LLM4SGG: Large Language Models for Weakly Supervised Scene Graph Generation

-CVPR 2024 Poster-

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, Chanyoung Park†

Presenter: Kibum Kim

Ph.D Student
Department of Industrial & Systems Engineering
KAIST

†Corresponding Author

CONTENT

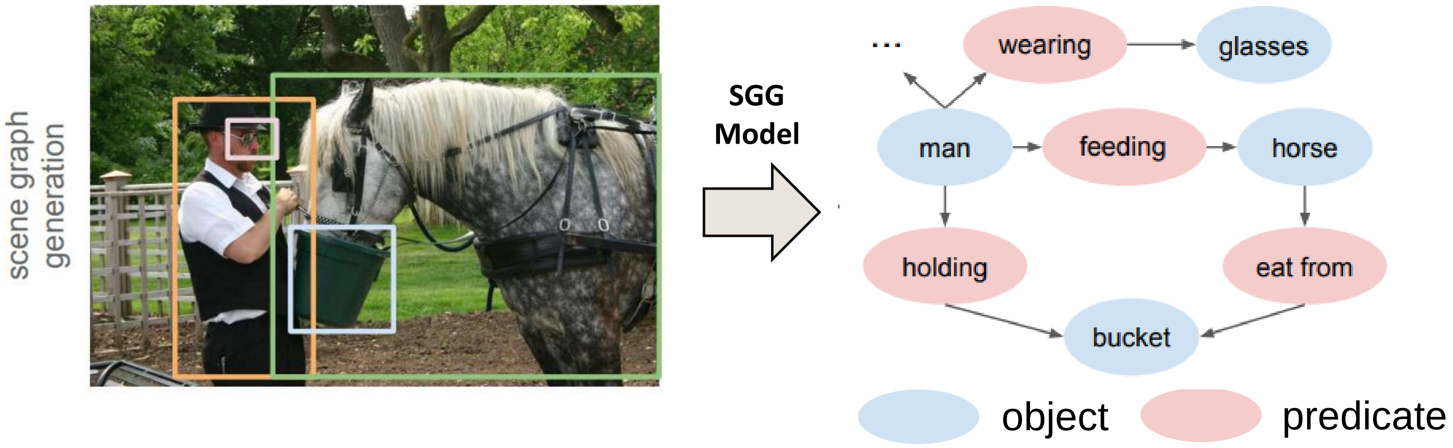
- **Scene Graph Generation**
- **Weakly Supervised Scene Graph Generation**
- **LLM4SGG: Large Language Models for Weakly Supervised Scene Graph Generation**
  - Motivation
  - Method
  - Experiment
  - Conclusion

SCENE GRAPH GENERATION (SGG)

- SGG aims to represent observable knowledges in an image in the form of a graph
- The knowledge includes 1) object information and 2) their relation information, which is mapped to a scene graph
  - E.g., Object information: {*man, horse, glasses, bucket*}
  - E.g., Relationship information between objects: {*feeding, wearing, …, holding, eat from*}

Input [1]
(image)

Output
(scene graph)

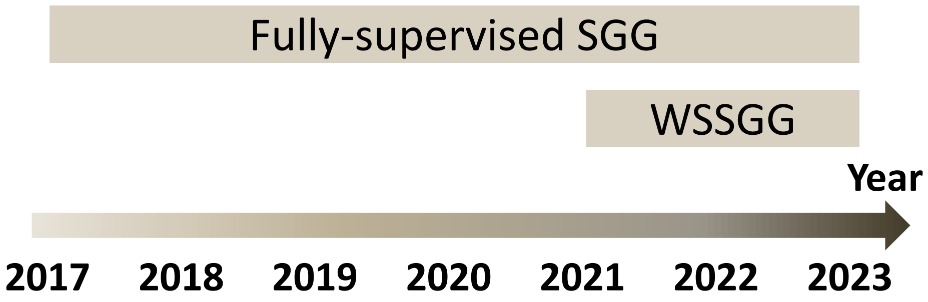


[1] Scene Graph Generation by Iterative Message Passing. Danfei Xu et al. CVPR'2017

