
COSINT-Agent: A Knowledge-Driven Multimodal Agent for Chinese Open Source Intelligence

Wentao Li^{1,2} Congcong Wang¹ Xiaoxiao Cui^{1,2} Zhi Liu^{1,2} Wei Guo^{1,2} Lizhen Cui^{1,2}

Abstract

Open Source Intelligence (OSINT) requires the integration and reasoning of diverse multimodal data, presenting significant challenges in deriving actionable insights. Traditional approaches, including multimodal large language models (MLLMs), often struggle to infer complex contextual relationships or deliver comprehensive intelligence from unstructured data sources. In this paper, we introduce COSINT-Agent, a knowledge-driven multimodal agent tailored to address the challenges of OSINT in the Chinese domain. COSINT-Agent seamlessly integrates the perceptual capabilities of fine-tuned MLLMs with the structured reasoning power of the Entity-Event-Scene Knowledge Graph (EES-KG). Central to COSINT-Agent is the innovative EES-Match framework, which bridges COSINT-MLLM and EES-KG, enabling systematic extraction, reasoning, and contextualization of multimodal insights. This integration facilitates precise entity recognition, event interpretation, and context retrieval, effectively transforming raw multimodal data into actionable intelligence. Extensive experiments validate the superior performance of COSINT-Agent across core OSINT tasks, including entity recognition, EES generation, and context matching. These results underscore its potential as a robust and scalable solution for advancing automated multimodal reasoning and enhancing the effectiveness of OSINT methodologies.

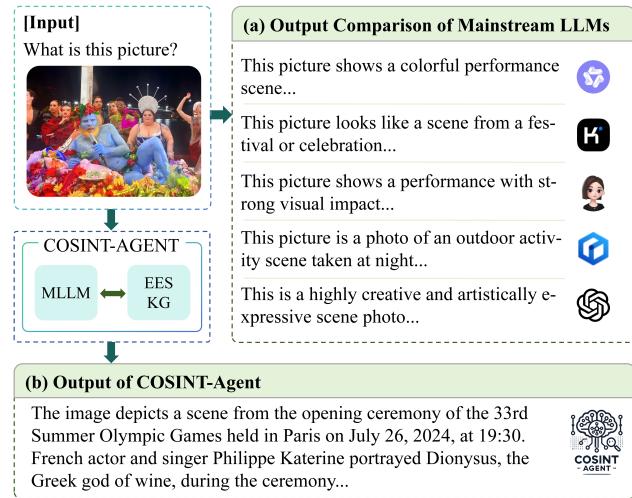


Figure 1. (a) The input image is processed by mainstream LLMs, which generate basic scene descriptions but lack in-depth entity, event, and context-level analysis. (b) In contrast, COSINT-Agent integrates the EES-KG to perform a more detailed analysis, identifying not just the scene, but also the events and context associated with the image. The English translations are provided for clarity and are not part of the original input or output.

1. Introduction

Open Source Intelligence (OSINT) refers to the systematic collection, analysis, and interpretation of publicly available data from a wide range of sources (Ghioni et al., 2024). The rapid proliferation of diverse digital content—including textual, visual, and social media data—has amplified the relevance of OSINT in various domains such as disaster response (Zhou et al., 2022), combating cybercrime (Nouh et al., 2019; Cartagena et al., 2020), enhancing cybersecurity threat awareness (Govardhan et al., 2023; Jesus et al., 2023; Riebe et al., 2024), sentiment analysis (Hernández et al., 2018), and risk assessment (Delavallade et al., 2017). The ability to efficiently extract actionable insights from vast, multimodal datasets offers significant potential for decision-making and situational awareness.

Despite its promising applications, the effective implementation of OSINT remains a challenging task due to the heterogeneity and enormous scale of data, as well as the necessity

¹School of Software, Shandong University, Jinan, China ²Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, China. Correspondence to: Zhi Liu <liuzhi@sdu.edu.cn>, Lizhen Cui <clz@sdu.edu.cn>.

of contextual reasoning. Existing OSINT analysis primarily relies on rule-based systems, traditional machine learning (Ndlovu et al., 2023), and deep learning methods (Zhou et al., 2022), which lack the adaptability and scalability required to handle complex and unstructured multimodal data. While multimodal large language models (MLLMs) offer promising potential for unifying text and image analysis, their application in OSINT remains largely unexplored due to two critical challenges. First, MLLMs are prone to hallucinations (Ji et al., 2023), generating factually incorrect outputs (Mondal et al., 2024), and lacking the capability to incorporate the latest domain-specific knowledge (Pan et al., 2024), which undermines their reliability in dynamic OSINT tasks. Second, these models are limited to generating superficial descriptions of images, lacking the ability to infer deeper contextual meanings such as the events depicted. For example, as illustrated in Figure 1, mainstream large language models—Tongyi Qianwen¹, Kimi², Doubao³, Wenxin Yiyuan⁴, and ChatGPT⁵—typically generate generalized scene descriptions such as "This is a colorful performance scene." While these outputs provide a surface-level understanding of the image, they fail to recognize the specific event or contextual details depicted. This limitation in deriving actionable insights from visual data significantly hinders their applicability and effectiveness in OSINT tasks.

To address these challenges, we propose COSINT-Agent, a knowledge-driven multimodal intelligent agent specifically designed for OSINT tasks. By integrating a fine-tuned MLLM with the Entity-Event-Scene Knowledge Graph (EES-KG), COSINT-Agent mitigates hallucinations and outdated knowledge, enhancing its reasoning capabilities with accurate, domain-specific context. Through two stages of fine-tuning, we improved the model's entity-type recognition and EES generation accuracy. COSINT-Agent ensures the incorporation of up-to-date information through the latest OSINT knowledge graph, providing a robust solution for dynamic intelligence scenarios. COSINT-Agent overcomes the limitation of generating superficial image descriptions by utilizing EES-Match, a framework that enables COSINT-MLLM to collaborate with the EES-KG. This allows COSINT-Agent to extract deeper semantic meanings from visual data, enabling the identification of the events depicted, their temporal and spatial contexts, and their broader implications. As shown in Figure 1, COSINT-Agent analyzes an image of the 33rd Summer Olympic Games opening ceremony, identifying not just the entities (Philippe Katerine), but also the event (the opening ceremony), its temporal and spatial contexts, and broader cultural impli-

cations. By combining advanced MLLM fine-tuning with knowledge graph integration, COSINT-Agent effectively bridges the gap between raw multimodal data and actionable intelligence, providing a comprehensive solution for OSINT applications.

