VieXplor: Towards an Interactive System for Vietnamese Multimodal Video Retrieval via Adaptive Query Understanding and Temporal Coherence

Trong-Nghia Nguyen[1, 3], **Gia-Minh Vo**[2, 3], **Tuan-Khanh Dao**[1, 3], **Thien-Nhan Nguyen**[1, 3], **Tinh-Anh Nguyen Nhu**[2, 3]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Ho Chi Minh City University of Technology, Ho Chi Minh City, Vietnam
[3]Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam
24521148@gm.uit.edu.vn, minh.vozamin@hcmut.edu.vn, 24520778@gm.uit.edu.vn, 24521235@gm.uit.edu.vn, anh.nguyennhu2306@hcmut.edu.vn

Abstract

The massive growth of video content and increasing complexity of user needs demand retrieval systems that surpass traditional limits. Current systems face three main challenges: (1) incomplete feature extraction, where keyframe sampling misses crucial moments and standard OCR struggles with Vietnamese scene text; (2) a semantic gap between user queries and video content; and (3) limited ability to reason over temporal event sequences. To address these, we introduce VieXplor, a comprehensive multimodal event retrieval system built on BEiT-3. It features three core innovations: (1) a hybrid keyframe extraction pipeline combining TransNetV2, CLIP, and Vintern-1B-V3.5 for accurate Vietnamese scene-text recognition with Proximal Refinement; (2) The Feedback-Informed Refinement module, which uses a GPT-5-powered Query Enhancement component, adapts to evolving user intent through feedback while reordering the results based on their similarity to the keyframes selected in the previous round; and (3) the BeamTRAKE algorithm, which ensures temporal consistency across video segments. By integrating advanced foundation models with specialized retrieval strategies, the system effectively bridges semantic, visual, and temporal understanding in large-scale video data.

Introduction

The rapid expansion of video content across digital platforms poses substantial challenges for information retrieval systems. Competitions such as Video Browser Showdown (VBS) (Rossetto et al. 2025), Lifelog Search Challenge (LSC) (Tran et al. 2025) and Ho Chi Minh City AI Challenge (HCMC AI Challenge) (Tang et al. 2025) have become key benchmarks for video retrieval innovation. Effective retrieval in this context requires models that can jointly reason over visual scenes, embedded text, and temporal information while maintaining contextual understanding at scale. In Vietnamese contexts, this challenge is further amplified by the complexity of the language and the diverse environmental conditions.

Despite recent progress in multimodal retrieval, existing systems remain limited by their inability to capture semantic, linguistic, and temporal coherence across video data. Traditional OCR-based methods for Vietnamese perform poorly when dealing with scene text that is curved, occluded, or distorted-conditions common in rural or mountainous regions (Nguyen et al. 2023). These methods rely primarily on pixel-level features and lack the ability to infer semantic meaning. Moreover, conventional retrieval systems tend to depend on exact lexical matching between queries and indexed text, leading to failures when users employ synonyms or complex phrasing (Rao et al. 2019).

Furthermore, existing keyframe-based video retrieval approaches further exacerbate this issue by treating frames or short clips independently, producing fragmented and often temporally incoherent results that fail to capture the continuous nature of real-world events. Such methods typically rely on static frame-level representations or aggregate global embeddings, which overlook the temporal dependencies and event transitions that are crucial for understanding video semantics (Gia et al. 2025). To address this problem, the HCMC AI Challenge has proposed the Temporal Retrieval and Alignment of Key Events (TRAKE) task as a more fine-grained and context-aware formulation of keyframe-based video retrieval. Instead of retrieving entire videos based solely on visual similarity, TRAKE focuses on locating and aligning semantically meaningful key events within a video timeline that correspond to a user's textual or multimodal query. By explicitly modeling temporal structure, event boundaries, and causal relationships, TRAKE enables retrieval systems to reason over when and how relevant actions occur, rather than merely identifying what appears in isolated frames. This temporal alignment not only enhances retrieval precision, but also lays the foundation for interactive and explainable video understanding systems, where users can navigate, visualize, and interact with event-level representations in a coherent and interpretable manner.

