21st IEEE International Conference on
Advanced Visual and Signal-Based Systems

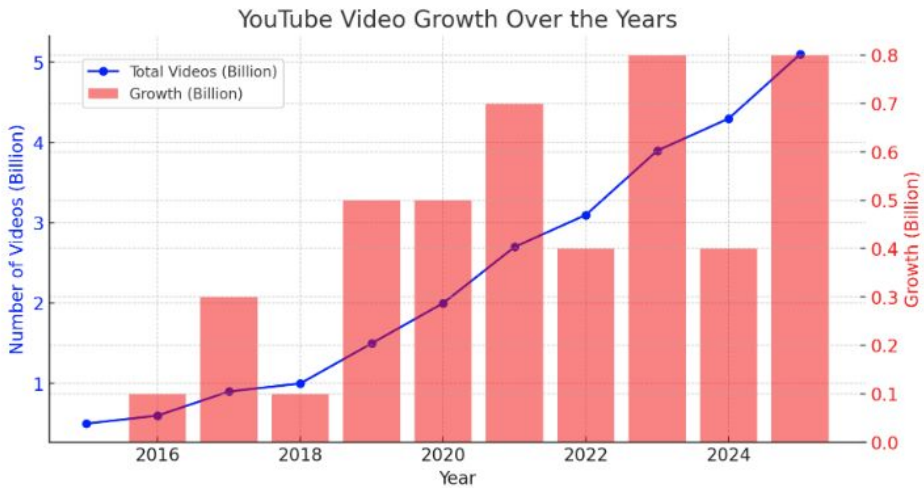
Fine-Grained Video Indexing and Retrieval with Vision-Language Models

Dong Gun Park[1], Soyeon Park[1], Chang Ha Lee[2], Hyunkyoo Choi[3], Charmgil Hong[1]
{**systec24**, eu2goo, charmgil}@handong.ac.kr, {hkchoi}@kisti.re.kr,{yielding}@gmdsoft.com
[1]Handong Global University
[2]GMDSOFT
[3]Korea Institute of Science and Technology Information

Content
s

1. Motivation

- **KUBiC (Korean Unification Bigdata Center)**
  - Search engine specialized on information related to North Korea
- Rapid growth of video content online
  - Number of videos hosted on YouTube stands at 5.1 billion in 2025, more than doubled in just a few years (2.2B in 2022) [SEO.ai Content Team., 2025]
- Video = Audio + Visuals
- What if we can pinpoint exact scenes/content inside a video?



SEO.ai Content Team. (2025, February 15). *How many videos are on YouTube?*

1. Motivation

- Early video indexing frameworks
  - Rely on basic visual cues (color, motion) [Hu et al., 2011]
  - Lack of high-level semantic understanding [Hu et al., 2011]
- Video indexing approaches that utilize AI
  - Restricted to retrieving a short portion of the entire video
- Need for an advanced indexing method that:
  - Capture high-level semantic meaning,
  - Enable accurate, content-based search and retrieval for the entire length of the video

Hu et al. (2011). *A Survey on Visual Content-Based Video Indexing and Retrieval.*

1. Motivation

- Early video indexing frameworks
  - Rely on basic visual cues (color, motion) [Hu et al., 2011]
  - Lack of high-level semantic understanding [Hu et al., 2011]
- Video indexing approaches that utilize AI
  - Restricted to retrieving a short portion of the entire video
- Need for an advanced indexing method that:
  - Capture high-level semantic meaning,
  - Enable accurate, content-based search and retrieval for the entire length of the video

We propose a new video indexing framework that utilizes AI techniques to transform raw video content into structured, searchable data

Hu et al. (2011). *A Survey on Visual Content-Based Video Indexing and Retrieval.*

2. Our Framework

- **"A unified multimodal indexing architecture that transforms raw video content into structured, searchable data"**
  - Audio transcription
  - Visual scene analysis
- **Goal**: Accurate, efficient, and fine-grained timestamp-level retrieval