Our contributions can be summarized as follows:

- We propose COSINT-Agent, the first MLLM-based Agent specifically designed for OSINT tasks. COSINT-Agent bridges the gap between multimodal data analysis and actionable intelligence, offering a novel solution for the challenges faced in open-source intelligence, particularly in processing and reasoning over diverse data types.
- We propose an innovative approach to integrating MLLMs with the EES-KG via EES-Match, enabling seamless collaboration between the model and the knowledge graph. This synergy significantly enhances the model's multimodal reasoning capabilities, allowing it to extract deeper contextual insights from complex visual and textual data.
- Through comprehensive experiments, we validate the effectiveness and value of COSINT-Agent in real-world OSINT scenarios. The results demonstrate the feasibility of our approach in improving entity-type recognition, image EES generation, and context matching via EES-Match in EES-KG, showcasing its potential as a powerful tool for OSINT applications.

2. Related Work

2.1. Multimodal Knowledge Graph

Multimodal Knowledge Graphs (MMKGs) enhance traditional knowledge graphs by integrating multimodal data, such as images and textual descriptions, to provide richer entity representations and support complex reasoning tasks. Prior work has demonstrated their effectiveness in various applications, including knowledge graph completion and triple classification by associating entity features with images (Xie et al., 2017; Mousselly-Sergieh et al., 2018), entity-aware image captioning through visual-entity relationships (Zhao & Wu, 2023), and recommendation systems by fusing structured knowledge with multimodal content (Sun et al., 2020). More recent approaches have explored MMKGs to improve large language model (LLM) reasoning. For instance, MR-MKG (Lee et al., 2024) introduces a multimodal reasoning framework that encodes structured knowledge using relation graph attention networks and aligns cross-modal semantics to bridge textual and visual information. While these works primarily focus on enriching LLMs with multimodal knowledge for specific tasks, our approach diverges by enabling real-time structured knowledge

¹<https://tongyi.aliyun.com/>

²<https://kimi.moonshot.cn/>

³<https://www.doubao.com/chat/>

⁴<https://yiyuan.baidu.com/>

⁵<https://chatgpt.com/>

retrieval and reasoning through EES-Match. Instead of passively integrating multimodal knowledge, COSINT-Agent actively matches and retrieves relevant contextual knowledge from the EES-KG, establishing a direct link between LLM-generated representations and structured event-driven intelligence.

2.2. Knowledge-Augmented LLMs

Recent advancements in knowledge-augmented LLMs have sought to mitigate hallucination and improve reasoning capabilities by integrating structured external knowledge. Traditional approaches focus on incorporating textual knowledge graphs into LLMs by retrieving relevant triples and transforming them into text-based prompts to enhance downstream tasks such as question answering and entity recognition. For instance, studies such as (Baek et al., 2023; Sen et al., 2023; Wu et al., 2023) utilize KG-to-text conversions to improve LLM reasoning without requiring re-training. However, direct injection of KG triples into LLM prompts often introduces noise due to irrelevant or extraneous context, limiting the effectiveness of such methods (Tian et al., 2024). More recent research, such as (Mondal et al., 2024), extends knowledge augmentation into multimodal settings, leveraging textual KGs to enhance reasoning in multimodal chain-of-thought tasks. (Wen et al., 2023) propose MindMap, a knowledge graph prompting technique that elicits structured reasoning pathways in LLMs, improving transparency and reducing hallucinations. While these methods enhance LLM performance, they primarily operate within textual knowledge representations, overlooking the structural and semantic richness of MMKGs. In contrast, our work integrates MMKGs into LLM inference via a novel EES-Match framework, enabling a structured and automated alignment between multimodal data and domain-specific structured knowledge. This approach ensures that LLMs can systematically retrieve, interpret, and reason over both textual and visual knowledge, leading to a more robust and context-aware OSINT analysis.

3. COSINT-Agent

3.1. Overview

This section provides an overview of COSINT-Agent, the core system designed for processing and reasoning over multimodal COSINT data. A visual representation of the workflow is shown in Figure 2. The COSINT-Agent system consists of three key components: EES-Recognition, EES-Match, and EES-KG. EES-Recognition employs a fine-tuned COSINT-LLM to extract EES from multimodal data, establishing a structured representation of the image content. EES-Match facilitates seamless alignment between multimodal reasoning and structured knowledge by linking extracted EES elements to EES-KG, a domain-specific

knowledge graph. This process enables the system to retrieve relevant contextual knowledge, enhancing its capacity for comprehensive data interpretation. By integrating structured reasoning with large-scale multimodal inference, EES-KG provides an essential contextual grounding, reinforcing COSINT-Agent’s ability to support OSINT tasks with more accurate, interpretable, and context-aware intelligence analysis.

3.2. Dataset

To support the development and evaluation of COSINT-Agent, we collected a total of 64,532 data records and 70,346 associated images, sourced from a variety of open-source platforms, as summarized in Appendix A. The dataset encompasses a diverse range of open-source information, enabling the development of robust and accurate models for OSINT tasks. To further enhance the system’s performance, two specialized instruction datasets were constructed, with the instruction data generated using ChatGPT4o and other MLLMs:

1. Entity(Type) Recognition Dataset:

This dataset is specifically designed to train and evaluate the recognition of entities and their associated types (e.g., Person, Organization, Location, Object, and others). It includes a total of 250,747 annotated entities for fine-tuning COSINT-LLM, significantly improving its ability to identify key entities relevant to OSINT. Additionally, a test set containing 62,242 entities was curated to assess the model’s generalization capabilities. The data underwent extensive cleaning and filtering processes, followed by manual verification to ensure high quality and relevance.

2. Multimodal EES Dataset:

The Multimodal EES Dataset focuses on extracting EES from images, which are critical for constructing the EES-KG and enabling seamless multimodal reasoning. This dataset comprises 3,225 annotated samples, with the data undergoing thorough cleaning, filtering, and final manual review to ensure consistency and reliability. It plays a crucial role in fine-tuning COSINT-LLM to enhance its ability to describe complex multimodal contexts.

