ICCK Transactions on Intelligent Systematics
http://dx.doi.org/10.62762/TIS.2025.610574

RESEARCH ARTICLE

Cross-Lingual Multimodal Event Extraction: A Unified Framework for Parameter-Efficient Fine-Tuning

Sheng Hong[1,2,*], **Xuanqi Wang**[3], **Zeyu Mei**[4] **and Thisura Bojitha Wickramaratne**[5]

[1] School of Cyber Science and Technology, Beihang University, Beijing 100191, China
[2] Nanchang University, Nanchang 330031, China
[3] School of Information Engineering, Nanchang University, Nanchang 330031, China
[4] International Business School, Beijing Foreign Studies University, Beijing 100089, China
[5] School of Cyber Science and Technology, Beihang University, Beijing 100191, China

Abstract

With the rapid development of multimodal large language models (MLLMs), structured event extraction (EE) has emerged as a critical intelligent information processing task, with increasing demand across multilingual and multimodal application scenarios. However, significant challenges remain in zero-shot multimodal and cross-language scenarios, including inconsistent cross-language outputs and the high computational cost of full-parameter fine-tuning. This study takes VideoLLaMA2 (VL2) and its improved version VL2.1 as the core models, and builds a multimodal annotated dataset covering English, Chinese, Spanish, and Russian (including 5,728 EE samples). It systematically evaluates the performance differences of zero-shot learning, and parameter-efficient fine-tuning (QLoRA) techniques. The experimental results show that for EE, QLoRA fine-tuning yields substantial gains across both models: VL2.1 achieves the highest trigger accuracy at 65.48%, while VL2 achieves the highest argument accuracy at 60.54%, with each figure representing the best performance obtained across the two fine-tuned models respectively. The study confirms that fine-tuning significantly enhances model robustness.

Keywords: event extraction, QLoRA, multimodal LLMs, multilingual NLP.

Academic Editor:
Javier Bajo

Submitted: 07 June 2025
Accepted: 11 August 2025
Published: 04 October 2025

Vol. 2, **No.** 4, 2025.
10.62762/TIS.2025.610574

***Corresponding author:**
Sheng Hong
shenghong@buaa.edu.cn

1 Introduction

The vigorous development of Multimodal Large Language Models (MLLMs) has brought profound changes to the field of Natural Language Processing (NLP), with recent surveys documenting rapid capability growth alongside persistent challenges in multilingual generalization, model alignment, and deployment efficiency that directly motivate research into cross-lingual and cross-modal Event Extraction (EE) [1]. However, current technologies still have application limitations: in zero-shot learning scenarios, their generalization ability for multilingual data and cross-modal transfer effects remain to be expanded, which has become an important research direction

Citation
Hong, S., Wang, X., Mei, Z., & Wickramaratne, T. B. (2025). Cross-Lingual Multimodal Event Extraction: A Unified Framework for Parameter-Efficient Fine-Tuning. *ICCK Transactions on Intelligent Systematics*, 2(4), 203–212.

for future technological iteration. This paper focuses on VL2 (based on Mistral-7B-Instruct-v0.2) [2] and VL2.1 (based on Qwen2-7B-Instruct), assessing their capabilities in EE across English (EN), Chinese (ZH), Spanish (ES), and Russian (RU), under zero-shot and fine-tuned conditions [3].

Initial research work has sufficiently proven the potential of MLLMs in extraction tasks. Lin et al. [4] demonstrated that few-shot learning with models like GPT-3 resulted in significant performance in structured NLP tasks, though multilingual performance varies. Conneau et al. [5] highlighted challenges in zero-shot cross-lingual transfer,indicating that models often struggle with languages lacking extensive pre-training data. For EE, Wadden et al. [6] proposed contextualized span representations, achieving high accuracy in English but less so in multilingual contexts. MLLMs, as surveyed by Wu et al. [7], show promise in integrating visual and textual cues, yet their zero-shot multilingual performance is limited by alignment issues. Based on this, our research evaluated VL2 and VL2.1, aiming to address the deficiencies in multilingual and multimodal extraction and explore the role of fine-tuning in improving performance.

2 Related Work

The primary objective of this research is to improve the structured extraction of events from multimodal, multilingual inputs using multimodal large language models. In essence, we aim to develop and evaluate techniques that enable an AI system to take in information from text, images, and videos in multiple languages (English, Chinese, Spanish, Russian) and output a well-structured representation of any events described expressed, all with high accuracy and efficiency. To achieve this goal, this section introduces the relevant working strategies.