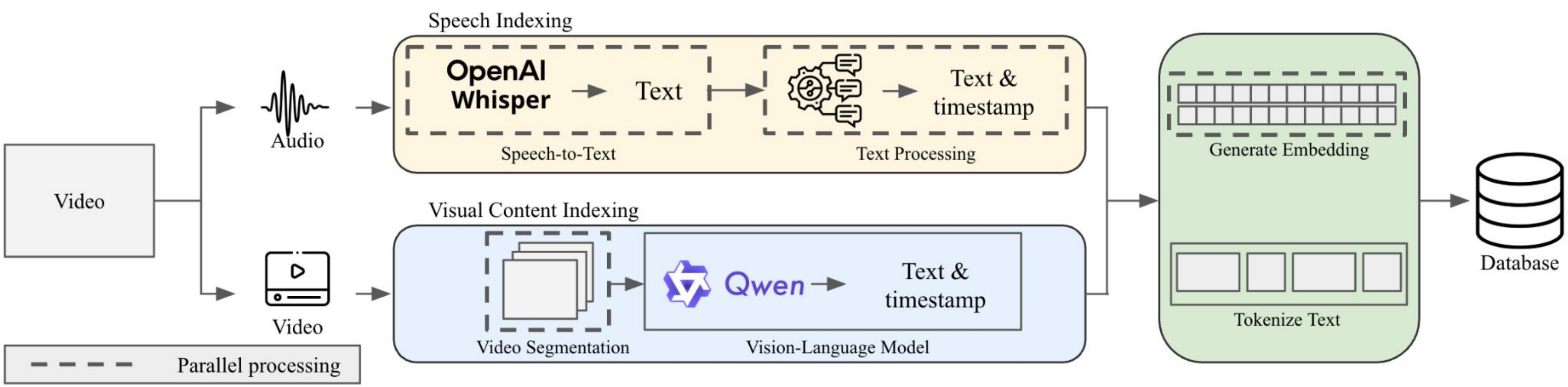


Figure 1. Video Indexing Flow

2. Our Framework

- 2.1. Speech Processing and Indexing
- 2.2. Visual Content Analysis and Indexing
- 2.3. Semantic Retrieval Strategy

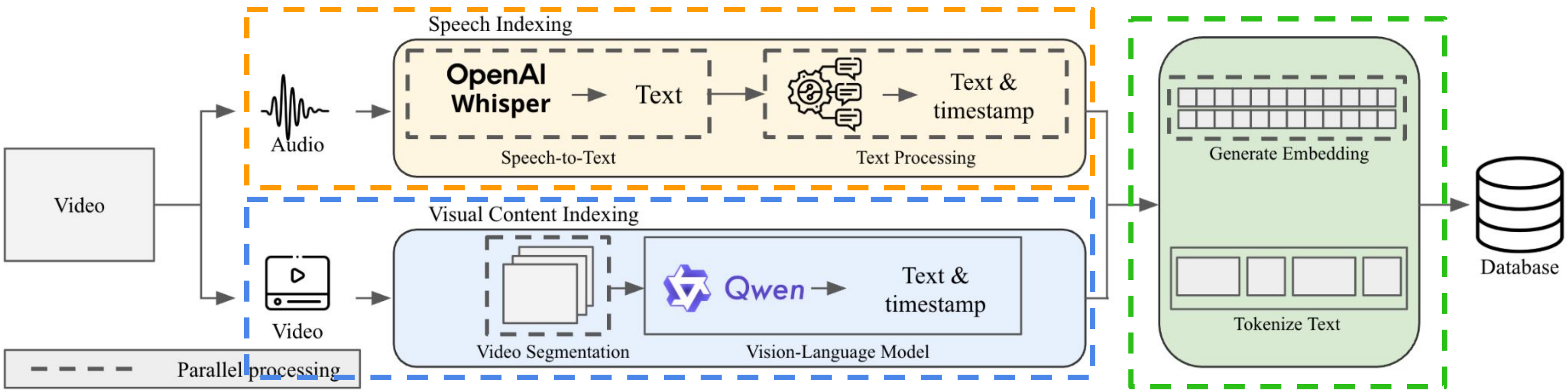
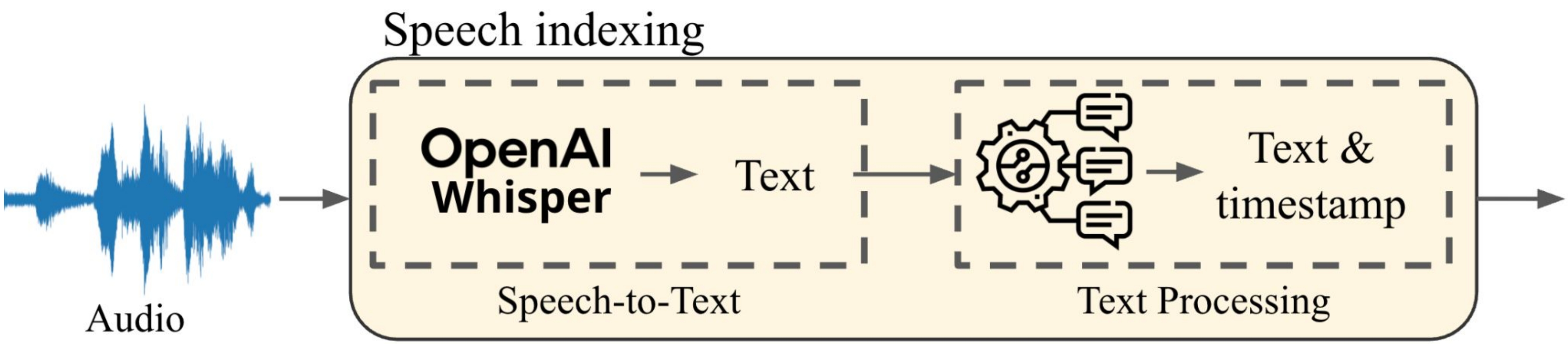


