

Aligning Vision to Language: Annotation-Free Multimodal Knowledge Graph Construction for Enhanced LLMs Reasoning

Supplementary Material

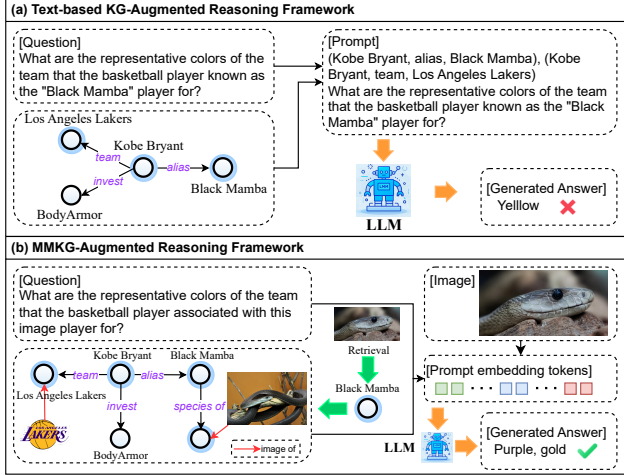


Figure 1. (a) The limited information contained in text-based KGs leads to inaccurate responses. (b) Leveraging MMKGs enables reasoning with enriched multimodal information to produce the correct answer.

A. Cross-Modal Reasoning Failures in Textual KGs

Multimodal learning, by virtue of its capability to synergistically integrate heterogeneous data modalities, establishes a comprehensive knowledge acquisition paradigm that significantly enhances reasoning robustness [3]. This principle extends to Multimodal Knowledge Graphs (MMKGs), where the semantic symbiosis between visual and textual modalities addresses the critical limitation of modal isolation inherent in conventional text-based KGs. As empirically demonstrated in Figure 1, pure textual KGs often induce hallucinated or incomplete responses due to their inability to resolve visual-textual semantic ambiguities. For instance, when queried about fine-grained visual attributes (e.g., spatial relationships or object properties absent in textual metadata), LLMs grounded solely on textual KG triples frequently generate plausible but factually inconsistent answers, as they lack access to cross-modal referential grounding. In contrast, MMKGs bridge this gap through bidirectional visual-textual entity linking, enabling LLMs to retrieve and reason over fused evidence from both modalities. Our qualitative analysis of the case in Figure 1 reveals that the multimodal reasoning path—leveraging both image-derived entities and textual relationships—is essential for deriving logically coherent and factually accurate conclusions.




Tweet posts			
	teachers take on top of Mount Sherman.	Sony announced a Bad Boys in the next few years.	Jackson is really my favorite.
Expected NER results	teachers take on top of [Mount Sherman LOC].	[Sony ORG] announced a [Bad Boys OTHER] in the next few years.	[Jackson PER] is really my favorite.
NER with text only	teachers take on top of [Mount Sherman OTHER].	[Sony ORG] announced a [Bad Boys PER] in the next few years.	[Jackson OTHER] is really my favorite.
MNER with previous methods	teachers take on top of [Mount Sherman PER].	[Sony PER] announced a [Bad Boys PER] in the next few years.	[Jackson OTHER] is really my favorite.

Figure 2. Three example social media posts with labelled named entities [1].

Type	#Chains	Mentions/Chain	Boxes/Chain
people	59766	3.17	1.95
clothing	42380	1.76	1.44
body parts	12809	1.50	1.42
animals	5086	3.63	1.44
vehicles	5561	2.77	1.21
instruments	1827	2.85	1.61
scene	46919	2.03	0.62
other	82098	1.94	1.04
total	244035	2.10	1.13

Table 1. Coreference chain statistics of Flickr30K-Entity. The number of mentions per chain indicates how salient an entity is. The number of boxes per chain indicates how many distinct entities it refers to.

