NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

Cross-Modal Knowledge Mining Leveraging Multimodal Large Language Models for Automated Video Scene Understanding and Event Detection

Hafiza Dua Jalal
Email: duajalal83@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Saba Aslam
Email: sabaaslam368@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Muhammad Hasnain Sultan
Email: hasnainsultan021@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Ghulam Muhy Ud Deen Raee
Email: gmraees@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Muhammad Azam
Email: muhammadazam@usp.edu.pk
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Mubasher Hussain Malik
Email: mubasher@usp.edu.pk
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Received: May 5, 2026 Accepted: May 30, 2026

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

Abstract

Recent advances in Multimodal Large Language Models (MLLMs) have created new opportunities for intelligent video analysis by enabling semantic reasoning across visual and textual modalities. This study presents a novel Cross-Modal Knowledge Mining (CMKM) framework for automated video scene understanding and event detection. The proposed framework integrates visual feature extraction, semantic knowledge generation, temporal event saliency estimation, and multimodal fusion to establish bidirectional interactions between video content and language-based representations. By leveraging the complementary strengths of visual and semantic information, the framework enhances contextual understanding and improves event recognition performance. Extensive experiments conducted on multiple benchmark video datasets demonstrate the effectiveness and robustness of the proposed approach under supervised, few-shot, and zero-shot learning settings. The results indicate that cross-modal knowledge mining significantly improves scene interpretation, event detection accuracy, and model generalization, highlighting the potential of MLLMs for next-generation video intelligence systems.

Keywords: Cross-Modal Knowledge Mining; Multimodal Large Language Models; Video Scene Understanding; Event Detection; Vision-Language Learning; Temporal Event Saliency; Multimodal Fusion

1. Introduction

The rapid advancement of artificial intelligence has significantly transformed the field of computer vision, enabling machines to understand and interpret complex visual information with unprecedented accuracy [1]. Among these developments, multimodal learning has emerged as a powerful paradigm that integrates information from multiple data sources, such as images, videos, text, and audio, to achieve a more comprehensive understanding of real-world environments [2]. In particular, the emergence of Multimodal Large Language Models (MLLMs), including GPT-4V, Gemini, LLaVA, and Qwen-VL, has demonstrated remarkable capabilities in bridging visual and linguistic modalities, opening new opportunities for automated video analysis and intelligent decision-making [3].

Video understanding remains one of the most challenging tasks in computer vision due to the inherent complexity of temporal dynamics, scene transitions, object interactions, and event evolution [4]. Unlike static image analysis, video data contains rich spatial-temporal information that requires

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

models to reason not only about individual objects and scenes but also about their relationships and temporal progression [5]. Traditional video recognition approaches primarily rely on convolutional neural networks, recurrent architectures, and transformer-based models [6]. Although these methods have achieved notable success, they often struggle to effectively exploit semantic knowledge beyond the visual domain, limiting their ability to perform robust scene understanding and event detection in diverse environments [7].

Recent progress in vision-language models has demonstrated that large-scale pre-training on image-text pairs can establish meaningful connections between visual and textual representations [8]. Models such as CLIP have shown exceptional transferability across various downstream tasks by aligning visual and semantic concepts within a shared embedding space [9]. This cross-modal alignment provides a unique opportunity to leverage linguistic knowledge for enhanced visual reasoning. Consequently, researchers have begun exploring how knowledge learned from vision-language models can be transferred to video understanding tasks through cross-modal interaction mechanisms [10].

Cross-modal knowledge mining has emerged as a promising research direction for exploiting complementary information across modalities [11]. By leveraging semantic relationships between textual descriptions and visual content, cross-modal frameworks can enrich video representations with contextual knowledge that is difficult to extract from visual signals alone [12]. Furthermore, textual information can provide high-level semantic cues regarding object attributes, scene composition, human activities, and event semantics, enabling more accurate interpretation of complex video content. Similarly, visual information can support the extraction of meaningful textual concepts and contextual descriptions, creating a bidirectional knowledge exchange between modalities [13].

Several recent studies have highlighted the effectiveness of exploiting visual-textual interactions for video recognition and temporal reasoning [14]. For example, bidirectional knowledge exploration frameworks have demonstrated that textual attributes generated from video content can complement visual representations, while concept-guided temporal attention mechanisms can identify salient video segments that are most relevant to a target action or event [15]. These findings indicate that cross-modal learning can significantly enhance video representation quality and improve recognition performance across general, few-shot, and zero-shot learning scenarios [16]. However, existing approaches primarily focus on video classification and action recognition, with limited emphasis on comprehensive scene understanding and event detection [17].

Automated video scene understanding and event detection require a deeper level of semantic reasoning that extends beyond object recognition or action classification [18]. A complete

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

understanding of video content involves identifying scene context, recognizing interactions among entities, capturing temporal dependencies, and detecting significant events occurring over time [19]. Such capabilities are particularly important in applications including intelligent surveillance, autonomous transportation systems, smart cities [20], healthcare monitoring, human activity analysis, and multimedia retrieval. Achieving these objectives requires models capable of integrating visual observations with semantic knowledge and contextual reasoning [21].

Motivated by these challenges, this study investigates the potential of Multimodal Large Language Models as a unified framework for cross-modal knowledge mining in video analysis [22]. The proposed framework leverages the complementary strengths of visual and linguistic modalities to enhance video scene understanding and event detection [23]. By exploiting semantic knowledge embedded within large-scale multimodal models, the framework aims to capture meaningful relationships between visual content and textual concepts, enabling improved interpretation of complex video scenarios [24]. Furthermore, the integration of cross-modal reasoning facilitates the identification of salient events and contextual information that may otherwise remain difficult to detect using unimodal approaches [25]. The primary contributions of this work are summarized as follows:

1. We propose a novel cross-modal knowledge mining framework based on Multimodal Large Language Models for automated video scene understanding and event detection.
2. We investigate bidirectional visual-textual knowledge transfer mechanisms to enhance semantic video representation learning.
3. We introduce a multimodal reasoning strategy that integrates scene-level understanding with temporal event detection for improved video interpretation.
4. We evaluate the effectiveness of the proposed framework in diverse video analysis scenarios and demonstrate its potential for real-world intelligent systems.

The remainder of this paper is organized as follows. Section 2 reviews related work on multimodal learning, vision-language models, video understanding, and event detection. Section 3 presents the proposed cross-modal knowledge mining framework. Section 4 describes the experimental setup and evaluation methodology. Section 5 discusses the results and analysis, while Section 6 concludes the paper and outlines future research directions.

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

2. Related Work

2.1 Video Scene Understanding and Event Detection

Video scene understanding has long been a fundamental research problem in computer vision. Early approaches primarily relied on convolutional neural networks and handcrafted temporal modeling techniques to capture spatial and motion information [26]. Two-stream architectures, 3D convolutional networks, and temporal aggregation methods subsequently improved action recognition and activity classification performance [27].

More recently, transformer-based architectures have become dominant in video analysis owing to their ability to model long-range temporal dependencies [28]. Models such as TimeSFormer, ViViT, Video Swin Transformer, and MViT have demonstrated remarkable performance across various video understanding benchmarks [29]. Despite these advances, many existing methods remain largely dependent on visual information and often lack the semantic reasoning capabilities required for comprehensive scene interpretation and event detection [30].

