Synergizing Multimodal Temporal Knowledge Graphs and Large Language Models for Social Relation Recognition

Haorui Wang[1], **Zheng Wang**[2], **Yuxuan Zhang**[1],
Bo Wang[1], **Bin Wu**[1]*
[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]China Mobile Research Institute, Beijing, China
{wang_harry_cn,yxzhang,wangbo,wubin}@bupt.edu.cn
wangzhengyjy@chinamobile.com

Abstract

Recent years have witnessed remarkable advances in Large Language Models (LLMs). However, in the task of social relation recognition, Large Language Models (LLMs) encounter significant challenges due to their reliance on sequential training data, which inherently restricts their capacity to effectively model complex graph-structured relationships. To address this limitation, we propose a novel low-coupling method synergizing **m**ultimodal **t**emporal **K**nowledge **G**raphs and **L**arge **L**anguage **M**odels (mtKG-LLM) for social relation reasoning. Specifically, we extract multimodal information from the videos and model the social networks as spatial Knowledge Graphs (KGs) for each scene. Temporal KGs are constructed based on spatial KGs and updated along the timeline for long-term reasoning. Subsequently, we retrieve multi-scale information from the graph-structured knowledge for LLMs to recognize the underlying social relation. Extensive experiments demonstrate that our method has achieved state-of-the-art performance in social relation recognition. Furthermore, our framework exhibits effectiveness in bridging the gap between KGs and LLMs. We release our code at https://github.com/HarryWgCN/mtKG-LLM.

1 Introduction

Social relation recognition is an emerging and essential task in the field of natural language processing and computer vision. It focuses on identifying and interpreting the social relations across various data forms. As a combination of different modalities, videos offer rich information from both linguistic and visual aspects simultaneously, allowing for a deeper understanding of semantic content. Thus, social relation recognition in videos has become a prominent research topic over the past decade. This task involves extracting social relations between characters in the video, as shown in Figure 1, and maintaining a social network along the timeline. Beyond its direct applications in social network extraction and analysis, it serves as an important foundation for other multimodal tasks, such as video summarization and video question answering. Therefore, recognizing social relations in videos is both crucial and fundamental for multimodal information extraction.

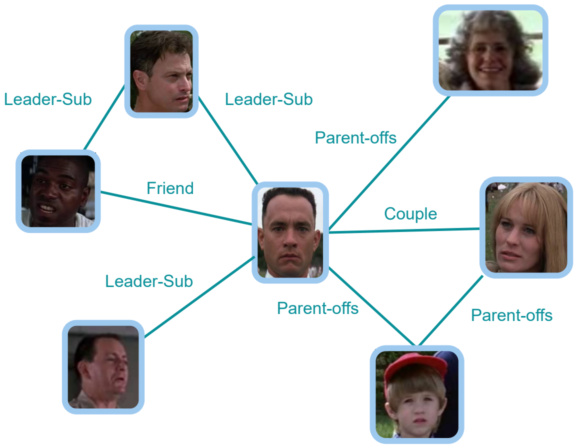


Figure 1: An example of the social network in the social relation recognition task.

In terms of data modeling in this task, previous works mainly utilized two types of methods to model the videos. Some methods (Teng et al., 2022; Dai et al., 2019; Kukleva et al., 2020; Xu et al., 2021; Teng et al., 2022) utilize visual and language encoders and perform reasoning on multimodal embeddings throughout the videos. Although these encoding and multimodal aggregation algorithms align with common patterns in Video Understanding tasks, the latent feature space is relatively too coarse for fine-grained and graph-structured social relations. Li et al. (2024) leverages LLMs for efficient multimodal understanding, achieving a breakthrough in this task. However, the exclusive application of LLMs for sequential processing is

*Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 4501–4520
November 4-9, 2025 ©2025 Association for Computational Linguistics

still inadequate for the graph-structured nature of social networks. (Dai et al., 2019; Liu et al., 2019; Wu et al., 2021; Yan et al., 2021; Wang et al., 2023; Lyu et al., 2024) model the scenes in the video as graphs. Yet, conventional graph operations are largely constrained to specific subgraphs due to limited hops. These methods also exhibit weak temporal reasoning capabilities due to inefficient temporal modeling, imposing several limitations on capturing complex temporal dynamics and long-range dependencies.

Large language models have succeeded across various fields in recent years, and Knowledge Graph has been regarded as an important graph-structured modeling method. As a result, there is growing interest in the integration of LLMs and KGs. Some methods (Wang et al., 2024; Huang et al., 2022; Jiang et al., 2023b; Choudhary and Reddy, 2023; Andrus et al., 2022; Luo et al., 2024, 2023; Chen et al., 2024a; Lee et al., 2024) render specific entities in the Knowledge Graph accessible to LLMs while resulting in a loss of contextual and structural features. Some methods (Chu et al., 2024; Cui et al., 2024; Shu et al., 2024b) for multimodal information extraction and mining exploit hypergraphs for KG processing. However, these structural operations are also constrained in subgraphs. In addition, summarizing the whole Knowledge Graph as proposed by Li et al. (2024) is also inappropriate, omitting critical details embedded within the graph structure.

Inspired by these observations, we propose a novel framework to synergize **m**ultimodal **t**emporal **K**nowledge **G**raphs and **L**arge **L**anguage **M**odels (mtKG-LLM) for social relation recognition. Firstly, we leverage Multimodal Large Language Models (MLLMs) to extract the multimodal information from each video scene, constructing spatial KGs. Secondly, we process each scene sequentially to update the temporal KGs along the timeline, which allows for continuous temporal reasoning of social relations. Thirdly, the temporal KGs are partitioned into communities for community-level contextual information. This approach preserves both global context and structural features within the KGs, facilitating multi-scale information retrieval for more reliable inference. Eventually, for each scene, we retrieve clues including detailed individual, interaction, and community-level contextual information. We utilize LLM without fine-tuning to recognize the social relations for characters based on available information.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to utilize the combination of KGs and LLMs for social relation recognition in videos.
- We propose the mtKG-LLM framework to construct multimodal temporal Knowledge Graphs and perform multi-scale information retrieval for low-coupling and effective synergy with Large Language Models.
- Our method achieves state-of-the-art (SOTA) performance with substantial improvements on major datasets.

2 Related Work

2.1 Social Relation Recognition

Over the years, researchers have explored diverse approaches to identify and classify social relations in videos. Lv et al. (2018) contributes to the SRIV dataset for further supervised learning and exploits TSN (Wang et al., 2016) for relation reasoning. Vicol et al. (2018) constructs the MovieGraphs dataset. Liu et al. (2019) proposes the ViSR dataset and MSTR model to extract relationships. Wu and Krahenbuhl (2021) constructs the LVU dataset for long-term video understanding. Some methods (Kukleva et al., 2020; Teng et al., 2022; Xu et al., 2021; Wu et al., 2021; Wang et al., 2023) utilize different modalities for more effective reasoning. Yet, these solutions fail to simulate the graph structure of social networks.

Subsequent works further explore the application of multimodal GNNs. Dai et al. (2019) proposes to fuse multimodal features via GNN. Yan et al. (2021) proposes MRR to predict via triples. Wu et al. (2021) models the video as heterogeneous graphs and implements multimodal processing. Wang et al. (2023) shifts graph operations for spatial relation modeling and performs temporal reasoning through a cumulative transformer. Lyu et al. (2024) proposes a hierarchical graph-based system for relation prediction. Although these GNN-based methods explicitly construct KGs, their spatial and temporal modeling capabilities are limited.

The studies of language models provide inspiration for Li et al. (2024) to introduce a vision foundation model for image-based social relation recognition. Nevertheless, the understanding and reasoning ability of this framework is not comparable with that of LLMs, and this framework only processes images without temporal information.

2.2 Multimodal and Temporal KGs

The development of knowledge graphs has undergone several key stages, evolving from early manually constructed ontologies to large-scale, automatically generated graphs of different types and forms (Jiang et al., 2023c). In recent years, the construction and inference of multimodal knowledge graphs (MMKGs) have enhanced the integration and understanding of diverse data modalities (Zhu et al., 2022). Addressing the dynamic nature of real-world data, Chen et al. (2023) designs the MSPT framework to balance existing knowledge with new information. Lee et al. (2024) employs a relation graph attention network and a cross-modal alignment module to enhance multimodal reasoning. Song et al. (2024) combines conventional knowledge engineering with LLMs to enrich scene understanding for real-world agents. These advancements highlights the growing importance in integrating and reasoning over complex multimodal data.

Temporal Knowledge Graphs (TKGs) have become a pivotal area of research, addressing the dynamic nature of real-world facts by incorporating temporal information into knowledge graphs. Lin et al. (2019) introduces an approach to temporal relation extraction using pretrained language models. Yan and Tang (2022) captures the temporal nature of events via KGs, allowing for the representation of chronological sequences, durations, and temporal overlaps. Liu et al. (2024c) introduce a dynamic hypergraph embedding approach for TKG reasoning, capturing high-order interactions between entities over time. These studies collectively advance the understanding and application of temporal dynamics in knowledge graphs, contributing to more accurate temporal inference models.

2.3 Unifying KGs and LLMs

Researchers have explored capabilities of LLMs in processing graph-structured data. KG related researches can be categorized into three types (Pan et al., 2024): KG-enhanced LLMs, LLMs-augmented KGs, and Synergized LLMs and KGs.

