Xây dựng dataset để đánh giá framework với baseline:

|  |  |
| --- | --- |
| Plan A: | Plan B: |
| Sử dụng chính dataset của HCM AI.  - Tận dụng phần keyframes đã được trích xuất ra từ trước, chia thành 3 tập, train, val và test có dạng (videoID-keyframeID).  - Tạo sinh query text từ các keyframes với sự đa dạng về nội dung và độ dài. Các query này sẽ được dùng để đánh giá sự khác biệt về hiệu suất. | Sử dụng MSR-VTT trong domain video captioning. Bộ dataset chứa các cặp (video clips, natural language descriptions).  - Trích xuất keyframes từ các clips, sau đó encode, sinh caption và hashtags giống như quá trình xây dựng database.  - Phần descriptions sẽ được sử dụng để tạo sinh query text từ các keyframes với sự đa dạng về nội dung và độ dài. Các query này sẽ được dùng để đánh giá sự khác biệt về hiệu suất. |
| Số lượng: 726 videos | Số lượng: 10000 videos |
| Ưu điểm:  Nhanh | Ưu điểm:  Phần description phù hợp để tạo sinh query.  Phần video có thể dùng để xây dựng database độc lập với tạo sinh query từ description.  MSR-VTT thường được sử dụng trong các paper về Video Retrieval. |
| Nhược điểm:  Có khả năng bị bias bởi vì quá trình xây dựng database thì tất các các keyframes đã được xử lý để sinh caption và hashtags bằng VLM LLaVa. 🡺 Có thể khắc phục bằng các tạo sinh query bằng một mô hình VLM độc lập khác. | Nhược điểm:  Làm lại từ đầu, tốn tài nguyên. |
| *MSR-VTT: 10,000 videos, 20 categories, 20 English sentences/video by Amazon Mechanical Turks. 29,000 unique words in all captions. 6,513 clips for training, 497 for validation, and 2,990 for testing.*  <https://paperswithcode.com/dataset/msr-vtt> | |
| **Metric**: R@1, R@5, R@10 | |