


# THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo: [https://youtu.be/\\_2d8icVe9E0](https://youtu.be/_2d8icVe9E0)
- Link slides:  
[https://github.com/minhducnguyen2602/CS519.O11/blob/main/%C4%90%E1%BB%A9c%20Nguy%E1%BB%85n%20Minh\\_CS519\\_Slide.pdf](https://github.com/minhducnguyen2602/CS519.O11/blob/main/%C4%90%E1%BB%A9c%20Nguy%E1%BB%85n%20Minh_CS519_Slide.pdf)

|  |   |
|--|---|
| <ul style="list-style-type: none"><li>• Họ và Tên: Nguyễn Minh Đức</li><li>• MSSV: 21520730</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>• Số câu hỏi QT của cả nhóm: 11</li><li>• Link Github:<br/><a href="https://github.com/minhducnguyen2602/CS519.O11">https://github.com/minhducnguyen2602/CS519.O11</a></li></ul> |
|--|---|

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

XÂY DỰNG HỆ THỐNG TRUY VẤN VIDEO HIỆU SUẤT CAO KẾT HỢP ĐA PHƯƠNG THỨC

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

DEVELOPING AN EFFICIENT MULTIMODAL VIDEO QUERY SYSTEM

## TÓM TẮT *(Tối đa 400 từ)*

Với sự tiên bộ và đa dạng của các loại thiết bị quay phim và sự phát triển mạnh mẽ của nhiều loại mạng xã hội, video ngày nay không chỉ là một dạng dữ liệu phổ biến, mà còn là nguồn thông tin đa chiều. Trong video, chúng ta không chỉ đơn thuần khai thác dữ liệu hình ảnh, mà còn có thể trích xuất dữ liệu âm thanh, các đối tượng và nhiều loại thông tin khác. Từ đó cho thấy, truy vấn thông tin từ những loại dữ liệu của video cũng sẽ trở nên phức tạp và đầy thách thức. Để truy vấn thông tin từ nhiều loại dữ liệu như vậy, cần có phương pháp hợp lý để đảm bảo hiệu suất tối ưu. Trước đây, đã có những nghiên cứu tương tự, nhưng thường chưa tận dụng triệt để nguồn dữ liệu đa dạng từ video. Do đó, đề tài này đã đề xuất xây dựng một hệ thống truy vấn video với hiệu suất cao, nhanh chóng, và tận dụng mọi thông tin có sẵn từ video (thông tin về hình ảnh, âm thanh, đối tượng trong ảnh, chữ trong ảnh) để đảm bảo kết quả truy vấn được chất lượng nhất. Đồng thời, ở đề tài này của nhóm cũng sẽ đề xuất xây dựng một ứng dụng hoàn chỉnh trên nền tảng web với những công nghệ mới, hiện đại để dễ dàng tương tác và sử dụng. Điều này giúp người dùng tiếp cận hệ thống một cách thuận tiện, đồng thời tận hưởng những tiện ích mà công nghệ truy vấn video hiệu suất cao mang lại.

## GIỚI THIỆU *(Tối đa 1 trang A4)*

Ngày nay, trên toàn cầu, có nhiều cuộc thi lớn chủ yếu tập trung vào thách thức của việc truy vấn thông tin từ video, như Video Browser Showdown (VBS), Lifelog Search Challenge (LSC). Ngay cả tại Việt Nam, cuộc thi HCM AI Challenge 2022 và 2023 cũng là một ví dụ điển hình. Trong các sự kiện này, các đội tham gia đều đặt ra nhiệm vụ xây dựng hệ thống tìm kiếm video từ một lượng dữ liệu video vô cùng lớn được cung cấp trước, với mục tiêu

đảm bảo tính nhanh chóng và chính xác của quá trình tìm kiếm.