KG-enhanced LLMs mainly focus on exposing Knowledge Graph information to LLMs. Wang et al. (2024), Huang et al. (2022), Jiang et al. (2023b) and Andrus et al. (2022) propose to provide LLMs with entity retrieval. Choudhary and Reddy (2023) and Tang et al. (2024) further retrieve neighbor entities within 3-hops. Lee et al. (2024) encodes specific sub-graphs as input tokens of LLMs. In these methods, the receptive field of LLMs is restricted to sub-graphs, ignoring contextual information. Chu et al. (2024), Cui et al. (2024), and Shu et al. (2024b) construct hypergraphs for LLM-based KG processing, while GraphRAG (Edge et al., 2024) proposes a community-wise multi-scale information retrieval. These studies motivate us to implement contextual information extraction at a moderate granularity, i.e., community level.

As for LLMs-augmented KGs, existing methods mainly utilize LLMs to construct Knowledge Graphs of higher qualities. Chen et al. (2024a) models the queries and facts as triples based on LLMs. Xu et al. (2024) and Wei et al. (2024) optimizes the entity and relation descriptions using LLMs. Shu et al. (2024a) improves link prediction tasks by guiding the LLM with relevant graph-structured knowledge and fine-tuning LLMs on knowledge graph-related tasks. Additionally, Chakraborty (2024) demonstrates the strong potential of LLMs in multi-hop reasoning over knowledge graphs, enhancing complex information retrieval and understanding that covers multiple entities and relations.

Luo et al. (2023) and Chen et al. (2024a) explore efficient integrations of LLMs and KGs without fine-tuning. As the entities and relations of KGs are generated by LLMs, the representation spaces will be largely aligned. Along this way, leveraging LLMs to understand and process such information is rational, inspiring us to construct a low-coupling and effective integration of KGs and LLMs.

3 Preliminaries

This section defines various mathematical notations for relevant concepts and properties.

1. $V \in \{V_C, V_D, V_I, V_B, V_{DT}, V_{IT}, V_{CM}, V_S\}$, where V_C, V_D, V_I, V_B, V_{DT}, V_{IT}, V_{CM}, and V_S represent the set of character, individual, interaction, background, temporal individual, temporal interaction, community, and social relation entities respectively.

2. $V_C = \{v_C^1, v_C^2, ..., v_C^n\}$, where v_C^i represents the i^{th} character entity in the KG.

3. ${V_S}^s = \{v_S^{1,2}, v_S^{2,1}, ..., v_S^{n,m}\}$, where $v_S^{i,j}$ represents the social relation for v_C^i and v_C^j.

4. $E = \{(v^i, r^{i,j}, v^j)\}$. Each edge in the KG represents the association between entities.

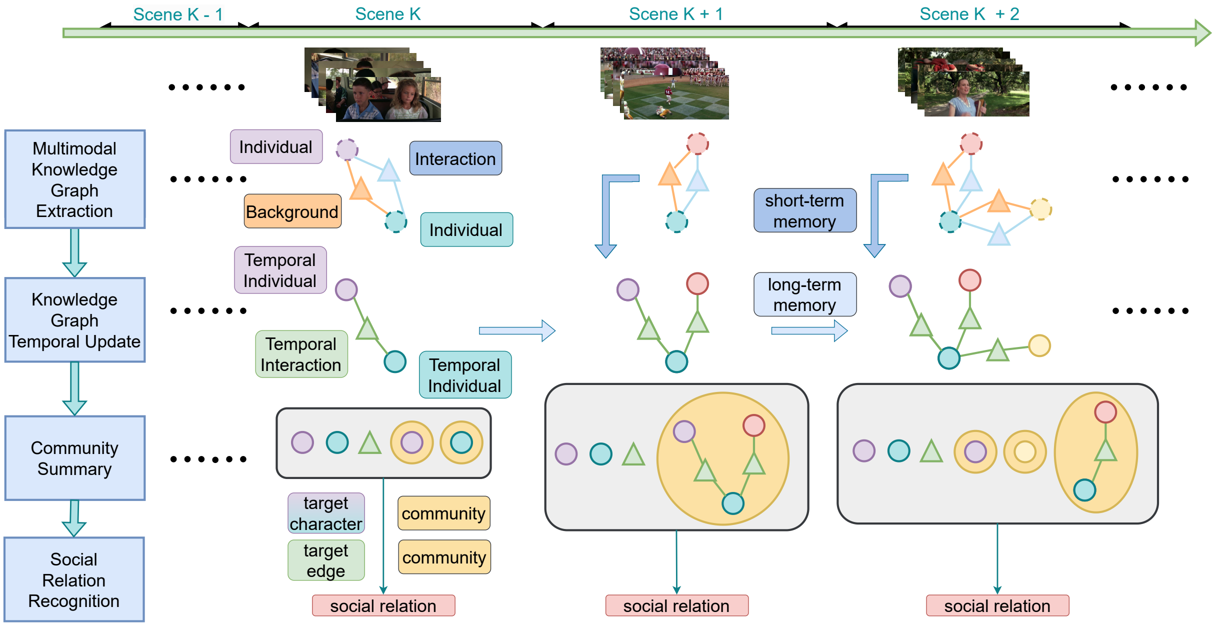


Figure 2: An overview of the mtKG-LLM framework. To provide a clearer representation, character entities, and temporal individual entities are integrated. In the visual depiction, dashed circles and solid circles represent V_D and V_{DT}. Additionally, V_I, V_B, and V_{IT} are triangles color-coded in blue, orange, and green, respectively.

5. $G = \{V, E\}$, where G is the Knowledge Graph with entities and edges.

It is worth noting that other definitions name the link in KGs as relations or relationships. To avoid confusion with social relations, we use the term edge as a substitute.

4 Method

4.1 Overview

Figure 2 demonstrates the overall framework of our proposed mtKG-LLM framework. Firstly, the Multimodal Knowledge Graph Extraction module models the scenes as multimodal spatial KGs. Secondly, the Knowledge Graph Temporal Update module constructs temporal KGs along the timeline. Thirdly, the Community Summary module summarizes communities in the temporal KGs. Finally, social relations are recognized via LLM with multi-scale KG information.

4.2 Multimodal Knowledge Graph Extraction

The Multimodal Knowledge Graph Extraction module constructs multimodal spatial KGs for each scene in the video. At the first step, we sample the frames within the scene as the visual content. The linguistic content consists of the conversations. We employ a Multimodal Large Language Model to directly extract the multimodal information, corresponding to the background entities V_B, individual entities V_D, and interaction entities V_I. Subsequently, we construct the spatial KG of each scene as shown in Figure 3, linking each $v_D^i \in V_D$ to corresponding $v_C^i \in V_C$, and each $v_I^{i,j} \in V_I$ to v_C^i and v_C^j. The background entity $v_D \in V_D$ is shared across all character pairs since all interactions occur in the same scene. The multimodal spatial KG construction is defined in Equation 1.

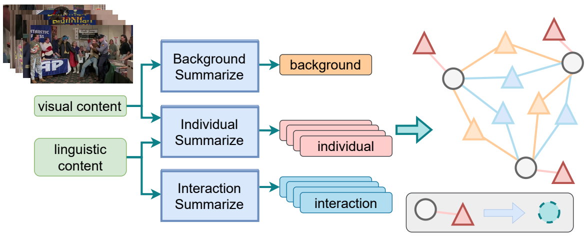


Figure 3: Multimodal Knowledge Graph Extraction. To provide a clearer representation, character entities, and individual entities are integrated.

$$V = \{V_C, V_D, V_I, V_B\} \tag{1}$$

It is worth noting that, our proposed framework mainly focuses on general videos that are not always covered by existing external knowledge sources. Due to this reason, instead of relying on exterior knowledge databases, our framework autonomously extracts relevant information. Although the datasets include movies that are famous

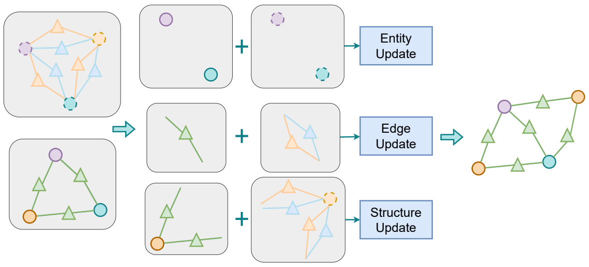


Figure 4: Knowledge Graph Temporal Update

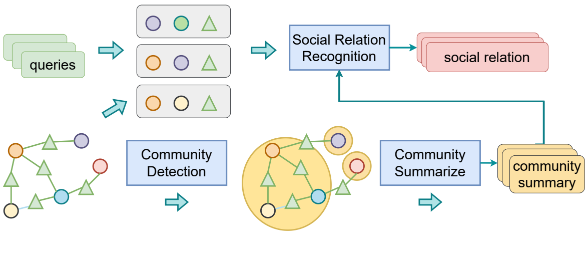


Figure 5: Community Summary and Social Relation Recognition.

(possibly in authorized KGs), we avoid using exterior validated knowledge to enhance the generalizability of our proposed method to general videos.

4.3 Knowledge Graph Temporal Update

The update procedure of the temporal KGs is illustrated in Figure 4. The KG is updated regarding entities and structures. Each temporal KG contains information from the previous temporal KG as long-term memory and the current spatial KG as short-term memory, shown in Figure 2. The spatial KG for the first scene is essentially a temporal KG due to the absence of prior information.

