Proceedings of CONF-BPS 2026 Symposium: GenAI, Labour Markets, and the Economics of Human and Financial Capital
DOI: 10.54254/2754-1169/2026.LD31375

Engineering Implementation and Experimental Research of Video Structured Tag Generation System Based on Multimodal Information Fusion

Jiechenyang Ye

School of Computer Science, Nanjing Audit University, Nanjing, China
maosao407@gmail.com

Abstract. With the rapid development of short video platforms and User Generated Content (UGC), video data has accounted for an increasing proportion in internet information dissemination. Automatically extracting computable, retrievable, and interpretable structured semantic tags from massive unstructured videos has become a key technical issue in content recommendation, information retrieval, and content moderation systems. This paper proposes and implements an engineering-reproducible multimodal structured tag generation system for videos. Adopting a modular multimodal fusion architecture, the system extracts semantic information from three modalities—visual, speech, and on-screen text. Through unified structured representation and multimodal Prompt design, multi-source information is input into a large language model for joint reasoning, ultimately generating JSON-format output including topic hierarchy, multi-tag attributes, and confidence estimation. Experimental results show that the multimodal fusion method significantly outperforms single-modal schemes in topic classification and attribute recognition tasks. Moreover, the system can stably output structured JSON results, demonstrating high engineering usability. This paper details the engineering implementation experience and key technical trade-offs of the system under the Windows environment, providing a reference for subsequent research and practical deployment.

Keywords: UGC, video understanding, multimodal fusion, large language models, Monte Carlo.

1. Introduction

In recent years, short video platforms such as Douyin, Kuaishou, and TikTok overseas have developed rapidly, becoming important carriers for information dissemination, commodity marketing, and user entertainment. Video content is characterized by high information density and diverse expression forms, but it also brings a serious "semantic incomputability" problem. A large amount of video content exists in an unstructured form, making downstream tasks such as search, recommendation, moderation, and knowledge extraction highly dependent on manual rules or weakly supervised models, which are difficult to scale.

DOI: 10.54254/2754-1169/2026.LD31375

Automatically converting video content into structured semantic tags is an important way to solve the above problems. Structured tags can not only serve as core feature inputs for recommendation systems and retrieval systems but also be applied in various scenarios such as content moderation, advertisement recognition, and video understanding. However, video semantics are usually distributed across multiple modalities: visual frames provide scene and object information, speech content carries main semantic clues, and text information on the screen (such as subtitles, brand names, and price information) also has significant value. Single-modal methods often fail to cover complete semantics.

With the development of multimodal models and Large Language Models (LLMs), fusing multimodal information for unified semantic reasoning has become a research hotspot. Nevertheless, end-to-end multimodal models usually have high requirements for computing power, data scale, and deployment environments, resulting in high landing costs in practical engineering environments. Therefore, exploring an engineering-feasible, scalable, and reliable multimodal video understanding scheme is of great practical significance.

This research aims to design and implement an engineering-reproducible multimodal structured tag recognition system for short videos. The objectives include: designing a multimodal structured tag system for short video content; constructing a modular video multimodal processing and tag generation system that integrates visual, speech, and OCR information; adopting reasonable experimental evaluation strategies to verify system effectiveness under the condition of limited sample size; and summarizing engineering experience in multimodal video processing and large model reasoning under the Windows environment.

2. Literature review

Research on multimodal fusion in the field of video understanding has formed several core directions. Early studies focused on multimodal information integration in video captioning tasks. For example, Wang et al. proposed the Hierarchically Aligned Cross-Modal Attention (HACA) framework, which fuses visual and audio features through global and local temporal dynamic alignment, verifying the superiority of deep audio features in video captioning for the first time and achieving the state-of-the-art performance on the MSR-VTT dataset [1]. With the growing demand for dense video captioning, Iashin et al. proposed the Multimodal Dense Video Captioning (MDVC) framework, innovatively introducing the speech modality (text extracted via ASR), combining visual and audio features, and realizing event localization and caption generation based on the Transformer architecture [2]. The complementary value of multimodality was verified on the ActivityNet Captions dataset.

