

ĐƠN VỊ TỔ CHỨC



SỞ THÔNG TIN VÀ TRUYỀN THÔNG
THÀNH PHỐ HỒ CHÍ MINH

ĐƠN VỊ PHỐI HỢP



ĐẠI HỌC QUỐC GIA
THÀNH PHỐ HỒ CHÍ MINH



SỞ KHOA HỌC VÀ CÔNG NGHỆ
THÀNH PHỐ HỒ CHÍ MINH



SỞ GIÁO DỤC VÀ ĐÀO TẠO
THÀNH PHỐ HỒ CHÍ MINH



THÀNH PHỐ HỒ CHÍ MINH



HỘI TIN HỌC THÀNH PHỐ HỒ CHÍ MINH

THƯỜNG TRỰC BTC



TRUNG TÂM PHÁT TRIỂN
KHOA HỌC VÀ CÔNG NGHỆ TRẺ

TẬP HUẤN THÍ SINH DỰ THI HỘI THI THỬ THÁCH TRÍ TUỆ NHÂN TẠO THÀNH PHỐ HỒ CHÍ MINH NĂM 2024



CHỦ ĐỀ: TRUY VẤN SỰ KIỆN TỪ VIDEO

BẢNG A: SINH VIÊN, THANH NIÊN

TP. Hồ Chí Minh, ngày 03 tháng 8 năm 2024

Giới thiệu thành viên

- ThS. Nguyễn Hải Đăng
- ThS. Đỗ Trọng Lễ
- CN. Nguyễn Quang Thức
- PGS. TS. Trần Minh Triết
Phòng thí nghiệm Công nghệ Phần mềm (SELab)
Trường Đại học Khoa học Tự nhiên, ĐHQG HCM.

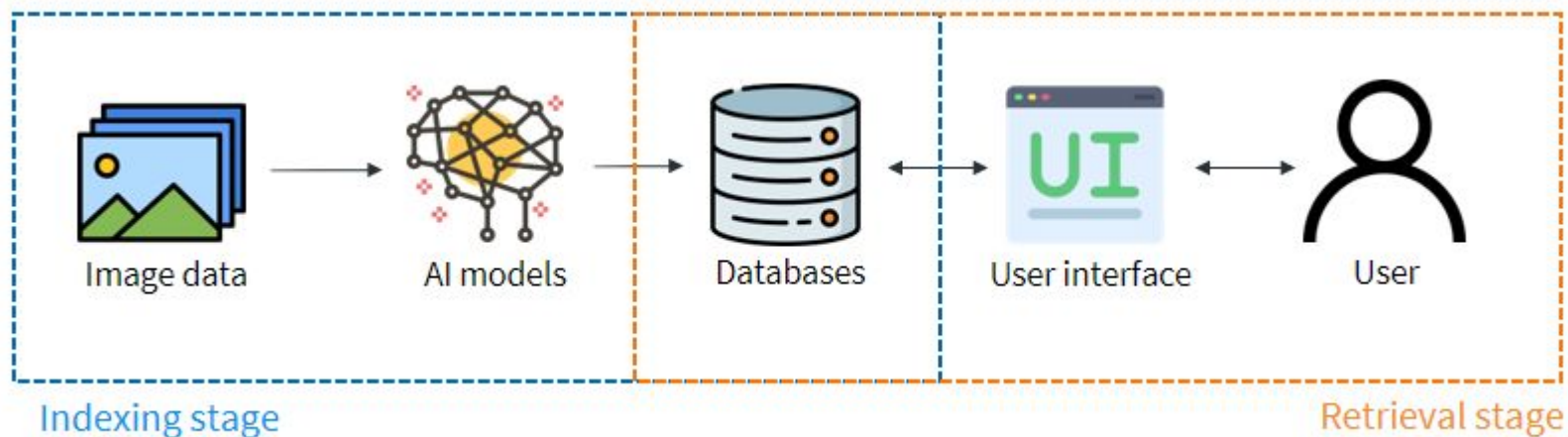


Tổng quan

- 1) **Hệ thống tìm kiếm video cần những gì?**
- 2) **Một số vấn đề cơ bản**
- 3) **Sơ lược về mô hình Thị giác-Ngôn ngữ**

Hệ thống tìm kiếm video

Hệ thống tìm kiếm video chỉ cần mô hình rút trích đặc trưng mạnh là đủ?



A decorative network diagram at the top of the slide, featuring a series of interconnected nodes and lines. A central node is highlighted with a dashed circle and contains the opening quotation mark.

“

*Mục tiêu của hệ thống tìm kiếm
là **tìm được**
video chúng ta mong muốn*

A decorative network diagram at the top of the slide, featuring a series of interconnected nodes and lines. A central node is highlighted with a dashed circle and contains a large opening quotation mark.

“

*Sẽ làm gì nếu
**mô hình truy vấn
không đủ tốt?***

A decorative network diagram at the top of the slide, featuring a series of interconnected nodes and lines. The nodes are represented by small circles, some of which are highlighted with a dashed border. The lines are thin and gray, creating a web-like structure that spans the width of the slide.

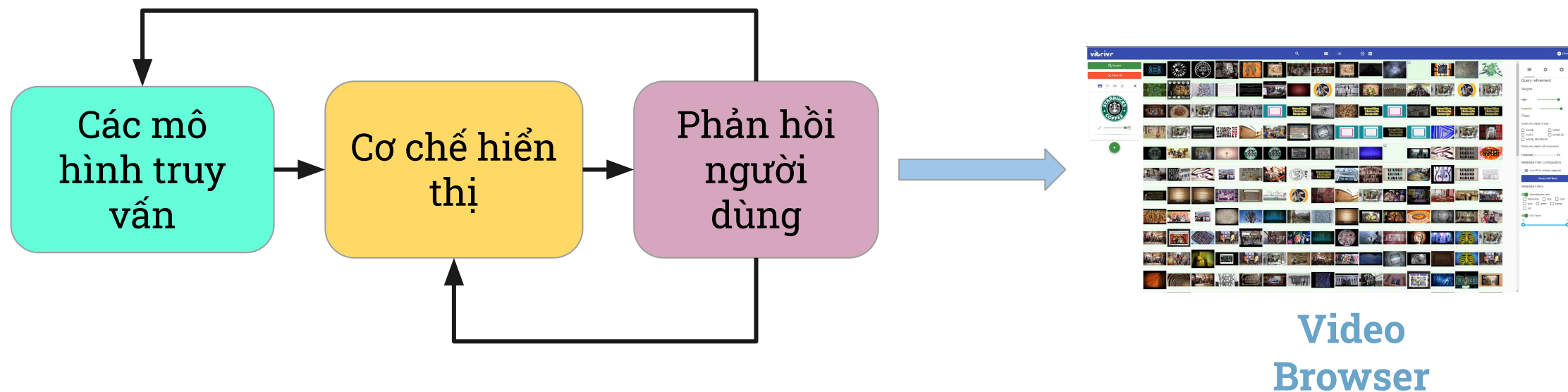
“

Cơ chế hiện thị

+

Phản hồi người dùng

Hệ thống tìm kiếm video





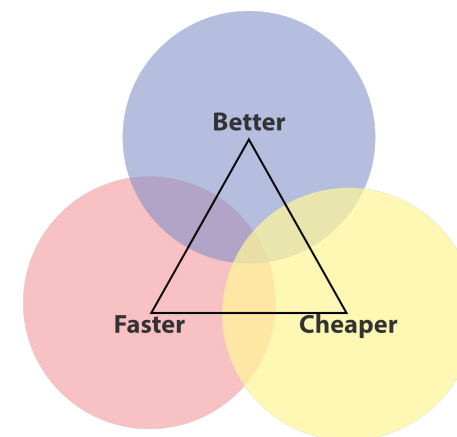
Các **mô hình** trong
bài toán **truy vấn video**
cần giải quyết yếu tố nào?