To overcome these limitations, we propose VieXplor, a context-aware multimodal video retrieval framework that integrates vision-language understanding and temporal reasoning. At its core, Vintern-1B-V3.5 (Doan et al. 2024), a large-scale Vietnamese vision-language model, replaces conventional OCR to allow conceptual text inference under challenging visual distortions. A Feedback-Informed Refinement module, powered by GPT-5, dynamically expands and reformulates user queries through semantic vari-

ants derived from relevance feedback and identifies additional frames similar to those selected in previous stages, improving retrieval robustness and precision. To preserve narrative coherence, the BeamTRAKE algorithm sequentially selects frames from the same video based on temporal order and relevance, maintaining consistency across the retrieved segments.

Our main contributions are as follows:

- **VieXplor:** A unified multimodal retrieval framework that integrates BEiT-3, TransNetV2, CLIP, Vintern-1B-V3.5, and the Proximal Refinement component for robust scene-text understanding and context-aware video retrieval.
- **Feedback-Informed Refinement:** We propose a GPT-5–driven mechanism that leverages user feedback to iteratively expand and refine search queries while also identifying additional frames similar to previously selected ones, thus enhancing retrieval accuracy.
- **BeamTRAKE:** This is a temporal retrieval algorithm that maintains both contextual and narrative continuity by sequentially linking the relevant frames within videos.

Related Work

Modern video retrieval systems commonly rely on keyframe extraction to improve computational efficiency and reduce redundancy in large-scale datasets. Early methods, such as distributed retrieval using color and texture descriptors (Patel 2012), depend heavily on low-level visual cues and treat each frame as an independent entity. Consequently, these techniques often fail to capture subtle yet semantically meaningful variations, especially in videos with gradual scene transitions, object movements, or lighting changes. To overcome such limitations, recent multimodal retrieval frameworks have introduced advanced modules for scene boundary detection, semantic filtering, and cross-modal feature fusion, allowing more context-aware understanding of video content.

For instance, TransNetV2 (Souček and Lokoč 2020) has been employed to achieve accurate scene boundary detection, ensuring that extracted keyframes represent meaningful transitions rather than arbitrary cuts. Meanwhile, CLIP (Radford et al. 2021) leverages vision-language alignment to facilitate semantic similarity-based frame selection, enabling retrieval systems to focus on frames that carry the most representative and contextually rich information. Furthermore, BEiT-3 (Wang et al. 2022) serves as a powerful vision-language encoder that enhances cross-modal alignment, ensuring that textual and visual representations are jointly optimized for retrieval accuracy. The integration of a proximal retrieval mechanism further helps preserve contextual relationships between frames, which are often lost during aggressive frame reduction.

In terms of scene-text understanding, conventional OCR models such as ASTER (Shi et al. 2019), Rosetta (Borisyuk, Gordo, and Sivakumar 2018), and CRNN (Shi, Bai, and Yao 2015) typically perform poorly on Vietnamese text, especially under real-world conditions involving poor lighting, cluttered backgrounds, or distorted typography. To address these challenges, the introduction of Vintern-1B-V3.5, a large-scale Vietnamese vision-language model, has marked a major advancement. By incorporating contextual semantics and leveraging large-scale multimodal training, Vintern-1B-V3.5 enables context-aware recognition of visual scenes and text, effectively handling curved, blurred, or occluded text that traditional OCR systems often fail to interpret.

Another key challenge lies in bridging the semantic gap between user queries and multimedia content. Traditional retrieval techniques, including static word embeddings such as Word2Vec (Mikolov et al. 2013) or Pseudo-Relevance Feedback (Pan et al. 2022), lack adaptability and struggle to dynamically align with user intent. In contrast, GPT-5 introduces a new paradigm for interactive retrieval by generating semantically equivalent query variants, refining them iteratively based on user feedback, and re-ranking retrieved results according to their semantic similarity to previously selected frames. This dynamic query optimization process substantially enhances retrieval relevance, precision, and robustness.

Lastly, while traditional frame selection methods such as Bag of Visual Words (BoVW) (Gidaris et al. 2020) efficiently compress video data, they often disregard temporal coherence, resulting in fragmented or contextually inconsistent retrieval outputs. The BeamTRAKE algorithm overcomes this issue by enforcing both temporal and contextual consistency: it selects an initial frame most relevant to the query, retrieves subsequent frames in chronological order, and ranks candidate sequences based on their average semantic similarity while pruning inconsistent frames. This design ensures coherent event-level retrieval and preserves the narrative flow of the video, effectively complementing multimodal and semantic retrieval modules for a more unified and meaningful retrieval experience.