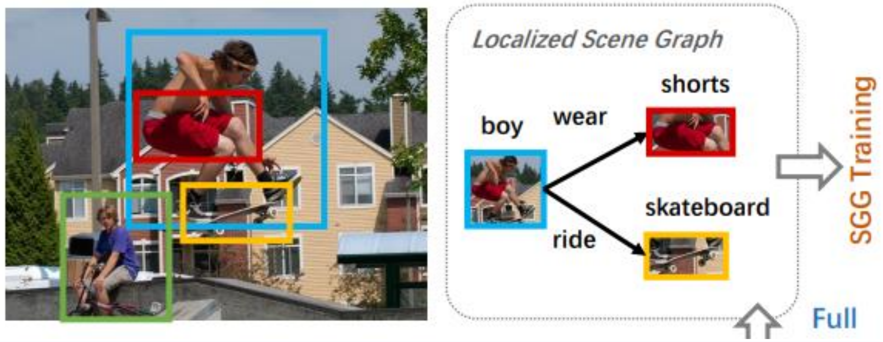
WEAKLY SUPERVISED SCENE GRAPH GENERATION

- **Weakly Supervised Scene Graph Generation (WSSGG)** aims to alleviate the issue of fully-supervised approach, which heavily relies on costly annotation.
  - Expensive Annotation: **1)** bounding box, **2)** entity class within bounding box, **3)** predicate class between entities
    Localized Scene Graph
  - Generating large-scale SGG data faces constraints due to the need for expensive human labor cost

- WSSGG studies generally utilize image-text pair datasets, which are readily accessible, for training the SGG model.



Timeline of fully supervised and weakly supervised SGG



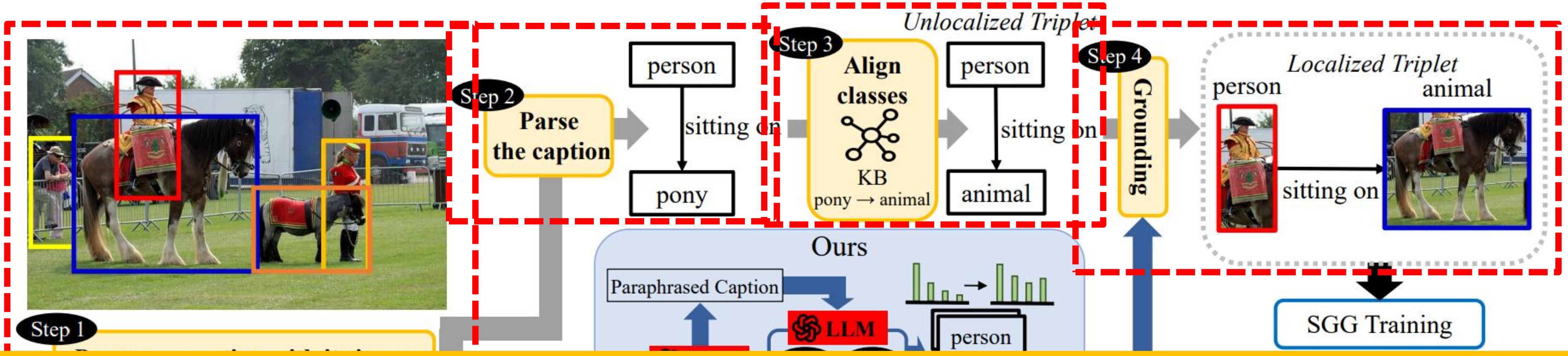
Expensive Annotation required for Fully supervised SGG [1]

[1] Integrating Object-aware and Interaction-aware Knowledge for Weakly Supervised Scene Graph Generation. Li et al. MM'22

LLM4SGG: Large Language Models for Weakly Supervised Scene Graph Generation

PIPELINE OF WSSGG

- Pipeline of training an SGG model with image caption datasets
  - **Step 1**: Preparing an image with its caption
  - **Step 2**: Parsing the image caption into <subject, predicate, object> triplets
  - **Step 3**: Aligning the entity/predicate classes of parsed triplets with the entity/predicate classes of target data (=Unlocalized Triplets)
  - **Step 4**: Grounding the unlocalized triplets with image regions (i.e., bounding boxes) extracted from pre-trained object detector

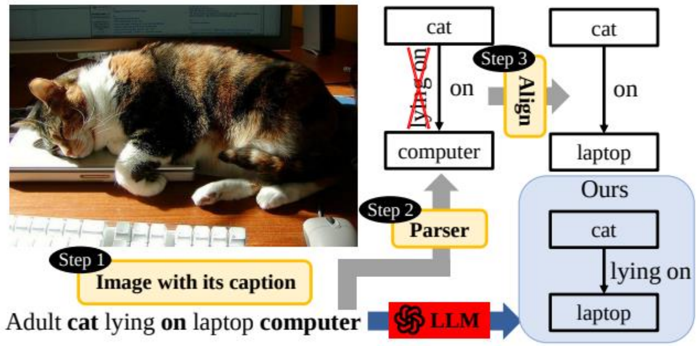


Existing WSSGG studies have mainly focused on grounding the unlocalized triplets (Step 4)

But! Do those unlocalized triplets have no issue? → Let's delve into it!