3.3. MLLM fine-tuning

COSINT-Agent was developed using the Qwen2-VL-7B model (Wang et al., 2024b) as the foundation for COSINT-MLLM. A structured two-stage fine-tuning process was conducted with LLaMA-Factory (Zheng et al., 2024), leveraging the previously constructed Entity(Type) Recognition and Multimodal EES Datasets. In the first stage, the model

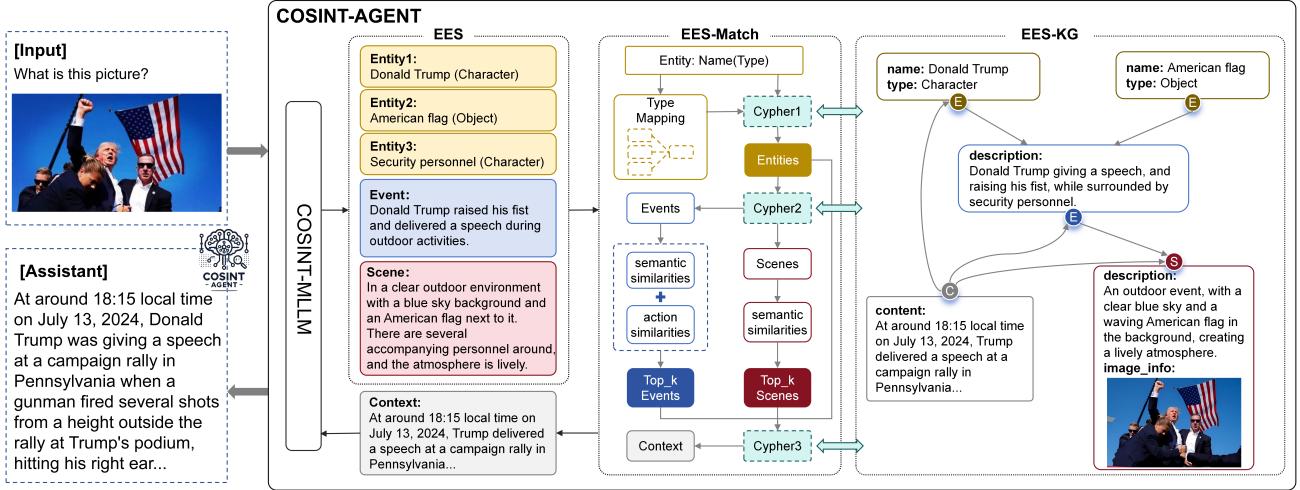


Figure 2. The image illustrates the inference process of COSINT-AGENT, where the system performs EES analysis on the image using COSINT-MLM, then matches the context with the knowledge graph through the EES-Match process, ultimately outputting the inferred result. Note that English is not a part of the input and output, where they are given for better illustration.

was fine-tuned using the Entity(Type) Recognition Dataset, aiming to enhance its capability to identify critical entities and their corresponding types within the OSINT domain. This stage involved instruction fine-tuning, which significantly bolstered the model’s foundational understanding of text-based entities and their roles, laying a strong groundwork for subsequent tasks. In the second stage, we further fine-tuned the model using the Multimodal EES Dataset to enable effective recognition and generation of Entities, Events, and Scenes from multimodal data. This stage optimized the model’s ability to perceive complex multimodal information and produce EES descriptions tailored for integration with the EES-Match framework. Key fine-tuning parameters included the use of LoRA (Hu et al., 2021) for efficient parameter updates, a learning rate of 5×10^{-5} , 3 training epochs, a batch size of 2, gradient accumulation steps set to 8, and the AdamW optimizer for improved convergence stability. This efficient setup ensured that the fine-tuning process was both resource-effective and highly performant. The effectiveness of this two-stage fine-tuning is demonstrated in the experimental results, with additional validation provided by the loss curves presented in Appendix B. The first phase refined the model’s text-based recognition capabilities, while the second phase focused on multimodal EES recognition, enhancing its reasoning and descriptive accuracy for OSINT tasks. Together, these stages significantly improved COSINT-Agent’s multimodal understanding and its ability to generate actionable intelligence, demonstrating its adaptability and robustness in processing both textual and visual data.

3.4. EES-KG

To demonstrate the effectiveness of the EES-KG, we use the Figure 2 as a concrete example. The hierarchical framework—Entity Level, Event Level, and Scene Level—systematically extracts, organizes, and represents multimodal knowledge, facilitating reasoning by MLLMs. Below, we describe each level, accompanied by relevant mathematical representations.

At the Entity Level, the goal is to identify and categorize key entities within the image. These entities are extracted from the image, focusing on semantic categories such as Location, Organization, Object, Person, Document, and others. Unlike traditional object detection, our method does not rely on bounding box annotations but instead emphasizes the semantic significance of each entity. Entities can be formally represented as:

$$E = \{e_1, e_2, \dots, e_n\} \quad (1)$$

Where E : Entity set. e_i : Individual entity node, represented as: $e_i = (id, type, name)$. id : Unique identifier for the entity. $type$: Entity type (e.g., Character, Object). $name$: Entity label (e.g., Donald Trump, American Flag). At this level, the entities serve as the building blocks for higher-level reasoning, forming the foundation for subsequent layers.

At the Event Level, the focus is on identifying and describing key events involving the primary entities in the image. These events capture the interactions, actions, or states of the entities, represented as textual descriptions. Unlike traditional event detection, which may rely on temporal markers or relationships, our approach emphasizes the content of the event itself, particularly how entities in the image are

involved in a specific situation or activity. Each event is represented by a description attribute, which provides a detailed account of the event taking place. For instance, the event describes Donald Trump giving a speech, and raising his fist, while surrounded by security personnel. Events can be mathematically represented as:

$$EV = \{(e_i, r_{ij}, e_j, d) | e_i, e_j \in E\} \quad (2)$$

Where EV : Event set. e_i, e_j : Entities participating in the event. r_{ij} : The action or interaction between the entities. d : A textual description of the event. Each event node can be further defined as: $ev = (id, entities, action, description)$. At this level, the events serve as the key drivers for understanding the dynamics of the scene, laying the groundwork for more complex contextual analysis in subsequent layers.