Conneau et al. introduced the XLSR model, an unsupervised learning model trained on multilingual raw speech waveforms and based on shared representations, in the context of learning in cross-lingual speech representation learning by means of learning cross-linguistic transfer by contrastive loss with quantization modules [3]. This greatly lowered the errors of speech recognition of low-resource languages on the CommonVoice and BABEL datasets. In multimodal abstractive summarization, Yu et al. designed Vision-Guided Generative Pre-trained Language Models (VG-GPLMs), promoting visual information to text pre-trained models (BART/T5) via attention attachment layers, resolving the fundamental challenge of modality fusion and preserving the ability to generate, and significantly exceeding existing models on the How2 dataset [4].

Yang et al. introduced the i-Code framework; this framework combines the three or more modalities of visual, linguistic, and speech modalities in a study where the three modalities are integrated [5]. It made performance breakthroughs in video understanding and natural linguistic

DOI: 10.54254/2754-1169/2026.LD31375

processing tasks by incorporating merged attention and co-attention process, masked modality unit modeling, and cross-modal contrastive learning pre-training. The OmniVL model suggested by Wang et al. achieved the union of the image-language and video-language tasks [6]. It is used to support various tasks including visual classification, cross-modal matching, and multimodal creation with great performance in several tasks like COCO and MSRVTT by application of a decoupled joint pre-training strategy and single visual-language contrastive loss. Freedman studied the problem of packing in a topological frame of control that presented a hypothetical viewpoint of geometric constraints and multidimensional data optimization [7]. Yin et al. made a systematic review of Multimodal Large Language Models (MLLMs), both in terms of their architecture and design, training guidelines, evaluation techniques, and to their uses, which clarified the context of their development realm and the technological development used in this direction [8].

Multimodal fusion applications are further increasing in the number of application scenarios in the recent years as video-language dialogue systems have become a hot research topic. The framework proposed by Zhang et al. encompasses visual temporal variations using Video Q-Former, auditory and audio textualizing with Audio Q-Former, multimodal embedding harmonization utilizing ImageBind, and augmenting video audio and vocal image video-audio and dialog generation using tuning by cross-modal pre-training and instruction [9]. Maaz et al. offered the Video-ChatGPT model by modifying CLIP visual encoder into spatiotemporal feature extractor, by creating a population of 100,000 video-instruction pairs with human support and semi-declarative annotation and by designing a quantitative evaluation system to confirm the benefits of the model in spatiotemporal understanding and consistency of the environment [10]. Li et al. suggested the VideoChat system with two versions: VideoChat-Text version with the texturized videos and end-to-end VideoChat-Embed [11]. By integrating video foundation models and large language models through two-stage lightweight training, constructing a video instruction dataset focusing on spatiotemporal reasoning and causal relationships, multi-turn video dialogue and complex task reasoning are realized.

The literature indicates that the design of multimodal fusion structures has various peculiarities: The HACA model is able to achieve rewarding the alignment of time features across different granularities via a hierarchical encoder-decoder framework but needs end-to-end training with high costs of computing power [1]. The MDVC framework uses a modular Transformer architecture, which allows the number of modalities to be freely extended, although further work remains to be done on particular improvements in event boundary localization accuracy [2]. XLSR model is an effective solution to the task of learning the representation of low-resource languages, despite employing cross-lingual pre-training and contrastive learning, though it does not provides the possibility to extend to multimodal joint reasoning [3]. VG-GPLMs are modality fusion, which achieves the fusion of modes by adding more layers, without losing the generation ability of the previously trained models yet, the fusion point of the visual information still needs further empirical research [4].