Figure 1. Video Indexing Flow

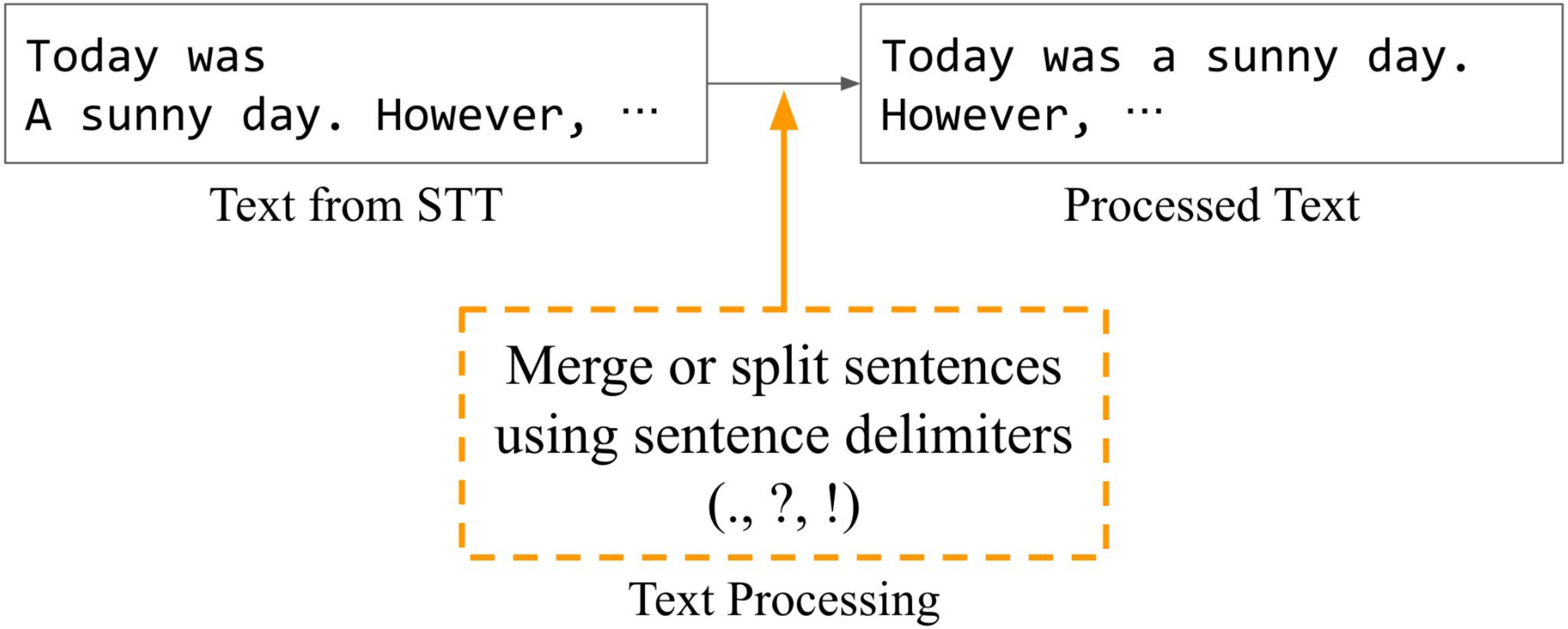
2.1. Speech Processing and Indexing

- Audio extraction from video
- Transcribe audio into text using speech-to-text modules
- Transcribed text exist in short phrases / series of sentences
  - Need to merge/split into single whole sentences
  - Split using sentence delimiters (period, question mark, exclamation mark)