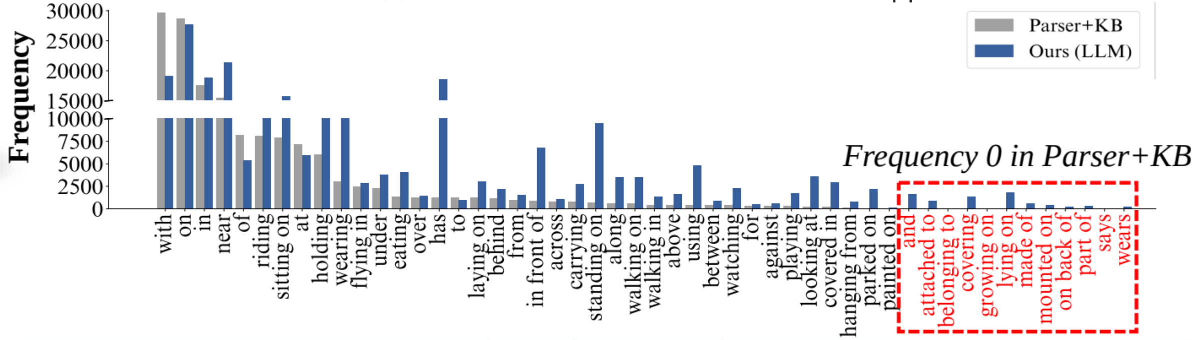
MOTIVATION: INHERENT ISSUES IN TRIPELT FORMATION PROCESS (1/2)

- **1. Issue in triplet formation process – Step 2**
  - Previous approach: Based on rule-based parser [1], existing works parse captions into triplets
    - Rule-based Parser [1] extract predicates without comprehending the context of captions.
  - ➢ **Semantic Over-simplification**: Informative predicates within captions are simplified into uninformative predicates.
    - Left figure: “lying on” within caption → “on”
- As a result, the long-tailed problem is exacerbated
  - Right figure: Predicate distribution from unlocalized triplets extracted by conventional approach (parser) and ours



Semantic Over-simplification in Step 2

※ Conventional approach: Parser + KB



Long-tailed problem in Conventional approach

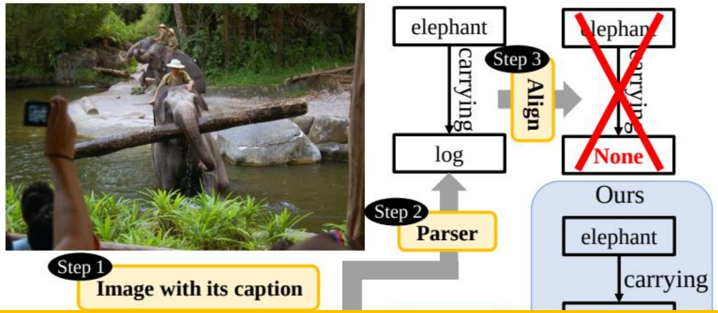
[1] Unified Visual-Semantic Embeddings: Bridging Vision and Language with Structured Meaning Representations. Wu et al. CVPR’19

MOTIVATION: INHERENT ISSUES IN TRIPELT FORMATION PROCESS (2/2)

- **2. Issue in triplet formation process – Step 3**
  - Previous approach: existing works align entity/predicate with those of target data based on knowledge base (e.g., WordNet [1])
    - Knowledge base (KB) fails to cover semantic relationship between a large number of words due to its static structured nature

      KB of synonyms, hypernyms, and hyponyms
  - ➢ **Low-Density Scene Graph**: Reduction in the number of triplets used for learning
    - ➢ <subject, predicate, object> triplet is discarded if alignment fails for any element in the triplet.
    - Left figure: A triplet is discarded since "log" is not aligned with predicate classes of target data (i.e., Visual Genome)

- As a result, insufficient supervision arises, leading to deterioration in the model's generalization