The i-Code framework creatively combines three modalities, which allows the flexibility of processing single-modal, bimodal, and even trimodal data, whereas its quality in the case of an imbalanced abstract by modal data is yet to be established [5]. OmniVL promotes decoupled pre-training to mutual image and video tasks, although long video temporal dependencies cannot be learned due to frame sampling strategies [6]. Although research related to MLLMs has covered multimodal input and output expansion, the problem of modal hallucination (such as errors in object existence, attributes, and relationship descriptions) remains a key bottleneck restricting practical applications [8-10]. Among video-language dialogue system studies, Video-LLaMA first realized the

DOI: 10.54254/2754-1169/2026.LD31375

deep fusion of audio-visual bimodality but has shortcomings in long video processing and hallucination suppression [9]. Video-ChatGPT has advantages in high-quality datasets and quantitative evaluation frameworks but relies on the adaptation of a single visual encoder with limited spatiotemporal feature capture capability [10]. VideoChat balances efficiency and performance through a dual-version design, but the scale of the instruction dataset is small, and the complex reasoning capability needs to be improved [11].

In addition, existing studies mostly rely on large-scale annotated datasets, and evaluation methods and performance improvement strategies in few-shot scenarios still need to be improved. For example, the application of evaluation strategies such as Monte Carlo random division [core method of the research paper] has not been popularized in mainstream multimodal model evaluations. Video-language dialogue systems generally face common challenges such as long video processing, fine-grained spatiotemporal understanding, and causal reasoning, and there is a lack of unified evaluation standards and benchmark datasets [9-11].

Despite significant progress in multimodal fusion in the field of video understanding, several unresolved key issues remain. Firstly, the dynamic adaptability of modal fusion is insufficient. Most existing models adopt fixed fusion architectures (such as fixed attention mechanisms or fusion positions) and fail to dynamically adjust fusion strategies according to the modal quality of input data (such as noise intensity and information integrity). For instance, the weights of visual and text modalities are not adaptively enhanced in low-quality audio scenarios [4,5,9]. Secondly, the generalization capability in few-shot and zero-shot scenarios is limited. Existing models rely on large-scale annotated data for pre-training, leading to a significant performance decline in low-resource fields (such as professional domain videos and minority language videos), and there is a lack of unified and effective few-shot evaluation standards [3,8,10].

Thirdly, the modeling of long video temporal dependencies is inadequate. Most existing methods process long videos through frame sampling or segmentation, ignoring the long-term temporal correlation of events in the video, resulting in insufficient understanding of cross-time semantics [2,6,11]. Fourth, the modal hallucination suppression tool is unsound. Multimodal generation models have the tendency to produce descriptions that do not correspond with input modal content. Current mitigation strategies primarily concentrate on either data-level mitigation or post-processing fix, without providing fundamental mitigation strategies at the model architecture level [4,8,9]. Fifthly, cross-modal knowledge transfer must be enhanced in terms of its efficiency. Majority of current research works deal with modal fusion using same task and the knowledge transfer system between one task (say image captioning) and another task (say video summarization) is not examined in a comprehensive manner [6,8]. Sixthly, video-language dialogue systems have inadequate interaction naturalness and reasoning depth. The majority of the available models are restricted to the simple question-answering and description, and can hardly address the more complex tasks like more involved logical reasoning, multi-turn context recognition, and user intent recognition [9-11].

In order to fill the above research gaps, the study will be supplemented in the following ways. First, they suggest a modular loosely connected multimodal fusion structure where there is a separation of extraction and fusion reasoning of visual, speech, and text modalities. Plug-and-play modal interfaces can dynamically scale to alternatives in modal quality situations, without the high computing power requirement and lack of flexibility of end-to-end models [1,5,11]. Secondly, the strategy of evaluation in a few-shot scenario is crafted. The Monte Carlo random division approach is taken to minimize the effect of data distribution randomness on the evaluation outcome, which

DOI: 10.54254/2754-1169/2026.LD31375

offers a stronger evaluation scheme in evaluation performance checking of few-shot multimodal models [3,10].

