Transforming Video Search: Leveraging Multimodal Techniques and LLMs for Optimal Retrieval

Binh Dinh, An Dao, Bao Trinh, Truong Dinh and Nguyen Vu

February 11, 2025

Transforming Video Search: Leveraging Multimodal Techniques and LLMs for Optimal Retrieval

Xuan-Binh Dinh-Thi[1¶], An Dao[2¶], Bao-Trinh Quoc[3¶], Truong-Dinh Nhat[3¶], and Hoang-Nguyen Vu[4*]

[1] FPT University
xuanbinh910711@gmail.com
[2] CMC Research Institute for Applied Technology
daoan1412@gmail.com
[3] University of Information and Technology
{22520125,22521575}@gm.uit.edu.vn
[4] AI Lab - AI VIETNAM
nguyenvuhoangwork@gmail.com

Abstract. The rapid growth of online video content has created an urgent need for efficient and accurate event-based video retrieval systems. Existing techniques, such as image-text retrieval, audio analysis, and text-based searches, frequently fail to cope with complex video data and extract useful information from multiple modalities. This paper presents the Multimodal Mapping and Retrieval System (MMRS-LMF), which uses Large Language Models and Multi-Stage Fusion. This novel system improves video retrieval by combining multimodal content (video, audio, and text) into a single, text-based format. It improves retrieval precision and recall by utilizing advanced text embedding techniques and multimodal fusion. The experimental results show significant improvements in retrieval accuracy across a variety of video datasets, demonstrating the system's ability to meet the needs of modern event-based video search applications.

Keywords: multimodal and multimedia retrieval · text-based image retrieval · interactive video retrieval · embedding-based search.

1 Introduction

The increasing expansion of internet video material poses obstacles to effective information retrieval. Content-based video retrieval, which employs textual searches to locate video frames, is an important study topic. As user expectations rise, there is a greater demand for faster, more accurate algorithms to locate specific frames in extensive video libraries.

* Corresponding author; ¶ equal contribution.
This research was partially supported by AI VIETNAM.

Motivated by the growing demand for advanced information retrieval solutions and inspired by prominent international video search competitions such as the Lifelog Search Challenge (LSC) [7,8] and the Video Browser Showdown (VBS), Vietnam founded the AI Challenge Competition, a national-level video search competition. This project necessitates querying events from a collection of around 300 hours that includes over 1,400 news videos. Queries can be used for various tasks, including image-based event descriptions, optical character recognition (OCR), and detecting multi-event interactions across frames.

In this work, we introduce our video search system which participated in the AI Challenge Competition. We improve on the Multi-User Video Search system, which combines embedding-based and text-based search to provide efficient image retrieval. To further improve the system, we add new functions such as image captioning, OCR, and image generating. These enhancements improve search results by extracting content, creating meaningful captions, and displaying potential outcomes. We additionally optimize for repeated user searches on the same query by avoiding overlapping search spaces and breaking inquiries down into smaller chunks, allowing searches around the pivot event across numerous frames.

2 Related Work

Effective video search has long been a difficulty for computer vision and information retrieval. As video content grows, more precise technologies are required. Modern multimedia complexity necessitates advancements beyond traditional keyword-based and object-recognition methods [4].

Tesseract OCR and Levenshtein algorithm offer effective text search inside massive, multilingual datasets, using a Bag of Words (BoW) model and scene-specific features for increased performance, especially with high-resolution photos (T-OCR, BoW). CLIP and interactive query reformulation improve video search accuracy by enabling zero-shot categorization and dynamic keyword refining [14]. The Perfect Match [15] approach enhances video indexing by categorizing visual information with YOLOv5 [9] and CLIP descriptors.

In image captioning, the widespread adoption of deep learning techniques has led to the dominance of sequence learning approaches, which typically combine Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) to generate sentences with flexible syntactic structures [2]. An end-to-end neural model based on Long Short-Term Memory (LSTM) [21] networks was developed to produce descriptive sentences for photographs. This method was further improved by including soft and hard attention processes, which enabled the model to dynamically focus on significant visual regions while creating relevant textual descriptions.