At the Scene Level, events are aggregated into a broader contextual representation, capturing high-level semantics and integrating temporal, spatial, and environmental context to form a comprehensive understanding of the image. By incorporating the surrounding context, including location, time, and additional background information, the Scene Level enhances the Agent's ability to comprehend complex multimodal data, improving its overall situational awareness and perceptual capabilities. For instance, consider the scene description: An outdoor event, with a clear blue sky and a waving American flag in the background, creating a lively atmosphere. This description captures the environmental context—clear blue sky, the waving American flag, and a lively atmosphere—providing a richer understanding of the scene. The scene can be mathematically represented as:

$$SC = f_{scene}(EV) \quad (3)$$

Where SC : Scene representation. f_{scene} : Scene aggregation function that clusters related events and contextualizes them. A scene node can be described as: $sc = (id, events, location, time, context)$. At this level, the combination of events, entities, and contextualization within the environment enhances multimodal reasoning, providing a comprehensive understanding of the scene and improving the Agent's situational awareness.

Building upon the EES Level, we introduce the Context node, which represents the broader context of the image within the OSINT. The Context node is linked to all Entity, Event, and Scene nodes, providing essential contextual information that relates to the image, such as geographical, temporal, or situational context drawn from the open-source data associated with the image. The Context node is connected to each of the nodes as follows:

$$C = \{(e_i, ev_j, sc_k) | e_i \in E, ev_j \in EV, sc_k \in SC\} \quad (4)$$

Where E : Set of Entity nodes. EV : Set of Event nodes.

SC : Set of Scene nodes. C : The Context node that is associated with each of the EES nodes.

The Context node is not initially part of the EES nodes but is dynamically retrieved through a successful match with the EES nodes. Let the process of matching be represented by the function $f_{EES-Match}$, which identifies the corresponding context for the image based on the identified Entity, Event, and Scene:

$$C = f_{EES-Match}(E, EV, SC) \quad (5)$$

Through the successful match of EES nodes with the Context node, the system can retrieve relevant background knowledge, allowing for a more holistic interpretation of the image's content and its relationship to broader events, locations, and timeframes.

3.5. EES-Match

The EES-Match module serves as a crucial link between COSINT-LLM with EES-KG, establishing a direct connection between the visual content and structured knowledge. This integration harnesses the full potential of MLLMs' multimodal understanding and generative capabilities, enabling more context-aware and insightful analysis of images. The EES-Match process is designed to integrate various multimodal data layers—Entities, Events, Scenes, and Context—into a comprehensive multimodal reasoning system. This process is executed in multiple stages, using type mappings and semantic similarity evaluations to link Entities with their corresponding Events and Scenes from the knowledge graph, ultimately retrieving the relevant Context, see Figure 2. The following steps outline the core components of the EES-Match process:

In the first stage, we begin with a structured analysis of the Entities in EES data by parsing their Name and Type attributes. Using both Name and Type ensures accurate identification of entities, as it helps avoid ambiguity in cases where the same name can refer to different entities across domains. For example, "Xiaomi" could refer to a food item or a technology company, and relying solely on the Name would lead to confusion. We then map similar Types, such as Object, Item, Entity or Location, Place, and Geographical Area, into corresponding categories to ensure that related entities are grouped together. The Name and Type are transformed into the Cypher1 query language format, enabling precise queries to the knowledge graph. The specific query process is detailed in Algorithm 1, presented in Appendix C.1, which outlines the steps for constructing and executing Cypher1 queries to retrieve relevant entities from the knowledge graph efficiently. This enables the retrieval of all matching entities based on both their Type and Name, ensuring a comprehensive match with the knowledge graph.

After retrieving matching entities, Cypher2 queries are gen-

Table 1. The table showcases the impressive performance of COSINT-MLLM after fine-tuning, as it outperforms other models in terms of precision, recall, and F1 score for the Entity Recognition task.

Model	Modality	Zh Support	Parameters	Precision	Recall	F1
Glm-4-9b-chat (GLM et al., 2024)	Text	Yes	9.40B	0.60	0.57	0.56
Qwen2.5-7B-Instruct (Team, 2024)	Text	Yes	7.07B	0.75	0.57	0.63
Meta-Llama-3-8B-Instruct (AI@Meta, 2024)	Text	Limited	7.50B	0.67	0.46	0.52
Glm-4v-9b (Wang et al., 2023)	Text, Image	Yes	13.91B	0.56	0.42	0.43
InternVL2-8B (Chen et al., 2024b)	Text, Image	Yes	8.08B	0.63	0.27	0.35
Qwen2-VL-7B-Instruct (Wang et al., 2024a)	Text, Image	Yes	7.07B	0.68	0.45	0.51
COSINT-MLLM	Text, Image	Yes	7.07B	0.81	0.79	0.79

erated to identify all related Events and Scenes from the knowledge graph. These queries utilize the previously matched entities to extract their associated events and scenes, forming the basis for multimodal reasoning. The detailed query process is outlined in Algorithm 2, as presented in Appendix C.2, which describes the structured approach to linking entities with their contextual events and scenes. For Events, we first analyze the description field of each event node to identify the key actions or relevant events involving the main figures. The first step is to calculate action feature similarity. This is computed using cosine similarity between the action vector representations of the event and the identified entities:

$$S_{action}(ev_i, ev_j) = \frac{\text{vec}(action_i) \cdot \text{vec}(action_j)}{\|\text{vec}(action_i)\| \|\text{vec}(action_j)\|} \quad (6)$$

Where $S_{action}(ev_i, ev_j)$ is the action similarity between the event ev_i and ev_j . $\text{vec}(action_i)$ and $\text{vec}(action_j)$ are the vector representations of the actions in the entity and event. Next, we calculate the semantic similarity of the event description, which measures the alignment between the entity’s context and the overall event description. This is also computed using cosine similarity:

$$S_s(ev_i, ev_j) = \frac{\text{vec}(desc_i) \cdot \text{vec}(desc_j)}{\|\text{vec}(desc_i)\| \|\text{vec}(desc_j)\|} \quad (7)$$