Tại sao lại tồn tại những cuộc thi như vậy? Điều này xuất phát từ thực tế rằng truy vấn thông tin từ lượng dữ liệu video lớn đã trở thành một thách thức quan trọng, xuất hiện trong nhiều lĩnh vực khác nhau của cuộc sống. Ví dụ, trong lĩnh vực An ninh giám sát, việc theo dõi và tìm kiếm người trong một lượng video lớn, phát hiện hành vi nguy hiểm, hoặc định vị hành lý thất lạc trong các khu vực công cộng đều đòi hỏi khả năng truy xuất thông tin hiệu quả từ video. Tương tự, trong giáo dục và giải trí, việc tìm kiếm thông tin cụ thể từ một lượng lớn video gốc cũng là một yêu cầu quan trọng.

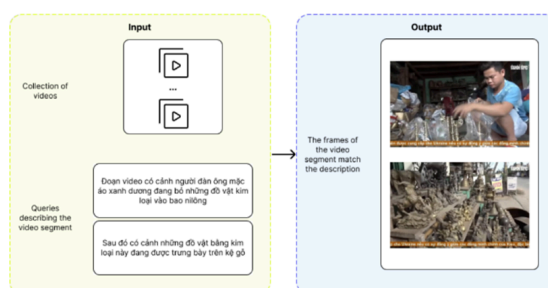
Từ những ví dụ trên, chúng ta rõ ràng thấy được sự quan trọng của bài toán truy vấn thông tin từ video. Đây không chỉ là một vấn đề cấp thiết mà còn đòi hỏi giải pháp tối ưu để đạt được hiệu suất tốt nhất trong quá trình xử lý. Trong những giải pháp đến từ các đội thi của các cuộc thi truy vấn video, thì hầu hết các đội tham gia sẽ sử dụng phương pháp chính là sử dụng các mô hình ngôn ngữ - hình ảnh như CLIP[1] để trích xuất đặc trưng hình ảnh và tìm kiếm các khung hình video phù hợp với mô tả của văn bản. Bên cạnh đó, có một số đội sẽ tận dụng thêm dữ liệu âm thanh, dữ liệu chữ trong ảnh và dữ liệu các đối tượng. Dựa vào những ý tưởng này, nhóm đề xuất xây dựng một hệ thống truy xuất linh hoạt và hiệu quả, có khả năng xử lý thông tin đầu vào một cách tối ưu với nhiều dạng dữ liệu khác nhau. Cụ thể input và output của bài toán như sau:

INPUT:

- Câu truy vấn dưới dạng văn bản (câu truy vấn có thể liên quan đến nội dung hình ảnh, âm thanh, các đối tượng,...) hoặc hình ảnh
- Tập hợp các video cần truy vấn

OUTPUT:

- Các khung hình phù hợp với câu truy vấn trong tập hợp các video đã cho



Hình 1: Ví dụ về mô tả đầu vào, đầu ra của bài toán.

## **MỤC TIÊU**

- Tìm hiểu tổng quan về bài toán truy vấn video và xác định hướng tiếp cận
- Xây dựng hệ thống truy vấn video bằng cách sử dụng và kết hợp nhiều phương thức truy vấn khác nhau như truy vấn dựa vào nội dung hình ảnh, nội dung âm thanh, các đối tượng, bằng hình ảnh cụ thể. Kết quả trả về của hệ thống phải có độ chính xác cao và thời gian xử lý nhanh (các kết quả trả về được xếp hạng theo độ liên quan từ cao đến thấp so với câu truy vấn, thời gian thực thi nhỏ hơn 500ms)
- Xây dựng ứng dụng với giao diện thân thiện với người dùng, dễ sử dụng.

## **NỘI DUNG VÀ PHƯƠNG PHÁP**

### **NỘI DUNG:**

- Nghiên cứu, tìm hiểu, tìm cách sử dụng các pretrained Vision-Language Models, Automatic Speech Recognition (ASR), Optical Character Recognition (OCR), Object detection để trích xuất các đặc trưng từ video nhằm mục đích tận dụng tối đa nguồn thông tin từ video.
- Nghiên cứu, đề xuất các thuật toán để kết hợp các phương thức tìm kiếm, tối ưu thời gian tìm kiếm, tối ưu kết quả tìm kiếm.
- Nghiên cứu, tìm hiểu các framework để xây dựng ứng dụng web đạt hiệu quả cao.