The AI Challenge (AIC) promotes innovation in data retrieval, advancing video search AI. Our research builds on these efforts, offering a framework that improves accuracy and retrieval efficiency in complex digital environments.

3 Data Preprocessing

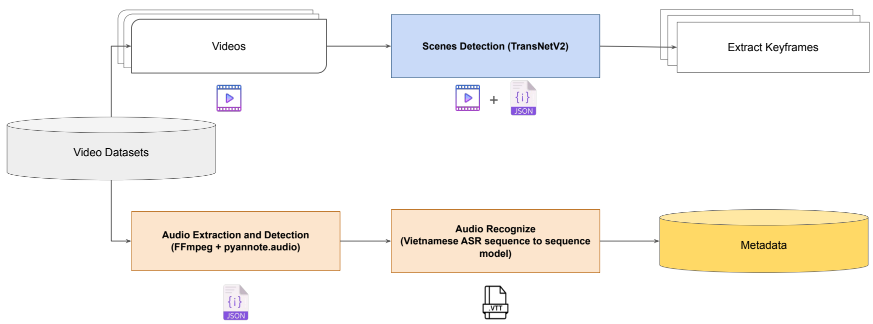

Fig. 1. The diagram shows a video analysis pipeline using TransNetV2 for keyframe extraction and FFmpeg with a Vietnamese ASR model for audio recognition, supporting object detection and semantic searches across video and audio data.

3.1 Video Preprocessing

We present a preprocessing system for a large-scale dataset of over 300 hours of diverse video footage. Due to its size and variability, an exhaustive search method is optimal. Our approach divides the video data into coherent scene components and extracts keyframes, denoted as J, where J_i refers to the keyframe at index i. Keyframes are extracted using the TransNetV2 model. For each video segment defined by frame indices $[p, j]$, we extract four keyframes, denoted as x_{frame} according to the following formula:

$$x_{\text{frame}} = Y_{(p+[j*(q-p)/3)])}, \forall j \in (0, 1, 2, 3) \tag{1}$$

where Y represent the keyframes extracted from a video segment, where p and q are the start and end frame indices, respectively, and $j \in 0, 1, 2, 3$ indicates each keyframe [20]. The segment between p and q is divided into three equal parts, selecting four keyframes. After extraction, a noise filter is applied to refine the final dataset.

3.2 Speech Recognition and Text Extraction from Video Data

This activity has two steps: audio extraction , detection and audio recognition. **Step 1: Audio Extraction and Detection:** We use FFmpeg to extract audio from videos, saving it as .wav for speech analysis. Pyannote.audio [3] detects speech segments, marking start and end times for transcription synchronization. **Step 2: Audio Recognition:** Detected segments are processed by a Vietnamese ASR model, designed for multi-talker scenarios ([18]), using WavLM [5] as the encoder and Bart-decoder (base) as the decoder.

4 Multi Modal Retrieval

Building on VISIONE [1], we create a text-based encoding framework that integrates object detection, color recognition, image tagging, captioning, image generation, and Google image crawling, advancing multimodal AI capabilities.