Các vấn đề trong **xây dựng mô hình truy vấn**

Làm sao truy vấn trên lượng **dữ liệu lớn hiệu quả** (Ví dụ: V3C1 có 1000 giờ nội dung video).

Cân bằng trong việc đánh đổi giữa

- **tốc độ chạy thuật toán**
- **sức mạnh thuật toán**
- **chi phí hệ thống**

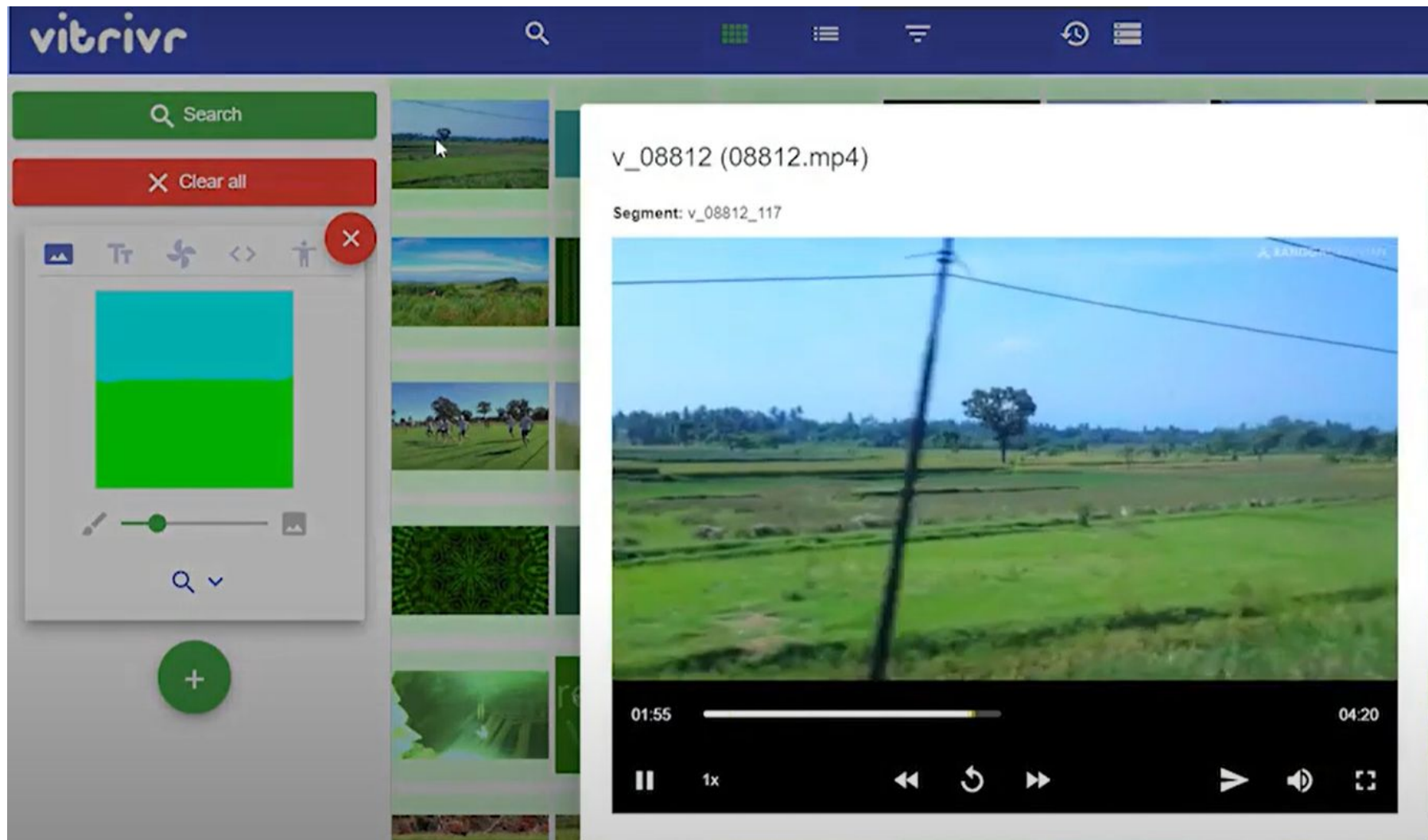




Các vấn đề trong **xây dựng mô hình truy vấn**

Liên kết giữa **thông tin do người dùng cung cấp** và **đoạn video không phải là duy nhất** (thông tin có thể khớp với nhiều đoạn video).

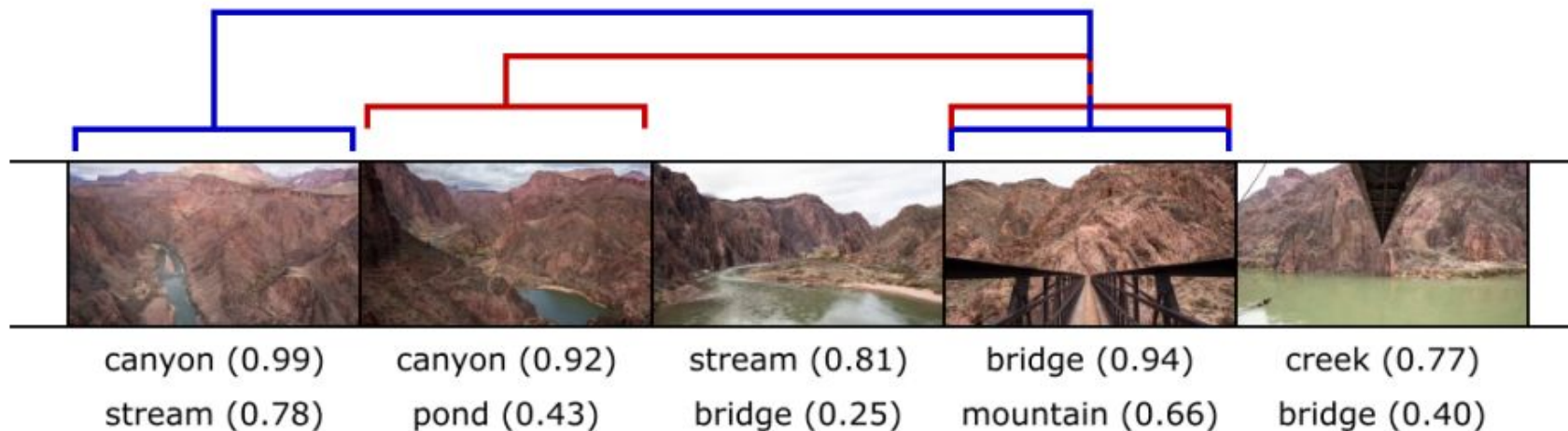
- Cơ chế **tăng thêm lượng truy vấn** để thu hẹp phạm vi tìm kiếm
- **Thay đổi phương thức truy vấn** để giúp liên kết chặt chẽ hơn (ví dụ như: truy vấn bằng ảnh sinh ra từ sketch)
- Cơ chế **kết hợp nhiều kiểu truy vấn**



vitrivr system

Các vấn đề trong xây dựng mô hình truy vấn

Làm sao để liên kết giữa các khung hình khi sử dụng các mô hình trên ảnh đơn?



Text query: “A slow pan up from a **canyon**, static shots of a **bridge** and redrock mountain.



• Vì sao **việc hiển thị** đóng
vai trò quan trọng?