Firstly, for characters that appear in both the previous temporal KG and the current spatial KG, the previous temporal entities V_{DT} and V_{IT} and current entities V_D, V_I, and V_B are summarized as updated V_{DT} and V_{IT} via LLMs according to Equation 2 and Equation 3. These update operations endow temporal entities V_{DT} and V_{IT} with updated temporal features based on current information. Thirdly, the non-overlapping structures (i.e., entities and edges) that are not involved in the update process are preserved in the temporal KG. This step maintains previous and current structural knowledge as contextual information.

$$V_{DT} = Summ(V_{DT}, V_D) \tag{2}$$

$$V_{IT} = Summ(V_{IT}, V_I, V_B) \tag{3}$$

4.4 Community Summary

In this module, we summarize the temporal Knowledge Graph at the community level for global contextual information retrieval as shown in Figure 5. In light of the clustering nature of social relations, we partition the whole KG into communities to preserve both global features and structural information simultaneously. Within each community, character entities exhibit stronger links to other character entities in the same community. This is to maximize Modularity (Newman and Girvan, 2004), defined in Equation 4. e_c is the actual number of edges in the community. cK_c^2 is the fraction of the sum of the degrees of the entities, and m denotes the total number of edges in the network.

$$H = \frac{1}{2m} \sum (e_c - \gamma \frac{K_c^2}{2m}) \tag{4}$$

Compared to the LLM-generated weight introduced by Edge et al. (2024), we calculate the number of interactions between characters as the weight of the edge for community detection. With communities detected, we generate summaries for each community via LLMs based on the entities involved. The community summary information is denoted as Equation 5 where COM is the set of communities detected. Along this way, community-level summaries form the global context with structure-related information preserved.

$$V_{CM} = \{summ(V_{DT}^c, V_{DI}^c), c \in COM\} \tag{5}$$

4.5 Social Relation Recognition

With all KGs generated as above, we retrieve relevant information for social relation inference. Regarding queries of a specific relation, we retrieve the corresponding entities from V_{DT} and V_{IT} in the temporal KGs constructed in the Knowledge Graph Temporal Update module. In addition, we gather all community entities V_{CM} constructed in the Community Summary module as supplementary contextual information. These encompass multi-scale knowledge in the KG. Eventually, social relations are recognized via LLMs as demonstrated in Figure 5 based on the knowledge retrieved. The LLMs are not fine-tuned on the datasets to lower the coupling between KGs and LLMs for a low hardware resource setting. The recognition can be formulated as follows:

$$v_S^{i,j} = Rec(v_{DT}^i, v_{DT}^j, v_{IT}^{i,j}, V_{CM}) \qquad (6)$$

5 Experiments

5.1 Settings

5.1.1 Dataset

We evaluate our framework on the following four video datasets: MovieGraphs (Vicol et al., 2018) dataset, HLVU (Curtis et al., 2020) dataset, ViSR (Liu et al., 2019) dataset and LVU (Wu and Krahenbuhl, 2021) dataset. Details of each dataset are provided in Appendix A. It is worth noting that, in order to evaluate the temporal reasoning performance on MovieGraphs dataset, we split the samples in an ordered manner. We assign and shuffle clips that come from the same movie to the same split to avoid data leakage. In addition, clips are processed in order with respect to the timeline.

5.1.2 Evaluation Protocol

Our framework recognizes the social relation regarding the scene-level pair-wise relation queries and video-level single-label queries. We calculate the accuracy of queries in each category as well as the average accuracy for MovieGraphs dataset and the overall mean average precision (mAP) for ViSR and LVU datasets. To follow the conventional settings in DVU challenge, we evaluate the performance on the HLVU dataset via recall metrics.

5.1.3 Baseline Methods

For experiments on MovieGraphs dataset, baseline methods include GCN (Kipf and Welling, 2016), PGCN (Liu et al., 2019), MSTR (Liu et al., 2019), LIReC (Kukleva et al., 2020), MRR (Yan et al., 2021), PMFL (Teng et al., 2022), OD-GCN (Hu et al., 2023) and SGCAT-CT (Wang et al., 2023). HLVU baselines involve GCN (Kipf and Welling, 2016), Multimodal (Yu et al., 2020), Graph-based (Lu et al., 2020), Joint Learning (Zhang et al., 2021) and SGCAT-CT (Wang et al., 2023). ViSR baselines include GCN (Kipf and Welling, 2016), TSN (Wang et al., 2016), PGCN (Liu et al., 2019), MSTR (Liu et al., 2019), HC-GCN (Wu et al., 2021) and SGCAT-CT (Wang et al., 2023). LVU experiment is conducted on OT (Wu and Krahenbuhl, 2021), OT++ (Xiao et al., 2022), VideoBERT (Sun et al., 2019), STAN-Large (Fish et al., 2022) and SGCAT-CT (Wang et al., 2023). Details of the baselines are provided in Appendix B

5.1.4 Implementation Details

Videos and clips are decomposed into scenes according to the similarity of consecutive frames with a threshold of 0.6. Within scenes, we sample frames at a rate of 2 frames per second. We employ pre-trained Faster R-CNN (Girshick, 2015) and ResNet18 (He et al., 2016) for character detection. Conversations are extracted via available subtitle info or Netease Jianwai. We prompt the LLMs to generate major relations for major relation recognition tasks. To avoid the impact of different multimodal processing, we employ the same GPT-4o-2024-11-20 model (Hurst et al., 2024) for multimodal information extraction. The community detection algorithm is Leiden (Traag et al., 2019).

All LLM invocations are completed through APIs. The specific versions of the LLMs are as follows: GPT-4-0613 (Achiam et al., 2023), Claude-3-5-sonnet-20240620 (Anthropic, 2024), Gemini-1.5-pro-001 (Team et al., 2024), Llama3.1-405b (Dubey et al., 2024), Doubao-1.5-pro-32k (Bytedance, 2025), Qwen-max (Yang et al., 2024), DeepSeek-V3-0324 (Liu et al., 2024a). The unified prompt system is demonstrated in Appendix D.

5.2 Results and Analysis

Table 1 illustrates the experiment results of different methods on MovieGraphs dataset. Our framework with GPT4 as the LLM component exhibits superior performance to existing methods in recognizing most social relations while scoring the second highest accuracy in both "Leader-sub" and "Opponent" relations. The overall mean accuracy increases by 15% over the previous SOTA. Statistics provide strong evidence for the superiority of our proposed method.

Existing works struggle with recognizing similar relations such as "Parent-offs", "Sibling", "Couple" and "Friend". The accuracy varies significantly across these relations. To some extent, characters of these relations exhibit similar interactions and character traits in comparable contexts. However, determining clues for recognition are subtle and distributed along the timeline. Consequently, due to limited training data, methods that mainly employ coarse feature embeddings for reasoning fail to capture nuanced yet crucial details for distinguishing between social relations. In contrast, our LLM-powered multimodal KGs capture valuable clues from both history temporal features and current scene information, thus generating robust and dis-

Method	Top-1 Accuracy								
	Leader-sub	Colleague	Service	Parent-offs	Sibling	Couple	Friend	Opponent	Average
GCN	0.295	0.365	0.132	0.325	0.280	0.167	0.391	0.158	0.264
PGCN	0.313	0.374	0.290	0.137	0.320	0.250	0.407	0.375	0.308
MSTR	0.409	0.392	0.342	0.407	0.434	0.326	0.373	0.357	0.380
LIReC	0.352	0.329	0.244	0.435	0.301	0.423	0.317	0.269	0.334
MRR	0.454	0.423	0.428	0.385	0.392	0.446	0.439	0.365	0.417
PMFL	0.363	0.484	0.485	0.333	0.161	0.368	0.440	0.386	0.401
OD-GCN	**0.463**	0.394	0.442	0.415	0.457	0.436	0.428	0.423	0.432
SGCAT-CT	0.387	0.573	0.415	0.382	0.459	0.406	0.509	**0.570**	0.463
mtKG-GPT4	0.43	**0.656**	**0.534**	**0.547**	**0.554**	**0.519**	**0.576**	0.429	**0.531**

Table 1: Experiment results on MovieGraphs dataset. The best results are highlighted.

Method	Recall
GCN	0.261
Multimodal	0.178
Graph-based	0.258
Joint Learning	0.385
SGCAT-CT	0.457
mtKG-GPT4	**0.598**

Table 2: Experiment results on HLVU dataset.

Method	mAP
GCN	0.435
TSN-ST	0.432
PGCN	0.447
MSTR	0.478
HC-GCN	0.487
SGCAT-CT	0.501
mtKG-GPT4	**0.574**

Table 3: Experiment results on ViSR dataset.

LLM	Top-1 Accuracy		
	s	s+t	s+t+c
GPT4	0.487	0.504	0.531
Doubao	0.479	0.494	0.528
Claude	0.480	0.501	0.527
Llama	0.456	0.504	0.517
Qwen	0.476	0.485	0.499
DeepSeek	0.453	0.472	0.494
Gemini	0.469	0.477	0.480

Table 5: Ablation studies on MovieGraphs dataset.