2.1 Baseline Evaluation

Baseline evaluation of Multimodal Large Language Models (MLLMs) in event extraction (EE) focuses on assessing their inherent capabilities in zero-shot scenarios, i.e., without task-specific fine-tuning [8]. This work builds on prior research highlighting the potential of MLLMs in integrating cross-modal and cross-lingual cues, while also addressing their limitations in generalization across low-resource languages and modalities. Our baseline evaluation targets VL2 and VL2.1, two state-of-the-art MLLMs,to quantify their zero-shot EE performance across four languages (English, Chinese, Spanish, Russian) and three modalities. This evaluation establishes a benchmark for trigger and argument identification accuracy,revealing gaps in cross-lingual consistency (e.g.,language mixing in outputs) and modality-specific weaknesses (e.g.,lower accuracy for video inputs) that inform subsequent fine-tuning strategies [17].

Furthermore, the effectiveness of task-specific model adaptation has been demonstrated in other domains involving complex cross-lingual structured prediction. For instance, Li et al. [12] proposed MT4CrossOIE, a multi-stage tuning framework for cross-lingual open information extraction that injects language-specific knowledge via mixture-of-LoRAs into a shared pre-trained model, highlighting the importance of tailored parameter-efficient adaptation for robust cross-lingual performance. This aligns with our objective of employing QLoRA to adapt MLLMs for cross-lingual event extraction.

2.2 QLoRA Fine-Tuning

Parameter-efficient fine-tuning (PEFT) techniques, such as QLoRA, address the high computational costs of full-parameter fine-tuning while enhancing model adaptation to specific tasks like cross-lingual multimodal EE. Open-source instruction-tuned models such as LLaMA 2 [9] have demonstrated the viability of this paradigm, serving as representative backbones for PEFT research. QLoRA extends LoRA (Low-Rank Adaptation) by combining low-rank matrix updates with 4-bit quantization, drastically reducing memory usage and enabling training on limited hardware [10]. In the context of EE, QLoRA fine-tuning is tailored to structured outputs(triggers and arguments) by adjusting low-rank matrix rank (r=128) and scaling factors (α=256) to balance adaptation strength and stability. By applying QLoRA to VL2 and VL2.1, this study aims to enhance cross-lingual consistency and EE accuracy while maintaining computational feasibility, addressing key limitations identified in baseline evaluations.

3 Methodology

This section outlines the methodology used to evaluate and optimize the VL2 and VL2.1 models for EE (end-to-end), covering language dimensions such as English, Chinese, Spanish, and Russian. The methodology aims to enhance the models' multimodal interaction capabilities and cross-lingual processing performance.

3.1 Dataset Construction

The development of a novel multimodal and multilingual dataset for EE addresses a significant gap in resources tailored for STI.Existing datasets often lack comprehensive coverage of text, image, and video modalities across multiple languages, particularly for specialized intelligence tasks. This dataset fills this void by integrating diverse data types including text, images, videos,and audio with a focus on four primary languages (i.e.,English, Chinese, Spanish, Russian) due to their geopolitical relevance and data availability. The dataset is designed to support robust EE tasks, ensuring relevance and reproducibility through a structured eight-step process: (1) Data Source Picking, (2) Data Collection, (3) Data Extraction, (4) Data Cleaning, (5) Data Deduplication, (6) Data Annotation, (7) Creation of a Multilingual and Multimodal Dataset, and (8) Conversion to VL2 Format. See Figure 1 for the Construction process of the dataset.

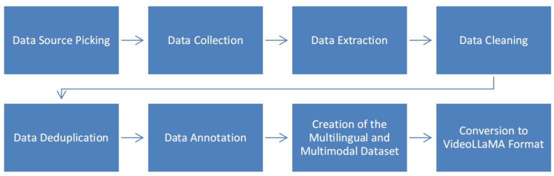


Figure 1. Construction process of the dataset.

3.1.1 *Data Source Picking*

The dataset was curated from open-source science and technology information websites worldwide, selected for data quality, legal compliance, and alignment with intelligence objectives. Sources encompass news articles, industry research reports, and enterprise dynamics,covering topics such as artificial intelligence, integrated circuits, and emerging technologies. Including eeworld.com.cn (Chinese, electronic technology developments), stdaily.com (Chinese, science and technology news), rand.org (English, research reports), cadence.com (English, system design solutions), new-science.ru (Russian, high-tech news), genbeta.com (Spanish, software applications), and hipertextual.com (Spanish, digital technology and science). These platforms ensure a rich, diverse corpus spanning text, images, videos. The dataset forms a parallel corpus with primary focus on English, Chinese, Spanish, and Russian for broader coverage. Sources were filtered to guarantee multimodal alignment and relevance to science and technology intelligence.

3.1.2 *Data Collection*

Data collection leveraged the 360 Data Collection Platform, an integrated system with modules for seed management, annotation, and structured data extraction. The Spider crawler tool systematically gathered raw web page data from selected sources, channeling it into a Kafka pipeline for scalable and reliable processing. This pipeline ensured efficient handling of multimodal data including text articles, images (e.g., diagrams, product photos), videos (e.g.,news clips, technology demos), and audio while maintaining traceability and integrity. The platform's modular design facilitated iterative refinement of collection parameters, optimizing coverage across languages and modalities. See Figure 2 for Data collection platform architecture.