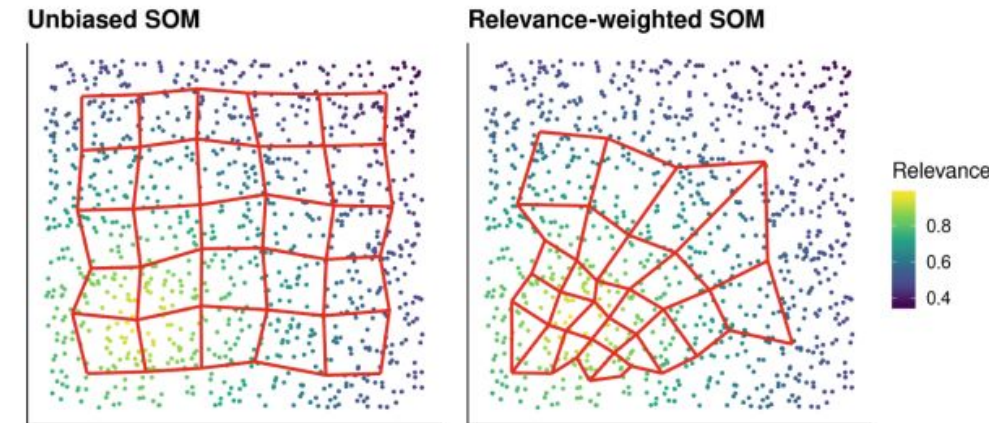
Các vấn đề trong cách **hiển thị kết quả truy vấn**

Số lượng **kết quả trả về** quá nhiều:

- Các frame gần giống nhau trong một video
- Số lượng video có liên quan để truy vấn

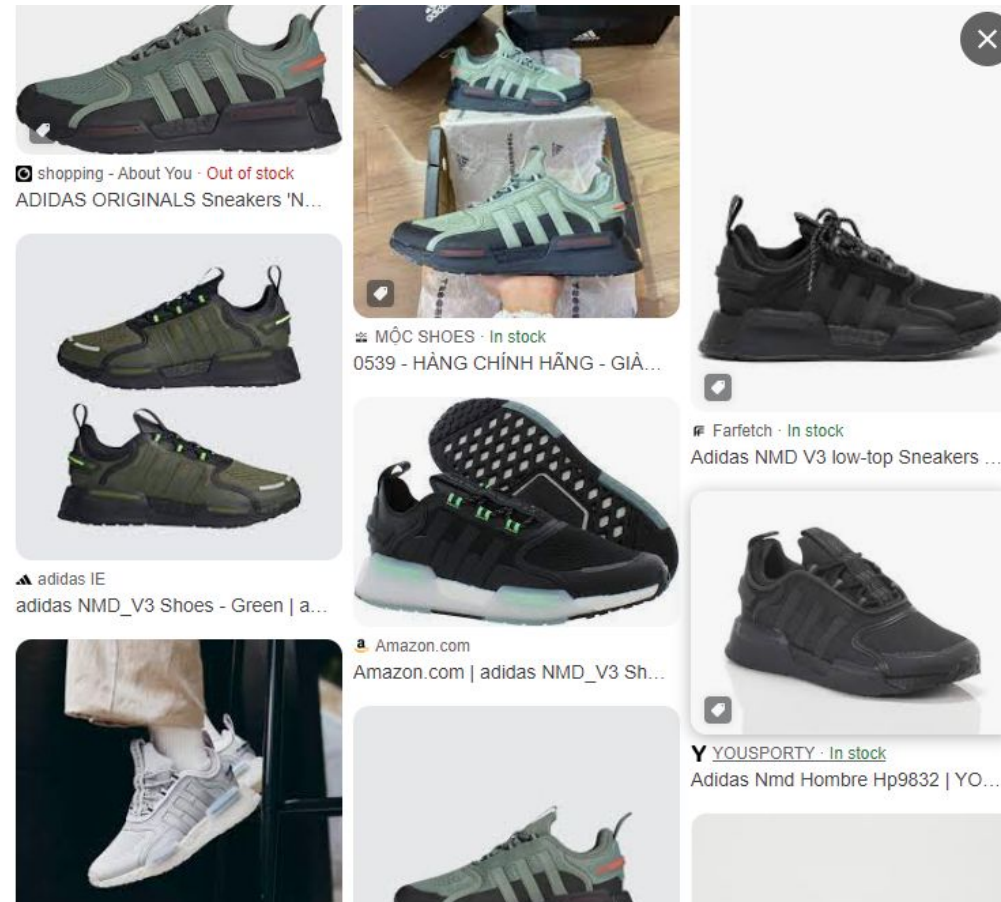
Chúng ta sẽ hiển thị gì

khi người dùng không biết bắt đầu từ đâu?



Các vấn đề trong cách **hiển thị kết quả truy vấn**

Ảnh truy vấn



Ảnh liên quan

Các vấn đề trong cách **hiển thị** kết quả truy vấn



VISIONE'23



Phản hồi
của người dùng
nên được xử lý ra sao?

Các vấn đề trong việc **xử lý phản hồi của người dùng**

Làm thế nào để hệ thống phản hồi của người dùng hoạt động cân bằng cho cả 2 trường hợp

- **Khám phá:** hiển thị các thông tin ít liên quan nhằm mục đích mở rộng phạm vi tìm kiếm (**khắc phục hạn chế** sự thiếu chặt chẽ cách biểu đạt của người dùng, sự thiếu chính xác của mô hình rút trích đặc trưng, ...)
- **Khai phá:** hiển thị các thông tin có mức độ liên quan cao, nhằm tách bạch các video có độ giống nhau cao.



Các vấn đề trong việc **xử lý phản hồi của người dùng**

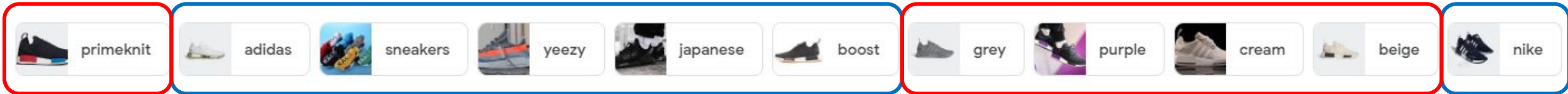
Làm thế nào để **hệ thống gợi ý những concept** liên quan từ truy vấn của người dùng

- Gợi ý các concept mà model nghĩ là **người dùng đang muốn truy vấn** (khám phá)
- Gợi ý các concept giúp model **giảm đi sự không chắc chắn** trong kết quả trả về (khai phá)

Các vấn đề trong việc **xử lý phản hồi của người dùng**

Câu truy vấn

nmd   



các concepts phục vụ cho việc **khám phá**



các concepts phục vụ cho việc **khai phá**



**Làm sao truy vấn
trên lượng dữ liệu lớn
hiệu quả**

A decorative network diagram at the top of the slide, featuring a series of interconnected nodes and lines. The nodes are represented by small circles, some of which are highlighted with a dashed border. The lines are thin and grey, creating a web-like structure that spans the width of the slide.

“

Hãy bắt đầu bằng cách tự nhiên nhất

.