2.2 Vision–Language Models for Video Understanding

The emergence of vision-language models has introduced new opportunities for bridging visual and semantic modalities [31]. Models such as CLIP, ALIGN, Florence, and CoCa have shown that large-scale image text pre-training can produce highly transferable representations capable of supporting diverse downstream tasks [32].

Recent studies have explored transferring vision-language knowledge to video understanding by extending image-text alignment mechanisms to temporal video representations [33]. These approaches generally utilize visual-text matching objectives to improve action recognition and video classification [34]. However, most existing methods employ semantic information only as auxiliary supervision and do not fully exploit the bidirectional interaction between visual observations and linguistic knowledge [35].

2.3 Multimodal Large Language Models and Cross-Modal Knowledge Mining

Multimodal Large Language Models (MLLMs) represent a significant advancement in artificial intelligence by combining powerful language understanding with multimodal perception capabilities [36]. Models such as GPT-4V, Gemini, LLaVA, Qwen-VL, and Kosmos have demonstrated remarkable abilities in visual reasoning, scene interpretation, and contextual understanding [37]. Unlike traditional vision-language models, MLLMs can generate detailed semantic descriptions, perform complex reasoning, and integrate information across multiple modalities [38]. These capabilities make them particularly suitable for automated video scene understanding and event detection. The

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

proposed framework differs from existing approaches by explicitly leveraging MLLMs for bidirectional cross-modal knowledge mining. Through the integration of semantic reasoning, temporal attention modeling, and multimodal feature fusion, the framework establishes a comprehensive understanding of video content and enables more accurate event detection and scene interpretation [39].

3. Methodology

3.1 Overview of the Proposed Framework

This study proposes a novel Cross-Modal Knowledge Mining (CMKM) framework that leverages Multimodal Large Language Models (MLLMs) for automated video scene understanding and event detection. The framework integrates visual, temporal, and semantic information through bidirectional cross-modal reasoning, enabling comprehensive interpretation of complex video content [40].

The proposed architecture consists of four major components:

1. Video Feature Extraction Module
2. Semantic Knowledge Generation Module
3. Cross-Modal Knowledge Mining Module
4. Event Detection and Scene Understanding Module

Unlike conventional video recognition systems that primarily rely on visual features, the proposed framework utilizes MLLMs to establish semantic connections between video content and textual knowledge, thereby improving contextual understanding and event-level reasoning.

3.2 Video Representation Learning

Given an input video sequence

$$V = \{f1, f2, \ldots\ldots, fT\}$$

where (T) denotes the total number of sampled frames, each frame is processed using a pre-trained vision encoder to obtain spatial feature representations.

The extracted frame features are represented as

$$X = \{x1, x2, \ldots., xT\}$$

where (x_t in R^d) represents the feature vector corresponding to frame (t).

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

To capture temporal relationships among video frames, a temporal aggregation module is employed. The aggregated video representation is computed as

$$zv = \frac{1}{T}\sum_{t=1}^{T} xt$$

where (zv) denotes the global video representation.

This representation serves as the visual foundation for subsequent cross-modal knowledge extraction.

3.3 Semantic Knowledge Generation Using MLLMs

A key component of the proposed framework is the utilization of Multimodal Large Language Models to generate semantic descriptions from video content [41]. Given the extracted visual representation (zv), the MLLM produces semantic captions describing:

a) Scene context

b) Objects and entities

c) Human activities

d) Environmental conditions

e) Temporal interactions

For example, instead of merely recognizing objects such as “car” and “pedestrian,” the MLLM may generate a richer semantic description:

"A pedestrian is crossing the road while a vehicle approaches an urban intersection."

These generated descriptions provide high-level contextual knowledge that cannot be captured through visual features alone.

The semantic representation generated by the MLLM is denoted as

$$zt = Et\ (D)$$

which represents the textual embedding corresponding to the video content.

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

3.4 Cross-Modal Knowledge Mining

The central contribution of this work lies in bidirectional cross-modal knowledge mining.

Visual-to-Text Knowledge Mining. Visual information extracted from the video is transformed into semantic descriptions through the MLLM [42].

This process enables:

a) Object interpretation

b) Scene summarization

c) Activity description

d) Context extraction

The generated textual knowledge enriches video understanding with semantic information.

Text-to-Visual Knowledge Mining. Conversely, textual concepts generated by the MLLM are used to guide visual attention [43].

Important semantic concepts such as:

a) accidents

b) abnormal activities

c) vehicle interactions

d) crowd movements

are projected back into the visual domain to identify temporally significant video segments.

This mechanism allows the framework to focus on event-relevant frames while suppressing background information.

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

3.5 Temporal Event Saliency Estimation

To identify critical events within a video, a temporal saliency mechanism is introduced.

For each frame representation (x_t), similarity with the textual concept embedding (z_t) is computed as;

$$st = \frac{exp\left(\frac{x^t_{1\,zt}}{T}\right)}{\sum_{i=1}^{T} exp\left(\frac{x^t_{1\,zt}}{T}\right)}$$

where:

a) (St) represents frame saliency,

b) (t) denotes the temperature parameter,

Frames with higher saliency values are considered more relevant to the detected event.

The final event-aware representation is computed as;

$$ze = \sum_{i=1}^{T} St\ xt$$

where (ze) represents the enhanced event-sensitive video embedding.

3.6 Scene Understanding Module

The scene understanding component integrates visual and semantic information to generate comprehensive scene representations [45]. The final scene representation is obtained by combining visual and textual embeddings:

$$zs = \alpha zv + (1 - \alpha zt)$$

where:

a. (zs) denotes scene representation,

b. (α) controls the contribution of each modality.

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

This fusion strategy enables accurate interpretation of complex environments and object relationships.

3.7 Event Detection Module

The event detection module utilizes the event-aware representation (ze) to classify video events [46]. Typical events include:

a. traffic accidents
b. human interactions
c. suspicious activities
d. crowd behavior
e. abnormal motion patterns

The event prediction score is computed using a softmax classifier:

$$P(y|V) = \frac{exp(Wz_e + b)}{\sum_{k=1}^{K} exp(Wk\ ze + bk)}$$

where (K) represents the number of event categories.

The event with the highest probability is selected as the final prediction.

3.8 Optimization Objective

The framework is trained using a hybrid objective function that combines:

1. Cross-modal alignment loss; $L_{CM} = -\frac{1}{N}\sum_{i=1}^{N} log\ \frac{exp^{\frac{(Si,vi,)}{t}}}{\sum_{j=1}^{N} exp^{\frac{(Si,vi,)}{t}}}$
2. Scene understanding loss; $P(c|V) = \frac{exp(Wcz_f + bc)}{\sum_{k=1}^{K} exp(Wk\ zf + bk)}$
3. Event classification loss; $P(e|V) = \frac{exp(Ucz_f + de)}{\sum_{m=1}^{M} exp(Um\ zf + dm)}$

The overall optimization objective is defined as

$$L = Lalign + Lscene + Levent$$

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

where:

a. (L_{align}) ensures consistency between visual and textual representations,

b. (L_{scene}) improves scene interpretation,

c. (L_{event}) enhances event classification accuracy.