Method	mAP
OB	0.531
OB++	0.524
VideoBERT	0.528
STAN-Large	**0.563**
SGCAT-CT	0.545
mtKG-GPT4	0.559

Table 4: Experiment results on LVU dataset.

criminative representations of the video. Therefore, our framework maintains a stable and outstanding recognition performance across similar relations.

Furthermore, the excellent performance of mtKG-LLM in these relations is also related to the clustering properties of relations. For instance, characters in "Colleagues" relations are normally working in the same place, while the company manager holds "Leader-sub" relations with all members of the colleague community. In addition, characters with "Parent-offs", "Sibling" and "Couple" relations constitute family scenarios. Thus, the community-level contextual information, which is comprised of both global and structural features, is crucial for recognizing such relations.

Contrary to being slightly behind for the "Learder-sub" relation, our method is outperformed by SGCAT by a significant amount for the "Opponent" relation. A potential explanation for this weakness is Covariance Shift. The definition of opponent is not as unified as other relations. In the vast training data of LLMs, the distribution of data containing the concept of "Opponent" deviates from the distribution of that in MovieGraphs dataset. Therefore, since we do not fine-tune our framework on the dataset, reasoning with LLMs potentially introduces bias in recognizing "Opponent" relations. Despite failing to take the leading position, our method still surpasses all other baseline methods, demonstrating its robustness.

The experiment on HLVU dataset further evaluates the capability of mtKG-LLM for pair-wise scene-level reasoning. As shown in Table 2, mtKG-LLM has improved the SOTA by 31%, validating the effectiveness of our proposed method. Additionally, Table 3 and Table 4 illustrate encouraging results of performing video-level major relation recognition. These results collectively reinforce

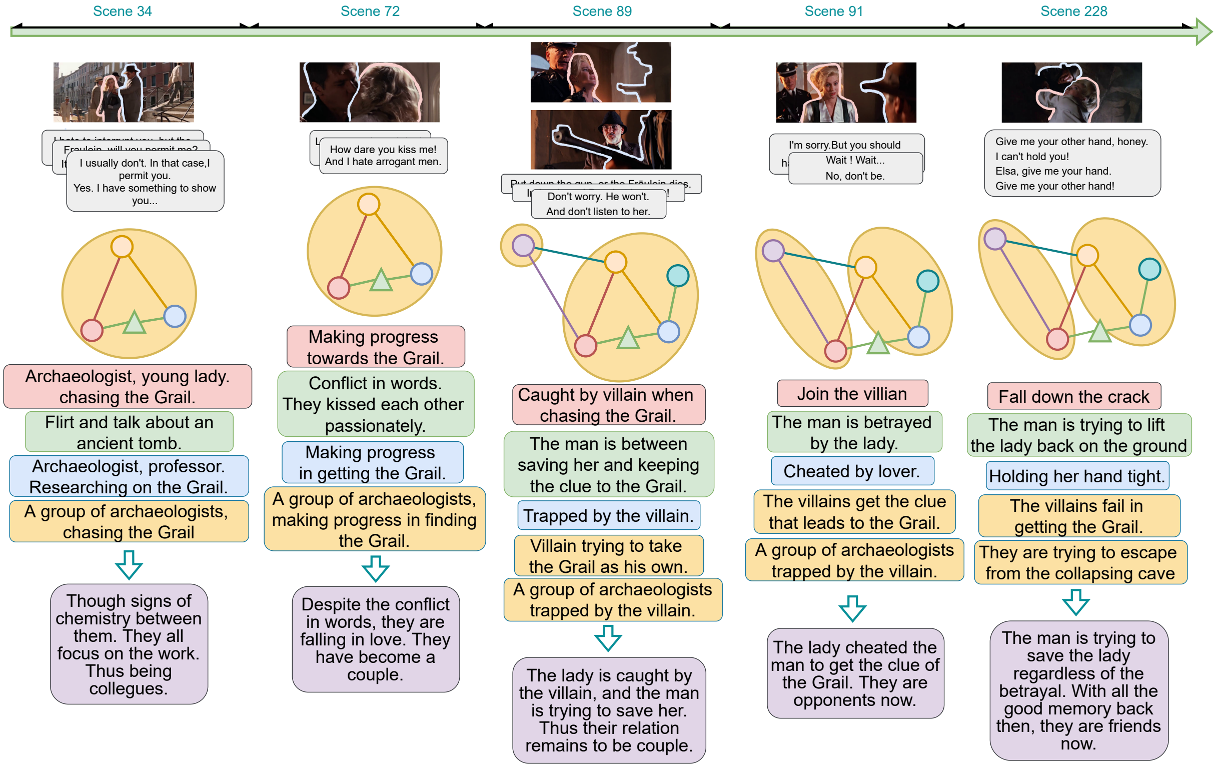

Scene 34
Scene 72
Scene 89
Scene 91
Scene 228
Fraulein, will you permit me?
I usually don't. In that case,I permit you.
Yes. I have something to show you...
How dare you kiss me!
And I hate arrogant men.
Put down the gun, or the Fräulein dies.
Don't worry. He won't.
And don't listen to her.
I'm sorry.But you should
Wait ! Wait...
No, don't be.
Give me your other hand, honey.
I can't hold you!
Elsa, give me your hand.
Give me your other hand!
Archaeologist, young lady. chasing the Grail.
Flirt and talk about an ancient tomb.
Archaeologist, professor. Researching on the Grail.
A group of archaeologists, chasing the Grail
Though signs of chemistry between them. They all focus on the work. Thus being collegues.
Making progress towards the Grail.
Conflict in words. They kissed each other passionately.
Making progress in getting the Grail.
A group of archaeologists, making progress in finding the Grail.
Despite the conflict in words, they are falling in love. They have become a couple.
Caught by villain when chasing the Grail.
The man is between saving her and keeping the clue to the Grail.
Trapped by the villain.
Villain trying to take the Grail as his own.
A group of archaeologists trapped by the villain.
The lady is caught by the villain, and the man is trying to save her. Thus their relation remains to be couple.
Join the villian
The man is betrayed by the lady.
Cheated by lover.
The villains get the clue that leads to the Grail.
A group of archaeologists trapped by the villain.
The lady cheated the man to get the clue of the Grail. They are opponents now.
Fall down the crack
The man is trying to lift the lady back on the ground
Holding her hand tight.
The villains fail in getting the Grail.
They are trying to escape from the collapsing cave
The man is trying to save the lady regardless of the betrayal. With all the good memory back then, they are friends now.

Figure 6: An example processing the evolution of relations. The target characters are denoted in red and blue temporal individual entities. The community entities are colored in yellow, covering relevant character entities. As for clear representations, we only depict relevant temporal interaction entities in green.

the superiority of our framework in handling complex social relation recognition tasks in videos.

5.3 Ablation Study

To verify the contribution of each part of the mtKG-LLM framework, we conduct an ablation study on MovieGraphs dataset. We divide the framework into three parts: spatial KG construction (s), temporal KG update (t), and community summary module (c). These modules are added incrementally to evaluate their individual impact. Owing to our low-coupling architecture, we are able to seamlessly switch between different LLMs, further demonstrating the framework's versatility and general capability in synergizing with various LLMs that demand no further adjustments.

As shown in Table 5, the addition of each module contributes to significant increases in the accuracy. The Knowledge Graph Temporal Update module enhances the framework by incorporating temporal dynamics for relations that require long-term memory. The Community Summary module further empowers the framework with both global and structural insights into the KGs. Combined with target entities retrieved from temporal KGs, multi-scale information is accessible for LLMs, facilitating more accurate reasoning. Notably, employing spatial KG modeling inclusively outperforms most methods, validating the effectiveness of the module in capturing multimodal features.

In addition, to further verify the universality of the proposed framework, we demonstrate the performance under resource-constrained settings and switch between alternative Multimodal LLMs.

We deploy a divers set of small-sized open-source models, including Vicuna-7B-v1.5 (Zheng et al., 2023), Qwen2.5-7B-Instruct (Team, 2024), Vicuna-13B-v1.5 (Zheng et al., 2023), Mistral-8B-Instruct-2410 (Jiang et al., 2023a), LLaMA-3.2-8B-Instruct (Dubey et al., 2024), and Qwen2.5-14B-Instruct (Team, 2024). The experimental results in Table 6 indicate that part of the LLMs underperform prior works. We attribute this to the general task solving nature of LLMs, which often trade off task-specific capabilities for broader task generalization. Additionally, prior models were specifically trained on the target dataset, giving them a distinct advantage. Nevertheless, the framework retains encouraging performance and significant flexibility. We view this as an opportunity for future work to incorporate fine-tuning or distillation techniques to improve both performance and efficiency in resource-constrained environments.

GPT4o + LLM	Top-1 Accuracy
Qwen2.5-7B-Instruct	0.403
vicuna-7b-v1.5	0.420
Llama-3.2-8B-Instruct	0.428
Ministral-8B-Instruct-2410	0.432
Qwen2.5-14B-Instruct	0.463
Vicuna-13B-v1.5	**0.468**

Table 6: Ablation Study Regarding LLMs on MovieGraphs dataset.

As for Multimodal LLM settings, we utilize different models including LLaVA-OneVision-72B (Liu et al., 2024b), InternVL2.5-78B (Chen et al., 2024b), GPT-4o (Hurst et al., 2024), Qwen2.5-VL-72B (Bai et al., 2025), Doubao-1.5-vison-pro (Guo et al., 2025) and Gemini-2-Pro (Google, 2024). As shown in Table7, altering Multimodal LLMs introduces performance variation. However, the model consistently outperforms prior baselines by a significant margin. This reveals the robustness of our approach.