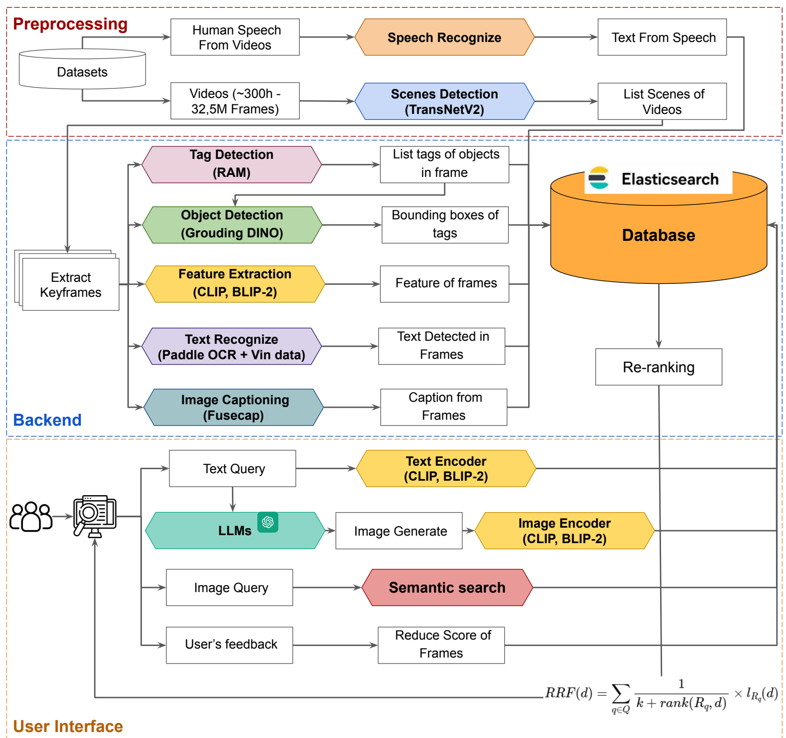

Fig. 2. The architecture comprises three main components: data management, retrieval system, and user interface. It facilitates effective multi-modal querying through fast data processing and model integration, enhancing accuracy and scalability for extensive video collections. The data management optimizes indexing and storage, while the retrieval system applies machine learning for complex searches. The user interface enables intuitive interactions for search refinement and quick results.

4.1 Feature Extraction

DFN5B-CLIP-ViT-H-14-384 [6], in conjunction with BLIP-2 [11], is a cutting-edge method for multimodal feature extraction. Conventional approaches frequently fail to capture fine-grained semantic links and involve substantial computing overhead. When processing large, high-dimensional datasets, our suggested CLIP variation performs better in terms of semantic alignment and computing efficiency. By capitalizing on cutting-edge multimodal learning breakthroughs, BLIP-2 integration improves text-visual coherence and domain adaptability.

4.2 Image Captioning

Image captioning is key to our system, enhancing the linguistic richness of images and aligning descriptions with user intent. We use the FuseCap framework [19], based on the BLIP model, which integrates outputs from vision modules (object detector, attribute recognizer, OCR) with LLM-generated captions, providing comprehensive and contextually accurate visual descriptions.

Algorithm 1 Keyframe Description Generation

Require: Keyframes $\{K_i\}$
Ensure: Descriptions $\{D_i\}$
Initialize model M, set threshold $T_s = 70\%$
for each keyframe K_i **do**
    $D_i \leftarrow M(K_i)$ if $i = 1$ or similarity$(K_i, K_{i-1}) \leq T_s$, else $D_i \leftarrow D_{i-1}$
    Save D_i
end for
return $\{D_i\}$

TransNetV2's keyframe creation technique results in overlaps and increases costs. To improve efficiency, we repeat captions with more than 70% similarity. Otherwise, FuseCap creates new captions for top-k searches using BLIP [19]. This strategy is effective for time series inquiries but not for single frame queries, and we used it in a competition for Vision Question Answering (VQA) assignments that required chronological data.

4.3 Optical Character Recognition

This study employs the open-source Paddle OCR framework for Vietnamese scene-text detection in two stages: detection and recognition. The detection phase combines ResNet50 with Differentiable Binarization (DB++) [12], while the recognition phase uses PP-OCRv3 [10] with the SVTR architecture for higher accuracy. Both models are fine-tuned with the VinText dataset [17] to improve real-world Vietnamese scene-text recognition, enhancing robustness and accuracy.

4.4 Tags and Objects Detection

The RAM model [22] analyzes semantic image information, such as destination, to create tags. Tags and confidence scores are encoded to ensure clarity. The

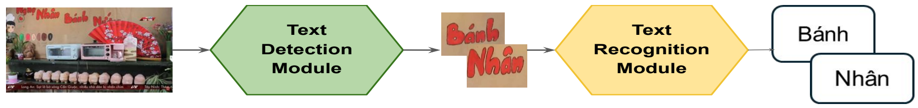

Fig. 3. End-to-end OCR pipeline for Vietnamese scene-text recognition, featuring text detection to localize regions and text recognition to extract and decode content.

method uses the Grounding DINO model [13], well-known for its high performance in object detection and reference expression interpretation, to forecast bounding boxes for image tags. It comprises a feature improvement component, linguistic question selection, and a multimodal decoder that improves linguistic integration over standard detectors.