B. Case Studies on Manual Annotation Overheads

The development of robust entity extraction models typically hinges on large-scale annotated corpora, yet the generalizability of these models remains intrinsically bounded by the semantic scope and granularity of their training datasets. Widely-adopted benchmarks such as Flickr30K-Entity [6] exemplify this constraint: while serving as de facto standards for evaluating visual-linguistic entity grounding, their construction necessitates labor-intensive manual annotations at scale. As illustrated in Figure 2, even high-quality annotations in such datasets often adopt a minimalist tagging paradigm—identifying only coarse-grained entities while neglecting fine-grained attributes and contextual relationships. This sparsity of semantic enrichment directly propagates to trained models, which consequently fail to capture the compositional semantics necessary for complex

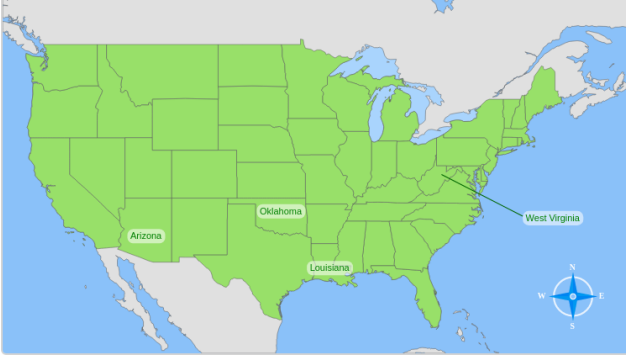


Figure 3. An example from the ScienceQA benchmark [5], illustrating multimodal question-answering scenarios that necessitate joint reasoning over textual prompts and visual evidence.

reasoning scenarios.

C. Case Studies on Visual Specificity Deficits in VLM-Generated Captions

As exemplified in Figure 3, vision-language models like BLIP-2 [4] tend to produce oversimplified textual descriptions that critically lack actionable visual-semantic signals. The VLM-generated caption (“A map of the united states with the location of the united states”) merely identifies coarse-grained scene semantics, failing to capture object-level attributes (color coding of regions), spatial relationships (border adjacency between Arizona and Mexico) and compositional context (compass orientation in lower-right corner). In contrast, human annotations (“This is a map of the United States. The main part of the country is shown in green, with several states labeled. Arizona is in the south-western part of the US, bordering Mexico. Oklahoma is in the central - southern region. Louisiana is located along the Gulf of Mexico in the southeastern part. West Virginia is in the eastern part of the country. There’s also a compass in the bottom - right corner to show directions.”) demonstrate essential characteristics for multimodal reasoning.

D. Retrieval Strategy in MMKG Construction

We adopt retrieval strategies based on the framework provided by LightRAG [2], which supports multiple modes:

- **local**: focuses on context-dependent information;
 - **global**: utilizes global knowledge;
 - **hybrid**: combines local and global retrieval methods;
 - **naive**: performs basic search without advanced techniques;
 - **mix**: integrates knowledge graph and vector retrieval;
- In our implementation, we rely on the **hybrid** retrieval mode, which balances the precision of local cues with the breadth of global knowledge. This strategy improves the relevance and completeness of retrieved information, which is crucial for high-quality MMKG construction.

Algorithm 1 MMKG Generation

Require: \hat{S} (refined description), T (external knowledge, optional)

Ensure: $\mathcal{G} = (\mathcal{E}, \mathcal{R})$ (knowledge graph)

- 1: $\mathcal{T} \leftarrow \hat{S} \oplus T$ ▷ Concatenate \hat{S} and T
- 2: $\mathcal{G} \leftarrow \text{LightRAG}(\mathcal{T})$ ▷ Generate graph via LightRAG
- 3: $(\mathcal{E}, \mathcal{R}) \leftarrow f_{\text{ERE}}(\mathcal{T})$ ▷ Extract entities and relations
- 4: **return** $\mathcal{G} = \{(h, r, t) \mid h, t \in \mathcal{E}, r \in \mathcal{R}\}$

LightRAG is an excellent project that effectively supports automatic MMKG construction, and its retrieval design plays a central role in our framework. Specifically, LightRAG introduces keyword-guided text chunking to expand the retrievable context. By leveraging both high-level and low-level keywords in combination with chunk-level vector retrieval, it enables more comprehensive knowledge access. In addition, the choice of the retrieval model is also important. Larger LLMs have slower retrieval speeds but better performance. In this experiment, we used Qwen2.5-7B for retrieval. We also tested the retrieval performance of 32B and 72B models, which showed a 1%-5% improvement in performance, but it also significantly increased the graph construction time. Therefore, we finally adopted a lightweight retrieval model. The details of the entire LightRAG are shown in Algorithm 1.