Dataset	How to annotate	# Triplet	# Image	
Fully-Supervised approach				
(a) Visual Genome	Manual	405K	57K	*7.1 triplets per img*
Weakly-Supervised approach				
(b) COCO Caption	Parser+KB	154K	64K	*2.4 triplets per img*
(c) COCO Caption	LLM	344K	64K	

To alleviate Semantic Over-simplification (Step 2) and Low-Density Scene Graph (Step 3) issues,

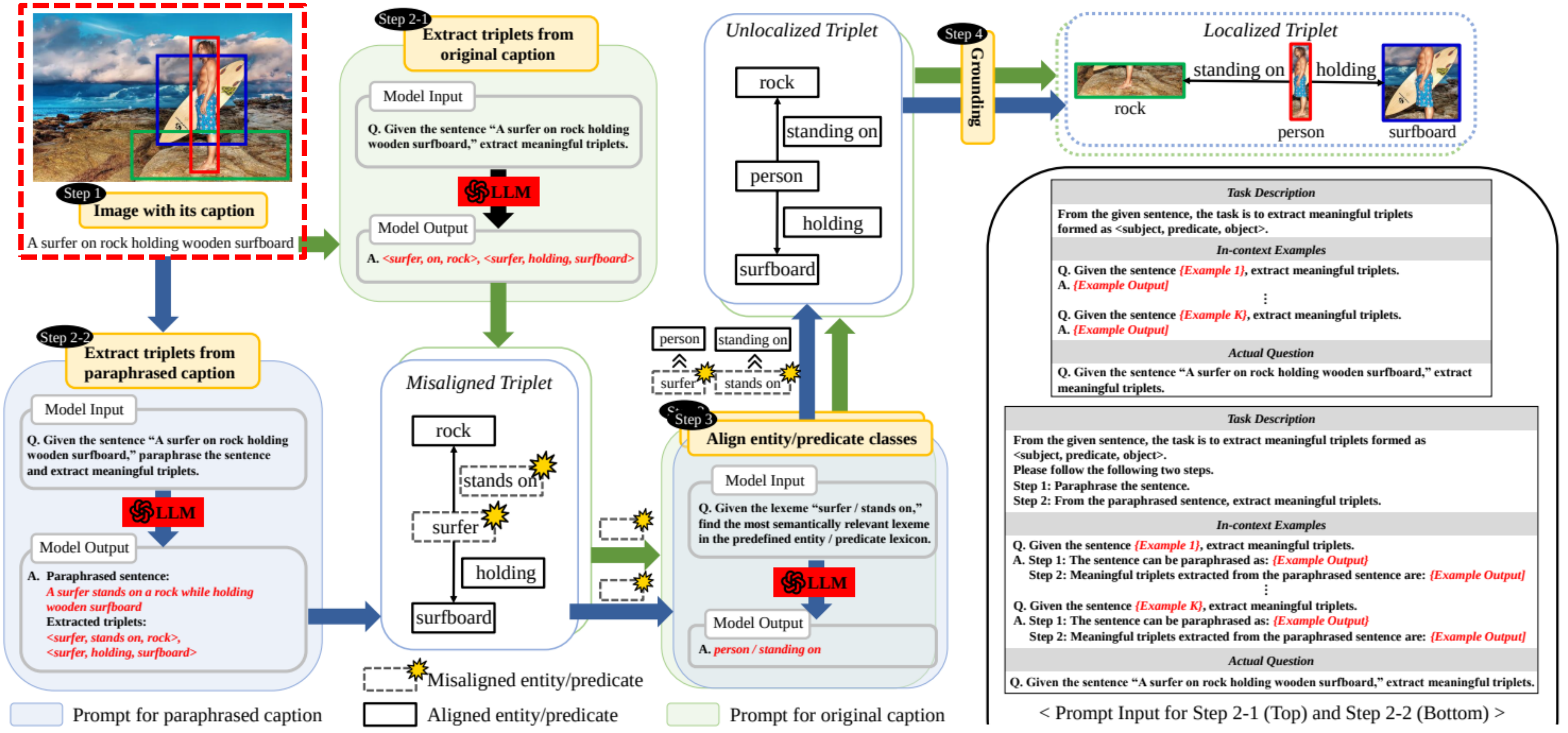
We introduce LLM for WSSGG task!