Data Preprocessing

Data processing constitutes a critical component of our retrieval pipeline, focusing on the extraction, encoding, and storage of relevant visual and textual features to enable efficient search and retrieval. The process is divided into three main stages: Frame Collection, Visual Feature Extraction, and Scene-Text Extraction.

Frame Collection

The frame collection process begins with scene segmentation using TransNetV2 to divide raw videos into coherent shots. Within each scene, frames are extracted at 5 frames per second (fps) and stored in a comprehensive Frames Database, which serves as an exhaustive repository for later neighbor-based retrieval. To construct an efficient index for initial retrieval, we apply a redundancy reduction step using CLIP ViT-B/32 to filter visually similar frames. The first frame in each sequence is selected as an anchor, and subsequent frames are compared against it, frames exceeding a predefined similarity threshold are discarded, while dissimilar ones replace the anchor. This procedure substantially reduces computational and storage overhead while preserving semantic diversity among representative keyframes, which are subsequently used for feature extraction.

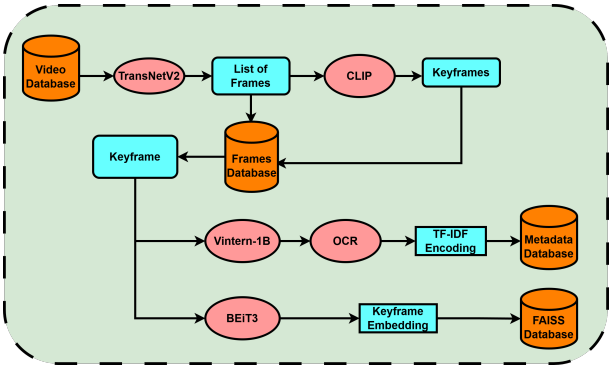

Figure 1: Data preprocessing pipeline: Videos are segmented into scenes using TransNetV2 and sampled at 5 fps. CLIP ViT-B/32 filters highly similar frames to produce representative keyframes, which are encoded by BEiT-3 for visual semantics and Vintern-1B-V3.5 for scene-text extraction. All embeddings are indexed with FAISS and stored for efficient multimodal retrieval.

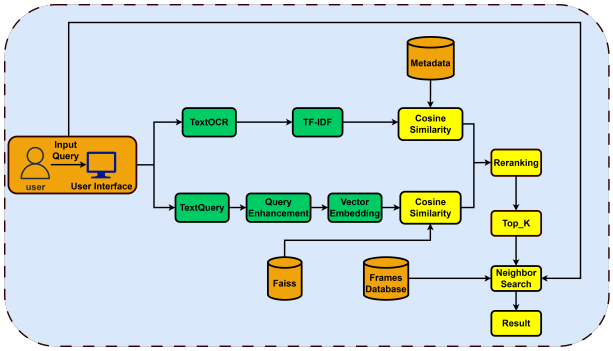

Figure 2: The system flowchart illustrates the hybrid retrieval and refinement process of VieXplor, which adopts a dual-branch architecture combining semantic search (vector embedding) and keyword search (TextOCR). Their results are fused in a Reranking module, followed by a Neighbor-based Search stage that refines outputs and compensates for keyframe sampling gaps by analyzing the complete frame database.

Visual Feature Extraction

To improve query understanding and support efficient retrieval, we employ BEiT-3 as the primary visual encoder. The model transforms keyframes into dense embedding vectors that capture high-level semantic information. A major advantage of BEiT-3 lies in its ability to learn strong cross-modal alignment between textual descriptions and their corresponding visual semantics, allowing for a more accurate interpretation of visual content in multimodal contexts. After feature extraction, all visual embeddings are indexed using FAISS, allowing for Approximate Nearest Neighbor (ANN) search to be performed efficiently at large scale.

Scene-Text Extraction

In addition to visual semantics, we incorporate a scene-text extraction module to enhance retrieval accuracy and support the reranking mechanism. This process plays a crucial role in identifying textual features embedded within video frames. We employ the Vintern-1B-V3.5 Vision–Language Model for scene text recognition. Taking advantage of large-scale vision–language pretraining, Vintern-1B-V3.5 demonstrates strong robustness in handling curved, occluded, or distorted text, particularly in Vietnamese linguistic contexts where diacritics and complex typography are prevalent. The extracted text is subsequently encoded using TF-IDF, and the results are stored in a metadata database. This textual representation serves as a complementary modality, facilitating multimodal retrieval and improving overall system robustness.