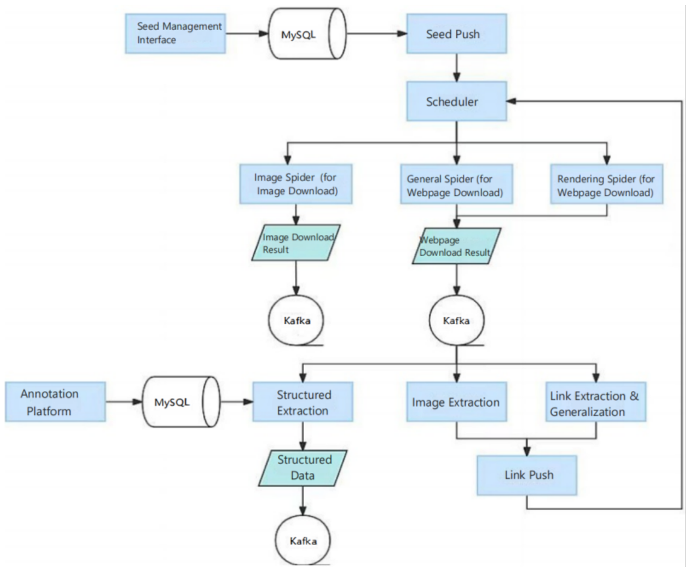


Figure 2. Data collection platform architecture.

3.1.3 *Data Extraction*

Structured extraction of multimodal content was achieved using XPath-based templates generated by the annotation platform. These templates target specific elements (i.e., text content, image URLs, video URLs, and audio URLs), within raw web pages, enabling precise isolation of relevant data. For example, text was extracted from article bodies, while video URLs were retrieved from embedded media players. Extracted data was stored in the Kafka pipeline and made accessible via a data request interface, streamlining downstream processing. This structured approach ensured consistency across modalities and minimized data loss during extraction.

3.1.4 *Data Cleaning*

To enhance data quality, a rigorous cleaning process was applied. Text data underwent removal of

HTML tags, stop words, and special characters to improve clarity and uniformity. Outliers, such as missing or incomplete values, were handled through interpolation (for minor gaps) or deletion (for irreparable entries), preserving dataset integrity. Cleaning ensured that subsequent annotation and modeling tasks operated on high-quality inputs, minimizing errors due to formatting inconsistencies or irrelevant content. This preprocessing discipline is particularly critical when preparing data for fine-tuning large language models, where input quality directly governs adaptation effectiveness [13].

3.1.5 Data Deduplication

Deduplication was performed to ensure uniqueness and prevent bias in downstream tasks. Whole-row deduplication eliminated identical entries, while text similarity filtering, using cosine similarity, removed near-duplicate content across languages and modalities. This step was critical for maintaining dataset diversity, particularly given the overlap often found in news and report summaries across sources. Deduplication enhanced the representativeness of the dataset, ensuring that models trained on it would generalize effectively.

3.1.6 Data Annotation

Annotation was conducted using Label Studio, an open-source platform supporting text, image, video, and audio labeling. Datasets were imported in JSON format, annotated by trained experts, and exported with entities, relationships, and attributes.

3.1.7 Creation of a Multilingual and Multimodal Dataset

The final dataset comprises 5,728 EE samples, covering text, image, and video modalities across four primary languages. The full dataset includes 3,449 texts, 1,971 image, and 308 video samples, with 1,197 English, 1,844 Chinese, 1,449 Russian, and 1,238 Spanish samples, as shown in Table 1.

Table 1. Dataset statistics.

	Event	Count
Modality	Image	1971
	Video	308
	Text	3449
	Total	5728
Language	English	1197
	Chinese	1844
	Russian	1449
	Spanish	1238
	Total	5728

3.1.8 Conversion to VL2 Format

To enable training with VL2 and VL2.1, the dataset was converted into VL2 format, structuring data as JSON objects with text tokens, CLIP-ViT or SigLIP image embeddings, and video frame sequences. Formats include text-only, text-image, text-video, and text-audio configurations.

3.2 Base Model Selection

Based on the primary focus on multimodality and multilinguality, we evaluated four MLLMs: Video-LLaVA (7B), Video-LLaMA (7B), VideoLLaMA2 (7B), and VideoLLaMA2.1 (7B). The comparison criteria included multimodal support (i.e., text, image, video), context length, multilingual performance in target languages (i.e., English, Chinese, Spanish, Russian), general performance on relevant benchmarks, and computational feasibility for training on available hardware (two A100-40GB GPUs). VL2 and VL2.1 were selected as the base models. VL2 (7B) employs a CLIP-ViT-Large-Patch14-336 encoder and a Mistral-7B-Instruct-v0.2 decoder, supporting a 32k context length. VL2.1 (7B), utilizes a SigLIP-So400mPatch14-384 encoder and a Qwen2-7B-Instruct decoder, enabling a 131k context length. These models were chosen over alternatives like Video-LLaVA [14] and Video-LLaMA [15] due to their superior context length capabilities and advanced decoders, enhancing multilingual and multimodal processing [16, 17]. Base Model comparison is given in Table 2.