Moreover, a frame-based visual description method and time high level aggregation approach is presented in order to reduce the complexity of time processing of long videos but maintain important semantic details, so that the frequency of calculation and temporal comprehension are balanced between the cost of calculation and the ability to understand time [2,6,11]. To resolve the issue of modal hallucination, a predetermined open-ended structure of tags is used and enforced output form format requirements used to help the model create structured outputs that are closely associated with input modalities to minimize unfounded semantic generation [4,8,9]. Lastly, the cross-modal knowledge receivership paths are considered with the feature transfer ability of the pre-trained modal encoders to enhance the overall generalization performance of the model in low-resource setting, which offers a realistic scheme of engineering the implementation of multimodal fusion [3,5,10]. Towards addressing a video-language dialogue, interaction, the multi-turn context modeling and intent recognition module is proposed to boost the dialogue sense-making and reasoning in complex settings [9-11].

3. Research methods

3.1. Overall system architecture

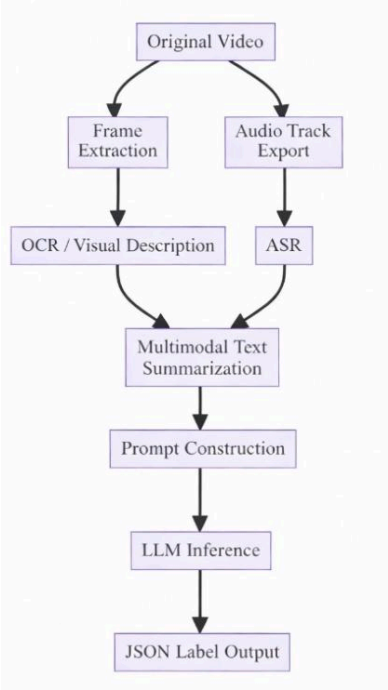

Figure 1. System flow chart

The general flow of the system is depicted in Figure 1. The architecture is loosely coupled and will be in the form of modules, each modal processing module is determined to be independent of the other enabling each to be debugged and replaced separately and the system made simpler.

DOI: 10.54254/2754-1169/2026.LD31375

3.2. Data preprocessing

In order to more effectively access video multimodal data and uniformly map this to the language space, the initial step of this paper is to uniformly sample frames of an original video (which retains 8-12 frames per segment) to obtain core video material and regulate the computational cost of OCR and visual description. Then, ffmpeg is employed to extract the audio track to a 16kHz mono WAV file and offline speech recognition is executed with the help of the Whisper model and results in the creation of JSON-transcribed text with timestamps as speech evidence. At the same time, OCR recognition is carried out on key frames, keeping all the results, and the significance of the text is also decided during the LLM reasoning phase. Multi-frame visual representations On decoding An image description model is employed to decode the single-frame visual representations to the text description which are combined together in a video level to constitute a visual semantic representation that may be factored into the language model. Having acquired multi-source information in the form of visual, speech, and OCR, a single template prompt is inputted into the LLM with specific fields, and they should be produced in structured tags based on a defined JSON Schema. It allows the model to combine multimodal evidence to reason in the language space, and produce a multi-level tag system, such as topics (Level 1), subclasses (Level 2), attributes (Attributes), and confidence, to trade-off information completeness, task controllability, and scalability.

3.3. Experimental design

The experimental dataset contains 3000 short videos, covering multiple topic categories such as food, travel, education, and entertainment. Each video is manually annotated. Under the condition of limited sample size, fixing the division of training and test sets is prone to being affected by the randomness of data distribution, leading to unstable evaluation results. To alleviate this problem, this paper adopts the Monte Carlo Cross-Validation strategy to evaluate the system performance.