System Overview

Our system is designed for efficient and accurate multimodal event retrieval from large-scale video datasets. It integrates advanced techniques for diverse query handling, user feedback incorporation, and embedding-based refinement. This section details the main components: the search process, feedback-informed refinement and TRAKE-specific optimizations.

VieXplor System

Embedding-Based Search: When a user submits a textual query, the system employs BEiT-3, a powerful multimodal foundation model, to encode the input into a unified embedding space. This process produces a query vector that preserves the semantic meaning of the original input, enabling direct comparison between modalities. The resulting query embedding is then matched against a FAISS-based index, which stores precomputed embeddings of representative keyframes extracted from the entire video corpus. Retrieval is performed using cosine similarity, returning the top-k candidate frames whose visual-semantic representations most closely align with the query.

OCR-Based Reranking: Although the top-k candidates obtained from the initial retrieval stage exhibit high semantic relevance, they may not always represent the most contextually accurate results. To refine retrieval precision, we introduce an OCR-based reranking mechanism that leverages textual metadata extracted directly from video frames. In parallel with embedding extraction, the Vintern-1B-V3.5 model - optimized for the Vietnamese language - is employed to detect and extract any textual content as mentioned in Section .

For each text-based query, the system computes two complementary similarity scores for every candidate frame:

- **QueryScore:** derived from cosine similarity between the query embedding and the keyframe embedding, measuring semantic alignment in the multimodal space; and
- **OCRScore:** obtained by representing both the user query and the extracted scene text as TF-IDF vectors and calculating their dot product, capturing lexical overlap.

The final ranking score is computed as a weighted combination of the two components:

$$FinalScore = (\lambda \times QueryScore) + (\beta \times OCRScore) \quad (1)$$

Through empirical evaluation, the optimal weights were determined to be $\lambda = 0.6$ and $\beta = 0.4$, providing a balanced trade-off between visual-semantic similarity and textual relevance.

Proximal Refinement: The proximal search refinement module addresses a key limitation of keyframes sampling: important frames may lie between selected keyframes and thus be missed. After Stage 2 identifies the highest-scoring frame and the user confirms the most relevant candidate, ViXplor re-examines the original video within a narrow temporal window (±20 frames). All frames in this range are re-encoded with BEiT-3 and re-ranked by semantic similarity. This refinement ensures retrieval of the most contextually relevant frame, even if it was initially excluded during keyframe filtering.

Feedback-Informed Refinement

User queries are often ambiguous, contain typographical or grammatical errors, or include unnecessary verbosity, making it difficult for the model to identify the main intent. To address this, we employ a Large Language Model (LLM) for query enhancement. Specifically, GPT-5 processes the original query to generate N semantically refined variants that emphasize the **main action, context, and key identifying elements**. During this stage, GPT-5 also standardizes spelling and grammar while removing noisy characters. This approach improves retrieval performance by: (1) diversifying query expressions to broaden semantic coverage, (2) highlighting salient features that better align with visual representations, and (3) producing cleaner, more coherent queries that reduce linguistic noise.

As illustrated in **Figure 3**, the workflow consists of two stages:

- **Enhancement Stage** – The system receives the user's query along with a tuned prompt; GPT-5 generates N semantically diverse query variants in parallel based on user's feedback.
- **Selection Stage** – The system presents these N candidates for the user to select or merge, forming the final enhanced query, which is then forwarded to the retrieval module.