VL2 is an advanced multimodal large language model designed for enhanced video and audio understanding, building upon the foundation of its predecessor with significant architectural improvements tailored for spatial-temporal modeling and multimodal integration. The model comprises a transformer-based language model paired with specialized vision and audio processing components, optimized for tasks such as video question answering, captioning, and EE as outlined [2].

3.3 Base Model Performance Evaluation

We scrutinized the baseline performance of VL2 and VL2.1 using zero-shot inference with temperature set to 0.2 and top_p to 0.9. Temperature controls output randomness, and top_p governs token selection diversity via nucleus sampling [10]. Low temperature (0.2) ensures deterministic, focused outputs, while high top_p (0.9) balances diversity and coherence. Together these settings optimize decoding for the structured nature of event extraction, where precision in trigger and argument identification is paramount—a requirement grounded in the dual-task formulation of trigger identification and argument role assignment established in prior EE research [6].

3.4 Model Fine-Tuning

QLoRA extends LoRA, a technique that updates model weights efficiently by introducing low-rank matrices, with 4-bit quantization to further reduce memory demands [11]. In LoRA, weight updates are formulated as:

$$W = W_0 + \Delta W \tag{1}$$

where $\Delta W = BA$, with $B \in R^{d\times r}$, $A \in R^{r\times k}$, and $r \ll d, k$. Only the low-rank matrices (B) and (A) are trained, significantly reducing the number of parameters updated compared to full fine-tuning, which modifies all weights (O(dk)). QLoRA enhances this by quantizing the base weights W_0 to 4-bit precision (W_0^{4bit}) and applying LoRA updates:

$$W = W_0^{4bit} + BA \tag{2}$$

Additionally, QLoRA incorporates double quantization, quantizing the quantization constants themselves to further optimize memory usage. This approach reduces memory requirements from approximately 100 GB for full fine-tuning of a 7B model to 10 GB for QLoRA, enabling training on two A100-40GB GPUs.

VL2 and VL2.1 were fine-tuned separately for EE using the full dataset with 5,728 samples on two A100-40GB GPUs. Separate models were trained to accommodate EE's structured JSON output, which includes triggers, event types, and arguments. The following are the parameters that have been used in the training.

Rank (r) = 128: This low-rank factorization of the update matrix ΔW forms the basis of LoRA. Here, A and B are the learned low-rank adaptation matrices with rank r=128, allowing the model to adapt efficiently to structured outputs like event triggers and arguments while keeping parameter count low. A lower rank was chosen for EE due to its constrained, structured output format, which requires less flexibility than generative tasks [18, 19].

$$\Delta W = AB,\ where\ A \in \mathbb{R}^{d\times r},\ B \in \mathbb{R}^{r\times k},\ and\ r \ll \min(d,k) \tag{3}$$

Alpha= 256: The final adapted weights are computed by scaling the LoRA updates ΔW by a factor α=256. This controls the strength of adaptation relative to the base model weights W, ensuring stable fine-tuning suited for EE. A moderate value suits EE's structured nature [19].

$$W' = W + \alpha \cdot \Delta W = W + \alpha AB \tag{4}$$

Max Tokens = 9,300: The model accommodates the complexity of event extraction (EE) inputs, which often involve long contexts with multiple triggers and arguments, by leveraging a large token limit of 9,300 and a high embedding dimension. This ensures sufficient capacity to process and encode lengthy documents while maintaining full contextual understanding. The size of each input sequence is calculated as the maximum token length multiplied by the model's embedding dimension, enabling robust handling of intricate, event-rich texts [19].

$$InputSize = MaxTokens \times EmbeddingDim \tag{5}$$

Epochs = 3: Given 3 epochs and a batch size of 2, the model performs 8,592 update steps during training. This ensures sufficient convergence without overfitting, considering the dataset size [19].

$$TotalSteps = Epochs \times \left(\frac{N}{BatchSize}\right) \tag{6}$$

Batch Size = 2: This estimate outlines the GPU memory footprint of the training process. By setting a batch size of 2 and quantizing weights to 4 bits, QLoRA enables memory-efficient training for high-context EE samples [19].

$$Memory \approx BatchSize \times MaxTokens \times Precision(bits) \tag{7}$$

Table 2. Base model comparison.