Where $S_s(ev_i, ev_j)$ is the semantic similarity between event description $desc_i$ and $desc_j$. $\text{vec}(desc_i)$ and $\text{vec}(desc_j)$ are the vector representations of the event description. The final Event similarity score is then calculated by combining the action similarity and semantic similarity with appropriate weights w_1 and w_2 , representing the relative importance of each similarity:

$$S_{EV}(ev_i, ev_j) = w_1 \cdot S_a(ev_i, ev_j) + w_2 \cdot S_s(ev_i, ev_j) \quad (8)$$

Where $S_{EV}(ev_i, ev_j)$ is the final event similarity score between event ev_i and ev_j . The top k events with the highest similarity scores are selected. For Scenes, we compute the

semantic similarity between the scene descriptions and the candidate scenes in the knowledge graph:

$$S_{scene}(sc_i, sc_j) = \frac{\text{vec}(scene_i) \cdot \text{vec}(scene_j)}{\|\text{vec}(scene_i)\| \|\text{vec}(scene_j)\|} \quad (9)$$

Where $S_{scene}(sc_i, sc_j)$ is the semantic similarity between scene descriptions sc_i and sc_j . $\text{vec}(scene_i)$ and $\text{vec}(scene_j)$ represent the vector representations of the scene descriptions. Similar to Event Matching, the top k scenes are selected based on the highest semantic similarity scores.

Finally, the aggregated Entities, Top-k Events, and Top-k Scenes are utilized to construct a Cypher3 query, designed to retrieve the corresponding Context node from the EES-KG. This query process establishes connections between the multimodal elements and their contextual representation in the knowledge graph. The detailed query process is outlined in Algorithm 3, as shown in Appendix C.3, providing a structured framework for context retrieval based on the EES nodes. The Context is defined as the broader background information that links together the relevant entities, events, and scenes within the multimodal image:

$$\text{Context} = f_{context}(\text{Entities}, \text{Top}_k \text{EV}, \text{Top}_k \text{SC}) \quad (10)$$

Where $f_{context}$ is a function that synthesizes the Top-k Events and Top-k Scenes to retrieve the appropriate Context. This process links the image’s components with broader contextual information, enhancing the model’s ability to interpret and reason about the image within a larger framework. The retrieved Context node provides crucial details, such as description, time, and location, further enriching the multimodal understanding.

4. Experiments

4.1. Setups

In our experiments, we designed three comprehensive tasks to evaluate the performance of COSINT-Agent: Entity

Table 2. Performance comparison of COSINT-MLLM (EES), COSINT-MLLM (ER+EES), and baseline models across Entity Recognition, Event Generation, and Scene Analysis tasks. Metrics include Precision, Recall, F1, ROUGE-1, ROUGE-L, and BLEU scores.

Model	Entity				Event			Scene	
	Precision	Recall	F1	ROU-1	ROU-L	BLEU	ROU-1	ROU-L	BLEU
Glm-4v-9b	0.42	0.54	0.45	0.38	0.35	0.37	0.26	0.21	0.14
InternVL2-8B	0.36	0.52	0.41	0.41	0.36	0.33	0.36	0.27	0.28
Qwen2-VL-7B	0.43	0.42	0.41	0.39	0.41	0.36	0.34	0.26	0.23
COSINT(EES)	0.67	0.71	0.68	0.65	0.61	0.58	0.57	0.42	0.41
COSINT(ER+EES)	0.71	0.75	0.72	0.66	0.63	0.59	0.59	0.43	0.42

Recognition, EES Generation, and Context matching via EES-Match in EES-KG. All experiments were conducted on a dual-A800 80GB GPU environment, ensuring efficient handling of LLMs and datasets. The experiments utilized Neo4j as the underlying knowledge graph framework, with Cypher (Francis et al., 2018) employed as the query language for interacting with the knowledge graph.

Entity Recognition experiment aimed to improve COSINT-MLLM’s ability to recognize entities and their corresponding types from text data, a fundamental requirement for OSINT tasks. The model was fine-tuned using the Entity(Type) Recognition Dataset, which contains diverse entities such as Person, Organization, Location, and Object. After fine-tuning, the performance of COSINT-MLLM was compared with other mainstream models in terms of accuracy, precision, recall, and F1 score for entity-type recognition.

To enhance COSINT-MLLM’s ability to perceive and describe multimodal data, the EES Generation experiment focused on extracting EES from text and images. The model was fine-tuned using the Multimodal EES Dataset, which includes multimodal instructions designed to evaluate the model’s understanding of complex multimodal relationships. We compared the output of COSINT-MLLM against other MLLMs based on metrics such as precision, recall, and F1 score for Entity extraction, as well as BLEU and ROUGE scores for Event and Scene descriptions.

Context Matching via EES-Match in the EES-KG experiment evaluated the accuracy of various LLMs in identifying the context of multimodal data through EES-Match. Specifically, the task involved matching the EES descriptions generated by different models to corresponding Context nodes in the EES-KG. The Context node represents the broader contextual meaning or scenario depicted by the multimodal data. Metrics such as Match accuracy, Hits@1, and Hits@2 were used to assess the models’ ability to identify the correct Context node, demonstrating the integration of multimodal reasoning and knowledge graph capabilities.

These experiments collectively evaluate COSINT-Agent’s capability in recognizing, generating, and reasoning over

multimodal data for OSINT tasks, highlighting its adaptability and robustness in integrating LLMs with structured knowledge graphs.

4.2. Entity Recognition

This experiment evaluates the ability of COSINT-MLLM to recognize entities and their associated types in textual data, compared to several mainstream LLMs with similar parameter sizes. Table 1 provides a detailed comparison, including each model’s modality type, Chinese language support, and parameter size, along with performance metrics such as Precision, Recall, and F1-score. The models evaluated include text-only models like GLM-4-9b-chat (GLM et al., 2024), Qwen2.5-7B-Instruct (Team, 2024), and Meta-Llama-3-8B-Instruct (AI@Meta, 2024), as well as multimodal models like GLM-4v-9b (Wang et al., 2023), InternVL2-8B (Chen et al., 2024b), and Qwen2-VL-7B-Instruct (Wang et al., 2024a; Bai et al., 2023). While text-only models generally exhibit limited adaptability to multimodal OSINT tasks, the multimodal models struggled to achieve comparable performance in entity recognition. COSINT-MLLM, fine-tuned using the Entity(Type) Recognition Dataset, achieved significant advancements, with a Precision of 0.81, Recall of 0.79, and F1-score of 0.79. While the fine-tuning process yielded noticeable improvements over text-only LLMs, the model exhibited substantial progress compared to other MLLMs, particularly in accurately identifying entities in Chinese text.