After each retrieval round, the retrieved results are re-ranked based on their similarity to the user-selected images from the previous iteration. Specifically, the selected candidates are encoded into embedding space and used as reference vectors to guide the ranking of new results based on cosine similarity, ensuring that subsequent retrievals remain aligned with user preferences. To effectively combine the textual and visual relevance signals, the final similarity score for each retrieved frame i is computed as:

$$Score(i) = \alpha \cdot Sim_{query}(i) + (1 - \alpha) \cdot Sim_{frame}(i) \quad (2)$$

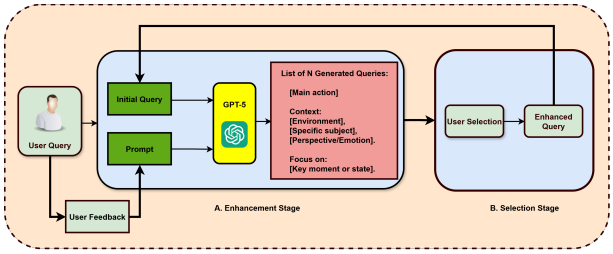

Figure 3: Overview of the two-stage Query Enhancement workflow: (A) GPT-5 generates N semantically diverse query variants, and (B) the user selects the most contextually appropriate query for retrieval.

where $Sim_{query}(i)$ denotes the similarity between frame i and the current query, and $Sim_{frame}(i)$ represents the average similarity between frame i and the user-selected frames from the previous iteration, computed as:

$$Sim_{frame}(i) = \frac{1}{|F_{sel}|} \sum_{j \in F_{sel}} Sim(i, j) \quad (3)$$

Here, $\alpha \in [0, 1]$ balances textual and visual relevance, with $\alpha = 0.7$ empirically providing the best performance. Meanwhile, the query itself is refined through user feedback, allowing GPT-5 to adjust its semantic focus toward the most relevant aspects of the user's intent. This iterative feedback loop continuously improves the retrieval precision by jointly adapting both the textual query and the visual relevance model over iterations.

BeamTRAKE

In the TRAKE task, the objective is to retrieve a sequence of temporally continuous events that belong to the same video. A naive approach—where each textual query describing an individual scene is processed independently to retrieve its most relevant frame—introduces a critical problem. Specifically, the frame with the highest similarity score for Query 1 might originate from Video A, while the best match for Query 2 could come from Video B. This result-scattering phenomenon disrupts the contextual and temporal coherence of the retrieved sequence, rendering it inconsistent with the original narrative flow of the video. To address this limitation and ensure intra-video consistency, we propose BeamTRAKE, a temporally aware retrieval algorithm designed to preserve contextual continuity within a single video.

In BeamTRAKE algorithm (**Figure 3**), retrieval begins by selecting the top-K keyframes with the highest semantic similarity to the text query. Subsequent frames are limited to those from the same video and occurring after the previously selected timestamps. When multiple keyframes originate from one video, the earliest frame serves as the temporal anchor, ensuring that the retrieved sequence follows the event's natural progression.

For each subsequent query, the system searches within this constrained space for the top K×2 candidate frames, ranking them by semantic similarity. Each beam (i.e., candidate frame sequence) is expanded by selecting the two most

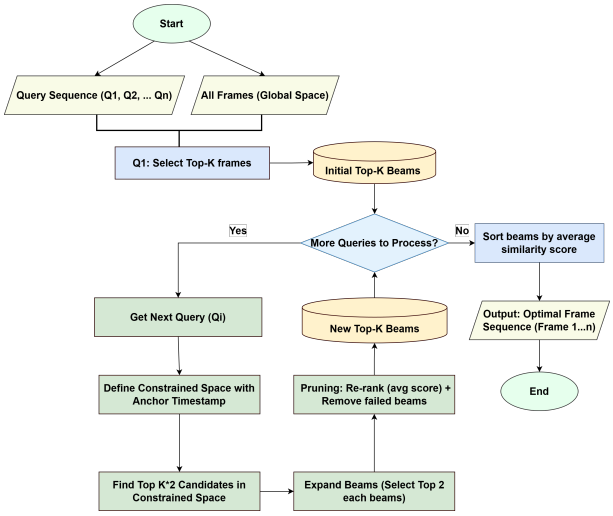

Figure 4: Overview of the BeamTRAKE algorithm. The method retrieves temporally coherent frame sequences by progressively expanding and pruning candidate beams across multiple query steps, ensuring that all selected frames maintain contextual consistency throughout the retrieval process.

similar timestamp-constrained frames, then re-ranked by the average similarity across the sequence to maintain semantic and temporal consistency. Beams violating alignment constraints - failing to obtain a sufficient number of valid candidate frames due to the selection stage - are pruned, and only the top-K beams with the highest aggregate scores are retained for the next iteration.

Through this iterative refinement, BeamTRAKE jointly optimizes semantic relevance and temporal coherence, mitigating inter-video scattering and yielding retrievals that preserve the chronological and contextual flow of the original event.