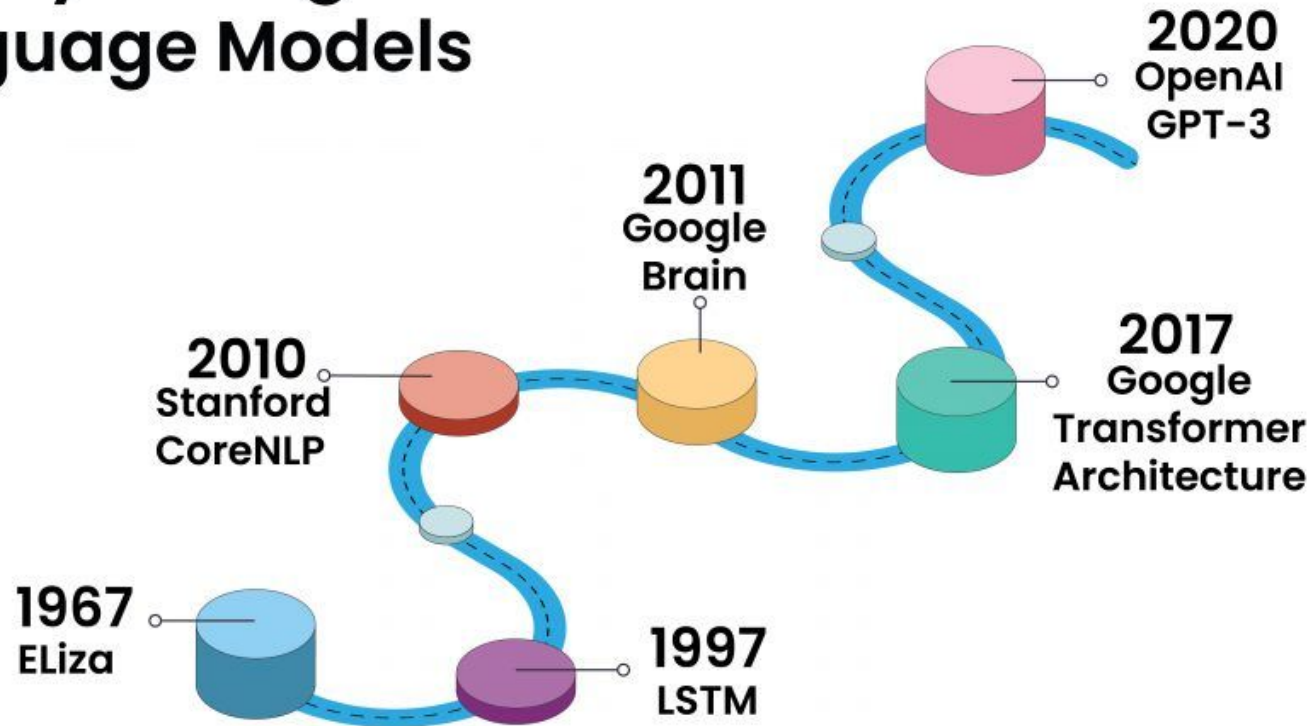
.

Mô tả những gì chúng ta thấy



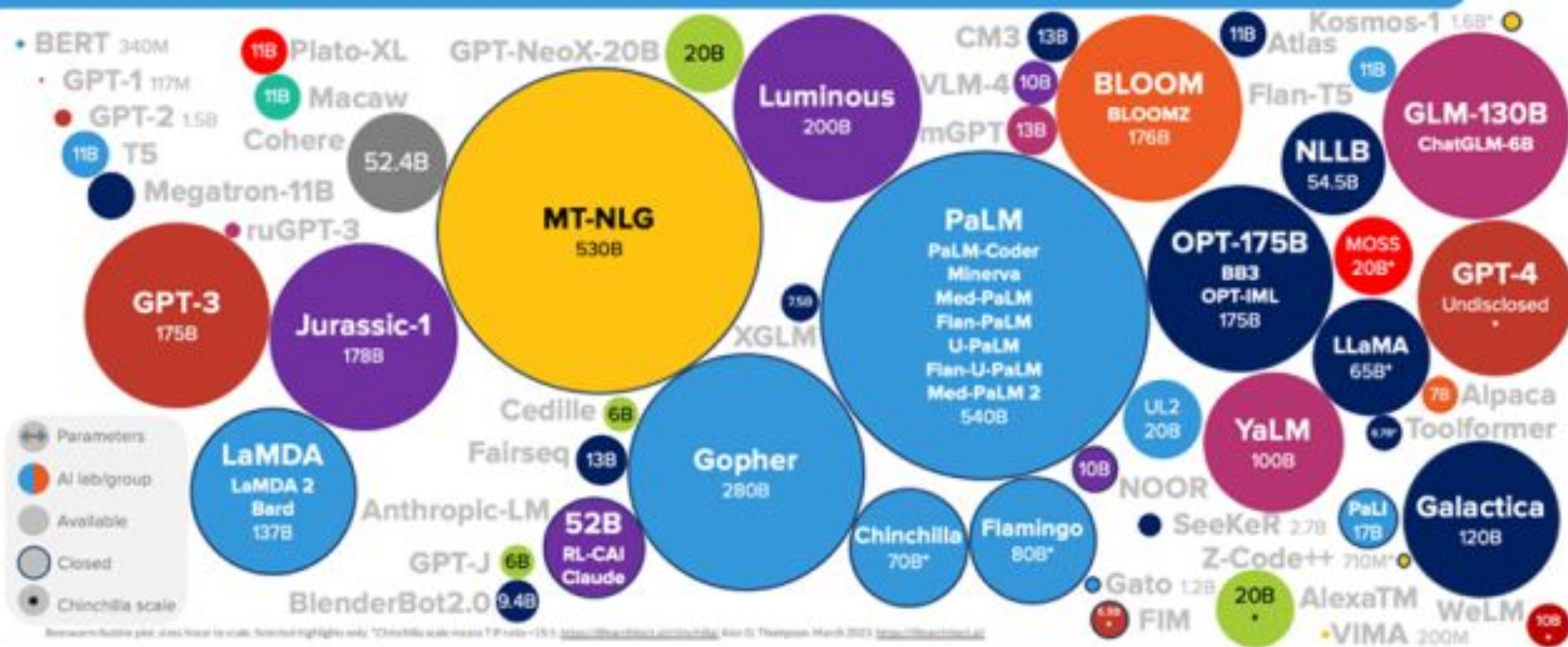
Sự phát triển của mô hình xử lý ngôn ngữ tự nhiên