MLLM + GPT4	Top-1 Accuracy
LLaVA-OneVision-72B	0.503
InternVL2.5-78B	0.529
GPT-4o	0.531
Qwen2.5-VL-72B	0.542
Doubao-1.5-vison-pro	0.548
Gemini-2-Pro	**0.599**

Table 7: Ablation Study Regarding MLLMs on MovieGraphs dataset.

5.4 Efficiency Analysis

As multiple LLM API calls are conducted throughout the video processing, we investigate the processing speed and computational cost of the proposed framework in comparison to previous neural network-based methods. According to the time consumption estimation in Appendix C, a 10-minute (600-second) video with about 5 scenes would roughly take 122.2 seconds for processing and require 5800 tokens. Time cost can be reduced to 12.2 seconds with a batch-size of 10 (greater batch-size may exceed the capacity of the API service). For comparison, we consider SGCAT-CT, a former state-of-the-art method that includes both visual and graph modules. On an RTX 3090 GPU with a batch size of 16, SGCAT-CT processes a 10-minute video in approximately 13 seconds.

Current approach is apparently more expensive than previous works, particularly for long videos. However, there is a trade-off between flexibility, performance, hardware requirements, cost, and inference delay. Our framework demonstrates capability and flexibility in understanding complex multimodal interactions. In addition, the use of LLMs offers greater transparency and interpretability compared to previous methods.

5.5 Case Study

We further conduct a case study depicted in Figure 6. The temporal KGs are displayed for each scene. The temporally updated entities and communities are summarized in boxes with corresponding colors. The reasoning for social relation recognition is attached at the bottom.

In the Indiana Jones movie, we focus on two characters, Elsa and Indiana Jones. Their social relations evolve from "Colleague" to "Couple", then from "Couple" to "Opponent", and finally from "Opponent" to "Friend". The reasoning process illustrated in the figure highlights the construction of temporal KGs upon spatial KGs, complemented by community summaries that provide contextual information. These modules facilitate a reliable inference of the transitions in social relations.

Cross-modal conflict is a typical issue as demonstrated in the second scene in Figure 6. Our proposed framework employs the powerful contextual understanding ability of LLMs, while leveraging both spatial and temporal information retrieved from KGs, to address the conflicting details that normally arise from a lack of context information.

6 Conclusion

In this paper, we propose the mtKG-LLM framework for recognizing social relations in videos. We implement spatial and temporal KGs via the Multimodal LLM to effectively capture the multimodal features both spatially and temporally. Additionally, we enhance LLM inference by incorporating a multi-scale Knowledge Graph information extraction mechanism, which expands the receptive field of LLM towards the entire Knowledge Graph. Comprehensive experiments demonstrate the effectiveness of our proposed framework.

Limitations

In our proposed framework, we adopt a non-invasive approach to synergize temporal KGs and LLMs. However, the low coupling pattern of the framework limits the depth of collaboration between KGs and LLMs to some degree. Thus, exploring possible strategies to optimize LLMs for efficient multi-scale Knowledge Graph processing is a direction for future work.

Deploying Large Language Models (LLMs) via APIs offers advantages in scalability and reduced infrastructure costs, as the service provider manages hardware and maintenance. However, this approach introduces network latency and ongoing usage expenses. Conversely, locally deploying traditional neural networks provides greater control over data privacy and can minimize latency, but it requires significant upfront investment in hardware and entails maintenance responsibilities.

Ethical Considerations

This research leverages the generalization capabilities of LLMs for social relation recognition tasks. However, we also acknowledge that LLMs potentially generate inaccurate results due to inherent biases in training data and evaluation metrics. Bias detection and debiasing algorithms are possible solutions to mitigate this problem.

Acknowledgments

This work is supported by the National Natural Science Foundations of China under Grant (62372060,61972047).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Berkeley R Andrus, Yeganeh Nasiri, Shilong Cui, Benjamin Cullen, and Nancy Fulda. 2022. Enhanced story comprehension for large language models through dynamic document-based knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10436–10444.

AI Anthropic. 2024. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3:6.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bytedance. 2025. Doubao-1.5-pro. https://console.volcengine.com/ark/model/detail?Id=doubao-1-5-pro-32k.

Abir Chakraborty. 2024. Multi-hop question answering over knowledge graphs using large language models. *arXiv preprint arXiv:2404.19234*.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, Ishaan Singh Rawal, Cheston Tan, Dongkyu Choi, Bo Xiong, and Bo Ai. 2024a. Llm-based multi-hop question answering with knowledge graph integration in evolving environments. *arXiv preprint arXiv:2408.15903*.

Xiang Chen, Jintian Zhang, Xiaohan Wang, Ningyu Zhang, Tongtong Wu, Yuxiang Wang, Yongheng Wang, and Huajun Chen. 2023. Continual multimodal knowledge graph construction. *arXiv preprint arXiv:2305.08698*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Nurendra Choudhary and Chandan K Reddy. 2023. Complex logical reasoning over knowledge graphs using large language models. *arXiv preprint arXiv:2305.01157*.

Zhixuan Chu, Yan Wang, Qing Cui, Longfei Li, Wenqing Chen, Zhan Qin, and Kui Ren. 2024. Llm-guided multi-view hypergraph learning for human-centric explainable recommendation. *arXiv preprint arXiv:2401.08217*.

Hejie Cui, Xinyu Fang, Ran Xu, Xuan Kan, Joyce C Ho, and Carl Yang. 2024. Multimodal fusion of ehr in structures and semantics: Integrating clinical records and notes with hypergraph and llm. *arXiv preprint arXiv:2403.08818*.

Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. Hlvu: A new challenge to test deep understanding of movies the way humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 355–361.

Pilin Dai, Jinna Lv, and Bin Wu. 2019. Two-stage model for social relationship understanding from videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1132–1137. IEEE.

Wenlong Dong, Qing Zhu, and Qirong Mao. 2025. Key clues guided video character social relationship recognition enhanced by llm. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Edward Fish, Jon Weinbren, and Andrew Gilbert. 2022. Two-stream transformer architecture for long video understanding. *arXiv preprint arXiv:2208.01753*.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Google. 2024. Gemini 2.0 pro. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Yibo Hu, Chenyu Cao, Fangtao Li, Chenghao Yan, Jinsheng Qi, and Bin Wu. 2023. Overall-distinctive gcn for social relation recognition on videos. In *International Conference on Multimedia Modeling*, pages 57–68. Springer.

Ningyuan Huang, Yash R Deshpande, Yibo Liu, Houda Alberts, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2022. Endowing language models with multimodal knowledge graph representations. *arXiv preprint arXiv:2206.13163*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Structgpt: A general framework for large language model to reason over structured data. *arXiv preprint arXiv:2305.09645*.

Xuhui Jiang, Chengjin Xu, Yinghan Shen, Xun Sun, Lumingyuan Tang, Saizhuo Wang, Zhongwu Chen, Yuanzhuo Wang, and Jian Guo. 2023c. On the evolution of knowledge graphs: A survey and perspective. *arXiv preprint arXiv:2310.04835*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. 2020. Learning interactions and relationships between movie characters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9849–9858.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. 2024. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*.

Wanhua Li, Zibin Meng, Jiawei Zhou, Donglai Wei, Chuang Gan, and Hanspeter Pfister. 2024. Socialgpt: Prompting llms for social relation reasoning via greedy segment optimization. *arXiv preprint arXiv:2410.21411*.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A bert-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 65–71.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. 2019. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3566–3574.

Xinyue Liu, Jianan Zhang, Chi Ma, Wenxin Liang, Bo Xu, and Linlin Zong. 2024c. Temporal knowledge graph reasoning with dynamic hypergraph embedding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15742–15751.

Yang Lu, Asri Rizki Yuliani, Keisuke Ishikawa, Ronaldo Prata Amorim, Roland Hartanto, Nakamasa Inoue, Kuniaki Uto, and Koichi Shinoda. 2020. Deep video understanding of character relationships in movies. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 120–129.

Linhao Luo, Jiaxin Ju, Bo Xiong, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Chatrule: Mining logical rules with large language models for knowledge graph reasoning. *arXiv preprint arXiv:2309.01538*.

Linhao Luo, Zicheng Zhao, Chen Gong, Gholamreza Haffari, and Shirui Pan. 2024. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. *arXiv preprint arXiv:2410.13080*.

Jinna Lv, Wu Liu, Lili Zhou, Bin Wu, and Huadong Ma. 2018. Multi-stream fusion model for social relation recognition from videos. In *MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24*, pages 355–368. Springer.

Yuanjie Lyu, Penggang Qin, Tong Xu, Chen Zhu, and Enhong Chen. 2024. Interactnet: Social interaction recognition for semantic-rich videos. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Mark EJ Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Penggang Qin, Shiwei Wu, Tong Xu, Yanbin Hao, Fuli Feng, Chen Zhu, and Enhong Chen. 2023. When i fall in love: Capturing video-oriented social relationship evolution via attentive gnn. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):5160–5175.

Dong Shu, Tianle Chen, Mingyu Jin, Chong Zhang, Mengnan Du, and Yongfeng Zhang. 2024a. Knowledge graph large language model (kg-llm) for link prediction. *arXiv preprint arXiv:2403.07311*.