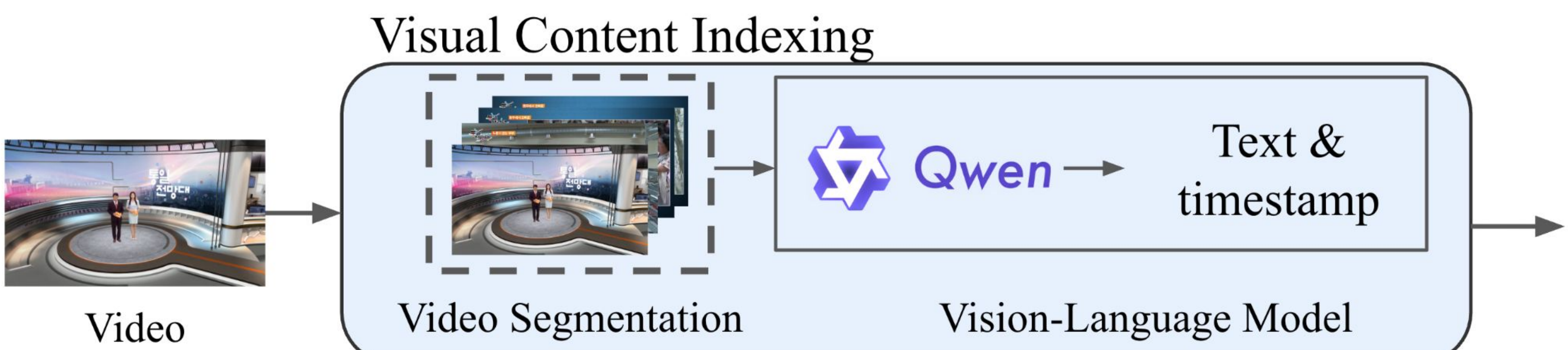
2.1. Speech Processing and Indexing

- Audio extraction from video
- Transcribe audio into text using speech-to-text modules
- Transcribed text exist in short phrases / series of sentences
  - Need to merge/split into single whole sentences
  - Split using sentence delimiters (period, question mark, exclamation mark)



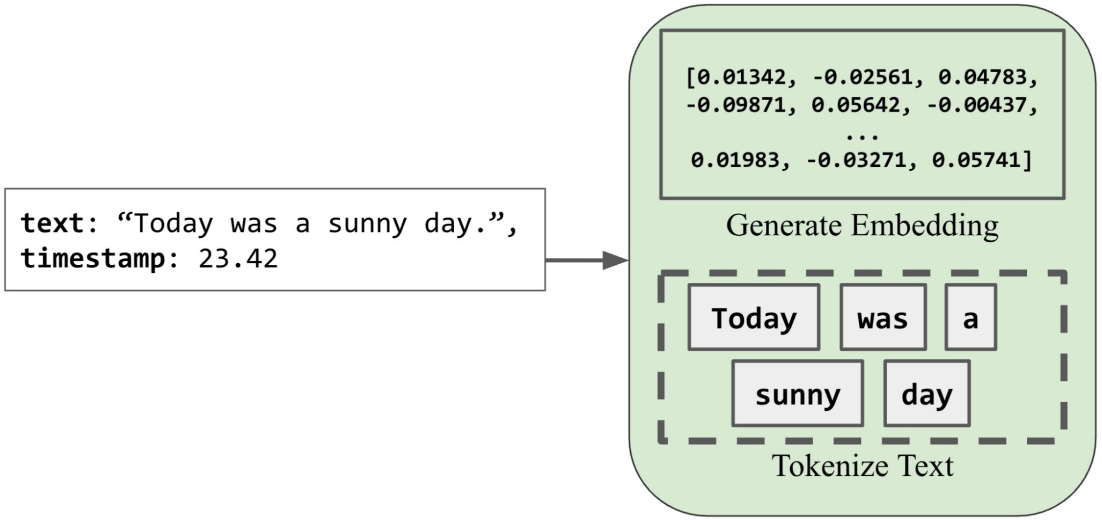
2.2. Visual Content Analysis and Indexing

- Video segmented into short scenes based on visual discontinuities
  - Each segment contains one scene at a time for fine-grained analysis
  - Parallel processing for efficiency
- Each segment processed by vision-language models (VLMs)
  - Generates descriptions of scenes in the video segments



2.3. Semantic Retrieval Strategy

- Two complementary strategies:
    - Keyword-based search (Best Matching 25 (BM25) [Robertson et al., 2009]) → precise lexical matching
    - Embedding-based search (vector similarity) → captures semantic meaning
- Generated text from audio/visual indexing are converted into vector embeddings
    - Enables retrieval based on meaning by calculating the vector similarity
- All text and embeddings are saved to a database along with its matching timestamps



Robertson et al. (2009). *The probabilistic relevance framework: BM25 and beyond.*

2.3. Semantic Retrieval Strategy

- Two complementary strategies:
  - Keyword-based search (Best Matching 25 (BM25) [Robertson et al., 2009]) → precise lexical matching
  - Embedding-based search (vector similarity) → captures semantic meaning

- Reciprocal-Rank Fusion (RRF) score [Cormack et al., 2009] is used to combine results of both methods

$$\mathrm{RRFscore}(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)}$$

- Additional score adjustments using video-level metadata (title, description)
  - Minor score boost to documents that contain metadata matching search query

Robertson et al. (2009). *The probabilistic relevance framework: BM25 and beyond.*
Cormack et al. (2009). *Reciprocal rank fusion outperforms condorcet and individual rank learning methods.*