5 System Overview

5.1 Data Management

Elasticsearch is the primary data management platform, allowing for quick text and image searches. Custom indexes contain text, OCR results, image descriptions, and dense vectors. A Vietnamese analyzer enhances text search accuracy by folding ASCII characters and removing stopwords. It enables efficient hybrid searches by utilizing fuzzy matching for text and HNSW [16] for vector searches.

5.2 Retrieval system

Embedding-based Searching. Our system offers two search methods: text query and image query. In text queries, users can enhance their search performance by alternating between the feature extraction models presented in the section 4.1. This transformation will take advantage of each model. In image queries, users can select any image from the web interface or image generated by LLMs and search for images similar to it by semantic search; the similarity calculation is done using HNSW algorithm.

Image Generation. Our experiments showed that CLIP and BLIP-2 struggled with time- and event-specific queries, limiting contextual understanding. To improve this, we integrated the GPT API for semantic analysis and image generation, enhancing complex query handling, image search accuracy, and bridging the gap between English comprehension and image retrieval.

Users Interactions. Using user queries to filter frames improves data quality but risks losing valuable information. To mitigate this, we implemented a feedback mechanism allowing users to adjust frame rankings. This retains less critical frames while excluding noisy ones. Details on the ranking process are in Section. 5.3.

Multi Modal Retrieval. To prevent cross-referencing in multi-model queries, we use Reciprocal Rank Fusion (RRF) to combine fields like tags, OCR results, and image captions for improved accuracy. RRF merges individual retrievals, each using specific methods (e.g., text search or HNSW for vectors), into a single ranking by calculating a final score based on ranks across results using the following formula:

$$\mathrm{RRF}(d) = \sum_{i=1}^{N} \frac{1}{k + r_i(d)} \cdot I(d \in \mathrm{result}(q_i)) \tag{2}$$

$\mathrm{RRF}(d)$ calculates the score for document d, based on its performance across N queries. Here, $r_i(d)$ represents d's rank in the results of the i-th query, with i ranging from 1 to N. The constant k stabilizes rankings, while $I(d \in \mathrm{result}(q_i))$ is 1 if d appears in query i's results, and 0 otherwise. This formula rewards documents that consistently achieve higher rankings across multiple queries.

5.3 System Usage and Features

The system's user interface is organized into three main components: A) Search Input and Configuration, B) Results Display and Interaction, and C) Filtered Results and Search History Management. Each component serves a specific function and works together to enhance user interaction and improve retrieval efficiency in the video search workflow.

Search Input and Configuration The primary input interface allows users to search in Vietnamese and English using natural language queries. It features a configurable picture reference panel for similarity searches, and filtering techniques like OCR and ASR to refine results based on text or speech in movies. Users can adjust the number of results, use clustering to remove duplicates, and access a large language model for image production, along with a search bar for quick image ID searches.

Results Display and Interaction The results display and interaction section enables visitors to interact with search results in four ways: selecting frames for similarity-based searches, playing videos to see the context of each frame, copying or sharing frame IDs, and exploring neighboring frames for better sequential understanding. The integrated video player allows for full replay and accurate frame movement, allowing for detailed content study and quick assessment and comparison of results.

Filtered Results and Search History Management The interface features a repository for filtered frames and search history, critical for organizing iterative searches and ensuring transparency throughout the process. Users may save filtered frames for rapid retrieval, providing a complete snapshot of their search progress. The design prioritizes efficiency and usability, creating a smooth experience across all components.