Hongji Shu, Chaojun Meng, Pasquale De Meo, Qing Wang, and Jia Zhu. 2024b. Self-supervised hypergraph learning for enhanced multimodal representation. *IEEE Access*.

Yaoxian Song, Penglei Sun, Haoyu Liu, Zhixu Li, Wei Song, Yanghua Xiao, and Xiaofang Zhou. 2024. Scene-driven multimodal knowledge graph construction for embodied ai. *IEEE Transactions on Knowledge and Data Engineering*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.

Siyang Sun, Xiong Xiong, and Yun Zheng. 2022. Two stage multi-modal modeling for video interaction analysis in deep video understanding challenge. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7040–7044.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2024. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–500.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yiyang Teng, Chenguang Song, and Bin Wu. 2022. Learning social relationship from videos via pretrained multimodal transformer. *IEEE Signal Processing Letters*, 29:1377–1381.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. 2018. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8581–8590.

Haorui Wang, Yibo Hu, Yangfu Zhu, Jinsheng Qi, and Bin Wu. 2023. Shifted gcn-gat and cumulative-transformer based social relation recognition for long videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 67–76.

Jiapu Wang, Kai Sun, Linhao Luo, Wei Wei, Yongli Hu, Alan Wee-Chung Liew, Shirui Pan, and Baocai Yin. 2024. Large language models-guided dynamic adaptation for temporal knowledge graph reasoning. *arXiv preprint arXiv:2405.14170*.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 806–815.

Chao-Yuan Wu and Philipp Krahenbuhl. 2021. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894.

Shiwei Wu, Joya Chen, Tong Xu, Liyi Chen, Lingfei Wu, Yao Hu, and Enhong Chen. 2021. Linking the characters: Video-oriented social graph generation via hierarchical-cumulative gcn. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4716–4724.

Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. 2022. Hierarchical self-supervised representation learning for movie understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9727–9736.

Derong Xu, Ziheng Zhang, Zhenxi Lin, Xian Wu, Zhihong Zhu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024. Multi-perspective improvement of knowledge graph completion with large language models. *arXiv preprint arXiv:2403.01972*.

Tong Xu, Peilun Zhou, Linkang Hu, Xiangnan He, Yao Hu, and Enhong Chen. 2021. Socializing the videos: A multimodal approach for social relation recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1):1–23.

Chenghao Yan, Zihe Liu, Fangtao Li, Chenyu Cao, Zheng Wang, and Bin Wu. 2021. Social relation analysis from videos via multi-entity reasoning. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 358–366.

Zhihua Yan and Xijin Tang. 2022. Hierarchical storyline generation based on event-centric temporal knowledge graph. In *International Symposium on Knowledge and Systems Sciences*, pages 149–159. Springer.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

YouTube. 2020. Movieclips. https://www.movieclips.com/.

Fan Yu, DanDan Wang, Beibei Zhang, and Tongwei Ren. 2020. Deep relationship analysis in video with multimodal feature fusion. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4640–4644.

yWorks. 2019. yed. https://www.yworks.com/products/yed.

Beibei Zhang, Fan Yu, Yanxin Gao, Tongwei Ren, and Gangshan Wu. 2021. Joint learning for relationship and interaction analysis in video with multimodal feature fusion. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4848–4852.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):715–735.

A Additional Introduction of Datasets

The details of the datasets are as follows:

MovieGraphs dataset is comprised of 7,637 clips from 51 movies collected from IMDB for long-term reasoning. A group of annotators went through a training phase and a cross-checking phase to ensure high-quality annotations. It includes multimodal annotations such as visual scenes, character interactions, dialogue transcripts, and situational attributes (e.g., emotions, relationships). We follow the same label set as in prior work (Yan et al., 2021) for pair-wise scene-level annotations. This dataset provides the most comprehensive evaluation of our method given its fine granularity and rich multimodal context.

HLVU dataset consists of 19 movies from public websites such as Vimeo and the Internet Archive. Movies are divided into scene segments. A group of human assessors were recruited to provide scene-level pair-wise labels via yEd (yWorks, 2019) graphing tools. It features multimodal inputs, including video frames, and subtitles, supporting holistic video understanding tasks.

ViSR dataset contains more than 8,000 videos from 200 movies which have a wide variety of types such as adventure, family, comedy, drama, and crime. Each video clip is labeled by at least five annotators by maximum voting to guarantee the quality. It provides video-level social relation labels along with aligned audio that contains subtitles, enabling joint vision-language reasoning.

LVU is a long video understanding dataset constructed on the publicly available MovieClips dataset (YouTube, 2020), containing 30K videos from 3K movies. It offers multimodal annotations from the description associated with each video, such as character identities, dialogue, and scene descriptions, facilitating large-scale video-language analysis.

MovieGraphs and HLVU datasets are annotated at character-pair level. ViSR and LVU datasets (ViSR and LVU) primarily label major social relations at the video level, leveraging multimodal signals (visual, textual, and auditory) for comprehensive video understanding.

B Additional Introduction of Baselines

Existing works recognize social relations in different manners. We divide them into two groups, scene-level pair-wise and video-level single-label. We select approaches that are suitable or can be modified to be suitable for corresponding objectives. Some methods cannot be applied to other datasets due to several reasons. Firstly, methods (Kukleva et al., 2020; Yan et al., 2021; Teng et al., 2022) with limited receptive fields cannot fully process video-level datasets. In addition, approaches (Wu and Krahenbuhl, 2021; Yu et al., 2020; Zhang et al., 2021; Sun et al., 2022; Xiao et al., 2022; Fish et al., 2022) for joint video understanding tasks cannot process other datasets without video understanding annotations. Besides, some video-level methods (Wu et al., 2021) performing video-level aggregations cannot process scene-level data. Furthermore, works (Xu et al., 2021; Lyu et al., 2024) with different category mappings or other settings are not involved. Due to the reason that we modify the structure of MovieGraphs dataset for better temporal reasoning and evaluation, some methods (Qin et al., 2023; Dong et al., 2025; Lyu et al., 2024) that either did not released the source code or replied to contacts will not be included in the experiment.

As for scene-level pair-wise comparisons, we conduct experiments on the MovieGraphs and HLVU datasets. GCN (Kipf and Welling, 2016) is a conventional graph convolution network. PGCN (Liu et al., 2019) adopts multi-scale graph modeling for social relation recognition. MSTR (Liu et al., 2019) combines PGCN and TSN (Wang et al., 2016) for spatial and temporal information aggregation. LIReC (Kukleva et al., 2020) presents a multimodal architecture to extract contextual features. MRR (Yan et al., 2021) proposes a multi-entity representation method to predict relationships. PMFL (Teng et al., 2022) adopts self-supervised learning on transformers for better representation. Multimodal (Yu et al., 2020) proposes a multimodal fusion method while Graph-based (Lu et al., 2020) further adopts GNNs for data modeling. Joint Learning (Zhang et al., 2021) applies joint learning of interaction and relationship. OD-GCN models the videos at overall-level, distinctive-level and global-level, enhancing multi-perspective representation. SGCAT (Wang et al., 2023) converts the conventional character-centric modeling to novel relation-centric modeling and employs a cumulative transformer for temporal reasoning.

As for video-level single-label comparisons, we conduct experiments on the ViSR and LVU datasets. Apart from the previously mentioned methods, HC-GCN (Wu et al., 2021) constructs social graphs at different levels and performs aggregations. Object Transformers (OT) (Wu and Krahenbuhl, 2021) adopts an object-centric design for video understanding. OT++ (Xiao et al., 2022) supplements (Wu and Krahenbuhl, 2021) with a self-supervised pre-trained contextualizer. VideoBERT (Sun et al., 2019) learns joint distributions over visual and linguistic data. STAN-Large (Fish et al., 2022) designs a two-stream transformer architecture to model static and contextual features.

C Efficiency Estimation

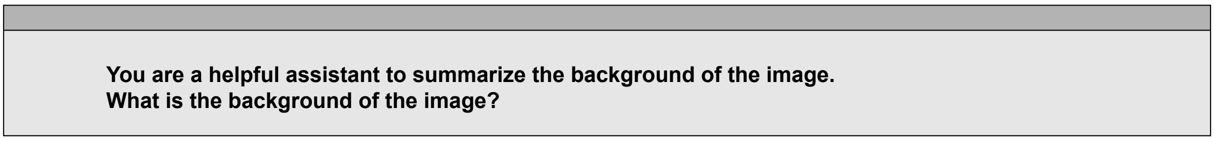
To estimate the delay and token usage of our framework, we consider a video for L seconds and average token cost at each operation recorded during experiments. Scenes typically last about 2 minutes (120 seconds), so for a video of length L seconds, the number of scenes is approximately $\frac{L}{120}$. For each scene, the multimodal LLM call to construct the knowledge graph uses 800 tokens with 22 seconds' delay. Temporal update operations between scenes consume about 300 tokens each with 2 seconds' delay, and about 400 tokens for community summary with 2.4 seconds' delay. The final LLM call for social relation recognition requires roughly 200 tokens with 1.8 seconds' delay. Thus, the total estimated token usage is approximately $\frac{L}{120}\times 800+(\frac{L}{120}-1)\times 300+400+200$ tokens. And the delay is about $\frac{L}{120}\times 22+(\frac{L}{120}-1)\times 2+2.4+1.8$ seconds.