3. Experiment

- Dataset
  - 300 YouTube videos from two Korean TV programs
    - Nam Buk-ui Chang (North-South Window), KBS (Korean Broadcasting System)[1]
      - 124 videos
      - Average 38.5 minutes runtime
    - Tongil Jeonmangdae (Unification Observatory), MBC (Munwha Broadcasting Corporation)[2]
      - 176 videos
      - Average 32.2 minutes runtimes

[1] YouTube: Nam Buk-ui Chang (North-South Window) - KBS
[2] YouTube: Tongil Jeonmangdae (Unification Observatory) - MBC

3. Experiment

- Models used:
  - Speech-to-Text:
    - Whisper Medium [Radford et al., 2023] OpenAI Whisper
  - Vision-Language Model:
    - Qwen2.5-VL-3B [Bai et al., 2025] Qwen
- Prompt used for VLM

"This video contains ***Korean text inside a box in the upper left corner*** *or is related to North Korea and inter-Korean relations. Based on the video,* ***describe the actions people are taking and provide an overall explanation of the events*** *occurring in the scene. Include* ***only the description of the video itself, without any additional information****."*

→ The TV show contains the **topic** it is screening in the **box on top left**. Guides the model to use that hint into understanding the scene better

→ Request for an **explanation** of the scene

→ **Prevent influence from external knowledge** and have it answer based on the scene it's given

Radford et al. (2023). *Robust speech recognition via large-scale weak supervision.*
Bai et al. (2025). *Qwen2.5-VL Technical Report.*

3. Experiment

- Models used:
  - Speech-to-Text:
    - Whisper Medium [Radford et al., 2023] OpenAI Whisper
  - Vision-Language Model:
    - Qwen2.5-VL-3B [Bai et al., 2025] Qwen
- Prompt used for VLM

Radford et al. (2023). *Robust speech recognition via large-scale weak supervision.*
Bai et al. (2025). *Qwen2.5-VL Technical Report.*
YouTube: Tongil Jeonmangdae (Unification Observatory) - MBC

3. Experiment

- Models used:
  - Speech-to-Text:
    - Whisper Medium [Radford et al., 2023] OpenAI Whisper
  - Vision-Language Model:
    - Qwen2.5-VL-3B [Bai et al., 2025] Qwen
- Prompt used for VLM

*"This video contains **Korean text inside a box in the upper left corner** or is related to North Korea and inter-Korean relations. Based on the video, **describe the actions people are taking and provide an overall explanation of the events** occurring in the scene. Include **only the description of the video itself, without any additional information**."*

→ The TV show contains the **topic** it is screening in the **box on top left**. Guides the model to use that hint into understanding the scene better

→ Request for an **explanation** of the scene

→ **Prevent influence from external knowledge** and have it answer based on the scene it's given

Radford et al. (2023). *Robust speech recognition via large-scale weak supervision.*
Bai et al. (2025). *Qwen2.5-VL Technical Report.*

3. Experiment

- Models used:
  - Speech-to-Text:
    - Whisper Medium [Radford et al., 2023] OpenAI Whisper
  - Vision-Language Model:
    - Qwen2.5-VL-3B [Bai et al., 2025] Qwen
- 166,351 generated text entries

	Number of Text Generated		
Number of Videos	STT	VLM	Total Text
300	72,629	93,722	166,351

Table 1. Number of Text Generated For 300 Videos

- 50 diverse, realistic search queries related to inter-Korean relations
  - Created using LLMs to prevent bias

Radford et al. (2023). *Robust speech recognition via large-scale weak supervision.*
Bai et al. (2025). *Qwen2.5-VL Technical Report.*

3. Experiment

- Metric:
  - On-Topic Rate (OTR) [Zheng et al., 2024]
    - A metric to measure the relevance of search results to a user's query
    - A LLM makes decisions on whether it is relevant or not

$$\mathrm{OnTopicRate}(q, d_i) = \begin{cases} 1 & \text{if the pair is relevant} \\ 0 & \text{otherwise} \end{cases}$$

  - OTR@K
    - For every query, select the top K returned documents
    - Calculate OTR@K by dividing the number of relevant query-document pairs by the total number (K) of results considered

$$\mathrm{OTR@}K = \frac{\sum_{i=1}^{K} \mathrm{OTR}(q, d_i)}{K}$$

Zheng et al. (2024). *Semantic Search Evaluation.*

3. Experiment

- Retrieval Accuracy (OTR@K [Zheng et al., 2024])

Search Strategy	OTR@10	OTR@20	OTR@30	OTR@40	OTR@50
Vector-only	0.895±0.006	0.858±0.011	0.837±0.012	0.807±0.012	0.793±0.011
BM25-only	0.910±0.003	0.905±0.004	0.900±0.003	**0.890±0.007**	**0.880±0.007**
Combined (RRF)	**0.957±0.007**	**0.927±0.007**	**0.905±0.007**	0.889±0.008	0.872±0.008

Table 2. Retrieval Performance Comparison on Search Strategy (On-Topic Rate at K ± std)