### **PHƯƠNG PHÁP:**

- Nghiên cứu, tìm hiểu, tìm cách sử dụng các pretrained Vision-Language Models ví dụ như CLIP, BLIP[2],... để trích xuất thông tin của hình ảnh và text dưới dạng vector một cách tốt nhất.
- Nghiên cứu, tìm hiểu, tìm cách sử dụng các mô hình OCR đạt hiệu quả cao trên những bộ dữ liệu lớn, bao quát được tối đa các trường hợp có thể có của dữ liệu. Một số mô hình ví dụ như Parseq[3], DBNet++[4],... Nếu như mô hình không có pretrained trên tiếng Việt thì sẽ finetune lại trên các tập dữ liệu tiếng Việt như VinText[5], hoặc trên bộ dữ liệu tự thu thập.
- Nghiên cứu, tìm hiểu, tìm cách sử dụng mô hình ASR để chuyển dữ liệu từ âm thanh sang text, ví dụ như Whisper[6].
- Nghiên cứu, tìm hiểu, tìm cách sử dụng một số mô hình Object detection như Co-DERT[7], Grounding Dino[8],... để trích xuất thông tin về các đối tượng.

- Nghiên cứu, tìm hiểu, tìm cách sử dụng FAISS để hỗ trợ tìm kiếm các vector sau khi được trích xuất từ CLIP, BLIP,... và Elasticsearch để tìm kiếm dữ liệu dạng text được trích xuất từ OCR, ASR, Object Detection.
- Xây dựng một số thuật toán để kết hợp các loại truy vấn, dựa vào score của từng loại kết quả truy vấn.
- Thiết kế giao diện của ứng dụng dễ sử dụng, không quá rắc rối đối với người dùng mới và việc thao tác phải dễ dàng. Giao diện bao gồm các vùng để hiển thị frame và video, cho phép tìm kiếm bằng text hoặc bằng hình ảnh.
- Sử dụng các công nghệ mới, hiện đại để xây dựng ứng dụng web. Nghiên cứu thử nghiệm sử dụng ReactJS để xây dựng giao diện, nginx để làm web server, FastAPI cho phần backend.

## KẾT QUẢ MONG ĐỢI

- Xây dựng được một hệ thống truy vấn từ video, dễ dàng thêm dữ liệu, xử lý kết quả nhanh và chính xác.
- Xây dựng được một ứng dụng web hoàn chỉnh có thể truy vấn trên video dựa vào mô tả dưới dạng text hoặc ảnh, kết quả trả về của các truy vấn phải được sắp xếp theo độ liên quan từ cao đến thấp (score dự kiến đạt được sẽ là  $R@1 = 0.6$  và  $R@10 = 0.9$  với  $R@K$  là tỷ lệ số lượng truy vấn mà hệ thống trả về kết quả chính xác trong top K đầu tiên được lựa chọn), thời gian thực hiện truy vấn nhanh, rơi vào khoảng 300-500ms.
- Hệ thống có thể được mang đi ứng dụng trong thực tế, ví dụ ở trường học, sân bay, ... để hỗ trợ các công việc liên quan đến tìm kiếm người, vật, sự kiện trong lượng lớn video.

## TÀI LIỆU THAM KHẢO

- [1] 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
- [2] 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

[3] Darwin Bautista, Rowel Atienza: Scene Text Recognition with Permuted Autoregressive Sequence Models. ECCV (28) 2022: 178-196

[4] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, Xiang Bai: Real-time Scene Text Detection with Differentiable Binarization. CoRR abs/1911.08947 (2019)

[5] Nguyen Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, Minh Hoai: Dictionary-Guided Scene Text Recognition. CVPR 2021: 7383-7392

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever: Robust Speech Recognition via Large-Scale Weak Supervision. ICML 2023: 28492-28518

[7] Zhuofan Zong, Guanglu Song, Yu Liu: DETRs with Collaborative Hybrid Assignments Training. ICCV 2023: 6725-6735

[8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. CoRR abs/2303.05499 (2023)