Model	Modalities	Context Length	Visual Encoder	Language Decoder
Video-LLaVA	Image, Video, Text	2K	CLIP Vision Encoder	Vicuna-7B
Video-LLaMA	Image, Video, Text, Audio	4k	ViT-G/14 + BLIP-2 Q-Former	LLaMA2-7B
VideoLLaMA2	Image, Video, Text, Audio	32k	CLIP-ViT-Large-Patch14-336	Mistral-7B-Instruct-v0.2
VideoLLaMA2.1	Image, Video, Text, Audio	131k	SigLIP-So400m-Patch14-384	Qwen2-7B-Instruct

The lower rank (r=128) aligns with EE's need for precise encoding of structured outputs, as the task's constrained trigger-argument structure requires less adaptation flexibility than open-ended generative tasks [20]. A higher max token count (with 9,300) accommodates complex inputs, such as multilingual texts or multimodal data with detailed event descriptions. Fewer epochs (with 3) were chosen to balance convergence and overfitting risks, given the dataset's size and diversity across four languages and modalities.

The following are the overall training arguments that are used in the EE training.

Optimizer: This is the AdamW optimization update rule used in both EE. AdamW, the default for QLoRA, was used with a learning rate of 1e-5 to ensure stable convergence across tasks [11].

$$\theta_{t+1} = \theta_t - \eta \cdot \left(\frac{m_t}{(\sqrt{v_t} + \varepsilon)} \right) + \lambda \cdot \theta_t \tag{8}$$

where m_t and v_t represent the first and second moment estimates, $\eta = 1e-5$ is the learning rate, and λ is the weight decay coefficient, ensuring convergence and regularization.

Hardware: Two A100-40GB GPUs enabled parallel training of EE models, leveraging QLoRA's memory efficiency

The EE focus on encoding allows for a more constrained specification. Within the context of maximum tokens, epoch size, and batch size, these parameters create an appropriate balance between memory efficiency and performance on smaller datasets, as QLoRA is explicitly designed to do [11, 19].

3.5 Evaluation Criteria

We evaluated performance using accuracy.We computed overall accuracy using a test set comprised 320 samples for EE, with 80 samples per language (English, Chinese, Spanish, Russian). Also calculated separate accuracies for trigger and event type identification (Tr Accuracy) and argument and role identification (Arg Accuracy), reflecting the task's dual components [6].

Event Extraction:Here calculated separate accuracies for trigger and event type identification (Tr Accuracy) and argument and role identification (Arg Accuracy). Separating Tr and Arg Accuracy captures EE's dual components such as trigger identification and argument role assignment, reflecting the task's complexity. Precision-focused metrics ensure accurate event structuring, vital for intelligence analysis where incorrect triggers or rolescould lead to significant misinterpretations.

Metrics: Trigger Accuracy (Tr Accuracy):

$$Tr\ Accuracy = \frac{\text{Correct triggers and roles}}{\text{Total triggers}} \tag{9}$$

A correct trigger matches the ground-truth trigger word/phrase and its associated event type.

Argument Accuracy (Arg Accuracy):

$$Arg\ Accuracy = \frac{\text{Correct arguments and roles}}{\text{Total arguments}} \tag{10}$$

A correct argument matches the ground-truth argument and its role.

4 Experimental Results

This section systematically evaluates the performance of VL2 and VL2.1 models in multilingual EE tasks, covering English, Chinese, Spanish, and Russian with text, image, and video as inputs. The assessment includes zero-shot learning, QLoRA fine-tuning performance. Accuracy metrics are summarized for and EE (trigger, Tr; argument, Arg), focusing on key evaluation dimensions for both tasks.

4.1 Zero-Shot Base Model Performance

VL2 achieved an overall Tr Accuracy of 28.40% and Arg Accuracy of 23.89%, while VL2.1 recorded 24.76% Tr Accuracy and 20.88% Arg Accuracy, as shown in Table 3 and Figure 3. These modest scores reflect the complexity of identifying event triggers and their associated arguments without task-specific training. Language-specific performance varied significantly. For VL2, English samples yielded 34.38% Tr Accuracy and 28.13% Arg Accuracy, followed by Chinese (31.77% Tr, 24.83% Arg), Spanish (29.13% Tr, 26.19% Arg), and Russian (18.34% Tr,16.42% Arg). VL2.1 demonstrated a notable strength in Chinese, achieving 44.8% Tr Accuracy and 36.25% Arg Accuracy, but struggled elsewhere, particularly in Russian (12.5% Tr, 9.88% Arg), with English (20.63% Tr, 18.19% Arg) and Spanish (21.14% Tr, 19.2% Arg) also lagging.Comparing the models, VL2 outperformed VL2.1 overall in Tr Accuracy by 3.64%, driven by better consistency across languages, whereas VL2.1's Chinese performance surpassed VL2 by 13.03% in Tr Accuracy, indicating a language-specific advantage possibly tied to its Qwen2-7B-Instruct decoder.

Table 3. Zero-shot event extraction performance.