[1] Wordnet: a lexical database for english. Miller et al. Communication of ACM'95

METHOD: PREPARING IMAGE & CAPTION (1/4)

- Step 1: Preparing an image with its caption (e.g., COCO caption dataset)

1. Preparing an image with its captions



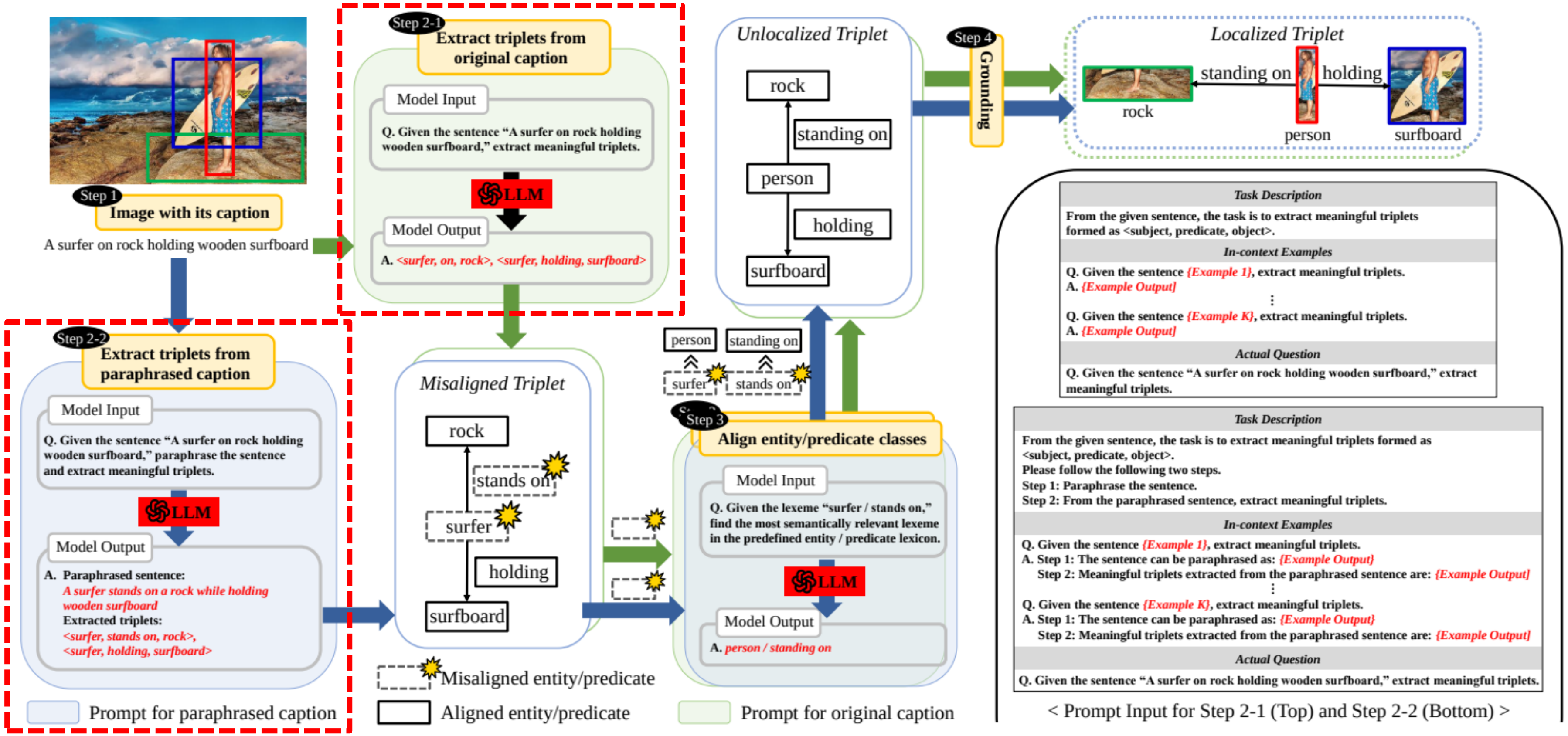
Pipeline of LLM4SGG

METHOD: TRIPLET EXTRACTION FROM CAPTIONS (2/4)

- Step 2-1: Extracting triplets from original captions via LLM, Step 2-2: <u>Extracting triplets from paraphrased captions</u> via LLM

To further alleviate Low-Density Scene Graph issue

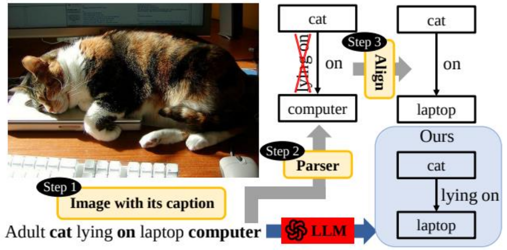
2-1. Extracting triplets from original captions