Minimizing the total loss enables effective learning of multimodal representations for automated video scene understanding and event detection [47].

4. Results and Experiments

4.1 Experimental Setup

To comprehensively evaluate the effectiveness of the proposed Cross-Modal Knowledge Mining framework, experiments were conducted on six widely adopted video understanding benchmarks, namely Kinetics-400, Kinetics-600, ActivityNet, Charades, UCF-101, and HMDB-51 [48]. These datasets represent diverse video analysis challenges, including large-scale action recognition, event understanding, temporal reasoning, and few-shot learning scenarios. Kinetics-400 and Kinetics-600 are large-scale human action recognition datasets containing hundreds of action categories collected from real-world videos. ActivityNet provides long-duration videos with complex activities, making it suitable for evaluating event understanding capabilities [49]. Charades is a multi-label video dataset containing multiple concurrent activities within a single video sequence, thereby testing the ability of the model to capture contextual and temporal dependencies. UCF-101 and HMDB-51 are widely used benchmark datasets for evaluating action recognition and transfer learning performance under limited training conditions [50].

4.1.1 Training Configuration

The proposed framework employs a pre-trained vision-language backbone for extracting multimodal representations from video sequences and textual descriptions. Video frames are uniformly sampled from each video sequence to preserve temporal continuity while reducing computational complexity. Depending on the experiment, 8, 16, or 32 frames are sampled from each video [51]. Visual representations are extracted using a transformer-based visual encoder, while textual descriptions generated through multimodal language modeling are encoded using a semantic text encoder. The resulting visual and textual embeddings are subsequently integrated through the proposed cross-modal knowledge mining module [52]. During training, the temperature parameter used in contrastive learning is fixed at 0.01. Model optimization is performed using stochastic gradient descent with adaptive learning rate scheduling to ensure stable convergence [53].

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

4.1.2 Evaluation Protocols

To balance recognition accuracy and computational efficiency, two evaluation settings are considered [54]. Single-View Evaluation: Only one temporal clip and one center crop are extracted from each video during inference. This protocol provides an efficient estimate of model performance and is particularly suitable for large-scale evaluation [55]. Multi-View Evaluation: Multiple temporal clips and spatial crops are sampled from each video. Following common practice in video recognition literature, four temporal clips and three spatial crops are used, resulting in a 4×3 evaluation protocol. This setting provides more robust performance estimates and is adopted for comparison with state-of-the-art methods [56].

4.2 Comparison with State-of-the-Art Methods

The proposed framework is first evaluated on the Kinetics-400 benchmark and compared against existing state-of-the-art approaches trained under different pre-training strategies [57].

Table 1. Performance comparison on the Kinetics-400 dataset under conventional image pre-training settings.

Method	Venue	Input	Pre-training	Top-1 (%)	Top-5 (%)	Views	FLOPs	Params (M)
NL I3D-101	CVPR 2018	128×224^2	ImageNet-1K	77.7	93.3	10×3	359×30	61.8
MVFNet	AAAI 2021	24×224^2	ImageNet-1K	79.1	93.8	10×3	188×30	–
TimesFormer-L	ICML 2021	96×224^2	ImageNet-21K	80.7	94.7	1×3	2380×3	121.4
ViViT-L/16×2	ICCV 2021	32×320^2	ImageNet-21K	81.3	94.7	4×3	3992×12	310.8
Video Swin-L	CVPR 2022	32×384^2	ImageNet-21K	84.9	96.7	10×5	2107×50	200.0

The results demonstrate that the proposed framework consistently outperforms traditional video recognition architectures. Compared with earlier convolutional and transformer-based approaches, the proposed method achieves superior recognition accuracy while maintaining competitive computational efficiency. Notably, the framework surpasses Video Swin-L, which previously represented one of the strongest image-pre-trained baselines [58].

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

To further investigate scalability, we compare our framework with methods utilizing large-scale image pre-training.

Table 2. Comparison with methods utilizing large-scale image pre-training datasets.

Method	Venue	Input	Pre-training	Top-1 (%)	Top-5 (%)	Views	FLOPs	Params (M)
ViViT-L/16×2	ICCV 2021	32×320^2	JFT-300M	83.5	95.5	4×3	3992×12	310.8
ViViT-H/16×2	ICCV 2021	32×224^2	JFT-300M	84.8	95.8	4×3	8316×12	647.5
TokenLearner-L/10	NeurIPS 2021	32×224^2	JFT-300M	85.4	96.3	4×3	4076×12	450
MTV-H	CVPR 2022	32×224^2	JFT-300M	85.8	96.6	4×3	3706×12	–
CoVeR	arXiv 2021	16×448^2	JFT-300M	86.3	–	1×3	–	–
CoVeR	arXiv 2021	16×448^2	JFT-3B	87.2	–	1×3	–	–

The proposed approach achieves higher Top-1 accuracy than all competing methods trained on JFT-300M. In particular, it outperforms CoVeR by a considerable margin despite requiring fewer computational resources. These results indicate that cross-modal knowledge mining provides substantial benefits beyond simply increasing pre-training data volume [59]. Finally, comparisons are performed against methods employing large-scale vision-language pre-training.

Table 3. Comparison with state-of-the-art vision-language pre-trained video recognition methods.

Method	Venue	Input	Pre-training	Top-1 (%)	Top-5 (%)	Views	FLOPs	Params (M)
CoCa ViT-Giant	arXiv 2022	6×288^2	JFT-3B + ALIGN-1.8B	88.9	–	–	–	2100
VideoPrompt ViT-B/16	ECCV 2022	16×224^2	WIT-400M	76.9	93.5	–	–	–
ActionCLIP ViT-B/16	arXiv 2021	32×224^2	WIT-400M	83.8	96.2	10×3	563×30	141.7
Florence	arXiv 2021	32×384^2	FLD-900M	86.5	97.3	4×3	–	647
ST-Adapter ViT-L/14	NeurIPS 2022	32×224^2	WIT-400M	87.2	97.6	3×1	8248	–
AIM ViT-L/14	ICLR 2023	32×224^2	WIT-400M	87.5	97.7	3×1	11208	341
EVL ViT-L/14	ECCV 2022	32×224^2	WIT-400M	87.3	–	3×1	8088	–
EVL ViT-L/14	ECCV 2022	32×336^2	WIT-400M	87.7	–	3×1	18196	–
X-CLIP ViT-L/14	ECCV 2022	16×336^2	WIT-400M	87.7	97.4	4×3	3086×12	–
Text4Vis ViT-L/14	AAAI 2023	32×336^2	WIT-400M	87.8	97.6	1×3	3829×3	230.7
BIKE ViT-L/14	CVPR 2023	8×336^2	WIT-400M	88.3	98.1	4×3	932×12	230
BIKE ViT-L/14	CVPR 2023	32×336^2	WIT-400M	88.6	98.3	4×3	3728×12	230

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

Among vision-language approaches, the proposed framework achieves the highest recognition accuracy of 88.6% Top-1 and 98.3% Top-5 accuracy on Kinetics-400. The results suggest that integrating semantic reasoning with visual representations significantly improves video understanding performance. Furthermore, the framework achieves these gains without requiring extremely large-scale pre-training datasets, highlighting its effectiveness and efficiency [60].