History of Large Language Models

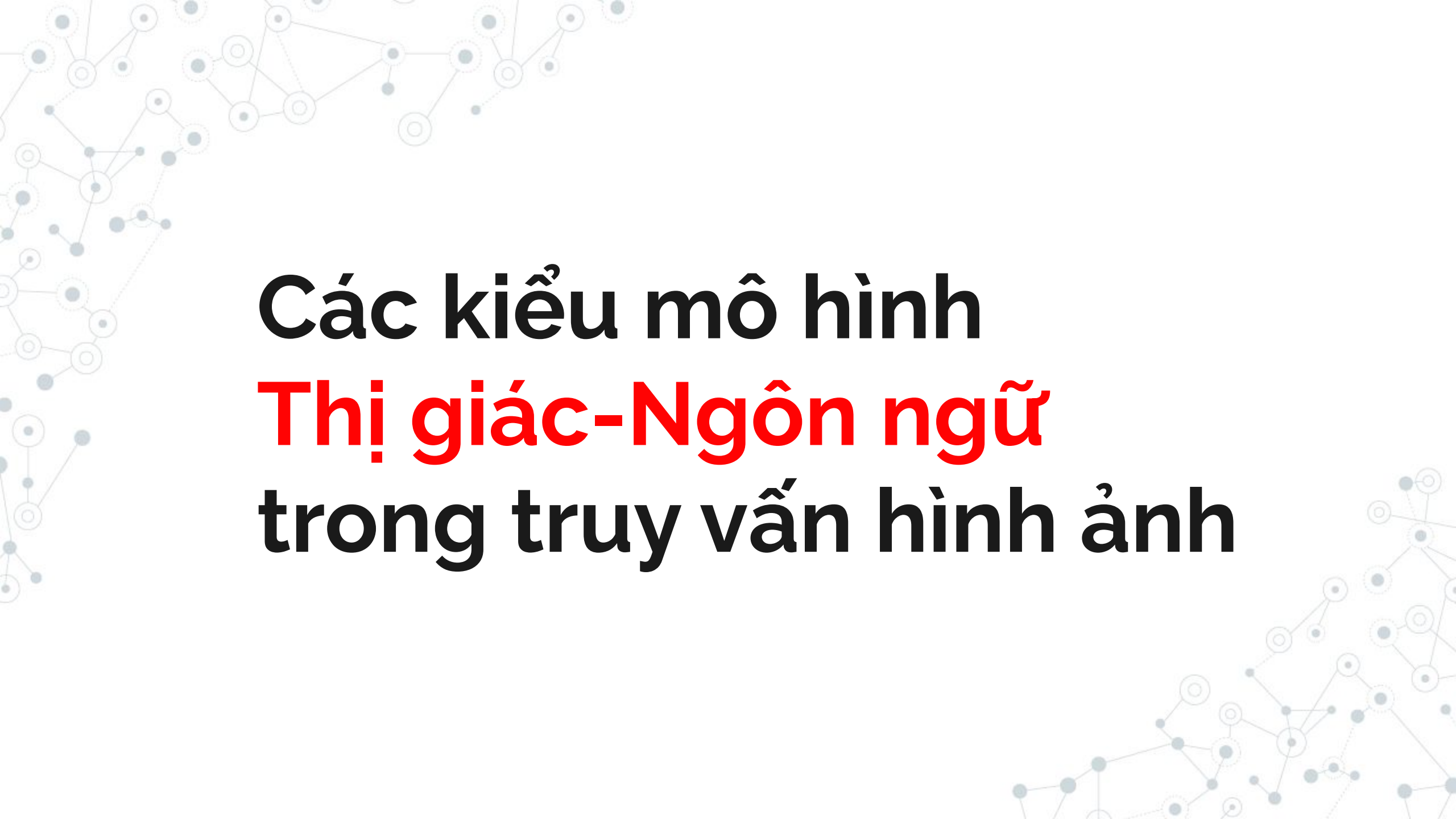


Nguồn: https://www.scribbledata.io/wp-content/uploads/2023/05/LLL_Evolution-02-1024x576.jpg

LANGUAGE MODEL SIZES TO MAR/2023



LifeArchitect.ai/models



Các kiểu mô hình **Thị giác-Ngôn ngữ** **trong truy vấn hình ảnh**

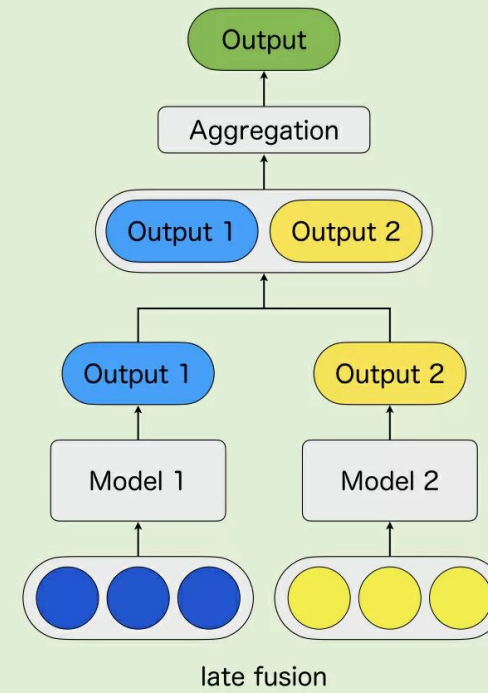
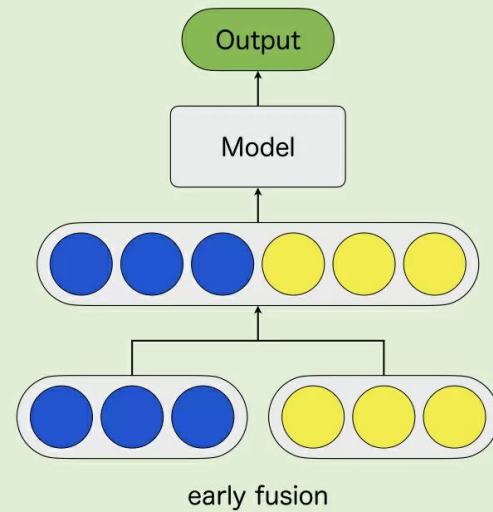
Sự ra đời của các mô hình thị giác-ngôn ngữ lớn

- Nguồn dữ liệu khổng lồ về cặp (ảnh, câu mô tả) *không cần gán nhãn*
- Các mô hình kiến trúc mới có khả năng tận dụng dữ liệu lớn (e.g transformers)
- Tài nguyên và hệ thống lớn để huấn luyện các mô hình này

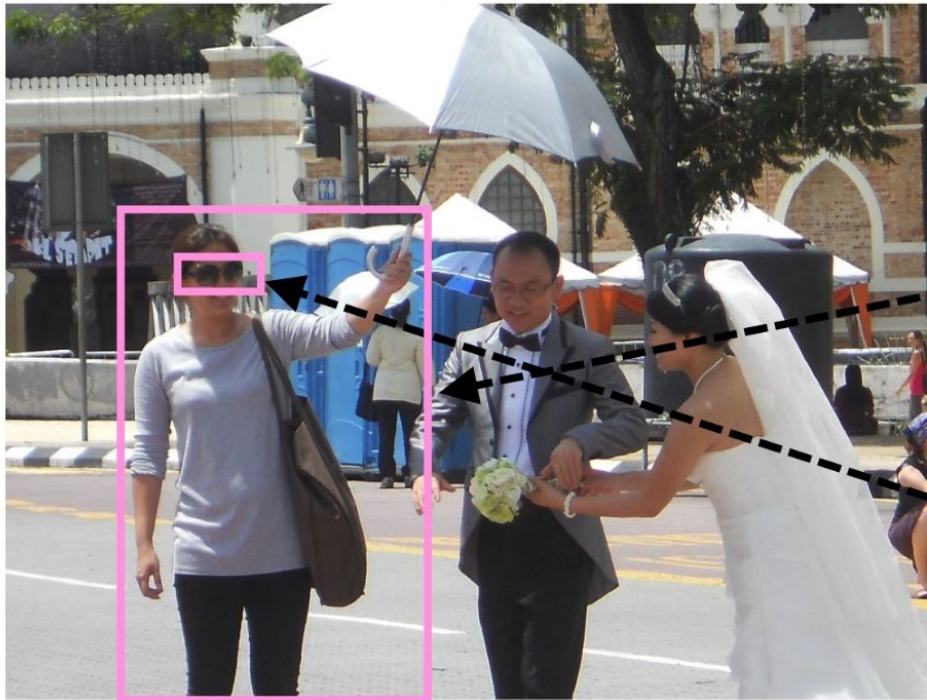


Tốc độ chạy
hay **sức mạnh**
thuật toán.