2-2. Extracting triplets from paraphrased caption

Based on comprehension of captions' context via LLMs, we extract triplets

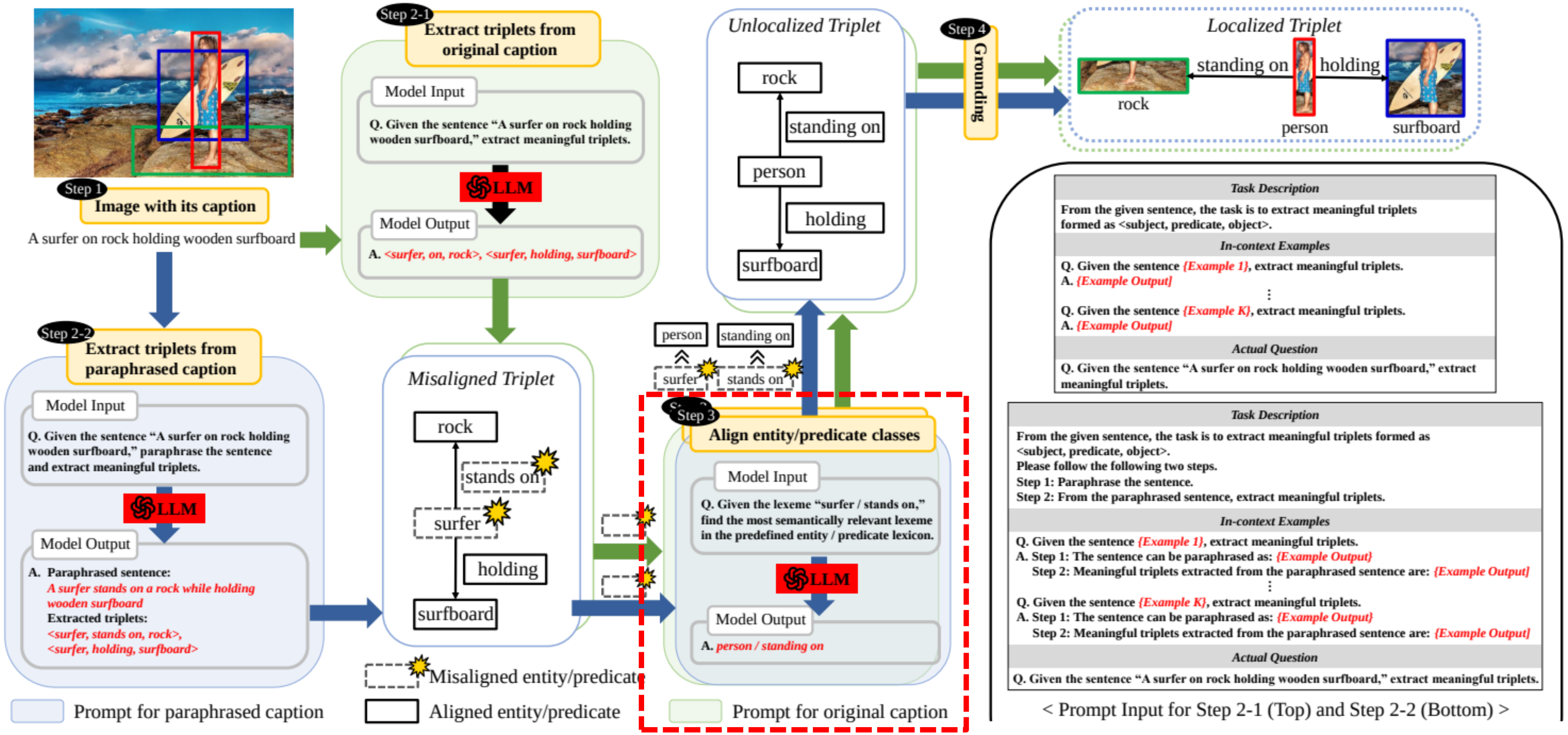
Alleviation of Semantic Over-simplification



Pipeline of LLM4SGG

METHOD: ALIGNMENT OF ENTITY / PREDICATE CLASSES (3/4)

- Step 3: Aligning the entities (subject, object) and predicate of misaligned triplets obtained in Step 2 with those of target data