Model	Language	Tr Accuracy (%)	Arg Accuracy (%)
VL2	English	34.38	28.13
	Chinese	31.77	24.83
	Spanish	29.13	26.19
	Russian	18.34	16.42
	Overall	28.40	23.89
VL2.1	English	20.63	18.19
	Chinese	44.80	36.25
	Spanish	21.14	19.20
	Russian	12.50	9.88
	Overall	24.76	20.88

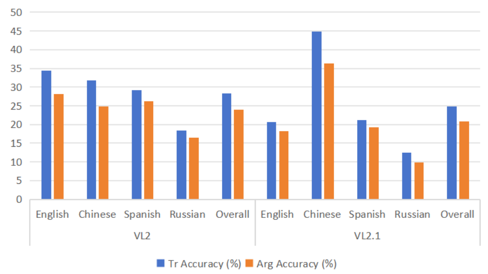


Figure 3. Tr & Arg Accuracy of VL2 and VL2.1 in different languages.

Error analysis reveals critical challenges. Both models exhibited language mismatches, such as VL2.1 producing Chinese triggers for Spanish inputs or English triggers for Russian samples. VL2 occasionally generated incorrect event types, mislabeling a "launch" event as a "policy change," reflecting confusion in semantic categorization. Incomplete outputs were prevalent, with VL2.1 showing a higher incidence of missing arguments, particularly in Russian, where only partial roles were extracted. A stylistic difference emerged: VL2 favored longer trigger phrases (e.g., "announced a new technology"), aligning with verbose contextual cues, while VL2.1 preferred concise keywords (e.g., "technology"), which sometimes sacrificed specificity [21]. These trends suggest that zero-shot event extraction struggles with multilingual coherence and structural fidelity, necessitating targeted improvements.

Overall, in zero-shot settings, both models demonstrate moderate capabilities in EE, with VL2.1 showing a language-specific advantage in Chinese and VL2 achieving more balanced performance across all four languages. However, language inconsistencies—such as generating Chinese triggers for English inputs or mixing languages in Spanish and Russian outputs—reveal limitations in cross-lingual generalization, consistent with prior work on multilingual LLMs [17]. VL2.1's stronger performance in English and Chinese suggests a bias possibly inherited from its Qwen2-7B-Instruct backbone, which is documented to have been pre-trained on substantially larger corpora in high-resource languages including Chinese and English [16]. Its weaker handling of Russian inputs, points to difficulties with morphologically complex languages, a recognized challenge in LLMs [5].

4.2 Fine-Tuned Model Performance

QLoRA fine-tuning markedly improved EE performance over zero-shot baselines (VL2: 28.40% Tr Accuracy, 23.89% Arg Accuracy; VL2.1: 24.76% Tr, 20.88% Arg). As shown in Table 4 and Figure 4, VL2 achieved 65.28% Tr Accuracy and 60.54% Arg Accuracy, while VL2.1 recorded 65.48% Tr Accuracy and 53.04% Arg Accuracy. Language-specific results reveal nuanced strengths. For VL2, English samples yielded 68.13% Tr and 63.96% Arg, Chinese 67.92% Tr and 65.08% Arg, Spanish 60.52% Tr and 59.58% Arg, and Russian 64.58% Tr and 53.54% Arg. VL2.1 excelled in Chinese with 78.54% Tr Accuracy and 56.77% Arg Accuracy,followed by English (64.38% Tr, 56.46% Arg), Spanish (61.5% Tr, 50.31% Arg), and Russian (57.5% Tr, 48.65% Arg). Comparing models, VL2 demonstrated balanced performance across languages, with Arg Accuracy consistently above 53%, whereas VL2.1's Chinese Tr Accuracy surpassed VL2 by 10.62 percent, likely due to its Qwen2-7B-Instruct decoder's pre-training strengths. However, VL2.1's Arg Accuracy lagged, particularly in Russian, trailing VL2 by 4.89 percent. See Table 4 and Figure 4, for Fine-tuned EE performance.

The gains are striking: VL2's Tr Accuracy surged from 28.40% (zero-shot) to 65.28%, a 36.88-point increase, and Arg Accuracy rose by 36.65 percentage points. VL2.1's Tr Accuracy improved by 40.72 points to 65.48%, and Arg Accuracy climbed from 20.88% to 53.04%, a 32.16-point gain. These improvements stem from QLoRA's ability to adapt model weights to task-specific patterns, eliminating key zero-shot errors [11]. Notably, fine-tuned models produced no language mismatches—unlike zero-shot runs where Chinese triggers appeared in Spanish samples. Outputs were complete, with all arguments and roles correctly identified, and JSON structures adhered to the specified format, resolving issues like missing event roles (e.g., "entity" in "innovation") seen in zero-shot settings. The challenge of incorrect event types (e.g., "launch" mislabeled as "policy change") was also mitigated, reflecting better semantic alignment. See Table 4 and Figure 4, for Fine-tuned EE performance.

Table 4. Fine-Tuned Model Performance.