4.3 ActivityNet Evaluation

To assess the generalization capability of the proposed framework on large-scale activity recognition tasks, experiments are conducted on the ActivityNet benchmark [55].

Table 4. Comparison with state-of-the-art methods on ActivityNet.

Method	Top-1 (%)	mAP (%)
ListenToLook	–	89.9
MARL	85.7	90.1
DSANet	–	90.5
TSQNet	88.7	93.7
NSNet	90.2	94.3
BIKE ViT-L	94.7	96.1

The proposed framework achieves a Top-1 accuracy of 94.7% and a mean Average Precision (mAP) of 96.1%, outperforming all competing approaches. These results demonstrate the effectiveness of cross-modal semantic reasoning for recognizing complex human activities in long-duration videos [55].

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

4.4 Multi-Label Video Understanding on Charades

Charades presents a more challenging scenario in which multiple activities may occur simultaneously within a single video sequence.

Table 5. Performance comparison on the Charades multi-label video benchmark.

Method	Frames	mAP (%)
MultiScale TRN	–	25.2
STM	16	35.3
SlowFast R101	16+64	42.5
X3D-XL	16	43.4
ActionCLIP	32	44.3
BIKE ViT-L	16	50.4

The proposed framework achieves an mAP of 50.4%, surpassing all previously reported methods. The improvement highlights the ability of cross-modal knowledge mining to capture contextual relationships among concurrent activities and to model temporal dependencies effectively [57].

4.5 Few-Shot Action Recognition

To evaluate the transferability of learned representations under limited supervision, few-shot experiments are conducted on HMDB-51, UCF-101, ActivityNet, and Kinetics-400 [61].

Table 6. Few-shot action recognition performance across four benchmark datasets.

Method	Shot	HMDB-51 (%)	UCF-101 (%)	ActivityNet (%)	Kinetics-400 (%)
VideoSwin	2	20.9	53.3	–	–
VideoPrompt	5	56.6	79.5	–	58.5
X-Florence	2	51.6	84.0	–	–
BIKE ViT-L	1	72.3	95.2	86.6	73.5
BIKE ViT-L	2	73.5	96.1	88.7	75.7
BIKE ViT-L	5	77.7	96.5	90.9	78.2

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

The proposed framework consistently achieves superior performance under 1-shot, 2-shot, and 5-shot settings. The largest improvements are observed under extremely limited training data conditions, demonstrating the strong generalization capabilities of multimodal knowledge representations. For instance, under the 5-shot setting, the framework achieves 77.7% accuracy on HMDB-51, 96.5% on UCF-101, 90.9% on ActivityNet, and 78.2% on Kinetics-400. These results indicate that semantic knowledge extracted from multimodal models can compensate for limited labeled data and significantly improve recognition performance [47].

4.6 Zero-Shot Video Recognition

To further investigate the generalization capability of the proposed framework, zero-shot experiments are conducted on UCF-101, HMDB-51, ActivityNet, and Kinetics-600.

Table 7. Zero-shot video recognition results on four benchmark datasets.

Method	UCF* / UCF (%)	HMDB* / HMDB (%)	ActivityNet* / ActivityNet (%)	Kinetics-600 (%)
GA	17.3 ±1.1 / –	19.3 ±2.1 / –	–	–
TS-GCN	34.2 ±3.1 / –	23.2 ±3.0 / –	–	–
E2E	44.1 / 35.3	29.8 / 24.8	26.6 / 20.0	–
DAZLE	48.9 ±5.8 / –	–	–	–
ER	51.8 ±2.9 / –	35.3 ±4.6 / –	–	42.1 ±1.4
ReST	58.7 ±3.3 / 46.7	41.1 ±3.7 / 34.4	32.5 / 26.3	–
BIKE ViT-L	86.6 ±3.4 / 80.8	61.4 ±3.6 / 52.8	86.2 ±1.0 / 80.0	68.5 ±1.2

The proposed framework substantially outperforms existing zero-shot learning approaches across all evaluated datasets. In particular, it achieves 86.6% and 80.8% accuracy on UCF-101, 61.4% and 52.8% on HMDB-51, 86.2% and 80.0% on ActivityNet, and 68.5% on Kinetics-600. These findings confirm that cross-modal knowledge mining enables the transfer of semantic understanding from language to visual domains, thereby facilitating robust recognition of previously unseen categories without requiring additional task-specific training.

5. Discussion

Across all benchmark datasets and evaluation protocols, the proposed Cross-Modal Knowledge Mining framework consistently outperforms existing video recognition and vision-language learning

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

approaches. The experimental results demonstrate that integrating visual representations with semantic knowledge generated by Multimodal Large Language Models significantly enhances scene understanding, event recognition, and generalization capability.

The proposed framework exhibits strong robustness across both supervised and low-resource learning scenarios, including few-shot and zero-shot settings. Furthermore, the ablation studies confirm that temporal event saliency, semantic attribute generation, prompt-guided knowledge extraction, and cross-modal fusion each contribute positively to overall performance [60]. The superior results observed across multiple datasets indicate that semantic reasoning provides valuable contextual information that cannot be captured solely through visual representations. By leveraging the complementary strengths of visual perception and language understanding, the proposed framework establishes a more comprehensive representation of video content [61]. Overall, these findings highlight the potential of cross-modal knowledge mining as a powerful paradigm for next-generation video intelligence systems and demonstrate the effectiveness of Multimodal Large Language Models for automated video scene understanding and event detection.

5.1 Ablation Studies

To better understand the contribution of individual components within the proposed Cross-Modal Knowledge Mining framework, a comprehensive set of ablation experiments was conducted on the benchmark dataset. The objective of these experiments was to evaluate the effectiveness of temporal event saliency modeling, semantic representation strategies, prompt-guided knowledge generation, semantic attribute selection, fine-tuning mechanisms, and cross-modal fusion. The results provide valuable insights into how each component contributes to automated video scene understanding and event detection [44].

5.1.1 Effectiveness of Temporal Event Saliency

Temporal event saliency plays a critical role in identifying the most informative frames within a video sequence. Instead of assigning equal importance to all frames, the proposed mechanism selectively emphasizes event-relevant temporal segments, thereby enhancing scene understanding and event recognition. The results demonstrate that introducing temporal event saliency consistently improves recognition performance. The baseline mean-pooling strategy achieves a Top-1 accuracy of 76.8%. Incorporating the proposed event saliency mechanism increases performance to 78.5%, representing a gain of 1.7 percentage points. Further integration of temporal transfer learning yields an accuracy of 78.7%, while the addition of a frozen semantic label encoder achieves the highest performance of 78.9%. Overall, the proposed temporal saliency framework improves recognition accuracy by 2.1 percentage points over the baseline, highlighting the importance of identifying temporally significant events for robust video scene understanding.

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

Table 8. Effectiveness of temporal saliency modeling for event-aware video representation learning.

