# Service-oriented Study on Slot-level Metadata Derived from Multimodal Analysis

Hyun-Jeong Yim
*Media Research Division*
*Electronics and Telecommunications Research Institute*
*Daejeon, South Korea*
hjyim@etri.re.kr

Bo-mi Lim
*Media Research Division*
*Electronics and Telecommunications Research Institute*
*Daejeon, South Korea*
blim_vrossi46@etri.re.kr

Jung Sun Um
*Media Research Division*
*Electronics and Telecommunications Research Institute*
*Daejeon, South Korea*
korses@etri.re.kr

Jae Hyun Seo
*Media Research Division*
*Electronics and Telecommunications Research Institute*
*Daejeon, South Korea*
jhseo@etri.re.kr

***Abstract*—This study proposes a service-oriented approach that structures information derived from multimodal analysis into slot-based metadata. By integrating video, audio, and subtitle analysis results, the metadata is organized into slot units, enabling personalized service delivery. Experimental validation confirms the feasibility of applying slot-based multimodal metadata to personalized media services.**

***Keywords—Multimodal Analysis, Metadata structure, Personalized media Service***

## I. INTRODUCTION

Multimodal content analysis is a core technology for extracting semantic information from video, audio, and subtitles, enabling applications such as summarization, search, and personalized services. However, most existing studies have focused solely on semantic extraction. Attempts to integrate the results into standardized data structures or to connect them with transmission and playback systems have been limited.

Multimodal AI research has achieved progress in high-dimensional semantic analysis, but the outcomes often remain confined to isolated databases or embedding vectors. Conversely, existing international multimedia transmission standards[1][2] have been designed primarily around physical representations such as geometry and animation, while semantic layers are treated only as supplementary annotations. As a result, there is no bridging layer that links AI-derived semantics with standardized formats, preventing meaningful information from being fully utilized in real-world services.

## II. RELATED WORK

Recent advances in multimodal media analysis have been driven by large-scale pre-trained models that jointly process video, audio, and text. VideoLLaMA2 [3], trained on extensive video–language datasets, has demonstrated strong performance in tasks such as scene understanding, summarization, and question answering, thereby enabling the extraction of complex semantic units. This represents a significant step beyond unimodal approaches, providing a foundation for metadata-driven content representation and semantic-level streaming. However, the high computational cost and limited real-time capability of such large models remain major challenges for practical deployment.

In speech processing, WhisperX [4] has achieved precise word-level time alignment, substantially improving the synchronization between spoken content and video or subtitle streams. This capability enhances the temporal coherence of multimodal semantic units, which is critical for network efficiency and the reliability of personalized media services. Meanwhile, recent studies have begun to extend beyond traditional QoS/QoE optimization by exploring semantic-level information as a core resource for adaptive delivery and service personalization.

## III. TIME-ALIGNED MULTIMODAL SLOT ARCHITECTURE

This study distinguishes itself from prior approaches by structuring semantic information around time-aligned slots rather than simply listing multimodal analysis results. A slot integrates diverse modalities—such as objects, events, utterances, and subtitles—into a single coherent unit along a common timeline, represented in a lightweight JSON key–value schema. This representation not only preserves interoperability but also provides direct mapping potential to higher-level standards such as scene graphs or 3D representation formats. Unlike conventional result storage, the proposed framework can be consistently applied from the stage of content analysis to transmission and playback, thereby linking semantic information naturally with the service delivery process. Furthermore, by explicitly defining semantic units at the content layer, the framework introduces a capability that existing QoS-centric transmission structures have struggled to address, offering a solid foundation for future personalized and intelligent streaming architectures.

## IV. SYSTEM IMPLEMENTATION

The proposed system defines multimodal analysis outputs in a slot-based metadata structure and integrates them into the transmission and playback pipeline. In the first stage, video, audio, and subtitle data are preprocessed using FFmpeg to ensure temporal alignment and segment-level synchronization. The video stream is then analyzed by VideoLLaMA2 to capture visual semantics such as object presence, scene context, and event-level information. In parallel, the audio stream undergoes