Fusion

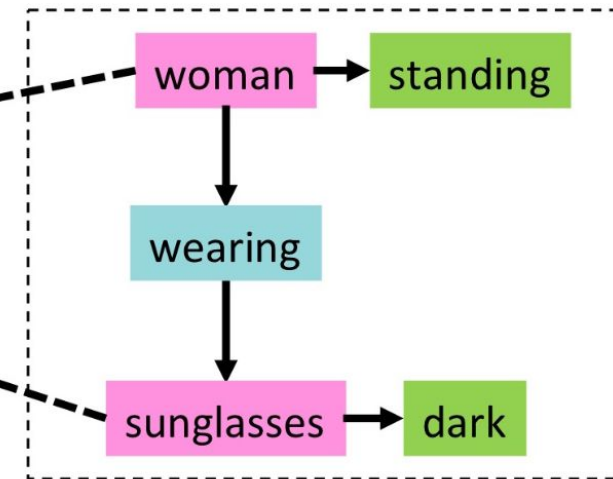


Scene Graph Grounding



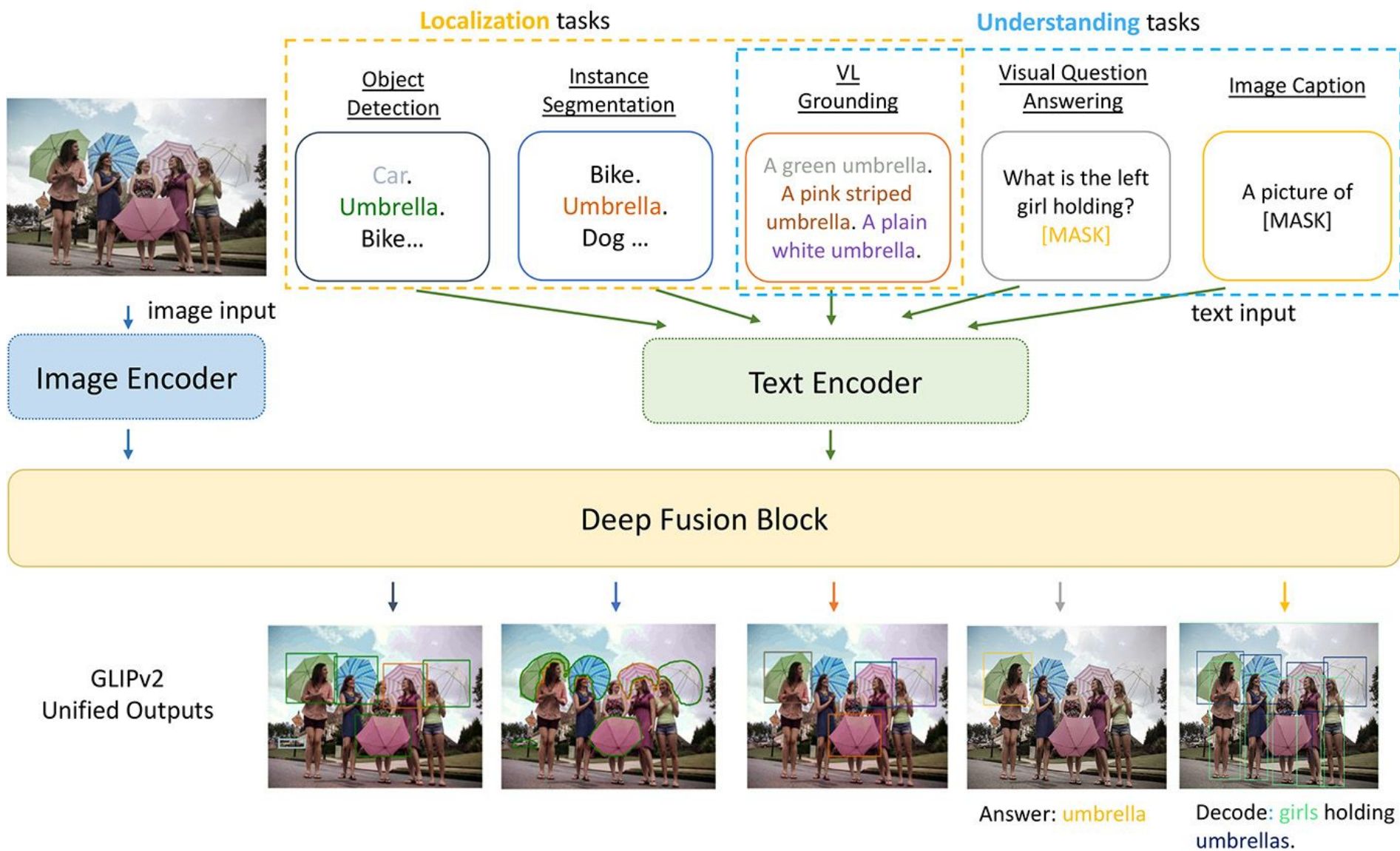
Energy Score: 0.05

Description: “A standing woman wearing dark sunglasses”



Scene Graph

objects attributes relationships



“person holding tray”



“girl in red coat”



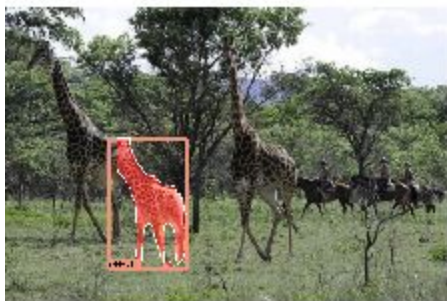
“darkest colored horse”



“lighter brown horse with head down”



“a small giraffe”



“giraffe to the far left”



“the slice of cake on the left”



“chocolate dessert cake on a plate”

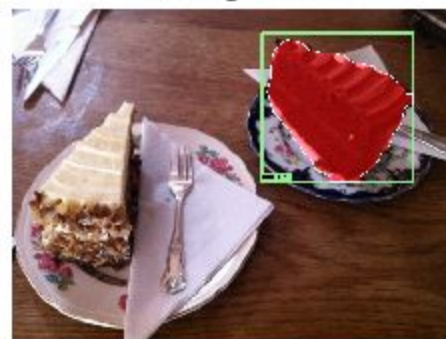
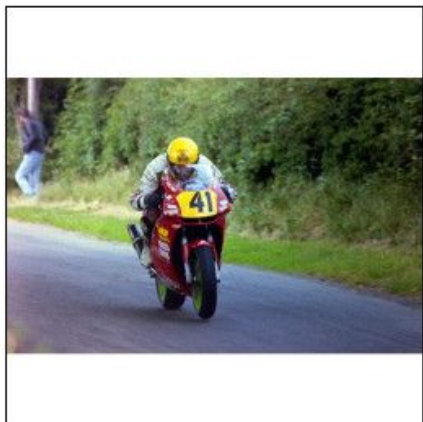


Figure 3: Sample results of objects referred by various query expressions.

Reference image

Relative caption

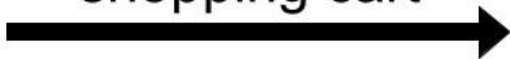
Target Images



is on a track and
has the front
wheel in the air



is shot from the
same angle and
is set inside a
shopping cart



has a dog of a
different breed and
shows a jolly roger





Kiểu kiến trúc Early-fusion

Kiểu kiến trúc này giúp **thông tin của ảnh và ngôn ngữ có thể bổ trợ cho nhau trong quá trình rút trích đặc trưng**. Có thể kể đến các mô hình tiêu biểu như GLIP, UNINEXT, ...