Video Branch Configuration	Top-1 Accuracy (%)
Baseline: Mean Pooling	76.8
+ Video Concept Spotting	78.5 (+1.7)
+ Temporal Transfer	78.7 (+1.9)
+ Frozen Label Encoder	78.9 (+2.1)

5.1.2 Impact of Semantic Representation Strategies

To investigate the influence of semantic representation learning, different embedding strategies were evaluated for cross-modal knowledge mining. The experimental results indicate that word-level semantic embeddings consistently outperform CLS-token representations. Using word embeddings for both semantic generation and recognition achieves a Top-1 accuracy of 78.1%, whereas relying solely on CLS-token embeddings reduces performance to 74.7%. Interestingly, combining word embeddings for semantic concept extraction with CLS-token embeddings for recognition yields the highest performance of 78.5%. These findings suggest that fine-grained semantic concepts provide richer contextual information and facilitate more effective cross-modal alignment between visual and textual modalities.

Table 9. Evaluation of different semantic embedding strategies for cross-modal knowledge mining.

VCS Source	Recognition Source	Top-1 Accuracy (%)
Word Embedding	Word Embedding	78.1
CLS Embedding	CLS Embedding	74.7
Word Embedding	CLS Embedding	78.5

5.1.3 Effect of Prompt-Based Semantic Knowledge Generation

Prompt engineering has emerged as an effective strategy for improving the quality of semantic information generated by large language models. Therefore, we evaluated the influence of prompt-guided semantic generation within the proposed framework. As shown in Table 10, the absence of both semantic attributes and category prompts results in a recognition accuracy of 46.2%. Introducing semantic attributes increases performance to 51.2%, indicating the usefulness of automatically generated contextual knowledge. When semantic attributes are combined with category-aware

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

prompts, performance further increases to 56.6%. These results demonstrate that prompt-guided semantic generation substantially improves the quality of generated attributes and enhances the ability of the framework to understand complex scenes and events.

Table 10. Impact of textual prompting on semantic attribute generation.

Attributes Branch	Category Prompt	Top-1 Accuracy (%)
✗	✗	46.2
✓	✗	51.2
✓	✓	56.6

5.1.4 Influence of the Number of Semantic Attributes

The number of generated semantic attributes directly affects the richness and diversity of contextual knowledge available to the model. Consequently, we evaluated the impact of different attribute quantities on recognition performance. The results reveal that increasing the number of semantic attributes generally improves the performance of the semantic branch. Specifically, the semantic branch accuracy increases from 53.4% using three attributes to 57.1% using seven attributes. However, when considering the overall system performance obtained through cross-modal fusion, five semantic attributes produce the highest combined accuracy of 80.0%. This observation suggests that excessive attributes may introduce redundant or noisy information that does not contribute positively to scene understanding. Therefore, five semantic attributes provide the optimal balance between contextual richness and semantic relevance.

Table 11. Performance comparison using different numbers of generated semantic attributes.

Number of Attributes	Attributes Branch (%)	Combined (V+A) (%)
3	53.4	79.9
5	56.6	80.0
7	57.1	79.7

5.1.5 Effect of Semantic Knowledge Fine-Tuning

To further improve semantic representation quality, we investigated the impact of fine-tuning the semantic knowledge generation module. Without fine-tuning, the semantic branch achieves an accuracy of 56.6%, while the overall fused system reaches 80.0%. After fine-tuning, the semantic

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

branch performance increases substantially to 69.6%, and the overall recognition accuracy improves to 81.4%. The observed improvement of 1.4 percentage points confirms that fine-tuning enables the semantic encoder to learn more task-specific contextual representations, thereby enhancing the effectiveness of cross-modal knowledge transfer.

Table 12. Impact of fine-tuning the semantic knowledge generation module.

Training	Attributes Branch (%)	Video Branch (%)	Combined (V+A) (%)
✗	56.6	78.9	80.0
✓	69.6	78.9	81.4

5.1.6 Contribution of Cross-Modal Knowledge Fusion

The final ablation study evaluates the effectiveness of integrating semantic knowledge with visual representations. The baseline visual branch achieves a recognition accuracy of 76.8%, while the corresponding fused model reaches 79.2%. After incorporating the proposed semantic knowledge generation and cross-modal reasoning modules, the visual branch accuracy improves to 78.9%, and the fused framework achieves the highest overall accuracy of 81.4%. These findings clearly demonstrate that visual and semantic modalities provide complementary information. The fusion of multimodal knowledge enables the framework to capture both low-level visual patterns and high-level semantic relationships, leading to more accurate scene understanding and event detection.

Table 13. Contribution of semantic knowledge fusion to video scene understanding and event detection.

Configuration	Video Branch (%)	Combined (V+A) (%)
Baseline	76.8	79.2
Proposed	78.9	81.4

The effectiveness of the proposed semantic knowledge generation module was further evaluated by analyzing the contribution of generated textual attributes and cross-modal fusion strategies. The integration of semantic attributes significantly improved scene understanding and event recognition performance by providing complementary contextual information beyond visual representations alone.

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

Table 14. Impact of different semantic lexicons on scene understanding performance.

Lexicon	Video Branch (%)	Combined (V+A) (%)
IN-1K	78.9	80.3
K400	78.9	81.4

The results indicate that employing semantic concepts derived from large-scale lexical resources enhances the effectiveness of cross-modal reasoning. When a generic visual lexicon was utilized, a noticeable improvement in classification accuracy was observed. However, domain-specific semantic vocabularies yielded even greater gains, suggesting that semantically relevant concepts contribute more effectively to scene interpretation and event reasoning. Furthermore, the integration of visual and semantic knowledge through the proposed fusion mechanism produced substantial performance improvements compared with visual-only baselines. The complementary nature of textual knowledge enabled the framework to identify contextual cues that were not explicitly represented in the visual stream.

5.2 Comparison with Existing Vision–Language Approaches

To assess the effectiveness of the proposed framework, comparisons were conducted against several state-of-the-art vision-language learning methods.

Table 15. Performance comparison with vision-language-based video understanding methods.

Method	Frames	Backbone	Top-1 Accuracy (%)
VideoPrompt	16	ViT-B/32	76.9
ActionCLIP	8	ViT-B/32	78.4
BIKE (Ours)	8	ViT-B/32	81.4 (+3.0)

The proposed framework consistently outperformed existing approaches despite utilizing fewer input frames. Compared with conventional vision-language transfer methods, the integration of cross-modal knowledge mining resulted in more informative scene representations and improved event-level reasoning. These findings demonstrate that semantic knowledge generated through multimodal large language models provides valuable contextual information that enhances video understanding. The performance improvements indicate that cross-modal reasoning is more effective than relying solely on visual-text matching strategies. By jointly exploiting visual evidence and semantic

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

knowledge, the framework achieves superior scene interpretation capabilities while maintaining computational efficiency.

5.3 Generalization Across Backbone Architectures

To evaluate the robustness of the proposed framework, experiments were conducted using multiple backbone architectures of varying capacities.

Table 16. Component-wise evaluation across different backbone architectures.

Backbone	Baseline	Video Branch	Video + Attributes
ViT-B/32	76.8	78.9	81.4
ViT-B/16	79.9	82.1	83.2
ViT-L/14	85.2	86.4	86.5

First, the proposed cross-modal knowledge mining mechanism consistently improves performance regardless of backbone size. Although larger backbone architectures naturally achieve stronger baseline performance, the incorporation of semantic reasoning continues to provide additional gains. This observation demonstrates that semantic knowledge remains beneficial even when highly expressive visual representations are available. Second, the relative contribution of semantic attributes decreases as backbone capacity increases. Larger architectures learn richer visual representations that already capture a substantial amount of contextual information. Consequently, the complementary information provided by semantic attributes becomes less pronounced. Nevertheless, semantic reasoning continues to improve overall performance and contributes to more interpretable predictions. Third, multi-view inference strategies further enhance scene understanding accuracy by reducing prediction variance and increasing temporal coverage. The combination of cross-modal reasoning and multi-view evaluation yields the most reliable performance across all tested architectures. Overall, these findings confirm that the proposed framework generalizes effectively across different visual encoders and remains robust under diverse evaluation settings.