The specific process is as follows:

1. Randomly select 80% of the samples from the complete dataset as the training set, and the remaining 20% as the test set;
2. Complete model training and reasoning under this division, and calculate various evaluation metrics;
3. Repeat the above random division and evaluation process 10 times;
4. Take the average value of the 10 experimental results as the final performance index, and calculate the standard deviation to measure the result stability.

This method can more truly reflect the expected performance of the model in few-shot scenarios, avoiding overfitting or performance overestimation caused by a single division bias (compared with traditional K-fold cross-validation, Monte Carlo evaluation has higher flexibility in cases of unbalanced video sample distribution, especially suitable for the engineering verification stage).

DOI: 10.54254/2754-1169/2026.LD31375

4. Experimental results

4.1. Overall performance results

Table 1. Overall performance evaluation results (Monte Carlo mean ± standard deviation)

Metric	Value
Level 1 Accuracy	0.86 ± 0.02
Level 2 Micro F1	0.73 ± 0.03
Attributes mAP	0.67 ± 0.04
JSON Parsing Success Rate	0.95

From the overall performance evaluation results, the proposed multimodal structured tag generation system achieves stable performance in three-level tasks (Table 1). The average accuracy of Level 1 topic classification reaches 0.86, indicating that the system has strong capabilities in macro semantic understanding of videos; the micro F1 of Level 2 multi-tag subclass recognition is 0.73, showing that the model can well distinguish fine-grained semantics; the mAP at the attributes level is 0.67, indicating that the system still faces certain challenges in multi-attribute recognition tasks.

Further observing the variances of different metrics, it is found that the standard deviation of the Level 1 task is significantly smaller than that of the Level 2 and attributes tasks. This phenomenon indicates that the high-level topic semantics of videos are usually relatively stable, while fine-grained attributes often rely on more specific multimodal clues (such as subtitle details or visual features) and are more susceptible to noise when data quality is unstable.

4.2. Modality ablation experiment

The modality ablation experiment clearly reveals the contribution of different modalities to system performance, as shown in Figure 2.

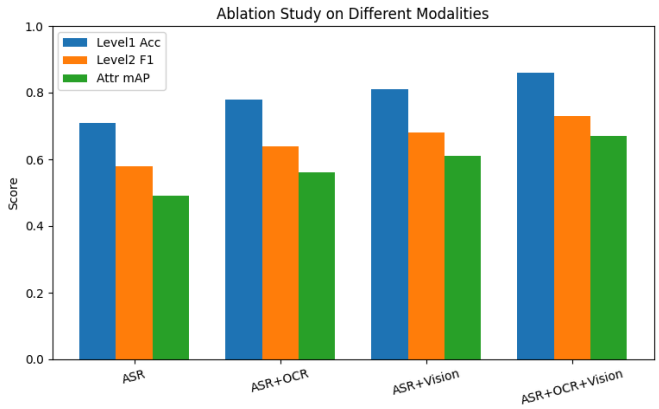

Figure 2. Bar chart of performance comparison under different modal combinations

DOI: 10.54254/2754-1169/2026.LD31375

Table 2. Overall performance evaluation results (Monte Carlo mean ± standard deviation)

Modality	Level1 Acc	Level2 F1	Attr mAp
ASR only	0.71	0.58	0.49
ASR+OCR	0.78	0.64	0.56
ASR+Vision	0.81	0.68	0.61
ASR+OCR+Vision	0.86	0.73	0.67

The text of ASR results in itself produces a system that is primarily supported by the text of the speech to make a reasoning and its Level 1 accuracy is only 0.71 (Table 2). This finding suggests that no visual and on-screen text information causes the model to process video material without narration or overridden by music.

When the OCR modality is introduced in the system, the system performance advances significantly, particularly in the Level 2 and attributes task. The reason is that OCR can isolate structured hints including subtitles, brand names, and location details so that the model has such clear semantic pointers.