E. Selection of Sensitivity Threshold τ

We select the sensitivity threshold τ empirically based on performance on the validation set. In practice, τ can be approximately determined by observing the token length distribution of captions: datasets with richer visual content and longer captions tend to benefit from a lower τ , while simpler datasets can tolerate a higher τ . This provides a practical way to adjust τ without extensive tuning.

In addition, we notice a key pattern when analyzing the relevance scores across windows. Around certain values of τ , the scores tend to cluster tightly on both sides of the threshold. As a result, even a small change in τ near these points can lead to a large change in the number of tokens being pruned. This indicates that the pruning process is especially sensitive around those points, and adjusting τ even slightly may have a big impact on the final token budget.

F. Construction Cost and Scalability

Construction cost is a complex issue, which we analyze from the perspectives of time and hardware requirements. Time-wise, the main components are CoE and LightRAG. While using APIs can significantly speed up the process, off-line deployment and inference are also feasible. For example, generating descriptions with Qwen2-VL-7B achieves around 60 tokens per second, processing one image ev-

ery 4 seconds. Thus, processing 1k images takes approximately 1.21 hours. Constructing a KG with Qwen2.5-7B yields about 196k tokens per hour, leading to a total of 1.33 hours for 1k images. The intermediate pruning step, accelerated by CLIP’s fast processing speed, is negligible. Overall, the cost is much lower than manual annotation or fine-tuning LLMs, making the method applicable to large-scale datasets. For resource-constrained users, deploying a lightweight VLM with CoE is comparable to or even more efficient than deploying a powerful VLM, further demonstrating the scalability of our approach.

G. Discussion on VLM Usage and Design Flexibility

Our observations on the number and type of VLMs used in CoE are consistent with the original conclusions drawn in the CoE paper [7]. Regardless of the specific VLM architecture, increasing the number of models N consistently improves performance up to a saturation point, after which further scaling yields diminishing returns. Moreover, we find that convergence is achieved more quickly when using lower softmax temperatures or simpler datasets. These factors reduce the ambiguity in model disagreement, allowing consensus to form more rapidly among the ensemble.

Interestingly, our results also show that using a single, strong VLM can achieve performance comparable to a cascade of smaller, lightweight models. This suggests a practical trade-off between model strength and ensemble size—while ensembling helps in reaching consensus across diverse weak learners, a single high-capacity model may suffice in many scenarios, especially when computational resources are limited.

In the original CoE method, the outputs from all VLM experts are first aggregated together, and then a selection process determines which expert descriptions to use. To save time in constructing the MMKGs with LLMs, we instead adopted a sequential strategy where the output of one expert is used as the prompt input for the next. We also evaluated the original aggregation and selection strategy on a smaller-scale dataset and found it to perform well, sometimes even surpassing the sequential approach. This confirms that CoE’s original design of aggregating all experts’ outputs before selecting which descriptions to use is effective and remains a strong baseline. However, correspondingly, using LLMs to construct MMKGs based on these aggregated descriptions requires significantly more time.

Additionally, while we apply pruning only at the final description step, pruning during intermediate steps may also yield good results depending on the dataset and task. There is no fixed rule for when or how to apply pruning, and our framework is designed to be flexible enough to accommodate different strategies. We emphasize that both our CoE framework and the SV step are intended to be adaptable, al-

lowing users to experiment freely and select the approach that best suits their needs.

There are various VLMs that can be used for pruning. Among them, we recommend CLIP due to its fast inference speed and pruning performance comparable to other VLMs. Given its efficiency and effectiveness, CLIP serves as a practical choice for pruning in many scenarios.

References

- [1] Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. Multi-modal named entity recognition with image attributes and image knowledge. In *Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part II* 26, pages 186–201. Springer, 2021. 1
- [2] ZIRUI GUO, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation, 2024. 2
- [3] Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. Multimodal reasoning with multimodal knowledge graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10767–10782, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1
- [4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 19730–19742, 2023. 2
- [5] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, pages 2507–2521. Curran Associates, Inc., 2022. 2
- [6] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1
- [7] Ziyang Xiao, Dongxiang Zhang, Yangjun Wu, Lilin Xu, Yuan Jessica Wang, Xiongwei Han, Xiaojin Fu, Tao Zhong, Jia Zeng, Mingli Song, and Gang Chen. Chain-of-experts: When LLMs meet complex operations research problems. In *The Twelfth International Conference on Learning Representations*, 2024. 3