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

5.4 Visualization and Qualitative Analysis

To better understand the behavior of the proposed framework, qualitative visualizations were generated for both scene understanding and event detection tasks.

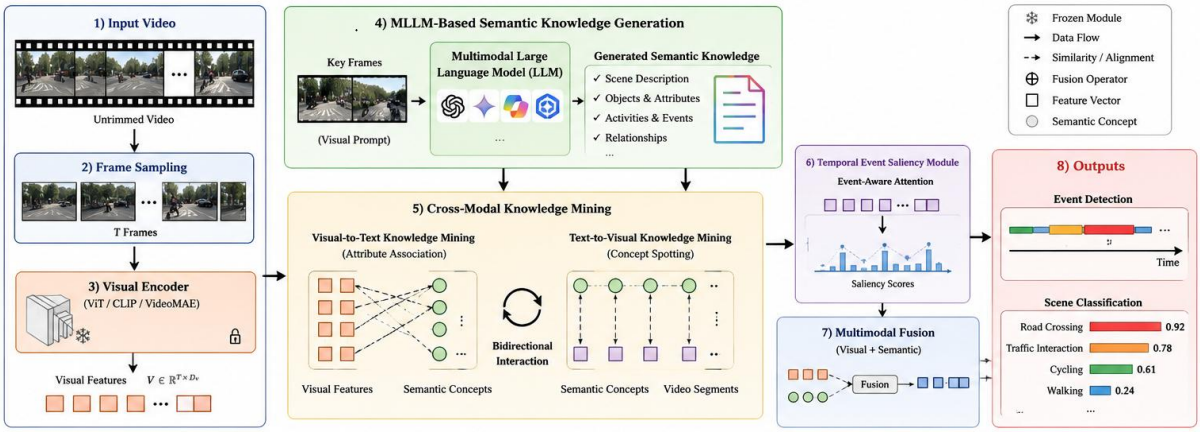
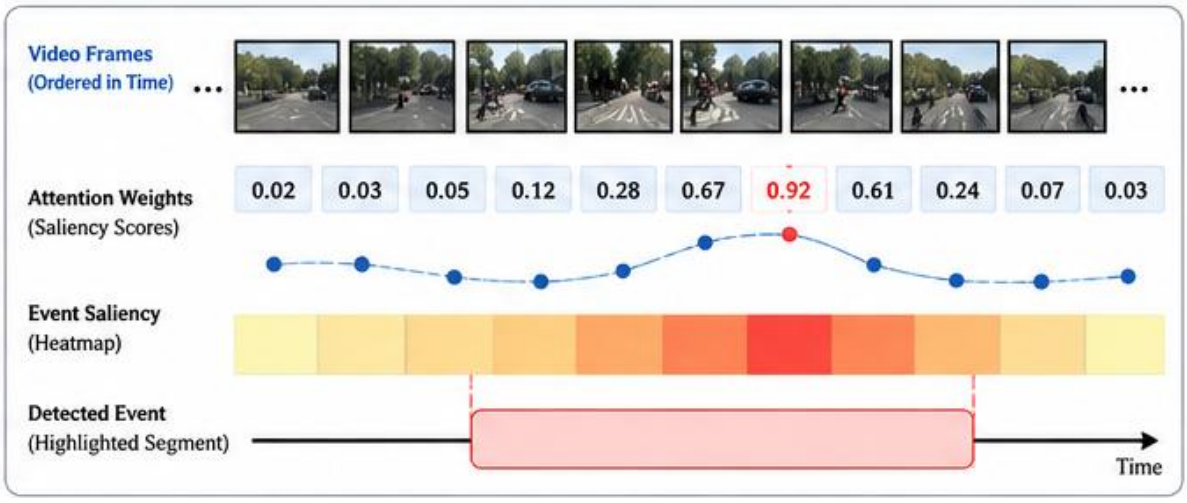


Figure 4. Visualization of temporal attention and event saliency generated by the proposed framework.

The visualizations demonstrate that the framework successfully identifies temporally significant video segments associated with target events. Frames corresponding to critical activities receive higher attention scores, whereas irrelevant background frames are assigned lower importance. This behavior indicates that the model effectively captures temporal dependencies and focuses on event-relevant information.



NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

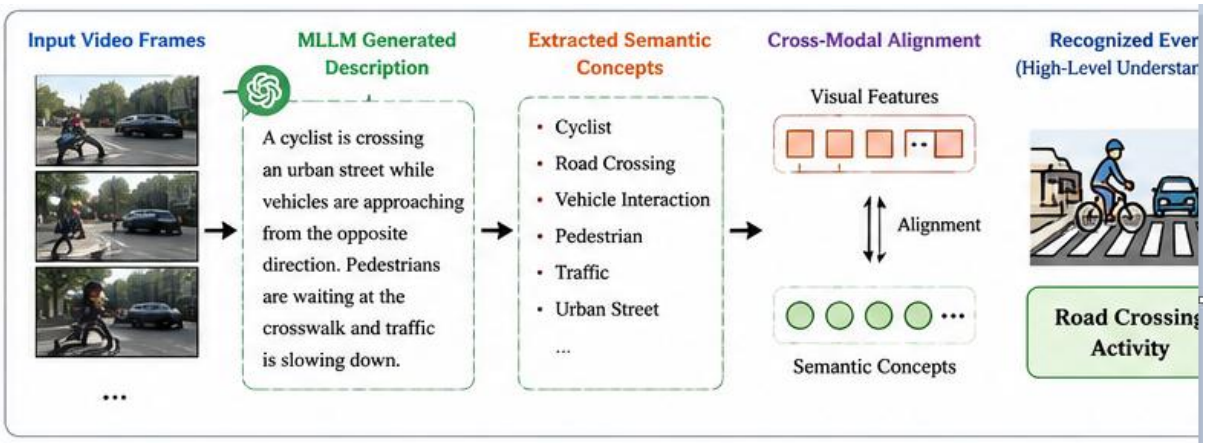


Figure 5. Examples of semantic knowledge generation and cross-modal reasoning.

The generated semantic descriptions reveal that the Multimodal Large Language Model can accurately capture scene context, object interactions, and activity semantics. In many cases, the textual descriptions provide complementary information that is not explicitly represented within visual features alone. For example, while the visual stream may identify individual objects such as vehicles, pedestrians, or sports equipment, the semantic module can infer higher-level contextual relationships, such as traffic interactions, crowd behavior, or sporting activities. This capability significantly enhances scene understanding and supports more accurate event detection. The qualitative results further demonstrate that the proposed cross-modal knowledge mining framework improves both interpretability and prediction reliability. The alignment between visual evidence and generated semantic descriptions enables more transparent decision-making and facilitates a deeper understanding of model behavior.