Model	Language	Tr Accuracy (%)	Arg Accuracy (%)
VL2	English	68.13	63.96
	Chinese	67.92	65.08
	Spanish	60.52	59.58
	Russian	64.58	53.54
	Overall	65.28	60.54
VL2.1	English	64.38	56.46
	Chinese	78.54	56.77
	Spanish	61.50	50.31
	Russian	57.50	48.65
	Overall	65.48	53.04

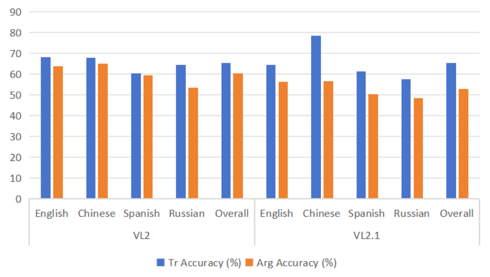


Figure 4. The Fine-tuned result of QLoRA.

Despite these advances, challenges persist. Russian Arg Accuracy remained lower,particularly for VL2.1 (with 48.65%), suggesting difficulties in capturing complex argument roles in morphologically rich languages. Training used two A100-40GB GPUs with parameters set at LoRA rank (r) = 128, LoRA alpha = 256, max tokens = 9,300, epochs = 3, and batch size = 2. VL2 required 182 minutes, while VL2.1 took 169.97 minutes, reflecting slight efficiency differences possibly tied to VL2.1's decoder architecture. These configurations balanced computational efficiency with performance, making QLoRA viable for resource-constrained settings.

Overall, Fine-tuning produced the notable improvements, eliminating language inconsistencies and incomplete outputs, a finding supported by studies on task-specific adaptation [4]. VL2's balanced EE performance suggests Mistral-7B's suitability for structured tasks. Training times,though feasible, highlight event extraction's computational complexity due to trigger-argument dependencies [6]. These findings suggest fine-tuned MLLMs could enhance real-time STI monitoring across diverse media.

5 Future Directions

However, there are still some limitations. The multimodal dataset–encompassing images, videos, and text–tends to heighten challenges in zero-shot scenarios. This is because cross-modal alignment remains a formidable task without fine-tuning, as noted in prior research [7]. Ongoing challenges persist, including the difficulty of achieving effective cross-modal alignment in zero-shot settings and the complexity of processing morphologically rich languages such as Russian. While the success of fine-tuned models across various modalities underscores the significance of adaptive tuning, the dataset's imbalanced language distribution-with, for instance, a larger number of Chinese event samples–might introduce potential biases.

To address these limitations, future research can pursue several directions, starting with model development. Testing larger models, such as VideoLLaMA2-72B or hybrid architectures with specialized visual encoders for video processing, could improve accuracy, particularly for video inputs, which lagged in this study [2]. Integrating advanced multimodal frameworks, like those combining CLIP-style vision encoders with transformer-based language decoders, may enhance cross-modal reasoning, pushing EE accuracies beyond the current 53–79% range observed across languages and models.

Dataset expansion is another critical avenue. Including additional languages, such as Arabic or Hindi, would broaden the system's applicability to diverse geopolitical contexts, addressing the intelligence community's need for global coverage. Established EE annotation frameworks such as ERE already define entity, relation, and event schemas that span multiple languages and domains [22], providing a methodological foundation for extending the dataset to additional target languages. Increasing video samples and balancing language distributions (e.g., equalizing Spanish and Chinese event samples) would mitigate biases and improve fairness across languages. Automated or semi-automated annotation tools, validated by human experts, could scale the dataset while maintaining quality, reducing reliance on labor-intensive manual processes [23].

Methodologically, adaptive prompting strategies [24], such as dynamic CoT that adjust reasoning steps based on input complexity [25, 26], could be explored for fine-tuned models to recover the benefits seen in zero-shot settings. Ablation studies varying QLoRA parameters (e.g.,LoRA rank, alpha, or quantization bits) would optimize efficiency and performance, addressing resource constraints. Additionally, modality-specific evaluations separating text, image, and video performance would clarify where models excel or falter, guiding targeted improvements.

Finally, applying these methods to other domains holds significant potential. EE could enhance policy analysis by tracking legislative changes or public sentiments across multimodal sources or improve social media monitoring by detecting emerging trends in global discourse. Integrating human-in-the-loop validation, where analysts refine model outputs in real-time, would tailor the system to intelligence tasks, ensuring reliability in high-stakes contexts [27]. These extensions would amplify the research's impact, positioning it as a versatile framework for automated intelligence processing across diverse applications.

6 Conclusion

This study presents a unified intelligent framework for cross-lingual multimodal event extraction, centering on VideoLLaMA2 (VL2) and VL2.1 with parameter-efficient fine-tuning (QLoRA) across English, Chinese, Spanish, and Russian. A novel 5,728-sample multimodal dataset spanning text, image, and video modalities was constructed to support systematic evaluation. Baseline zero-shot assessment revealed cross-lingual output mismatches, incomplete argument extraction, and performance gaps in morphologically complex languages. QLoRA fine-tuning substantially resolved these limitations, eliminating cross-lingual inconsistencies and producing complete, structured outputs with VL2.1 reaching 65.48% trigger accuracy and VL2 reaching 60.54% argument accuracy. These results confirm that parameter-efficient adaptation is a viable strategy for deploying intelligent multimodal systems under real-world resource constraints, offering a scalable blueprint for automated multilingual information extraction in intelligent systematics applications.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported by the National Key Research and Development Program under Grant 2022YFB3103602.