- Ưu điểm:
 - **Tăng tính lý giải của hệ thống** khi sử dụng các mô hình cho bài toán tạo liên kết chính xác giữa ngôn ngữ và hình ảnh/video (visual grounding).
- Nhược điểm
 - Cần yêu cầu **chạy lại toàn bộ mô hình** với mỗi truy vấn khác nhau

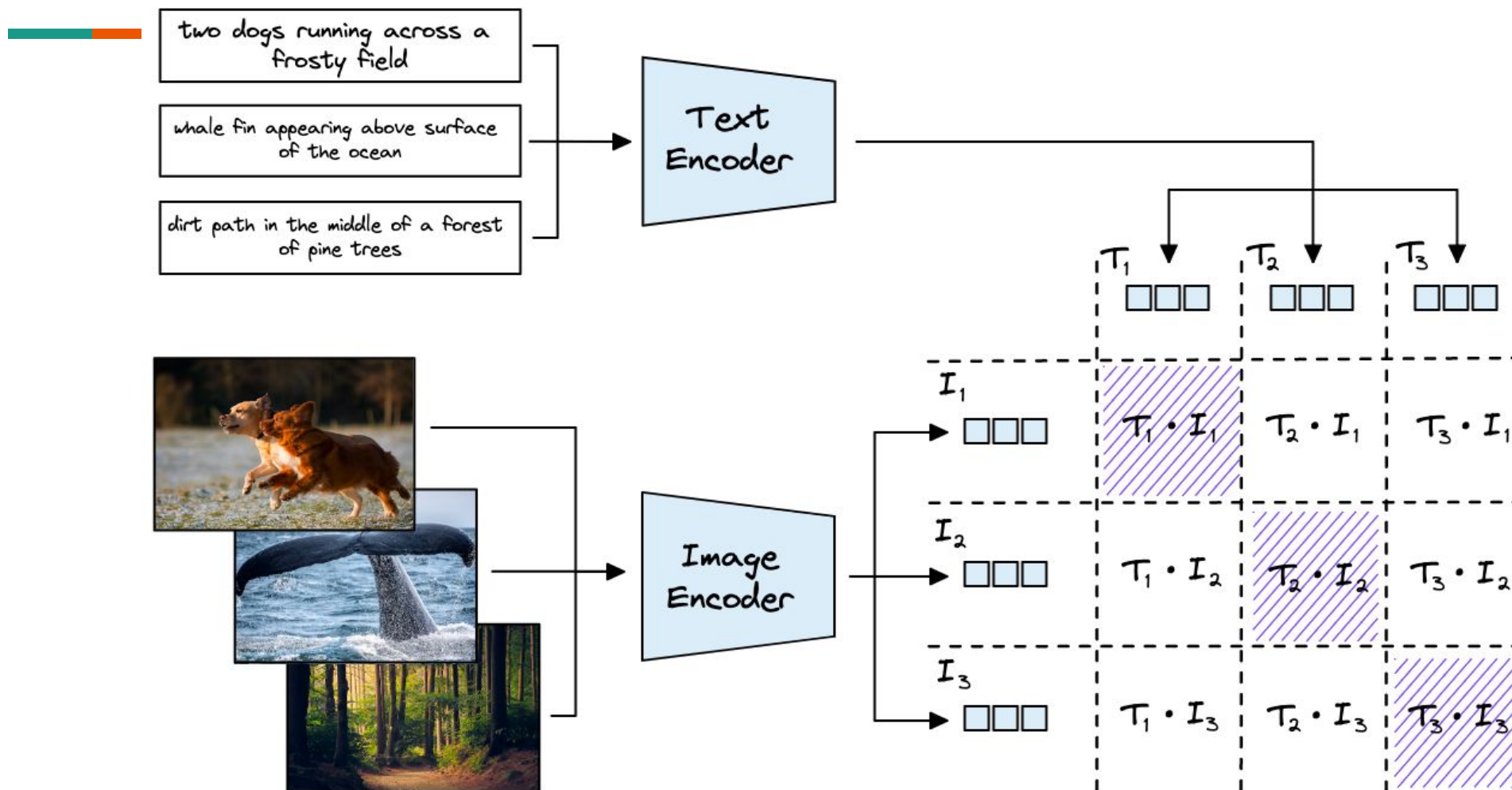
Khi nào cần dùng Early-fusion model

“

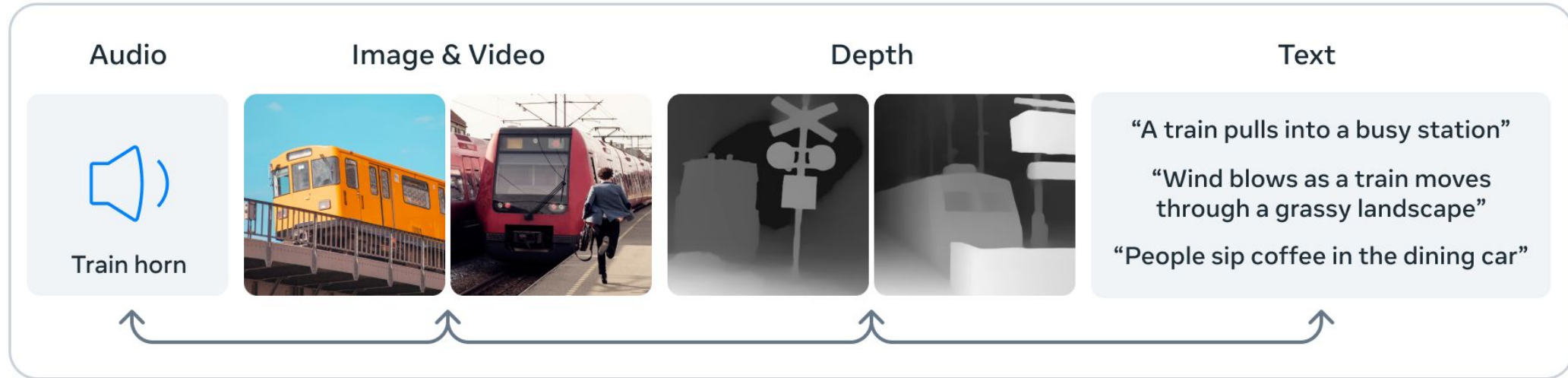
phù hợp chạy với số lượng dữ liệu thấp



các bước cuối cùng của quá trình truy vấn



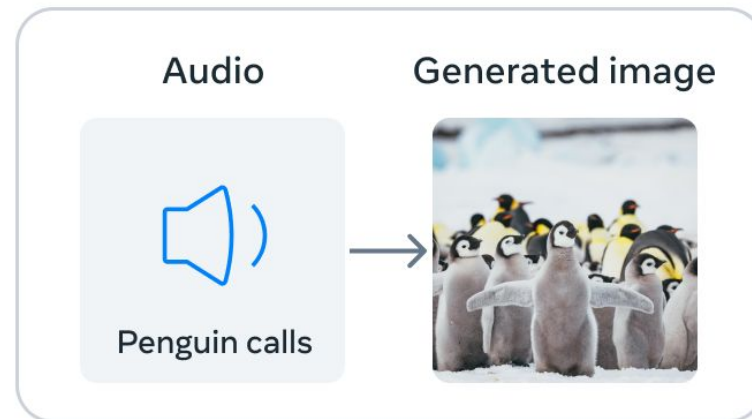
Cross-modal retrieval



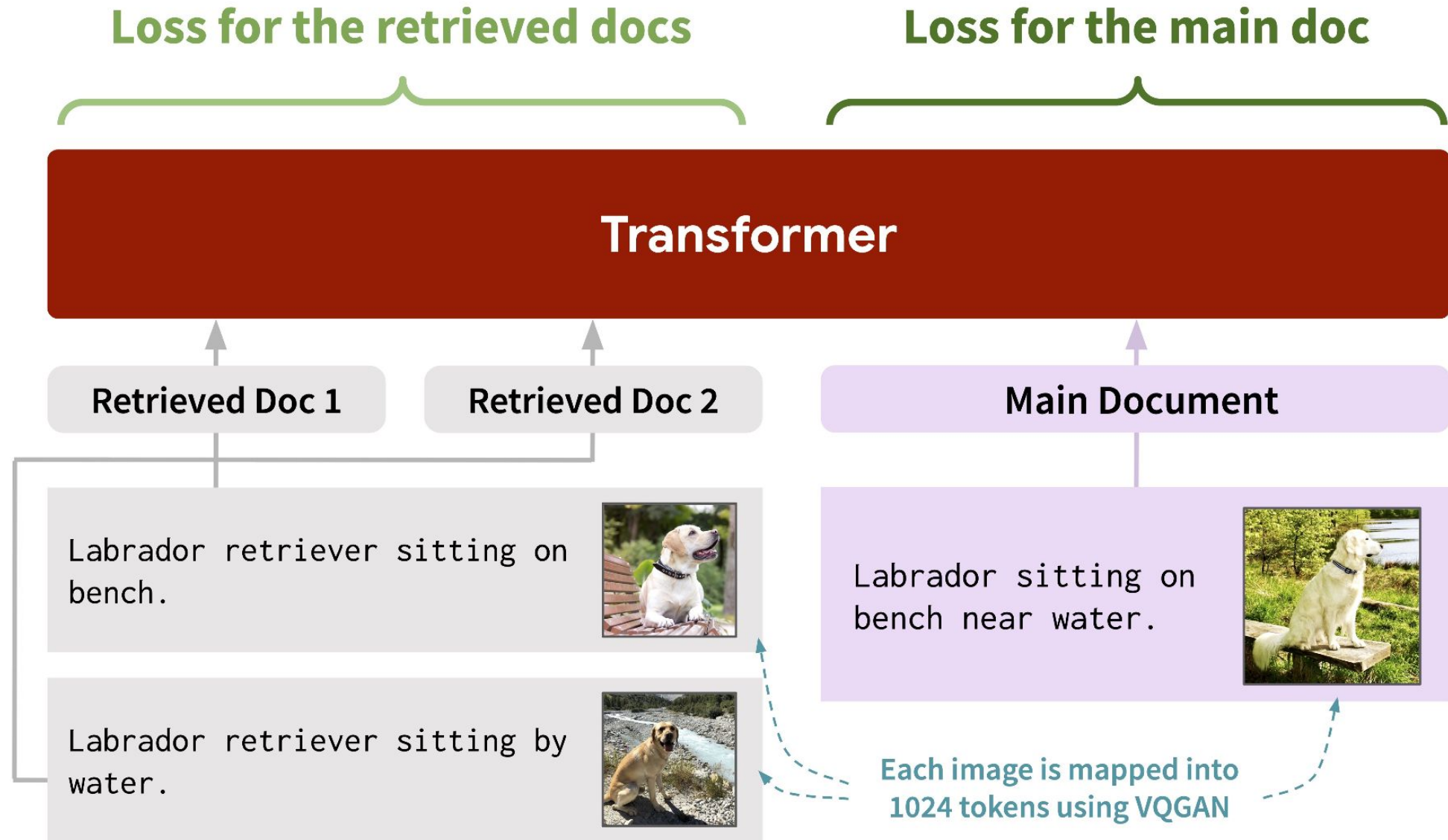
Embedding-space arithmetic

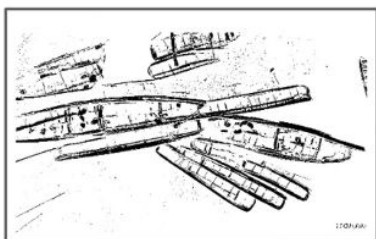


Audio to image generation



Retrieval-augmented Multimodal Generator





Sketch image

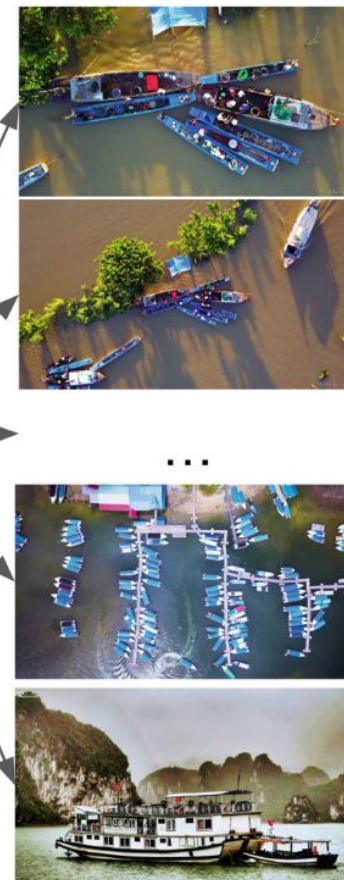
ControlNet



**image
encoder**

query

Database



(Top k images)

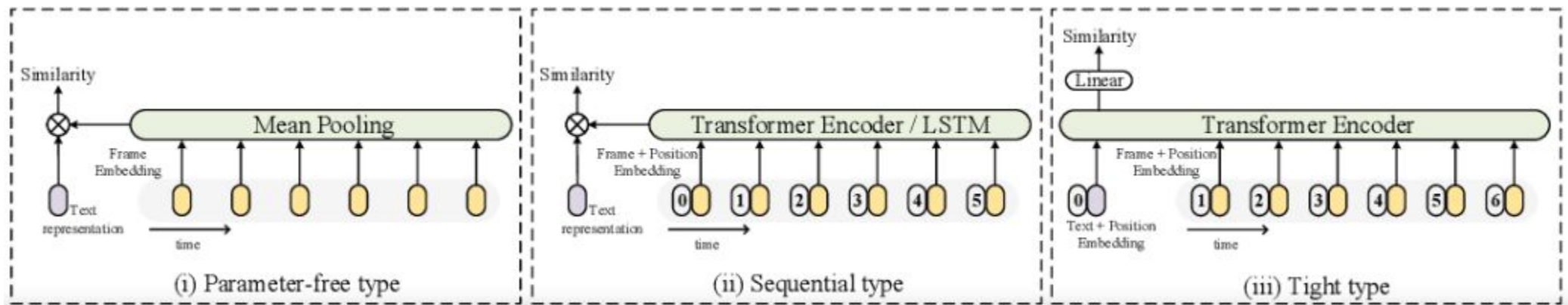
*a group of boats floating on
top of a river, cambodia,
myanmar, vietnam, fishing
boats, 3 boat in river, aerial
footage,*

Image caption

Các cách tiếp cận cơ bản sau khi fusion

Late-fusion **không phải chỉ có tính độ tương đồng** sau khi kết hợp

Các bước sau càng phức tạp cho ra độ chính xác cao, nhưng đổi lại **thời gian chạy lâu** (ở giữa cách tiếp cận early-fusion và late fusion)