In our framework, we employ a combination of different LLM API services (both language-only and multimodal), and the inference delay can vary depending on the network environment, geographic location, and backend load. Therefore, the reported latency values are intended for qualitative analysis only.

D LLM Prompts

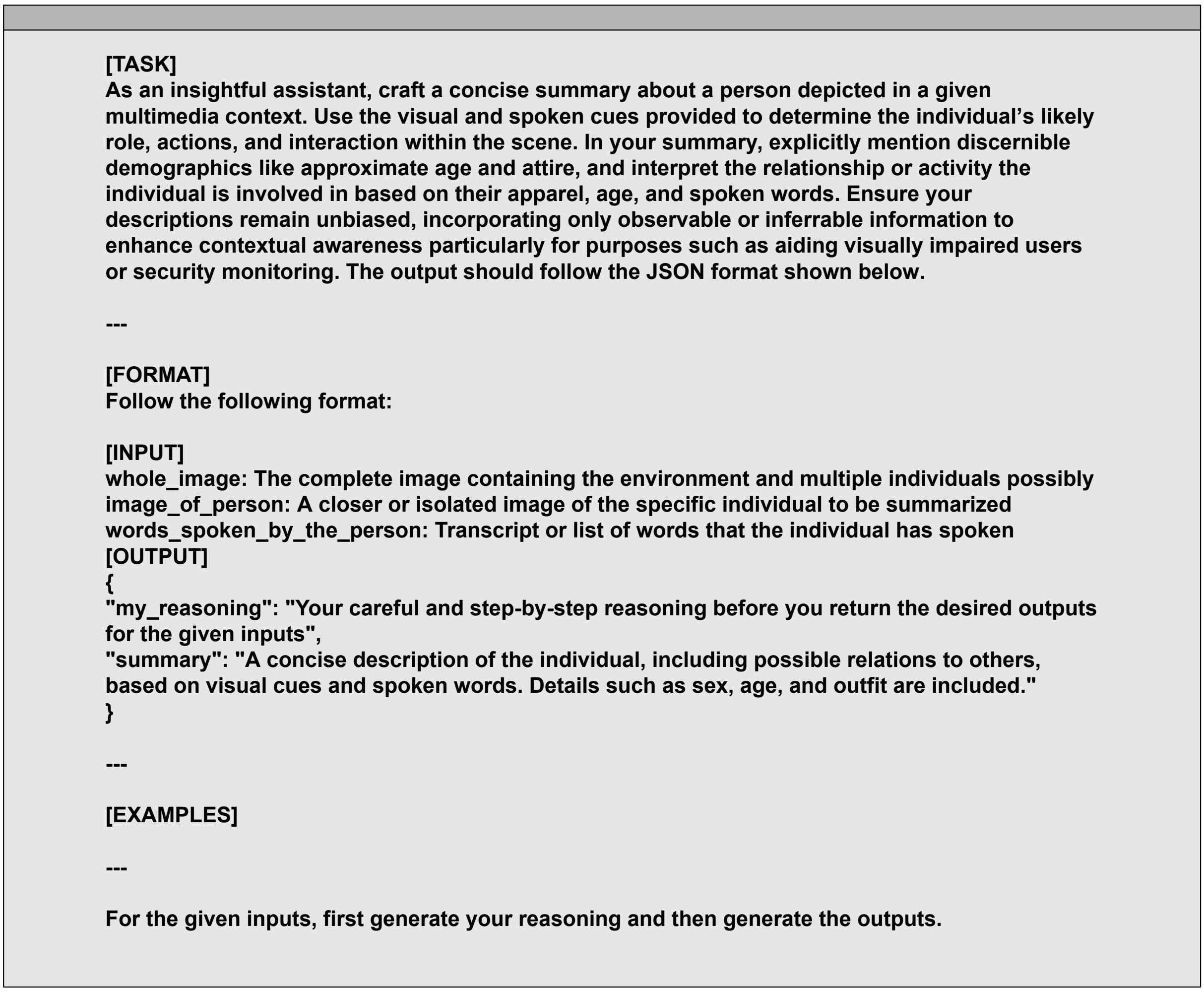
We use a unified prompt system for various operations implemented via LLMs. Due to limited space, the examples involved in the prompts are omitted and will be accessible in the released code after acceptance. The prompts of each operation are shown below.

The background summary prompt is shown in Figure 7. The individual summary prompt is shown in Figure 8. The interaction summary prompt is shown in Figure 9. The temporal update prompt is shown in Figure 10. The community summary prompt is shown in Figure 11. The social relation recognition prompt is shown in Figure 12.



You are a helpful assistant to summarize the background of the image.
What is the background of the image?

Figure 7: Prompt for background information extraction.



[TASK]
As an insightful assistant, craft a concise summary about a person depicted in a given multimedia context. Use the visual and spoken cues provided to determine the individual’s likely role, actions, and interaction within the scene. In your summary, explicitly mention discernible demographics like approximate age and attire, and interpret the relationship or activity the individual is involved in based on their apparel, age, and spoken words. Ensure your descriptions remain unbiased, incorporating only observable or inferrable information to enhance contextual awareness particularly for purposes such as aiding visually impaired users or security monitoring. The output should follow the JSON format shown below.

[FORMAT]
Follow the following format:

[INPUT]
whole_image: The complete image containing the environment and multiple individuals possibly
image_of_person: A closer or isolated image of the specific individual to be summarized
words_spoken_by_the_person: Transcript or list of words that the individual has spoken
[OUTPUT]
{
"my_reasoning": "Your careful and step-by-step reasoning before you return the desired outputs for the given inputs",
"summary": "A concise description of the individual, including possible relations to others, based on visual cues and spoken words. Details such as sex, age, and outfit are included."
}

[EXAMPLES]

For the given inputs, first generate your reasoning and then generate the outputs.

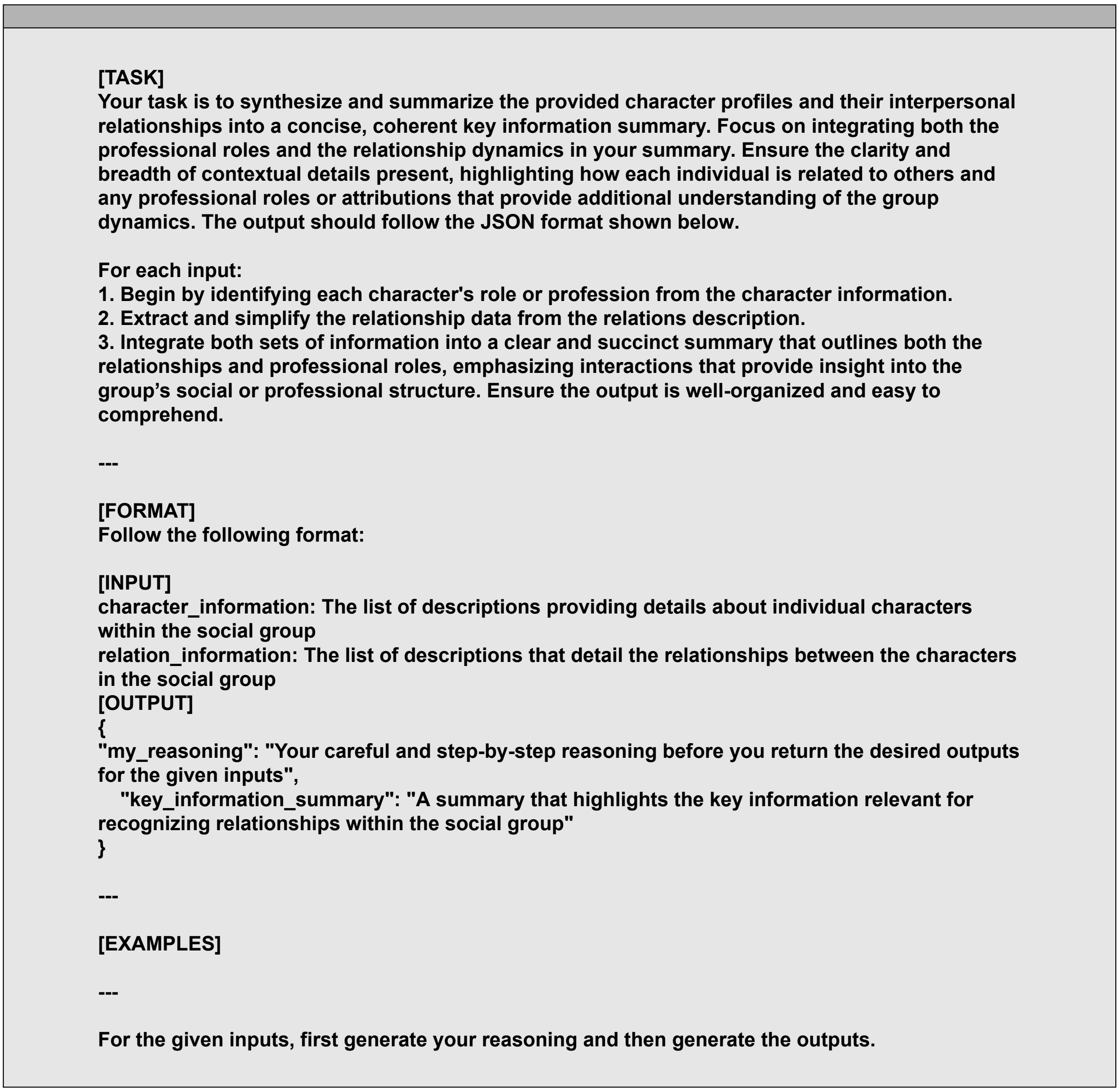
Figure 8: Prompt for individual information extraction.

```
[TASK]
As a perceptive assistant, you are expected to synthesize both visual and textual information to provide a concise and insightful summary of the interaction between two people depicted in the dataset. Evaluate the setting, the appearance and positioning of individuals, and their conversation to deduce their relationship, emotional state, and the purpose of their interaction. Use context clues from the whole image, detailed images of each person, and the conversation snippet provided. In the absence of a piece of information, use informed assumptions based on available data to complete your summary effectively. Make sure your summary encapsulates the essence of their interaction, elucidating on the underlying dynamics, possible intentions, and roles of the people involved. Your narrative should be brief yet comprehensive, providing clear insight into the nature of their interaction. The output should follow the JSON format below.

---

[FORMAT]
Follow the following format:

[INPUT]
whole_image: the entire image showing both people
image_of_first_person: image focusing on the first person
image_of_second_person: image focusing on the second person
conversation: recorded or text-based conversation between the two people
[OUTPUT]
{
"my_reasoning": "Your careful and step-by-step reasoning before you return the desired outputs for the given inputs",
"interaction_summary": "summary of the interaction between the two people based on the given images and conversation"
}

---

[EXAMPLES]

---

For the given inputs, first generate your reasoning and then generate the outputs.
```