Conflicts of Interest

Sheng Hong served as an Associate Editor of the *ICCK Transactions on Intelligent Systematics* at the time of manuscript submission. To ensure the integrity of the peer-review process, Sheng Hong was not involved in the editorial handling, peer review, or decision-making process for this manuscript, which was handled independently by another editor. The remaining authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Mohammed, A., & Kora, R. (2025). A Comprehensive Overview and Analysis of Large Language Models: Trends and Challenges. *IEEE Access, 13*, 95851-95875. [Crossref]

[2] Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., ... & Bing, L. (2024). Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

[3] Huang, K. H., Hsu, I. H., Natarajan, P., Chang, K. W., & Peng, N. (2022, May). Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 4633-4646). [Crossref]

[4] Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., ... & Li, X. (2022, December). Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 9019-9052). [Crossref]

[5] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[6] Wadden, D., Wennberg, U., Luan, Y., & Hajishirzi, H. (2019). Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

[7] Wu, J., Gan, W., Chen, Z., Wan, S., & Yu, P. S. (2023, December). Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 2247-2256). IEEE. [Crossref]

[8] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

[9] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[10] Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

[11] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, *36*, 10088-10115.

[12] Li, T., Wang, Z., Chai, L., Yang, J., Bai, J., Yin, Y., ... & Li, Z. (2024). Mt4crossoie: Multi-stage tuning for cross-lingual open information extraction. *Expert Systems with Applications*, *255*, 124760. [Crossref]

[13] Parthasarathy, V. B., Zafar, A., Khan, A., & Shahid, A. (2024). The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*.

[14] Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., & Yuan, L. (2023). Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

[15] Zhang, H., Li, X., & Bing, L. (2023). Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

[16] Yang, A., et al. (2024). Qwen2 technical report. arXiv. [Crossref]

[17] Chirkova, N., & Nikoulina, V. (2024). Zero-shot cross-lingual transfer in instruction tuning of large language models. *arXiv preprint arXiv:2402.14778*.

[18] Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., ... & Zhao, T. (2023). Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*. [Crossref]

[19] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

[20] Xiang, W., & Wang, B. (2019). A survey of event extraction from text. *IEEE Access*, *7*, 173111-173137. [Crossref]

[21] Marchisio, K., Ko, W. Y., Bérard, A., Dehaze, T., & Ruder, S. (2024, November). Understanding and mitigating language confusion in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 6653-6677). [Crossref]

[22] Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., ... & Ma, X. (2015, June). From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the 3rd workshop on EVENTS: Definition, detection, coreference, and representation* (pp. 89-98). [Crossref]

[23] Siriborvornratanakul, T. (2025, May). From Human Annotators to AI: The Transition and the Role of Synthetic Data in AI Development. In *International Conference on Human-Computer Interaction* (pp. 379-390). Cham: Springer Nature Switzerland. [Crossref]

[24] Kamesh, R. (2024). Think Beyond Size: Adaptive Prompting for More Effective Reasoning. *arXiv preprint arXiv:2410.08130*.

[25] Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., & Smola, A. (2023). Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

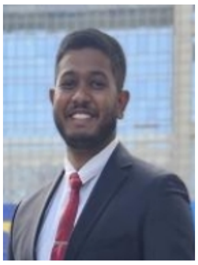
[26] Zhang, X., Wang, Z., & Li, P. (2023, June). Multimodal Chinese Event Extraction on Text and Audio. In *2023 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE. [Crossref]

[27] Butcher, B., Zilka, M., Hron, J., Cook, D., & Weller, A. (2024). Optimising human-machine collaboration for efficient high-precision information extraction from text documents. *ACM Journal on Responsible Computing*, *1*(2), 1-27. [Crossref]

Sheng Hong is an Associate Professor and Doctoral Supervisor at Beihang University, and a Professor and Doctoral Supervisor at Nanchang University, China. He is a Leading Scholar in the Safety Discipline of Beijing and serves as the Chief Technology Officer (CTO) and Deputy Commander of the Major Project on Smart Manufacturing Safety. (Email: shenghong@buaa.edu.cn)

Xuanqi Wang is currently a postgraduate student majoring in Artificial Intelligence at the School of Information Engineering, Nanchang University, Nanchang, China. His research interests include artificial intelligence and its applications. (Email: 15894886025@163.com)

Zeyu Mei is currently a student at the International Business School, Beijing Foreign Studies University, Beijing, China. (Email: 202220102010@bfsu.edu.cn)



Thisura Bojitha Wickramaratne is a postgraduate student majoring in Network and Information Security at the School of Cyber Science and Technology, Beihang University, Beijing, China. (Email: thisurawz1@gmail.com)