Experiments

Experimental Setup

Dataset: The HCMC AI Challenge 2025, a competition focused on event retrieval in videos in Vietnam, provides a benchmark dataset consisting of short news videos with durations ranging from approximately 15 to 20 minutes each. In this work, we utilize this benchmark dataset to evaluate the effectiveness of our proposed video retrieval and temporal alignment framework, enabling a standardized and realistic assessment under Vietnamese multimedia conditions.

Overall UI

Section A: As illustrated in **Figure 5**, the panel A serves as the query configuration area of the VieXplor system, where users can select the retrieval mode (KIS, QA, or TRAKE), input a textual query, and adjust parameters such as the Top-K value. It also supports OCR-based reranking by incorporating textual content extracted from frames. Once configured, the *Search* button initiates the retrieval process, providing flexible control over query formulation and reranking.

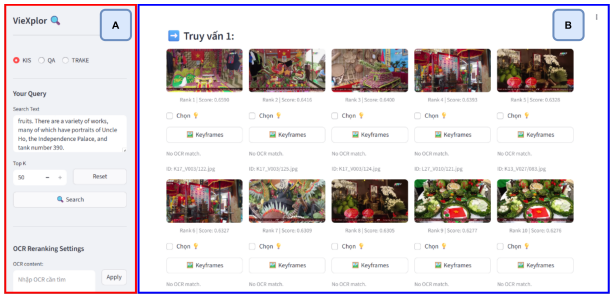

Figure 5: User interface of the VieXplor system, showing the query control panel (A) and the ranked result display (B) with interactive options for refinement and proximal search.

Section B: This panel displays the ranked retrieval results with representative frames, ranks, scores, and identifiers. Users can select frames via the checkbox to perform reranking in subsequent iterations or use the *Keyframes* button to activate the Proximal Search mechanism, retrieving nearby frames within the same video to enhance contextual coherence. These interactions together enable an iterative, context-aware retrieval process integrating textual, visual, and OCR information.

OCR Reranking and Proximal Refinement

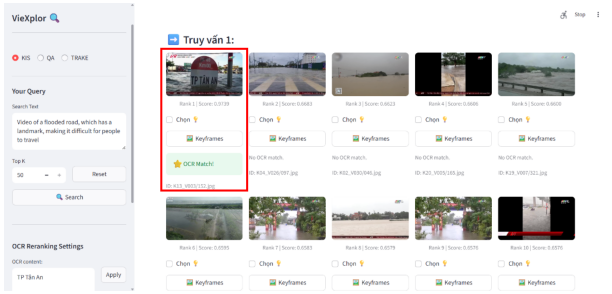

Figure 6: OCR Reranking: refining retrieval by reprioritizing candidates matching user-provided OCR text cues.

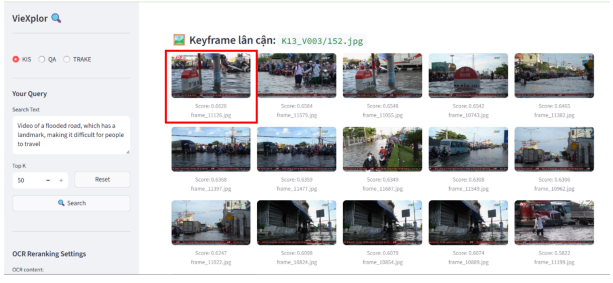

Figure 7: Proximal Search: expanding retrieval by exploring temporally adjacent frames for contextual coherence.

Our system incorporates two key refinement modules - OCR Reranking and Proximal Refinement - to enhance multimodal retrieval beyond the initial embedding-based search. The OCR Reranking module, demonstrated in **Figure 6**, acts as an intelligent post-filter: after the initial semantic retrieval (e.g., "flooded road and people struggling to travel"), users can input additional textual cues (e.g., "TP Tân An") to reprioritize candidates containing matching OCR-extracted text, improving alignment between visual and textual contexts.

As shown in **Figure 7**, the Proximal Refinement module, activated via the "Keyframes" button, mitigates sparse keyframe sampling. Once a candidate is selected, the system retrieves temporally adjacent frames, re-encodes them with the BEiT-3 model, and re-ranks them against the query. This process captures subtle yet crucial frames, yielding more coherent and semantically precise retrieval results.