- Retrieval Latency
  - Average query response: **0.35 sec**
- Indexing Efficiency
  - STT: ~70 seconds
  - Video Segmentation: ~489 seconds, VLM: ~ 1,299 seconds
  - Entire indexing process **12% faster** than video duration

Zheng et al. (2024). *Semantic Search Evaluation.*

3. Experiment

- Query Example: “North Korea Tourist Attractions”

Search Query	Rank	Related	Result (Translated from Korean to English)
North Korean Tourist Attractions	1	True	North Korean media have recently been consecutively featuring major **tourist attractions**, including **Kumgangsan**, **Chilbosan**, and **Monggeumpo**, among others.
	2	True	Some also have been highlighted as **trendy tourist destinations** that have ventured onto North Korean soil.
	3	True	It was a time when **tourism** to **Kumgangsan** was taking place.
	…		
	9	False	Changgangwon, a sports facility that is representative of North Korea.
	10	True	Every summer, North Korea showcased various **summer resorts**, stoking a vigorous effort to **attract tourists**.

Table 3. Experiment Results for the Search Query "*North Korea Tourist Attractions*"

- Demonstrates semantic relevance and accurate segment retrieval despite variation in wording
- Results precisely linked to timestamps in videos

3. Experiment

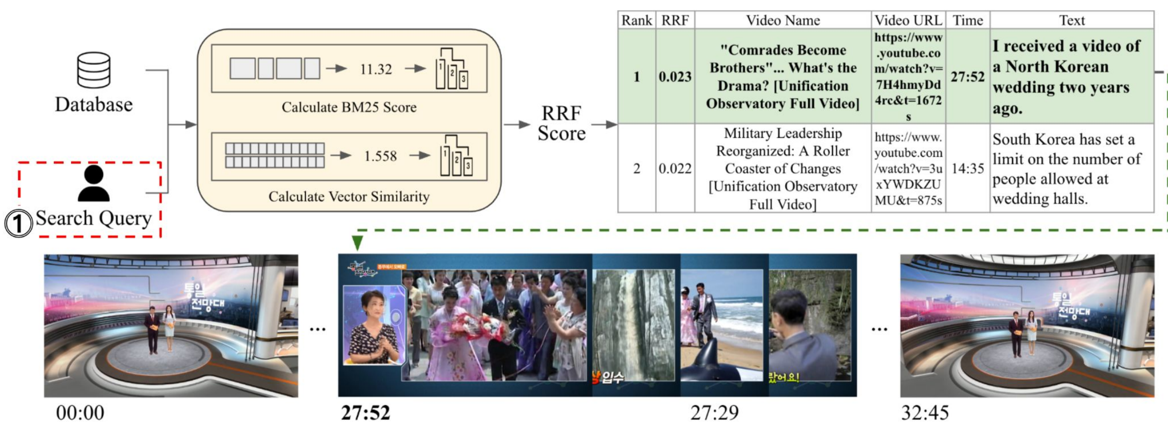
- Query Example: “North-South Dialect Difference”

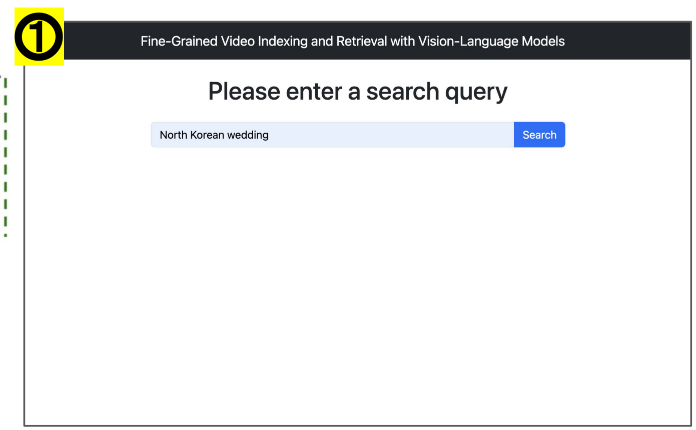
Search Query	Rank	Result (Translated from Korean to English)
North-South Dialect Difference	1	The **language difference** between North and South Korea seems to be most prominent in vocabulary.
	2	The video shows a situation where there is a **pronunciation difference** between Korean and North Korean. Three people are sitting at a table, and one person **pronounces 'hada' as 'heoda'**. This seems to illustrate the **language difference** between North and South Korea. The video appears to be discussing this pronunciation difference.
	3	An example like 'the dialect of Pyeongan-do is not as strong as that of Hamgyeong-do' is confirmed in the dialect of Pyeongan-do.
	4	The Japanese word 'gu-ja,' which means 'position,' is interpreted differently in North and South Korea.
	5	According to the 2016 domestic survey on North-South language awareness, there has been a significant reduction in the sense of rejection towards people using dialects such as Gyeongsang-do or Jeolla-do.
	6	Through dramas, there is an emphasis on using the Pyongyang standard language.
	7	The video deals with the linguistic and cultural differences between North and South Korea, focusing on the terms used to refer to the North-South border.
	8	North Korea's national opera has undergone a change in vocal technique, unlike our traditional changgeuk.
	9	However, North Korean cultural language has some pronunciation differences compared to ours.
	10	Upon closer examination, there are also differences in grammar and speech style.