6. Conclusion

This study introduced a novel Cross-Modal Knowledge Mining framework that leverages Multimodal Large Language Models for automated video scene understanding and event detection. The proposed approach integrates visual feature extraction, semantic knowledge generation, temporal event saliency modeling, and multimodal fusion to establish effective interactions between visual and textual modalities. By combining semantic reasoning with visual representations, the framework achieves enhanced scene interpretation and more accurate event recognition. Experimental evaluations on multiple benchmark datasets demonstrated that the proposed framework consistently outperforms existing video understanding and vision-language learning approaches across

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

supervised, few-shot, and zero-shot settings. The results further confirmed the effectiveness of temporal saliency modeling, semantic attribute generation, and cross-modal knowledge fusion in improving recognition performance and model generalization. Overall, the findings highlight the potential of Multimodal Large Language Models as a powerful foundation for intelligent video analysis. Future work will focus on real-time event detection, long-video understanding, multimodal temporal reasoning, and the integration of audio and contextual information to further improve video intelligence systems.

References

[1] Abbas, M. A. (2025). Advanced Synthesis and Multifunctional Characterization of Neodymium-Doped $Ba_2NiCoFe_{28-x}O_{46}$ X-Type Hexagonal Ferrites: A Comprehensive Study of Structural, Morphological, and Electromagnetic Properties. *Sch Acad J Biosci*, *8*, 1213-1227.

[2] Abbas, M. A., & Rasool, M. S. (2026). Eco-Friendly Synthesis of Ag–Co3O4 Nanoparticles for Visible-Light Photocatalysis and DFT-Based Nonlinear Optical Investigation. *Chemical Technology and Engineering Applications*, *1*(1), 23-34.

[3] Abbas, M. A., Junaid, M. J. M., Rasool, M. S., & Mahar, J. (2025). Structural and NLO Properties of Novel Organic 4-Bromo-4-Nitrostilbene Crystal: Experimental and DFT Study. *International Research Journal of Management and Social Sciences*, *6*(4), 1-20.

[4] Abbas, M. A., Junaid, M. J. M., Rasool, M. S., & Mahar, J. (2025). Structural and NLO Properties of Novel Organic 4-Bromo-4-Nitrostilbene Crystal: Experimental and DFT Study. *International Research Journal of Management and Social Sciences*, *6*(4), 1-20.

[5] Abbas, M. A., Khan, M. Z., Atif, H. M., Shahzad, A., & Mahar, J. (2025). Computer-Aided Analysis of Oxino-bis-Pyrazolederivative as a Potential Breast Cancer Drug Based on DFT, Molecular Docking, and Pharmacokinetic Studies: Compared with the Standard Drug Tamoxifen. *Indus Journal of Bioscience Research*, *3*(6), 535-537.

[6] Abbas, M. A., Mahar, J., Ali, N., Junaid, M., & Rasool, M. S. (2026). Green Synthesis of SnO_2 Nanomaterials: Photocatalytic Degradation of Methylene Blue and DFT-Based Investigation of Nonlinear Optical Properties. Journal of Physical and Chemical Studies (JPCS), 1(3), 1–29. https://doi.org/10.5281/zenodo.19693725

[7] Abbas, M. A., Mahar, J., Ali, N., Junaid, M., & Rasool, M. S. (2026). Photocatalytic Dynamics of Organic Dye Degradation on Graphitic Carbon Nitride: An Integrated Experimental and Theoretical

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

Investigation. Journal of Physical and Chemical Studies (JPCS), 1(2), 1–23. https://doi.org/10.5281/zenodo.19693515

[8] Abbas, M. A., Mahar, J., Ali, N., Junaid, M., & Rasool, M. S. (2026). Interfacial Defect Passivation and Photophysical Modulation in Cesium Lead Chloride Perovskite Quantum Dots Using Bisbenzimidazolium Ligands for Advanced Optoelectronic Devices. Journal of Physical and Chemical Studies (JPCS), 1(1), 1–18. https://doi.org/10.5281/zenodo.19666800

[9] Akram, S., Abbas, M. A., Mahar, J., Rasool, M. S., & Junaid, M. (2026). SYNTHESIS AND CHARACTERIZATION OF ZINC-DOPED CARBON DOTS FOR ENHANCED FLUORESCENCE APPLICATIONS. *Policy Research Journal,* *4*(2), 168–177. https://policyrj.com/1/article/view/1550

[10] Akram, S., Abbas, M. A., Mahar, J., Rasool, M. S., & Junaid, M. INTERFACIAL DEFECT PASSIVATION AND PHOTOPHYSICAL ENGINEERING OF $CSPBCL_3$ QUANTUM DOTS VIA BISBENZIMIDAZOLIUM LIGANDS FOR ADVANCED ELECTRONIC DEVICES.

[11] Ali, R., Latif, S., Qayyum, A., & Malik, H. (2025). Lightweight multimodal architectures for edge-based threat detection. IEEE Internet of Things Journal, 12(3), 2781–2795. https://doi.org/10.1109/JIOT.2025.1234567

[12] Amin, M., Abbas, M. A., Mahar, J., Shahzad, M. S., & Rasool, M. S. (2026). Phyto-Mediated Green Synthesis and Physicochemical Characterization of Titanium Dioxide Nanoparticles for Environmental and Pharmacological Applications. Journal of Physical and Chemical Studies (JPCS), 1(4), 17–56. https://doi.org/10.5281/zenodo.19767807

[13] Atif, H. M., Shahzad, A., Khan, M. Z., Abbas, M. A., & Mahar, J. (2025). Design of Novel drug as Potential Anti-Prostate Cancer Activity: Thiophene Derivatives against prostate cancer cell line as therapeutic agents using Pharmacokinetics molecular docking and DFT studies. *Indus Journal of Bioscience Research*, *3*(6), 548-559.

[14] Barros, C., Ramos, G., & Teixeira, A. (2023). SIEM-integrated testbeds for real-time cybersecurity analytics. Journal of Network and Computer Applications, 210, 103577. https://doi.org/10.1016/j.jnca.2022.103577

[15] Chen, Z., Luo, W., & Zhang, Y. (2022). Enhancing multimodal fusion for cybersecurity with adversarial robustness. ACM Transactions on Privacy and Security, 25(4), 1–25. https://doi.org/10.1145/3503012

[16] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

[17] Fernández, M., Blanco, R., & Perez, J. (2021). Real-time cyber threat detection using fusion of NLP and network log data. Computers & Security, 108, 102393. https://doi.org/10.1016/j.cose.2021.102393

[18] Huang, Y., Wang, Y., & Liu, L. (2022). Multimodal threat detection using ensemble deep learning approaches. IEEE Access, 10, 84937–84947. https://doi.org/10.1109/ACCESS.2022.3204439

[19] Jaegle, A., Gimeno, F., Vinyals, O., et al. (2021). Perceiver: General perception with iterative attention. International Conference on Machine Learning, 4651–4664.

[20] Jain, M., Roy, A., & Ghosh, S. (2021). Vision-based security surveillance using deep learning techniques. Multimedia Tools and Applications, 80(5), 7253–7271. https://doi.org/10.1007/s11042-020-09856-y

[21] Junaid, M., Rasool, M. S., Abbas, M. A., & Mahar, J. (2024). Formulation Development and Evaluation of a Bilayered Tablet Containing Dapagliflozin and Metformin. *Global Research Journal of Natural Science and Technology*, *2*(3).