4.3. EES Generation

As shown in Table 2, COSINT-MLLM (ER+EES) demonstrates consistent improvements over COSINT-MLLM (EES) and other mainstream multimodal large language models (MLLMs) across all tasks and metrics. For the Entity Recognition task, COSINT-MLLM (ER+EES) achieves Precision (0.71), Recall (0.75), and F1 (0.72), compared to Precision (0.67), Recall (0.71), and F1 (0.68) for COSINT-MLLM (EES). These results indicate that incorporating entity recognition in the fine-tuning process provides additional benefits, improving the model’s understanding of entities and their types. In the Event Generation

Table 3. Performance of different models in context matching via EES-Match.

Model	Match	Hit@1	Hit@2
Glm-4v-9b	0.76	0.44	0.47
InternVL2-8B	0.73	0.34	0.37
Qwen2-VL-7B	0.77	0.61	0.66
COSINT(EES)	0.91	0.81	0.87
COSINT(ER+EES)	0.96	0.83	0.89

task, COSINT-MLLM (ER+EES) achieves slightly better scores, with ROUGE-1 (0.66), ROUGE-L (0.63), and BLEU (0.59), compared to ROUGE-1 (0.65), ROUGE-L (0.61), and BLEU (0.58) for COSINT-MLLM (EES). Similarly, in the Scene Analysis task, COSINT-MLLM (ER+EES) achieves ROUGE-1 (0.59), ROUGE-L (0.43), and BLEU (0.42), slightly outperforming COSINT-MLLM (EES), which achieves ROUGE-1 (0.57), ROUGE-L (0.42), and BLEU (0.41). While the improvements between COSINT-MLLM (ER+EES) and COSINT-MLLM (EES) are incremental, they highlight the advantages of combining entity recognition and multimodal EES fine-tuning in refining the model’s overall performance. Both configurations significantly outperform baseline models like Qwen2-VL-7B, InternVL2-8B, and Glm-4v-9b, which show much lower scores across all metrics. For instance, InternVL2-8B, the best-performing baseline, achieves ROUGE-1 (0.41), ROUGE-L (0.36), and BLEU (0.33) for Event Generation, and ROUGE-1 (0.36), ROUGE-L (0.27), and BLEU (0.28) for Scene Analysis.

4.4. Context Matching via EES-Match in EES-KG

For the Context Matching via EES-Match in the EES-KG experiment, we directly used the EES Generation task dataset, where the model-generated data, including entities, events, and scenes, was matched to the knowledge graph using the EES-Match module. We utilized BGE-M3 (Chen et al., 2024a) for vectorization and zh-core-web-sm for extracting key actions and event-related information. The matching process was evaluated using Hits@1 and Hits@2, comparing the performance of EES-Match with other baseline models. This approach enables the automated retrieval of relevant context, enhancing multimodal reasoning for OSINT applications. This table compares the ability of different models to match the Context node in the knowledge graph through EES-Match. The Match column represents the overall probability of successfully matching the Context node, while Hit@1 and Hit@2 show the probabilities of the Context node being ranked in the first and second positions, respectively. The results demonstrate that COSINT-MLLM (with and without Entity Recognition) achieves significantly higher performance in context matching than other MLLMs.

Table 4. Ablation study results for EES-Match, comparing different configurations (ES, EE, EES) and the impact of action similarity in Event matching.

Method	Similarity	Match	Hit@1	Hit@2
ES	semantic	0.96	0.48	0.59
EE	semantic	0.96	0.50	0.59
	action	0.96	0.38	0.47
	semantic+action	0.96	0.54	0.64
EES	semantic	0.96	0.78	0.82
	action	0.96	0.48	0.59
	semantic+action	0.96	0.83	0.89

As shown in Table 3, COSINT-MLLM (with and without Entity Recognition and EES fine-tuning) outperforms other mainstream models in matching the Context node. In particular, COSINT(ER+EES) achieves a Match score of 0.96, a Hit@1 score of 0.83, and a Hit@2 score of 0.89, highlighting the strength of the proposed framework in improving the accuracy of multimodal context retrieval. These results confirm the feasibility and effectiveness of leveraging EES-Match to integrate multimodal reasoning with knowledge graph-based context matching.

4.5. Ablation Study

To analyze the effectiveness of EES-Match, we conducted an ablation study focusing on three configurations: ES (Entity-Scene), EE (Entity-Event), and EES (Entity-Event-Scene). Furthermore, for Event matching, we evaluated the impact of incorporating action similarity alongside semantic similarity. As shown in Table 4, the following observations were made: The EES configuration consistently outperformed the ES and EE methods across all metrics. Specifically, EES with combined semantic and action similarity achieved the highest Hit@1 (0.83) and Hit@2 (0.89), demonstrating the importance of integrating all three elements (Entity, Event, and Scene) for context matching. Adding action similarity to the Event matching process improved the Hit@1 and Hit@2 scores for both EE and EES methods. For example, in the EES configuration, using both semantic and action similarity resulted in a significant increase in Hit@1 from 0.78 (semantic-only) to 0.83 (semantic + action), highlighting the importance of detailed action-level reasoning in multimodal context retrieval. These results validate the design choices of EES-Match and demonstrate that the combination of semantic and action similarity, along with the integration of Entity, Event, and Scene information, significantly enhances the performance of context matching within the knowledge graph.

5. Conclusion

We proposed COSINT-Agent, a knowledge-driven multimodal intelligent agent designed to address OSINT challenges by integrating COSINT-MLLM with the EES-KG through the innovative EES-Match framework. Fine-tuned using specialized datasets for entity recognition and EES generation, COSINT-Agent demonstrates superior performance in identifying entities, generating EES descriptions, and retrieving context-specific insights. Experimental results highlight its ability to bridge multimodal analysis and structured reasoning, establishing a scalable and precise solution for OSINT tasks. Looking forward, we anticipate the expanded application of MLLMs in OSINT, with future efforts focusing on integrating real-time data, domain-specific knowledge, and advanced reasoning techniques to further enhance their adaptability and impact in complex intelligence scenarios.