Kiểu kiến trúc Late-fusion

Kiểu kiến trúc này hướng đến việc **rút trích đặc trưng từng loại dữ liệu một cách độc lập**. Các mô hình tiêu biểu có thể kể đến như CLIP, OWL-ViT, ...

- Ưu điểm:
 - Tiết kiệm thời gian lúc truy vấn do dữ liệu của ảnh đã được rút trích đặc trưng sẵn.
- Nhược điểm
 - Rất **khó kiểm soát được kết quả trả về** do phụ thuộc hoàn toàn vào sức mạnh của mô hình rút trích đặc trưng.

Khi nào cần dùng Late-fusion model

“

Khi truy vấn trên dữ liệu lớn

Mức **độ phức tạp** ở bước cuối cùng **phụ thuộc vào độ lớn dữ liệu**

Prompt Engineering và Ensembling

“

“zero-shot performance can be significantly improved by customizing the prompt text to each task.”

Alec Radford et. al.

A decorative network pattern in the top-left corner, consisting of interconnected nodes and lines. Some nodes are highlighted with black outlines, and there are several solid black dots scattered within the network.

Hỏi đáp

A decorative network pattern in the bottom-right corner, similar to the one in the top-left, with interconnected nodes and lines, some highlighted with black outlines and solid black dots.