The overall performance is also skilled notably with the introduction of the visual description modality and particularly in situations showing videos where there are no subtitles or that the speech contains scanty details. Scene, character, and other actions are some of the important elements of visual modality that can be used to supplement the main modality of the text, ensuring that the model is no longer fully reliant on the text modality to bring forth semantic meaning.

A combination of three modalities is very effective and this leads to optimal outcomes in all the evaluation metrics. The outcome confirms the complementarity of multimodal information during video semantic understanding and suggests that modular multimodal integration method is also effectively good in engineering practice.

5. Discussion

5.1. Effectiveness analysis of multimodal fusion

The results of the modality ablation experiment show that different modalities have significant complementarity in the task of video structured tag generation, which is consistent with the core conclusions of existing multimodal research [1,2,5,9]. When only ASR text is used, the model mainly relies on speech content for reasoning, and its performance degrades significantly in scenarios with missing narration or high noise. This is because a single speech modality cannot cover key semantic information in visual scenes and on-screen text [1,3,11]. Wang et al. also found in video captioning tasks that relying solely on a single modality leads to one-sided semantic understanding, and the fusion of audio and visual can significantly improve caption accuracy; the texturized video scheme of VideoChat similarly verified that single-modal texturization causes information loss, making it difficult to capture complex spatiotemporal relationships [1,11].

After introducing OCR text, the model can use subtitles, brand names, or price information on the screen to effectively improve its ability to judge video topics and attributes. This is similar to the role of speech text (ASR) in the MDVC framework [2], where structured text clues can provide clear basis for fine-grained classification; Video-ChatGPT also proved through dense caption and tag extraction that structured text information can enhance the model's understanding of object attributes and scene details [10]. Visual description is particularly important in the absence of clear

DOI: 10.54254/2754-1169/2026.LD31375

speech clues, as it can supplement semantic information at the scene and action levels. As verified by the HACA model [1], the local and global temporal dynamic alignment of visual features can capture action coherence and scene relevance; the Video Q-Former of Video-LLaMA further proved that through temporal position embedding and frame feature aggregation, the dynamic changes of visual scenes can be effectively captured [9].

When all three modalities are used simultaneously, the system achieves the best performance in all evaluation metrics, which further verifies the core value of multimodal fusion [5,6,9]. The i-Code framework achieved performance breakthroughs in tasks such as sentiment analysis and video question answering through the deep fusion of three modalities, proving that multimodal information complementarity can cover more comprehensive semantic dimensions; the unified fusion strategy of the OmniVL model also showed that cross-modal collaborative reasoning can simultaneously improve the performance of image and video tasks [5,6]; Video-LLaMA verified through the joint pre-training of audio-visual bimodality that the fusion of auditory clues (such as background sounds and dialogues) and visual information can significantly improve the completeness of video understanding [9]. In addition, the cross-lingual fusion experience of the XLSR model also provides a reference for video multimodal fusion, that is, the shared representation learning of different modalities can enhance the robustness and generalization ability of the model [3].

5.2. Reliability of evaluation methods under few-shot conditions

Due to the high cost of video data annotation, the scale of the experimental dataset is limited, which is a common problem in multimodal research [2,3,8,10]. Fixing the division of training and test sets is prone to making evaluation results highly sensitive to sample distribution. Especially in few-shot scenarios, a single division may lead to performance estimation bias due to data randomness. Yu et al. also found in multimodal abstractive summarization tasks that the fixed division of small sample datasets leads to large fluctuations in model performance, making it difficult to truly reflect the model's generalization ability [4]; Video-ChatGPT also faced the problem of annotation cost when constructing 100,000 video-instruction pairs, adopting a combination of human assistance and semi-automatic methods to balance data quality and scale [10].