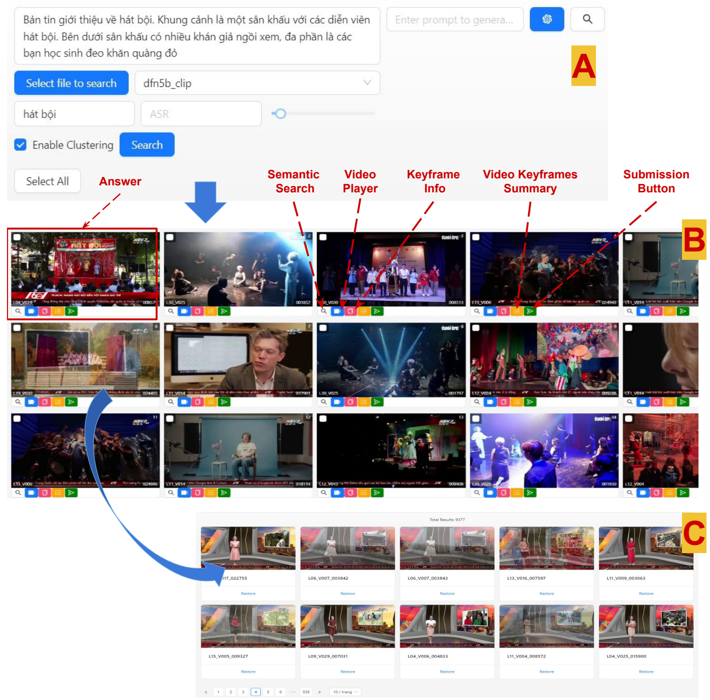

Fig. 4. The interface has three main parts: Part A lets users input data via text, generate images from a large language model, and search images by ID using a search bar. Part B displays results, allowing users to review and select answers, with features like a video player and nearest frame display for detailed checks. Part C stores noisy keyframes, enabling users to restore filtered keyframes.

6 Experimental Results

The experimental evaluation of the proposed system in Textual Known-Item Search (KIS) emphasized its multilingual capability, retrieval precision, and filter effectiveness. Using the AI Challenge 2024 private dataset with Vietnamese and English queries, the study tested the system's flexibility and accuracy with models like CLIP and BLIP-2, and advanced filters such as OCR and ASR. Results demonstrated the system's ability to deliver accurate, diverse outcomes across a wide range of search criteria.

The first query, titled "A shot of the cockpit of a vehicle moving at sea in blue," describes a cockpit scene with a control screen displaying values between 210 and 212. The initial search through the first thirty frames yielded no results. However, applying the OCR filter from the search input significantly improved accuracy, bringing the target frame to the top. Additionally, using frame tags in the search input enhances retrieval precision. For short video clips, incorporating specific frame-level keywords helps position the target image among the top

Fig. 5. Comparison of retrieval methods for the query: 'A shot of the cockpit of a vehicle moving at sea in blue.' Vision-Language Embedding retrieves the target at index 63, adding Semantic Tagging improves to index 61, while our method achieves the best result at index 32.

ten results. Enabling clustering before the search further enhances retrieval by ensuring diversity in similar keyframes.

Upon execution, the system ranks results using the formula outlined in Section 5.2. Users can iteratively remove noisy frames and re-query to refine the ranking accuracy if the desired image is not found. Irrelevant frames are permanently removed from the database, as shown in the Filtered Results and Search History Management component.

Table 1. Performance comparison of different methods based on accuracy and score

Method	Accuracy	Score
Vision-Language Embedding	0.84	43.6
Vision-Language Embedding + Semantic Tagging	0.97	56.6
Ours	**0.97**	**58.2**

The score function evaluates the rank r_i of the target image's position within the top-k results for each query, where k corresponds to the milestone images defined by five specific thresholds. This function quantifies cumulative performance across multiple preset thresholds, comprehensively evaluating the ranking system's effectiveness. The detailed methodology is as follows:

$$\text{score} = \frac{\sum_k \text{top}(r_i, k)}{5}, \quad \text{where } \text{top}(r_i, k) = \begin{cases} 1, & \text{if correct}(r_i) = 1 \text{ and } i \leq k, \\ & k \in \{1, 5, 20, 50, 100\} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

7 Conclusion

This work presented an improved video search system designed for large-scale content-based retrieval, integrating features like image captioning, OCR, and image generation for more accurate and relevant results. The system effectively manages repeated user searches by optimizing query decomposition. Future work will focus on refining multimodal integration, enhancing real-time search capabilities, and adapting to diverse languages. We also aim to incorporate user feedback for more personalized and robust retrieval.