3. Experiment

① Enter Search Query

- The embedding model loads when the page opens
- After the user enters a query and clicks **Search**, the system generates the query's embedding

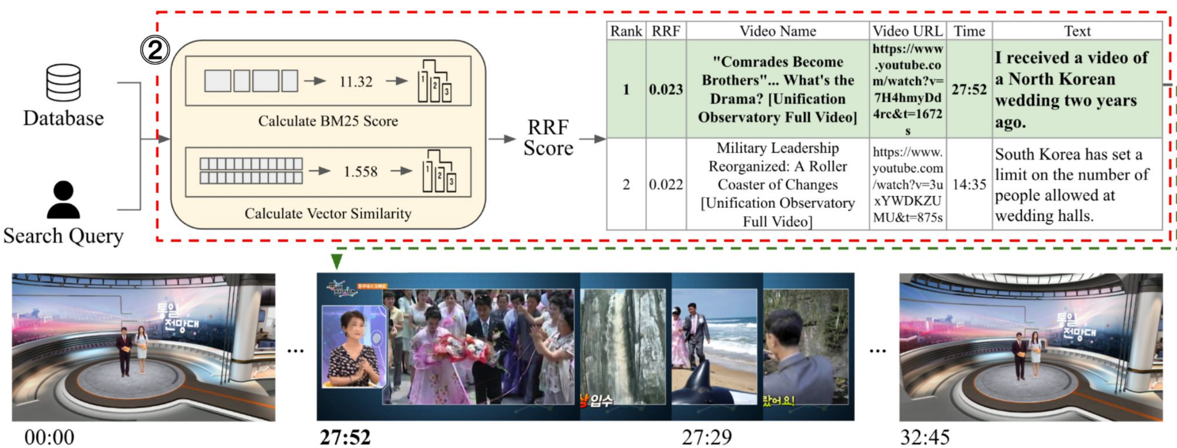


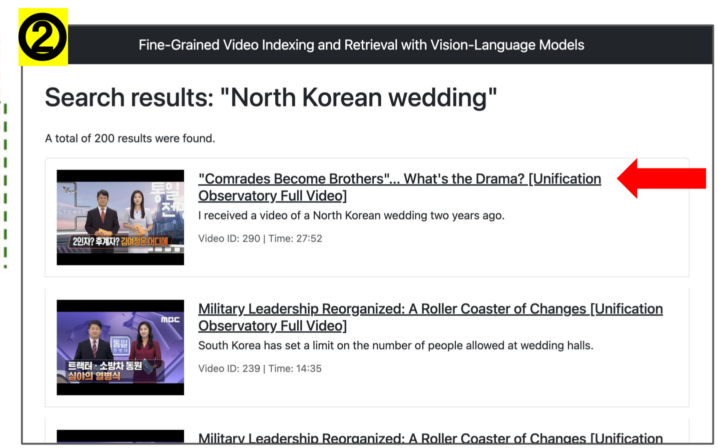


3. Experiment

② Calculate Scores & Display Results

- Vector similarity is computed between the query embedding and stored text embeddings
- The query is also tokenized for BM25-based retrieval using the database engine
- Metadata matching is applied to the retrieved documents
- Final results are ranked and displayed based on combined RRF scores