Feedback-Informed Refinement Usage

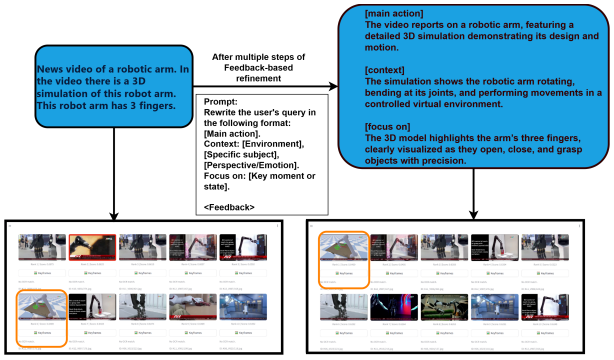

Figure 8: An illustration of structured query enhancement: The raw multi-clause query is refined by GPT-5, significantly improving retrieval rank for the 'robotic arm' ground truth.

Firstly, our initial verbose query — "News video of a robotic arm. In the video there is a 3D simulation of this robot arm. This robot arm has 3 fingers." — yielded suboptimal results, as its multi-clause structure diluted key semantic cues and caused the ground truth to rank low. Following this initial retrieval, we provided feedback to GPT-5 while simultaneously selecting the most relevant keyframes from the retrieved candidates. Guided by this feedback, GPT-5 generated refined, structured query variants that emphasized the core object (robotic arm) and its distinctive attribute (three fingers) within the specified context (3D simulation). Upon re-querying, combined with a re-ranking mechanism based on similarity to the previously selected keyframes, the ground truth was retrieved at rank 1, demonstrating the effectiveness of our feedback-driven query enhancement and similarity-based reranking strategy. It is shown in **Figure 8**.

BeamTRAKE Usage

For the TRAKE task, the system sequentially processes a user-defined sequence of text queries, where each query describes a distinct event. In the provided example, a 3-query

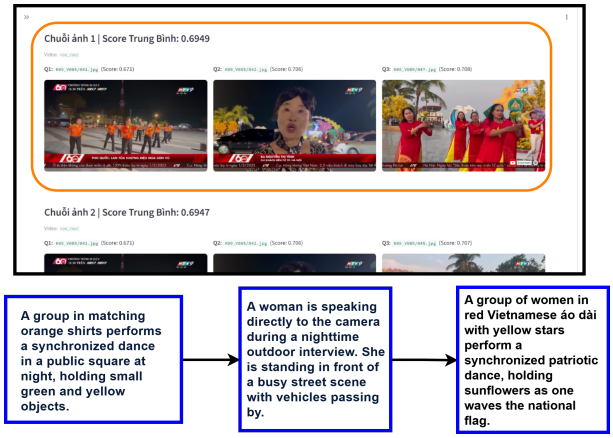

Figure 9: Illustration of a successful TRAKE query. A multi-query sequence (Q1, Q2, Q3) retrieves a single, temporally consistent video ('Best Result') ranked by its average score.

sequence was used to identify a specific scenario, as shown in **Figure 9**.

Upon submission, the system processes the query sequence and displays a ranked list of "Image Sequences". Each sequence represents a single candidate video and is ranked by an "Average Score," which reflects the overall fit of the entire sequence to the user's queries. The interface clearly displays the extracted frames matching each respective query. This allows the user to immediately identify a single video containing the correct events in the desired temporal order by examining the high-scoring sequences.

Conclusion

Recent advancements in event retrieval research have promoted continuous innovation and further benchmarked multimodal understanding using diverse, real-world datasets, particularly through complex tasks like Temporal Retrieval and Alignment of Key Events (TRAKE). Contributing to this growing progress, we present VieXplor, a comprehensive multimodal retrieval system designed to tackle three key challenges: (1) inefficient keyframe sampling and weak scene-text extraction, (2) semantic disparity between user queries and corresponding video content, and (3) lack of temporal consistency in event retrieval and alignment.