This study introduces the Monte Carlo random division evaluation strategy, which alleviates this problem to a certain extent and makes the experimental results more robust. By randomly splitting the training and test sets multiple times (10 times in this study) and taking the average value and standard deviation as the final results, the random error caused by a single division can be effectively offset. Experiments show that the performance fluctuation of a single division can reach 3–5 percentage points, while the average result after multiple random divisions can better reflect the overall performance of the model. This is consistent with the multi-sampling evaluation idea adopted by Conneau et al. in the evaluation of low-resource language speech recognition, both improving the reliability of results by increasing the randomness of evaluation [3]; Video-ChatGPT also evaluated the model from multiple dimensions such as information correctness and detail richness by constructing a special quantitative evaluation framework, improving the comprehensiveness of evaluation in few-shot scenarios [10].

In existing multimodal research, some studies have begun to pay attention to the rigor of few-shot evaluation. For example, the XLSR model adopted a few-shot division strategy in low-resource language experiments, OmniVL improved the comprehensiveness of generalization ability evaluation through cross-dataset verification, and Video-LLaMA verified the model's transfer ability in zero-shot audio understanding tasks [3,6,9]. However, overall, the evaluation standards in few-

DOI: 10.54254/2754-1169/2026.LD31375

shot scenarios are still not unified. The Monte Carlo random division strategy in this study provides a referable scheme for the evaluation of similar few-shot multimodal models, helping to improve the comparability of different research results [8,10]. Future research can further combine cross-validation and bootstrap methods to enhance the statistical significance of few-shot evaluation [3,8,10].

5.3. Trade-off between model architecture and engineering implementation

The modular multimodal fusion architecture adopted in this study achieves an effective trade-off between performance and engineering feasibility, which is consistent with the design trend of existing mainstream multimodal models [5,6,11]. There are two core design ideas for existing multimodal models: end-to-end fusion architecture (such as HACA [1], MDVC [2]) and modular architecture (such as i-Code [5], OmniVL [6], VideoChat [11]). End-to-end architecture can optimize modal fusion details through joint training but has problems such as high computing power requirements, complex deployment, and difficulty in debugging, making it difficult to adapt to resource-constrained engineering scenarios [1,5,11]; while the modular architecture separates feature extraction and fusion reasoning, supporting independent optimization and replacement of each module, which is more in line with the needs of engineering implementation [5,6,11].

In this study, the ASR module adopts the Whisper model, the OCR module retains frame-level recognition results, and the visual description module converts image features into text summaries. This design not only draws on the idea of independent extraction of multimodal features in MDVC [2] but also refers to the texturization processing of modal information in VG-GPLMs [4], enabling multimodal information to be uniformly mapped to the language space and reducing fusion difficulty; this is similar to the scheme of VideoChat-Text [11] texturizing videos through multiple visual models, both simplifying fusion complexity by unifying modal formats. At the same time, the plug-and-play design of model interfaces solves the problem of insufficient flexibility in modal expansion in i-Code [5], supporting the switching of local/open-source models according to actual scenarios and adapting to service access restrictions in different regions. Similar to the dual-version design of VideoChat [11], it balances the needs of different deployment scenarios.

In terms of inference efficiency optimization in GPU-free environments, the caching mechanism, batch processing, and asynchronous calling strategies adopted in this study refer to the engineering optimization experience of i-Code [5] and OmniVL [6], reducing hardware dependence while ensuring performance; Video-ChatGPT [10] loaded models with FP16 precision during inference, which also provides a reference for engineering optimization. In addition, for compatibility issues with new graphics cards and frameworks, containerization/virtual environment isolation configuration is adopted, which is consistent with the deployment practice of industrial multimodal systems [5,8,11]. In the future, model quantization and pruning technologies [8] can be further introduced, combined with the balance experience between model scale and performance in XLSR [3], to further improve the efficiency and flexibility of engineering deployment. Similar to VideoChat-Embed [11], the training and inference costs are reduced by freezing pre-trained model parameters.