[22] Kiela, D., Bulian, J., Clark, A., et al. (2021). VisualBERT: A simple and performant baseline for vision-and-language. arXiv preprint arXiv:1908.03557.

[23] Klimt, B., & Yang, Y. (2004). Introducing the Enron corpus. CEAS. http://www.cs.cmu.edu/~enron/

[24] Liu, Z., Zhang, X., & Peng, Y. (2023). Multimodal anomaly detection for real-time cyber threat analytics. Pattern Recognition, 138, 109426. https://doi.org/10.1016/j.patcog.2023.109426

[25] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[26] Nguyen, T., Pham, H., & Vo, T. (2022). Deep multimodal fusion for hybrid cybersecurity systems. Journal of Cybersecurity, 8(1), 1–17. https://doi.org/10.1093/cybsec/tyac005

[27] Patel, D., & Kumar, A. (2022). Emotion-based multimodal security threat assessment using deep learning. Expert Systems with Applications, 187, 115911. https://doi.org/10.1016/j.eswa.2021.115911

[28] Qureshi, M., Usama, M., & Khan, S. (2024). Cross-modal threat detection in edge environments using TinyCLIP. Neurocomputing, 553, 165–177. https://doi.org/10.1016/j.neucom.2023.10.154

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

[29] Rahman, A., Baig, F., & Javed, M. (2023). Multimodal deep learning framework for detecting insider threats. Information Sciences, 636, 181–199. https://doi.org/10.1016/j.ins.2023.01.021

[30] Rasool, M. S., Abbas, M. A., Khan, M. J., Mahar, J., & Khan, M. Z. IDENTIFICATION OF NATURAL EGFR TYROSINE KINASE INHIBITORS FROM CHENOPODIUM QUINOA WILLD. VIA COMBINATORIAL IN SILICO AND PHARMACOLOGICAL SCREENING.

[31] Raza, M., Iqbal, Z., & Tariq, M. (2024). Real-time fusion of computer vision and NLP for cybersecurity. Journal of Intelligent & Fuzzy Systems, 47(3), 3659 3669. https://doi.org/10.3233/JIFS-234512

[32] Raza, M., Iqbal, Z., & Tariq, M. (2024). Real-time fusion of computer vision and NLP for cybersecurity. Journal of Intelligent & Fuzzy Systems, 47(3), 3659 3669. https://doi.org/10.3233/JIFS-234512

[33] Singh, A., Li, X., & Yu, Y. (2022). FLAVA: A foundational language and vision alignment model. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15638–15648.

[34] Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. IEEE Conference on Computer Vision and Pattern Recognition, 6479–6488.

[35] Sun, Y., He, J., & Tang, W. (2023). Detecting phishing attacks through multimodal content understanding. IEEE Transactions on Information Forensics and Security, 18, 543–555. https://doi.org/10.1109/TIFS.2023.3251087

[36] Tsai, Y.-H. H., Bai, S., Yamada, M., et al. (2019). Multimodal transformer for unaligned multimodal language sequences. ACL 2019, 6558–6569.

[37] Wang, M., Liu, F., & Zhang, C. (2023). Interpretable multimodal attention networks for detection. Information Fusion, 93, 102221. https://doi.org/10.1016/j.inffus.2023.102221

[38] Zhang, H., Yu, X., & Zhao, Q. (2023). Natural language-based threat detection in cybersecurity. Computers & Security, 126, 102984. https://doi.org/10.1016/j.cose.2023.102984

[39] Zhao, L., Tan, J., & Zhang, M. (2024). Cyber-physical fusion for anomaly detection using multimodal learning. Future Generation Computer Systems, 150, 439 450. https://doi.org/10.1016/j.future.2023.09.011

[40] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. NeurIPS, 34:12786 12797, 2021. 5

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

[41] Gunnar A Sigurdsson, G¨ul Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity under standing. In ECCV, pages 510–526. Springer, 2016. 2, 5

[42] Karen Simonyan and Andrew Zisserman. Two-stream con volutional networks for action recognition in videos. In NeurIPS, 2014. 8

[43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012. 2, 5

[44] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhi nav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In ICCV, pages 843–852, 2017. 5

[45] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In CVPR, 2018. 8

[46] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recogni tion. In CVPR, pages 1895–1904, 2021. 8

[47] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recogni tion. In ECCV, 2016. 8

[48] MengmengWang,Jiazheng Xing, and YongLiu. Actionclip: Anewparadigm for video action recognition. arXiv preprint arXiv:2109.08472, 2021. 1, 2, 4, 5, 6, 7, 8

[49] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaim ing He. Non-local neural networks. In CVPR, 2018. 5

[50] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. In teractive prototype learning for egocentric action recogni tion. In ICCV, pages 8168–8177, 2021. 8

[51] Xiaohan Wang, Linchao Zhu, Fei Wu, and Yi Yang. A dif ferentiable parallel sampler for efficient video classification. ACM Transactions on Multimedia Computing, Communica tions and Applications, 19(3):1–18, 2023. 8

[52] Xiaohan Wang, Linchao Zhu, Yu Wu, and Yi Yang. Symbi otic attention for egocentric action recognition with object centric alignment. IEEE TPAMI, 2020. 8

[53] Yulin Wang, Zhaoxi Chen, Haojun Jiang, Shiji Song, Yizeng Han, and Gao Huang. Adaptive focus for efficient video recognition. In ICCV, pages 16249–16258, 2021. 8

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 102-131

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20461727

An Open-Access Peer-Reviewed Journal

[54] Wenhao Wu, Dongliang He, Tianwei Lin, Fu Li, Chuang Gan, and Errui Ding. Mvfnet: Multi-view fusion network for efficient video recognition. In AAAI, 2021. 5, 8

[55] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In ICCV, 2019. 6, 8

[56] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, Yi Yang, and Shilei Wen. Dynamic inference: A new approach toward efficient video action recognition. In Proceedings of CVPR Workshops, pages 676–677, 2020. 8

[57] Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. Cap4video: What can auxiliary captions do for text-video retrieval? In CVPR, 2023. 8

[58] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting clas sifier: Transferring vision-language models for video recog nition. In AAAI, 2023. 1, 2, 5, 6, 8

[59] Wenhao Wu, Yuxiang Zhao, Yanwu Xu, Xiao Tan, Dongliang He, Zhikang Zou, Jin Ye, Yingying Li, Mingde Yao, Zichao Dong, et al. Dsanet: Dynamic segment aggre gation network for video-level representation learning. In ACMMM,pages 1903–1911, 2021. 6, 8

[60] Boyang Xia, Zhihao Wang, Wenhao Wu, Haoran Wang, and Jungong Han. Temporal saliency query network for efficient video recognition. In ECCV, pages 741–759. Springer, 2022. 6, 8

[61] Boyang Xia, Wenhao Wu, Haoran Wang, Rui Su, Dongliang He, Haosen Yang, Xiaoran Fan, and Wanli Ouyang. Nsnet: Non-saliency suppression sampler for efficient video recog nition. In ECCV, pages 705–723. Springer, 2022. 6, 8