VieXplor integrates TransNetV2, CLIP, and Vintern-1B-V3.5 to achieve robust and context-aware visual–textual representation, while a GPT-5-based query expansion module and the BeamTRAKE algorithm further enhance semantic alignment, temporal reasoning, and overall retrieval coherence. VieXplor effectively demonstrates the strength and flexibility of combining advanced pretrained models with structured retrieval mechanisms, underscoring artificial intelligence's growing potential in large-scale media analysis, digital education, and knowledge discovery. This work reinforces the importance of integrating multimodal understanding, temporal structure, and semantic intelligence for the next generation of intelligent retrieval systems, paving the way for practical deployments in the near future.

References

Borisyuk, F.; Gordo, A.; and Sivakumar, V. 2018. Rosetta: Large Scale System for Text Detection and Recognition in Images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '18, 71–79. ACM.

Doan, K. T.; Huynh, B. G.; Hoang, D. T.; Pham, T. D.; Pham, N. H.; Nguyen, Q. T. M.; Vo, B. Q.; and Hoang, S. N. 2024. Vintern-1B: An Efficient Multimodal Large Language Model for Vietnamese. arXiv:2408.12480.

Gia, B. T.; Khanh, T. B. C.; Thanh, T. L. T.; Tran, K.; Trong, H. H.; Doan, T. T.; Le, K.; Do, T.; Le, D.-D.; and Ngo, T. D. 2025. Addressing Ambiguous Queries in Video Retrieval with Advanced Temporal Search. In Buntine, W.; Fjeld, M.; Tran, T.; Tran, M.-T.; Huynh Thi Thanh, B.; and Miyoshi, T., eds., *Information and Communication Technology*, 167–180. Singapore: Springer Nature Singapore. ISBN 978-981-96-4291-5.

Gidaris, S.; Bursuc, A.; Komodakis, N.; Pérez, P.; and Cord, M. 2020. Learning Representations by Predicting Bags of Visual Words. arXiv:2002.12247.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Nguyen, H.; Ta, C.-H.; Le-Nguyen, P.-T.; Tran, M.-T.; and Le, T.-N. 2023. Ensemble Learning for Vietnamese Scene Text Spotting in Urban Environments. In *2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 177–182. IEEE.

Pan, M.; Liu, Y.; Pei, Q.; Mao, H.; Jin, A.; Huang, S.; and Yang, Y. 2022. A Multi-Dimensional Semantic Pseudo-Relevance Feedback Information Retrieval Model. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 866–872.

Patel, B. V. 2012. Content Based Video Retrieval Systems. *International Journal of UbiComp*, 3(2): 13–30.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Rao, J.; Liu, L.; Tay, Y.; Yang, W.; Shi, P.; and Lin, J. 2019. Bridging the Gap between Relevance Matching and Semantic Matching for Short Text Similarity Modeling. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5370–5381. Hong Kong, China: Association for Computational Linguistics.

Rossetto, L.; Schoeffmann, K.; Gurrin, C.; Lokoč, J.; and Bailer, W. 2025. Results of the 2025 Video Browser Showdown. arXiv:2509.12000.

Shi, B.; Bai, X.; and Yao, C. 2015. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. arXiv:1507.05717.

Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2019. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9): 2035–2048.

Souček, T.; and Lokoč, J. 2020. TransNet V2: An effective deep network architecture for fast shot transition detection. arXiv:2008.04838.

Tang, Z.; Wang, S.; Anastasiu, D. C.; Chang, M.-C.; Sharma, A.; Kong, Q.; Kobori, N.; Gochoo, M.; Batnasan, G.; Otgonbold, M.-E.; Alnajjar, F.; Hsieh, J.-W.; Kornuta, T.; Li, X.; Zhao, Y.; Zhang, H.; Radhakrishnan, S.; Jain, A.; Kumar, R.; Murali, V. N.; Wang, Y.; Pusegaonkar, S. S.; Wang, Y.; Biswas, S.; Wu, X.; Zheng, Z.; Chakraborty, P.; and Chellappa, R. 2025. The 9th AI City Challenge. arXiv:2508.13564.

Tran, A.; Bailer, W.; Dang-Nguyen, D.-T.; Healy, G.; Hodges, S.; ór Jónsson, B.; Rossetto, L.; Schoeffmann, K.; Tran, M.-T.; Vadicamo, L.; and Gurrin, C. 2025. The State-of-the-Art in Lifelog Retrieval: A Review of Progress at the ACM Lifelog Search Challenge Workshop 2022-24. arXiv:2506.06743.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; and Wei, F. 2022. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. arXiv:2208.10442.