References

- AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- Baek, J., Aji, A., and Saffari, A. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.
- Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- Cartagena, A., Rimmer, G., van Dalsen, T., Watkins, L., Robinson, W. H., and Rubin, A. Privacy violating open-source intelligence threat evaluation framework: a security assessment framework for critical infrastructure owners. In *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0494–0499. IEEE, 2020.
- Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024a.
- Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.
- Delavallade, T., Bertrand, P., and Thouvenot, V. Extracting future crime indicators from social media. *Using Open Data to Detect Organized Crime Threats: Factors Driving Future Crime*, pp. 167–198, 2017.
- Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., Plantikow, S., Rydberg, M., Selmer, P., and Taylor, A. Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 international conference on management of data*, pp. 1433–1445, 2018.
- Ghioni, R., Taddeo, M., and Floridi, L. Open source intelligence and ai: a systematic review of the gelsi literature. *AI & society*, 39(4):1827–1842, 2024.
- GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Rojas, D., Feng, G., Zhao, H., Lai, H., Yu, H., Wang, H., Sun, J., Zhang, J., Cheng, J., Gui, J., Tang, J., Zhang, J., Li, J., Zhao, L., Wu, L., Zhong, L., Liu, M., Huang, M., Zhang, P., Zheng, Q., Lu, R., Duan, S., Zhang, S., Cao, S., Yang, S., Tam, W. L., Zhao, W., Liu, X., Xia, X., Zhang, X., Gu, X., Lv, X., Liu, X., Liu, X., Yang, X., Song, X., Zhang, X., An, Y., Xu, Y., Niu, Y., Yang, Y., Li, Y., Bai, Y., Dong, Y., Qi, Z., Wang, Z., Yang, Z., Du, Z., Hou, Z., and Wang, Z. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.
- Govardhan, D., Krishna, G. G. S. H., Charan, V., Sai, S. V. A., and Chintala, R. R. Key challenges and limitations of the osint framework in the context of cybersecurity. In *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, pp. 236–243. IEEE, 2023.
- Hernández, M., Hernández, C., Díaz-López, D., García, J. C., and Pinto, R. A. Open source intelligence (osint) as support of cybersecurity operations: Use of osint in a colombian context and sentiment analysis. *Revista Vínculos Ciencia, tecnología y sociedad*, 15(2), 2018.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Jesus, V., Bains, B., and Chang, V. Sharing is caring: Hurdles and prospects of open, crowd-sourced cyber threat intelligence. *IEEE Transactions on Engineering Management*, 71:6854–6873, 2023.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- Lee, J., Wang, Y., Li, J., and Zhang, M. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*, 2024.

- Mondal, D., Modi, S., Panda, S., Singh, R., and Rao, G. S. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18798–18806, 2024.
- Mousselly-Sergieh, H., Botschen, T., Gurevych, I., and Roth, S. A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 225–234, 2018.
- Ndlovu, L., Mkuzangwe, N., De Kock, A., Thwala, N., Mokoena, J., and Matimatjatji, R. A situational awareness tool using open-source intelligence (osint) and artificial intelligence (ai). In *2023 IEEE International Conference on Advances in Data-Driven Analytics And Intelligent Systems (ADACIS)*, pp. 1–6. IEEE, 2023.
- Nouh, M., Nurse, J. R., Webb, H., and Goldsmith, M. Cyber-crime investigators are users too! understanding the socio-technical challenges faced by law enforcement. *arXiv preprint arXiv:1902.06961*, 2019.
- Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.
- Riebe, T., Bäumler, J., Kaufhold, M.-A., and Reuter, C. Values and value conflicts in the context of osint technologies for cybersecurity incident response: A value sensitive design perspective. *Computer Supported Cooperative Work (CSCW)*, 33(2):205–251, 2024.
- Sen, P., Mavadia, S., and Saffari, A. Knowledge graph-augmented language models for complex question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pp. 1–8, 2023.
- Sun, R., Cao, X., Zhao, Y., Wan, J., Zhou, K., Zhang, F., Wang, Z., and Zheng, K. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 1405–1414, 2020.
- Team, Q. Qwen2.5: A party of foundation models, September 2024. URL <https://qwenlm.github.io/blog/qwen2.5/>.
- Tian, Y., Song, H., Wang, Z., Wang, H., Hu, Z., Wang, F., Chawla, N. V., and Xu, P. Graph neural prompting with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19080–19088, 2024.
- Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.
- Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.
- Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., and Tang, J. Cogylm: Visual expert for pretrained language models, 2023.
- Wen, Y., Wang, Z., and Sun, J. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*, 2023.
- Wu, Y., Hu, N., Bi, S., Qi, G., Ren, J., Xie, A., and Song, W. Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering. *arXiv preprint arXiv:2309.11206*, 2023.
- Xie, R., Liu, Z., Luan, H., and Sun, M. Image-embodied knowledge representation learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2017.
- Zhao, W. and Wu, X. Boosting entity-aware image captioning with multi-modal knowledge graph. *IEEE Transactions on Multimedia*, 2023.
- Zheng, Y., Zhang, R., Zhang, J., Ye, Y., Luo, Z., Feng, Z., and Ma, Y. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL <http://arxiv.org/abs/2403.13372>.
- Zhou, B., Zou, L., Mostafavi, A., Lin, B., Yang, M., Gharaibeh, N., Cai, H., Abedin, J., and Mandal, D. Victimfinder: Harvesting rescue requests in disaster response from social media with bert. *Computers, Environment and Urban Systems*, 95:101824, 2022.

A. Dataset

Table 5. This table summarizes the sources and quantities of the collected open-source intelligence data, including the number of text-based data entries and associated images.