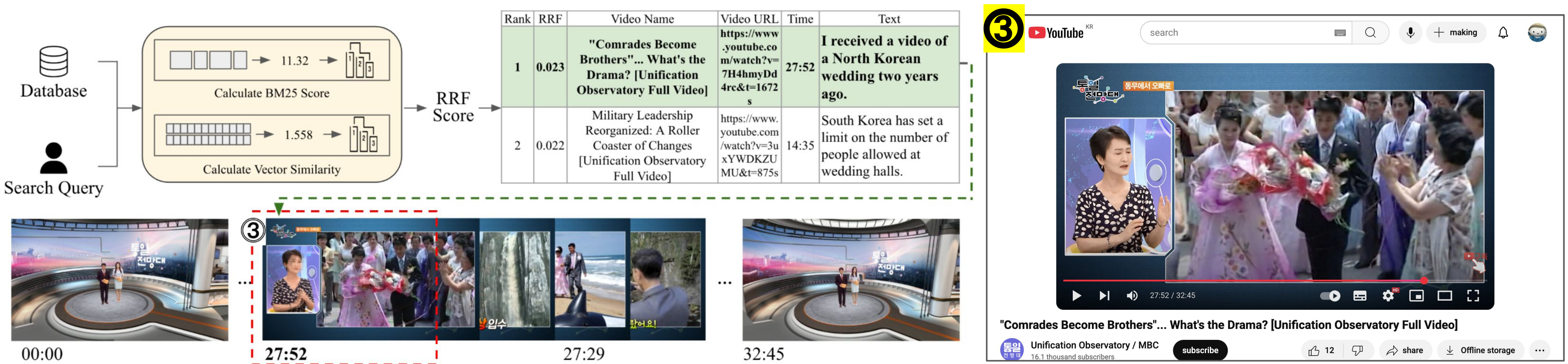




3. Experiment

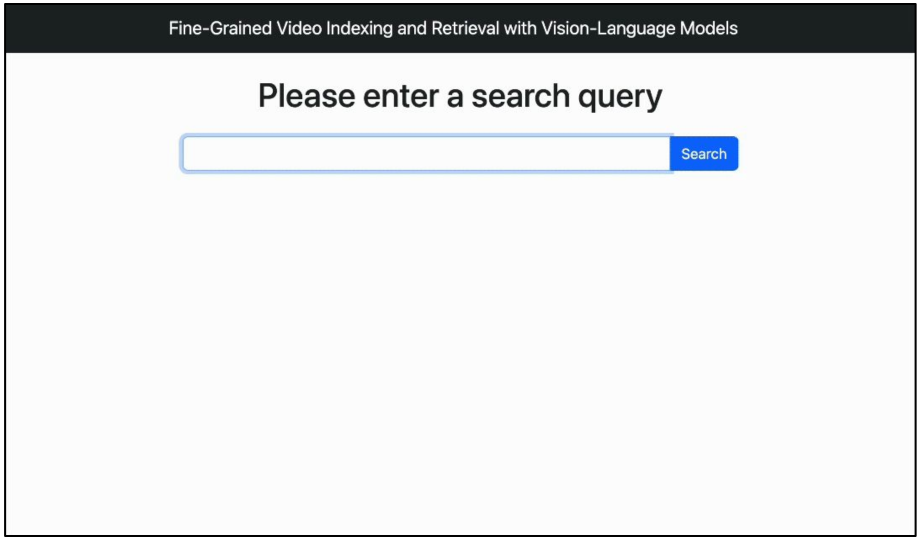
③ Plays the Exact Timestamp of the Result Selected

- Users are able to click on the results to watch the video
  - For this case, all videos are from YouTube, hence the user is redirected to the YouTube video
- The video starts playing at the exact timestamp shown on the result page



3. Experiment

- Example clip of the search service working
  - User enters query
    - *“North Korean Wedding”*
  - Results that match the user query are displayed in order of the score
  - When user clicks the top result, user is redirected to the corresponding YouTube video, playing from the timestamp written at 27:52



4. Conclusion

- Proposed a multimodal indexing framework that integrates speech-to-text and vision-language models for comprehensive speech and visual content processing
- Achieved high-accuracy, fine-grained retrieval at the timestamp level
  - OTR@10: 95.7%
- Demonstrated fast retrieval performance with an average query time of 0.35 seconds
- Scalable and applicable to large-scale multimodal video archives

Future Work

- Test on expanded datasets
- Extend the framework to support multilingual video collections
- Expand to different domains for real-world deployments

5. References

G. V. Cormack, C. L. Clarke, and S. Buettcher. *Reciprocal rank fusion outperforms condorcet and individual rank learning methods*. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pages 758–759, 2009.

W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank. *A survey on visual content-based video indexing and retrieval*. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41(6):797–819, 2011.

J. Jung, S. Park, H. Kim, C. Lee, and C. Hong. *Artificial intelligence-driven video indexing for rapid surveillance footage summarization and review*. In K. Larson, editor, Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, pages 8687–8690. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Demo Track.

M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. *Content-based multimedia information retrieval: State of the art and challenges*. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2(1):1–19, 2006.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. *Visual instruction tuning*, 2023.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. *Robust speech recognition via large-scale weak supervision*. In International conference on machine learning, pages 28492–28518. PMLR, 2023.

S. Robertson, H. Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval, 3(4):333–389, 2009.

C. Zheng, J. Wang, S. A. Zhang, A. Kishore, and S. Singh. Semantic search evaluation, 2024. 5

SEO.ai Content Team. (2025, February 15). *How many videos are on YouTube?.* https://seo.ai/blog/how-many-videos-are-on-youtube.

Thank you for your attention

Title	**Fine-Grained Video Indexing and Retrieval with Vision-Language Models**	
Presenter	Dong Gun Park	(systec24@handong.ac.kr)
Advisor	Charmgil Hong	(charmgil@handong.ac.kr)