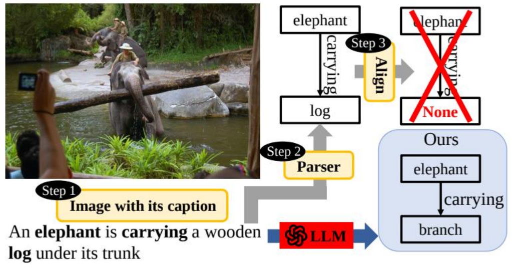


3. Alignment of Entity/Predicate with those of target data

Pipeline of LLM4SGG

Alignment based on semantic reasoning within LLMs

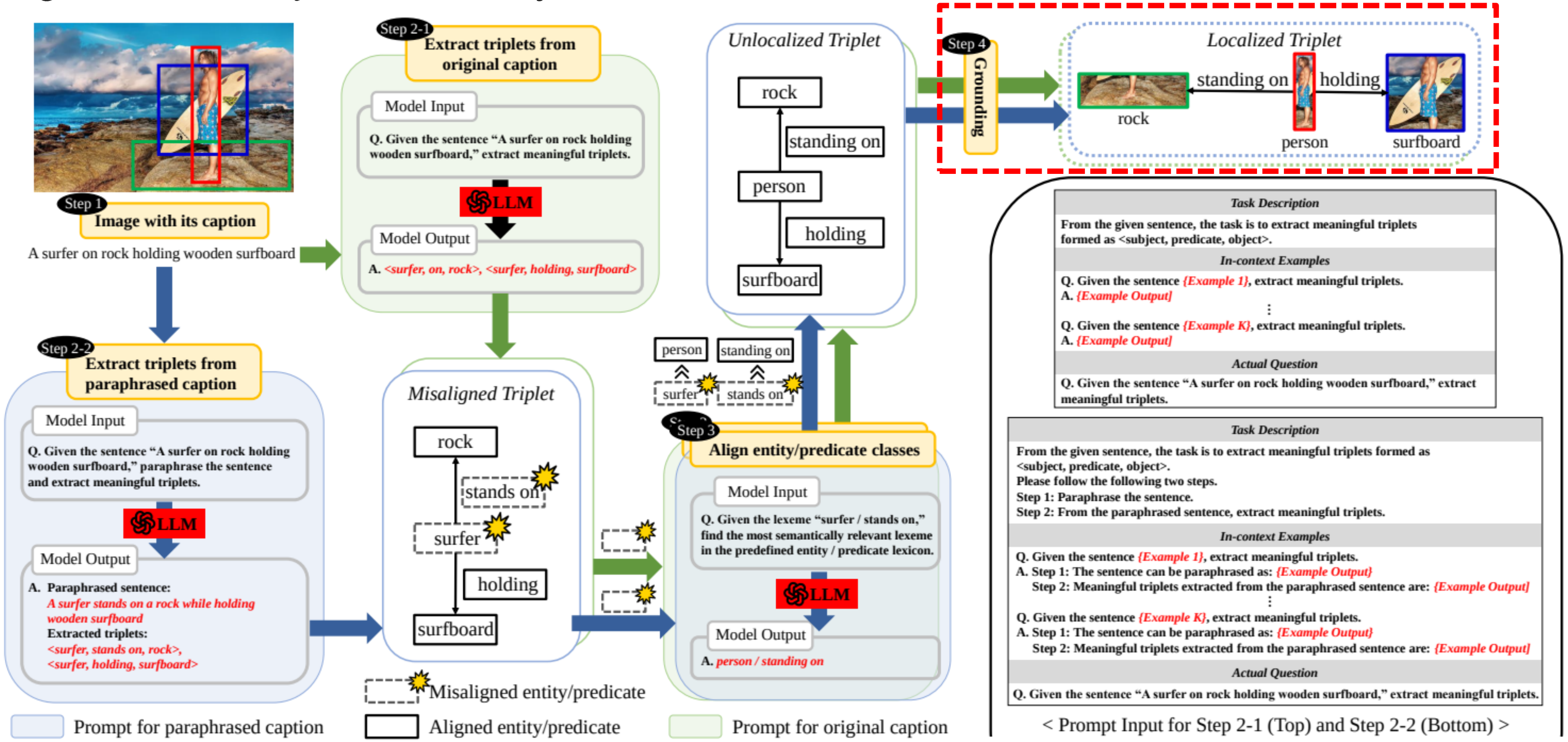
Alleviation of Low-Density Scene Graph



METHOD: GROUNDING OF UNLOCALIZED TRIPLETS (4/4)