Data Sources	Data Counts	Image Counts
Fenghuang ⁶	1052	3324
Huanqiu ⁷	11010	19515
Guofangbu ⁸	4564	4010
81cn ⁹	836	930
Tencent ¹⁰	887	3163
Sohu ¹¹	1771	6229
Wangyi ¹²	237	348
Xinhua ¹³	216	249
Yangshi ¹⁴	261	292
Weibo ¹⁵	43698	32286
All	64532	70346

The collected dataset is categorized into three distinct types based on the sources of open-source information and their credibility:

Official Open Source Information: This category includes data from trusted and authoritative sources, such as government websites and officially sanctioned platforms (e.g., Fenghuang, Huanqiu, and Guofangbu). These sources provide highly credible and comprehensive information, making them an ideal foundation for constructing reliable datasets. Due to its high credibility and completeness, this category is predominantly used for constructing datasets for Entity(Type) recognition and multimodal EES generation experiments.

Non-Official Open Source Information: This category encompasses data from reputable news platforms and commercial media outlets, such as Tencent, Sohu, and Wangyi. While the information in this category is generally reliable, occasional biases or inaccuracies may exist. Most of this data is utilized for Entity(Type) recognition experiment dataset construction, with a smaller portion incorporated into the multimodal EES generation dataset.

Personal and Social Media Open Source Information: This category primarily includes data from social media platforms, such as Weibo. Although these sources provide significant data volume, their reliability is lower due to potential biases, personal opinions, and unverified content. This data is mainly employed for Entity(Type) recognition dataset construction, with a small fraction used for the multimodal EES generation dataset.

By leveraging the diverse range of sources in this hierarchical structure, the dataset ensures a balanced representation of information with varying degrees of credibility. This structure allows for targeted use in different experimental tasks, enhancing the robustness and reliability of the COSINT-Agent in multimodal reasoning over heterogeneous open-source data.

¹<https://news.ifeng.com/>

²<https://mil.huanqiu.com>

³<http://www.mod.gov.cn>

⁴<http://www.81.cn/>

⁵<https://news.qq.com>

⁶<https://www.sohu.com>

⁷<https://war.163.com/index.html>

⁸<http://www.news.cn/milpro/index.htm>

⁹<https://military.cctv.com/index.shtml>

¹⁰<https://weibo.com>

B. LLM fine-tuning loss

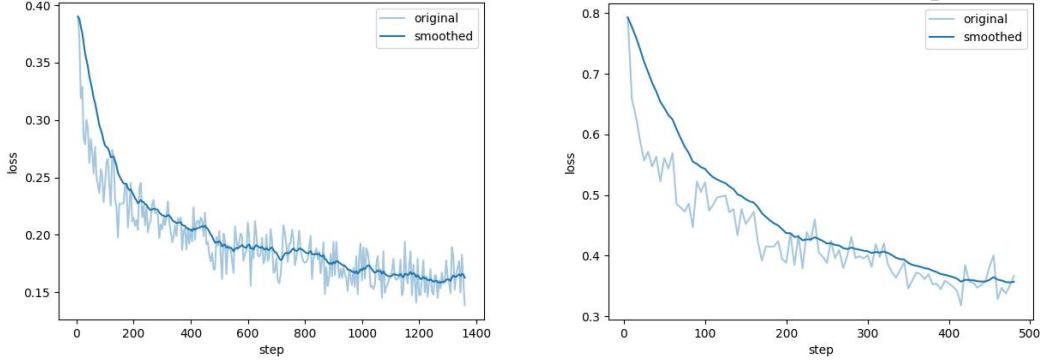


Figure 3. The left plot shows the loss curve of COSINT-MLLM during fine-tuning with the Entity(Type) Recognition Dataset, illustrating both the original and smoothed loss curves. The right plot presents the loss curve of COSINT-MLLM during fine-tuning with the Multimodal EES Dataset, similarly displaying the original and smoothed loss curves.

C. EES-Match Algorithm

C.1. Entity-Match Algorithm

Algorithm 1 Cypher1: Entity Query Process

```

1: if entity_type is not empty then
2:   Cypher1  $\leftarrow$  MATCH (e:Entity)
3:   WHERE e.name AND e.type
4:   RETURN e.name AS Name, e.type AS Type
5: else
6:   Cypher1  $\leftarrow$  MATCH (e:Entity)
7:   WHERE e.name
8:   RETURN e.name AS Name, e.type AS Type
9: end if
10: result  $\leftarrow$  graph.run(Cypher1).data()
11: if result is empty AND entity_type is not empty then
12:   Cypher1  $\leftarrow$  MATCH (e:Entity)
13:   RETURN e.name AS Name, e.type AS Type
14:   all_entities  $\leftarrow$  graph.run(Cypher1).data()
15:   if all_entities is empty then
16:     break
17:   end if
18:   candidates  $\leftarrow$  [item.Name for item in all_entities]
19:   result  $\leftarrow$  sim_func(name, candidates, all_entities)
20: end if

```

C.2. Queries on Events and Scenes

Algorithm 2 Cypher2: Queries on Events and Scenes

```
1: for each entity in matched_entities do
2:   entity_name, entity_type ← entity.Name, entity.Type
3:   if entity_type is None then
4:     Cypher2 ← MATCH (e:Entity {name:entity_name})
5:     OPTIONAL MATCH (e)-[:INVOLVED_IN]->(ev:Event)
6:     OPTIONAL MATCH (ev)-[:OCCURS_IN]->(s:Scene)
7:     RETURN ev.description AS Event, s.description AS Scene
8:   else
9:     Cypher2 ← MATCH (e:Entity {name:entity_name, type:entity_type})
10:    OPTIONAL MATCH (e)-[:INVOLVED_IN]->(ev:Event)
11:    OPTIONAL MATCH (ev)-[:OCCURS_IN]->(s:Scene)
12:    RETURN ev.description AS Event, s.description AS Scene
13:   results ← graph.run(Cypher2).data()
```

C.3. Queries on Context

Algorithm 3 Cypher3: Queries Context Based on Entity, Event, and Scene

```
1: function QUERYCONTEXT(entity_name, entity_type, event_description, scene_description)
2:   Cypher3 ← MATCH (e:Entity {name:entity_name, type:entity_type})
3:   MATCH (ev:Event {description:event_description})
4:   MATCH (s:Scene {description:scene_description})
5:   MATCH (c:Context)
6:   WHERE (c)-[:DESCRIBES]->(e) AND (c)-[:DESCRIBES]->(ev) AND
(c)-[:DESCRIBES]->(s)
7:   RETURN c
8:   results ← graph.run(Cypher3).data()
9:   return results
10: end function
```