References

1. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: Visione 5.0: Enhanced user interface and ai models for vbs2024. In: International Conference on Multimedia Modeling. pp. 332–339. Springer (2024)
2. Bhatnagar, P., Mrunaal, S., Kamnure, S.: Enhancing image captioning with neural models (2023)
3. Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., Gill, M.P.: pyannote.audio: neural building blocks for speaker diarization (2019)
4. Carós, M., Garolera, M., Radeva, P., i Nieto, X.G.: Automatic reminiscence therapy for dementia. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. pp. 383–387. ICMR '20, ACM, New York, NY, USA (2020). https://doi.org/10.1145/3372278.3391927
5. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing **16**(6), 1505–1518 (Oct 2022). https://doi.org/10.1109/jstsp.2022.3188113
6. Fang, A., Jose, A.M., Jain, A., Schmidt, L., Toshev, A., Shankar, V.: Data filtering networks (2023)
7. Gurrin, C., Jónsson, B.T., Schöffmann, K., Dang-Nguyen, D.T., Lokoč, J., Tran, M.T., Hürst, W., Rossetto, L., Healy, G.: Introduction to the fourth annual lifelog search challenge, lsc'21. In: Proceedings of the 2021 International Conference on Multimedia Retrieval. p. 690–691. ICMR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3460426.3470945
8. Gurrin, C., Le, T.K., Ninh, V.T., Dang-Nguyen, D.T., Jónsson, B.T., Lokoč, J., Hürst, W., Tran, M.T., Schöffmann, K.: Introduction to the third annual lifelog search challenge (lsc'20). In: Proceedings of the 2020 International Conference on

Multimedia Retrieval. p. 584–585. ICMR '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3372278.3388043
9. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., TaoXie, Michael, K., Fang, J., imyhxy, Lorna, Wong, C., Yifu), V, A., Montes, D., Wang, Z., Fati, C., Nadar, J., Laughing, xylieong: ultralytics/yolov5: v6.2 - yolov5 classification models, apple m1, reproducibility, clearml and deci.ai integrations (Aug 2022). https://doi.org/10.5281/zenodo.7002879
10. Li, C., Liu, W., Guo, R., Yin, X., Jiang, K., Du, Y., Du, Y., Zhu, L., Lai, B., Hu, X., Yu, D., Ma, Y.: Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system (2022)
11. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023)
12. Liao, M., Zou, Z., Wan, Z., Yao, C., Bai, X.: Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
13. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection (2023)
14. Lokoč, J., Vopálková, Z., Dokoupil, P., Peška, L.: Video search with clip and interactive text query reformulation. In: Dang-Nguyen, D.T., Gurrin, C., Larson, M., Smeaton, A.F., Rudinac, S., Dao, M.S., Trattner, C., Chen, P. (eds.) MultiMedia Modeling. pp. 628–633. Springer International Publishing, Cham (2023)
15. Lubos, S., Rubino, M., Tautschnig, C., Tautschnig, M., Wen, B., Schoeffmann, K., Felfernig, A.: Perfect match in video retrieval. In: Dang-Nguyen, D.T., Gurrin, C., Larson, M., Smeaton, A.F., Rudinac, S., Dao, M.S., Trattner, C., Chen, P. (eds.) MultiMedia Modeling. pp. 634–639. Springer International Publishing, Cham (2023)
16. Malkov, Y.A., Yashunin, D.A.: Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs (2018)
17. Nguyen, N., Nguyen, T., Tran, V., Tran, T., Ngo, T., Nguyen, T., Hoai, M.: Dictionary-guided scene text recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
18. Nguyen, T.B., Waibel, A.: Synthetic conversations improve multi-talker asr. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 10461–10465 (2024). https://doi.org/10.1109/ICASSP48485.2024.10446589
19. Rotstein, N., Bensaïd, D., Brody, S., Ganz, R., Kimmel, R.: Fusecap: Leveraging large language models for enriched fused image captions. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5677–5688 (2024). https://doi.org/10.1109/WACV57701.2024.00559
20. Trinh Xuan, K., Nguyen Khoi, N., Luong-Quang, H., Hoa-Xuan, S., Nguyen-Luong-Nam, A., An, M.H., Nguyen, H.P.: Multi-user video search: Bridging the gap between text and embedding queries. In: Proceedings of the 12th International Symposium on Information and Communication Technology. p. 923–930. SOICT '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3628797.3628957
21. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator (2015)
22. Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., Guo, Y., Zhang, L.: Recognize anything: A strong image tagging model (2023)