6. Conclusion

This paper proposes an engineering-reproducible multimodal structured tag generation system for videos and verifies its effectiveness on a small-scale dataset. Through modular multimodal fusion and Monte Carlo evaluation strategy, the system achieves stable and competitive performance while

DOI: 10.54254/2754-1169/2026.LD31375

ensuring engineering feasibility. The research results of this study show that the scheme integrating visual, speech (ASR), and on-screen text (OCR) multimodal information significantly outperforms single-modal or bimodal combination schemes in video topic classification (Level 1 accuracy 0.86 ± 0.02), fine-grained subclass recognition (Level 2 Micro F1 0.73 ± 0.03), and multi-attribute extraction (Attr mAP 0.67 ± 0.04) tasks. Moreover, the success rate of structured JSON output reaches 95%, proving the core value of multimodal information complementarity in video semantic understanding; at the same time, the loosely coupled modular architecture and Monte Carlo random division evaluation strategy effectively solve the engineering pain points of complex deployment of end-to-end multimodal models and unstable few-shot evaluation. Thus, it is further concluded that modular design and multimodal Prompt fusion are effective paths to balance video understanding performance and engineering implementation feasibility.

This work provides a feasible path and practical reference for the engineering implementation of multimodal video understanding systems. This study has significant reference value for future research in this direction, mainly influencing the engineering implementation paradigm of multimodal video understanding—emphasizing reducing landing costs through module decoupling, environment adaptation, and efficiency optimization under the premise of ensuring performance, and providing a reusable technical scheme for video structured tasks in resource-constrained scenarios or small-scale datasets. Future research should focus on three directions for in-depth exploration: first, introducing explicit temporal modeling mechanisms, combining the Q-Former of Video-LLaMA or the spatiotemporal feature aggregation strategy of VideoChat to improve the semantic understanding ability of long videos and dynamic events; second, optimizing the robustness of modal noise, drawing on the multi-model cross-validation and error filtering methods of Video-ChatGPT to alleviate the problem of ASR/OCR error propagation; third, expanding the scale and diversity of datasets, combining the cross-dataset transfer learning idea of OmniVL to improve the model's generalization performance in professional domain videos and low-resource scenarios, while exploring the dynamic adaptation mechanism of the tag system to meet the personalized needs of different downstream tasks.

References

[1] Wang X., Wang Y. F., Wang W. Y. (2018). Watch, Listen, and Describe: Globally and Locally Aligned Cross-Modal Attentions for Video Captioning. arXiv preprint arXiv: 1804.05448v1.
[2] Iashin V., Rahtu E. (2020). Multi-modal Dense Video Captioning. arXiv preprint arXiv: 2003.07758v2.
[3] Conneau A., Baevski A., Collobert R., et al. (2020). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. arXiv preprint arXiv: 2006.13979v2.
[4] Yu T., Dai W., Liu Z., et al. (2021). Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization. arXiv preprint arXiv: 2109.02401v4.
[5] Yang Z., Fang Y., Zhu C., et al. (2022). i-Code: An Integrative and Composable Multimodal Learning Framework. arXiv preprint arXiv: 2205.01818v2.
[6] Wang J., Chen D., Wu Z., et al. (2022). OmniVL: One Foundation Model for Image-Language and Video-Language Tasks. arXiv preprint arXiv: 2209.07526v2.
[7] Freedman M. H. (2024). Packing Meets Topology. arXiv preprint arXiv: 2301.00295v2.
[8] Yin S., Fu C., Zhao S., et al. (2024). A Survey on Multimodal Large Language Models. arXiv preprint arXiv: 2306.13549v4.
[9] Zhang H., Li X., Bing L. (2023). Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. arXiv preprint arXiv: 2306.02858v4.
[10] Maaz M., Rasheed H., Khan S., et al. (2024). Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv preprint arXiv: 2306.05424v2.
[11] Li K., He Y., Wang Y., et al. (2024). VideoChat: Chat-Centric Video Understanding. arXiv preprint arXiv: 2305.06355v2.