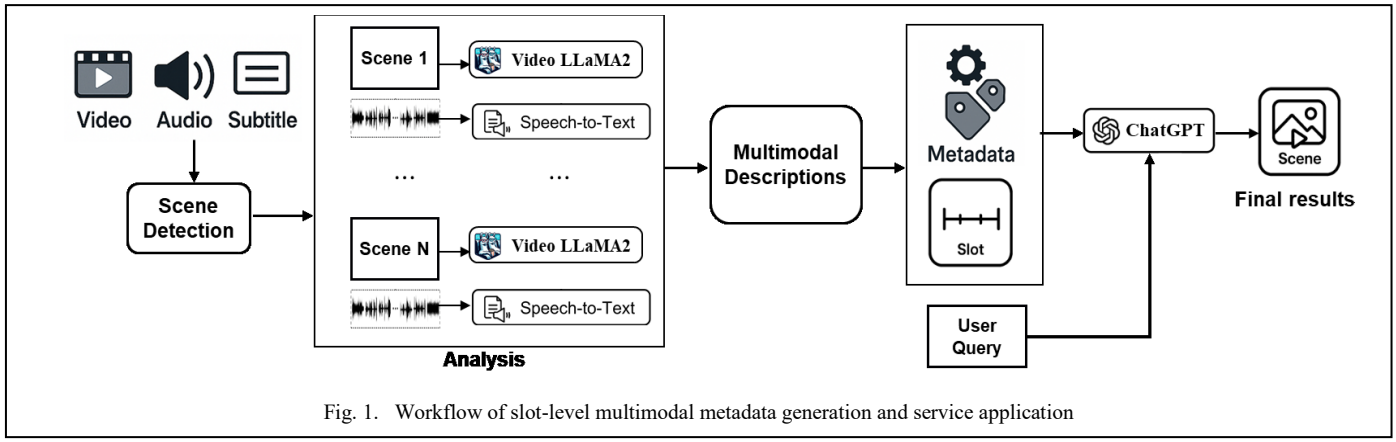


Fig. 1. Workflow of slot-level multimodal metadata generation and service application

automatic speech recognition using models such as WhisperX, producing utterance-level transcripts that are further aligned with subtitle tracks when available. The outcome of this stage is a set of low-level multimodal analysis results that serve as the foundation for subsequent integration.

In the second stage, these heterogeneous analysis outputs are organized into slot units, where each slot corresponds to a defined temporal segment. Within a slot, semantic entities from different modalities—objects and events detected from video, utterances derived from speech recognition, and aligned subtitle segments—are associated along both temporal and spatial dimensions. Through this process, the system produces a unified multimodal description, transforming raw recognition results into a semantically coherent representation of the content.

Finally, in the metadata management stage, the integrated slot units are encoded into a JSON-based metadata structure. Each slot encapsulates multimodal entities and their contextual relations, which are stored as an independent metadata layer decoupled from the raw media stream. This design enables flexible linkage with transmission formats and playback engines, supporting adaptive reconstruction of semantic narratives during playback as well as personalized and efficient delivery strategies tailored to user intent and network conditions.

Through this pipeline, we verified that multimodal analysis results can be unified within a standardized structure and effectively applied to media service scenarios such as search, summarization, and personalized streaming.

## V. MAIN RESULTS

In this study, we conducted experiments using Korean news videos with a total duration of 27 minutes. The entire video was segmented into 393 scenes, each ranging from 1 to 30 seconds in length, with most scenes falling within 2 to 5 seconds. When testing topic-based summarization on these segments, the system achieved accuracies of 97.6% for politics, 83.8% for North Korea, 100% for economy, 71.2% for accidents, and 89.1% for weather. These results demonstrate that the system is capable of generating a certain level of summarization and reorganization of fully produced content according to user requests. In particular, it was able to effectively condense full-length news videos into topic-specific summaries.

TABLE I. TOPIC RELEVANCE ANALYSIS RESULTS

| Topic | politics | North Korea | economy | weather |
|---|---|---|---|---|
| Relevant | 97.6% | 83.8% | 100% | 89.1% |
| Irrelevant | 0.3% | 0% | 0% | 0% |

## VI. CONCLUSION

This work proposes a system that structures semantic information extracted from video, audio, and subtitles into a slot-based metadata framework, enabling integration into transmission and playback processes. Unlike prior approaches that remain at the level of raw analysis results, our method defines semantic units along temporal and spatial axes, ensuring their applicability to service pipelines. The proposed system demonstrates practical feasibility. Future work will extend the framework to represent higher-level constructs such as events or narrative context, and evaluate its effectiveness in real service environments through integration with standardized delivery formats such as MPEG-DASH.

## ACKNOWLEDGMENT

This work was supported by internal fund/grant of Electronics and Telecommunications Research Institute(ETRI). [25YC1100, Development of fundamental technology for next-generation media coding and transmission standards]

## REFERENCES

[1] ISO/IEC, Information technology Coded representation ofimmersive media Part 14: Scene description, ISO/IEC 23090-14,2023, https://www.iso.org/standard/80900.html

[2] Kim, Jae‐Gon, et al. "Multimodal approach for summarizing and indexing news video." ETRI journal 24.1 (2002): 1-11.

[3] Cheng, Zesen, et al. "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms." arXiv preprint arXiv:2406.07476 (2024).

[4] Bain, Max, et al. "Whisperx: Time-accurate speech transcription of long-form audio." arXiv preprint arXiv:2303.00747 (2023).