Figure 9: Prompt for interaction information extraction.

```
[TASK]
As a narrative synthesis specialist, your task is to create cohesive summaries that seamlessly
bridge past conditions with current updates, focusing on progression and resolutions. For each
input:

1. Analyze the historical and current information provided.
2. Combine these details to illustrate a clear transition from the initial problem to the subsequent
improvements or solutions implemented.
3. Write a summary that integrates both the history and current updates, emphasizing the
effectiveness and results of the actions taken.
4. Provide your reasoning that led to this summary. Your narrative should serve as a benchmark
for clear communication in fields requiring regular updates, such as customer relationships,
academic research, and healthcare.

Each response should aim to engage and inform stakeholders by clearly demonstrating how
past challenges have been addressed, ensuring the narrative is valuable for real-world
applications and decision-making. Ensure your outputs adhere to the JSON structure provided,
featuring both your reasoning and the synthesized history. Your output should follow the JSON
format shown below.

---

[FORMAT]

[INPUT]
History_information: Previous contextual data provided by an expert
Current_information: Latest data or updates provided by an expert
[OUTPUT]
{
  "my_reasoning": "Careful and step-by-step reasoning before returning the outputs for the
given inputs"
  "New_history_information": "A summary that combines both the historical and current
information while emphasizing progress, maintaining chronological integrity and relevance for
decision-making"
}

---

[EXAMPLES]

---

For the given inputs, first generate your reasoning and then generate the outputs.
```

Figure 10: Prompt for temporal KG update.



[TASK]
Your task is to synthesize and summarize the provided character profiles and their interpersonal relationships into a concise, coherent key information summary. Focus on integrating both the professional roles and the relationship dynamics in your summary. Ensure the clarity and breadth of contextual details present, highlighting how each individual is related to others and any professional roles or attributions that provide additional understanding of the group dynamics. The output should follow the JSON format shown below.

For each input:
1. Begin by identifying each character's role or profession from the character information.
2. Extract and simplify the relationship data from the relations description.
3. Integrate both sets of information into a clear and succinct summary that outlines both the relationships and professional roles, emphasizing interactions that provide insight into the group's social or professional structure. Ensure the output is well-organized and easy to comprehend.

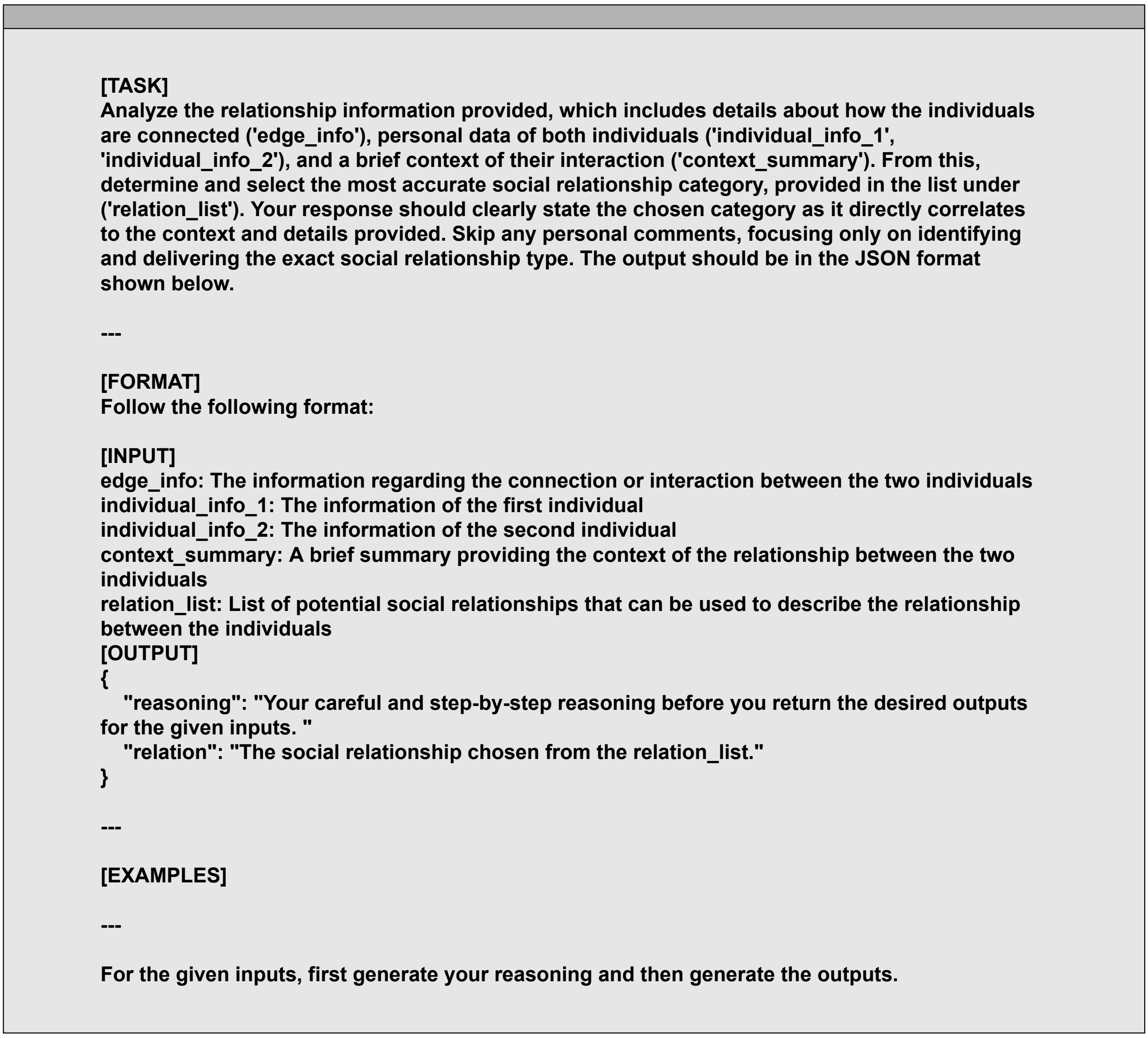
[FORMAT]
Follow the following format:

[INPUT]
character_information: The list of descriptions providing details about individual characters within the social group
relation_information: The list of descriptions that detail the relationships between the characters in the social group
[OUTPUT]
```
{
"my_reasoning": "Your careful and step-by-step reasoning before you return the desired outputs for the given inputs",
    "key_information_summary": "A summary that highlights the key information relevant for recognizing relationships within the social group"
}
```

[EXAMPLES]

For the given inputs, first generate your reasoning and then generate the outputs.

Figure 11: Prompt for community summary extraction.



```
[TASK]
Analyze the relationship information provided, which includes details about how the individuals
are connected ('edge_info'), personal data of both individuals ('individual_info_1',
'individual_info_2'), and a brief context of their interaction ('context_summary'). From this,
determine and select the most accurate social relationship category, provided in the list under
('relation_list'). Your response should clearly state the chosen category as it directly correlates
to the context and details provided. Skip any personal comments, focusing only on identifying
and delivering the exact social relationship type. The output should be in the JSON format
shown below.

---

[FORMAT]
Follow the following format:

[INPUT]
edge_info: The information regarding the connection or interaction between the two individuals
individual_info_1: The information of the first individual
individual_info_2: The information of the second individual
context_summary: A brief summary providing the context of the relationship between the two
individuals
relation_list: List of potential social relationships that can be used to describe the relationship
between the individuals
[OUTPUT]
{
   "reasoning": "Your careful and step-by-step reasoning before you return the desired outputs
for the given inputs. "
   "relation": "The social relationship chosen from the relation_list."
}

---

[EXAMPLES]

---

For the given inputs, first generate your reasoning and then generate the outputs.
```

Figure 12: Prompt for social relation recognition.