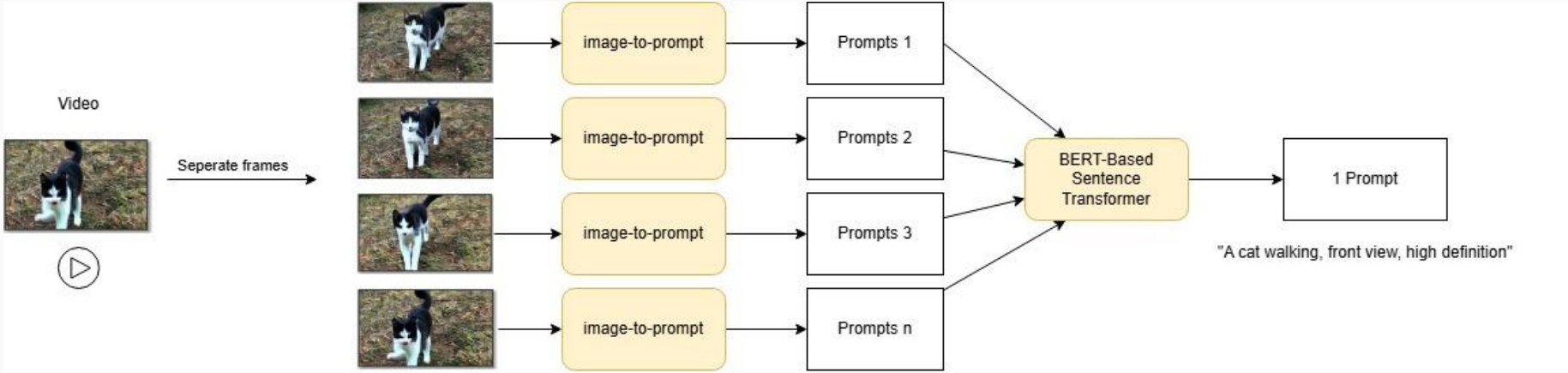
“A cat walking, front view, high definition” →

Video

# Nội dung và Phương pháp

Thiết kế thuật toán:

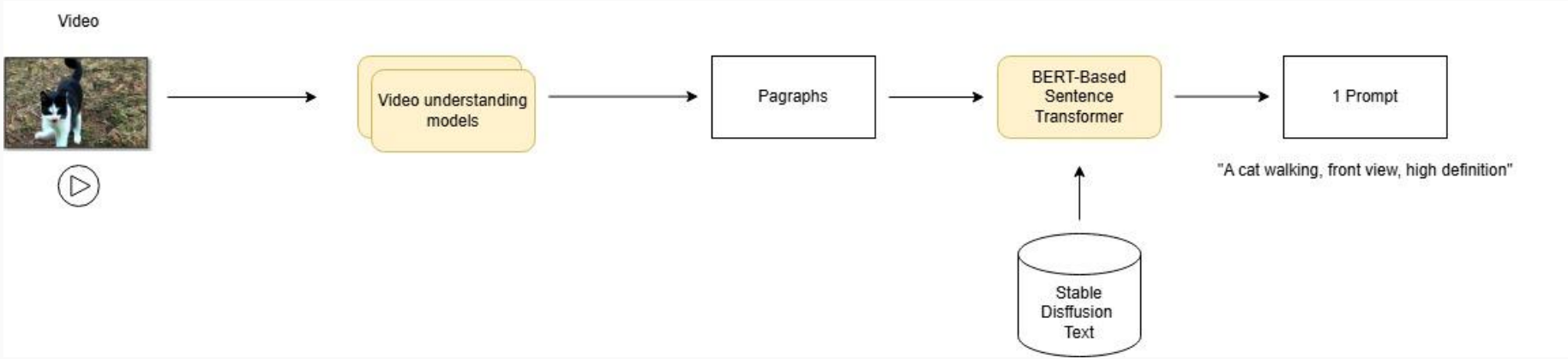
- Sử dụng từ mô hình image-to-text [2] hiện có



# Nội dung và Phương pháp

Thiết kế thuật toán:

- Xây dựng dựa trên các mô hình đọc hiểu Video:



# Nội dung và Phương pháp

Kiểm tra và đánh giá:

- So sánh prompt được dự đoán và prompt thực tế (BLEU, ROUGE, Word2Vec),...
- Sử dụng mô hình CLIP để đánh giá sự tương đồng giữa video được sinh ra từ prompt dự đoán và video gốc.
- Phân tích sự các yếu tố của video ảnh hưởng tới việc dự đoán prompt

=> Xây dựng ứng dụng từ mô hình đạt được

# Kết quả dự kiến

- Xây dựng được một mô hình cho phép chuyển đổi từ video sang prompt với hiệu suất ổn định:
  - Độ chính xác trên **60%**
  - Thời gian dự đoán ngắn dưới **1 phút**.
- Xây dựng được 1 ứng dụng sử dụng mô hình được đào tạo ở trên, nhằm đánh giá chất lượng của mô hình trong thực tế với tập người dùng thật.

# Tài liệu tham khảo


- [1]. Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, Tim Salimans: *Imagen Video: High Definition Video Generation with Diffusion Models*. CoRR abs/2210.02303 (2022)
- [2]. Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Mubarak Shah: *Reverse Stable Diffusion: What prompt was used to generate this image?* CoRR abs/2308.01472 (2023)
- [3]. Wenhao Wang, Yi Yang: VidProM: *A Million-scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models*. CoRR abs/2403.06098 (2024)