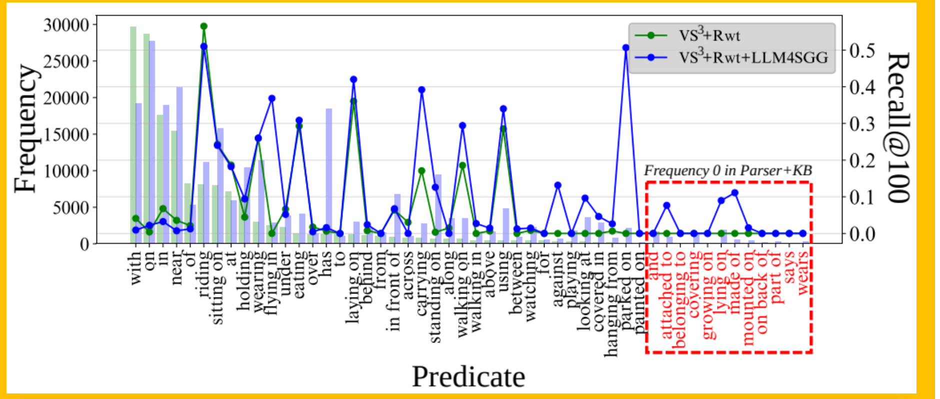
- Step 4: Grounding the unlocalized triplets to image regions using the grounding method of existing WSSGG works
  - Image region: Extracted by Pre-trained object detector

4. Grounding based on existing methods



Pipeline of LLM4SGG

EXPERIMENT: COMPARISON WITH BASELI

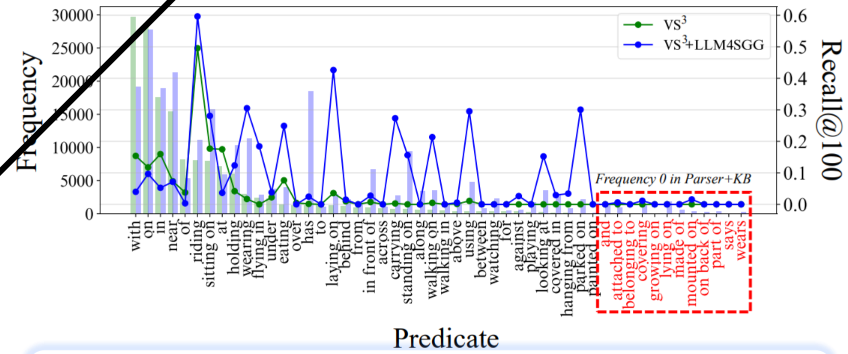


Per class comparison
when applying reweighting method

- Training dataset: COCO Caption (64K) / Test dataset: Visual Genome
- Grounding method
  - SGNLS [1] , VS^3 [2]
- Observation
  - 1) Performance enhancement in terms of mR@K → Alleviation of long-tailed problem for the first time (See right figure)
  - 2) Further Improvement on mR@K when applying reweighting method → it operates effectively since the number of tail predicate classes are increased

Method	R@50	R@100	mR@50	mR@100	F@50	F@100
Motif *(CVPR'18)* - Fully-supervised	31.89	36.36	6.38	7.57	10.63 / 12.53	12.53
LSWS *(CVPR'21)*	3.29	3.69	3.27	3.66	3.28	3.67
SGNLS *(ICCV'21)*	3.80	4.46	2.51	2.78	3.02	3.43
SGNLS *(ICCV'21)*+LLM4SGG	5.09 +1.29	5.97 +1.51	4.08 +1.57	4.49 +1.71	4.53 +1.51	5.13 +1.70
Li et al *(MM'22)*	6.40	7.33	1.73	1.98	2.72	3.12
VS^3 *(CVPR'23)*	6.60	8.01	2.88	3.25	4.01	4.62
VS^3 *(CVPR'23)*+LLM4SGG	**8.91** +2.31	**10.43** +2.42	7.11 +4.23	8.18 +4.93	**7.91** +3.90	**9.17** +4.55
VS^3 *(CVPR'23)+Rwt*	4.25	5.04	5.17	5.99	4.67	5.47
VS^3 *(CVPR'23)+Rwt*+LLM4SGG	5.10 +0.85	6.34 +1.30	**8.42** +3.25	**9.90** +3.91	6.35 +1.69	7.73 +2.26

Performance comparison with baselines



Per class comparison

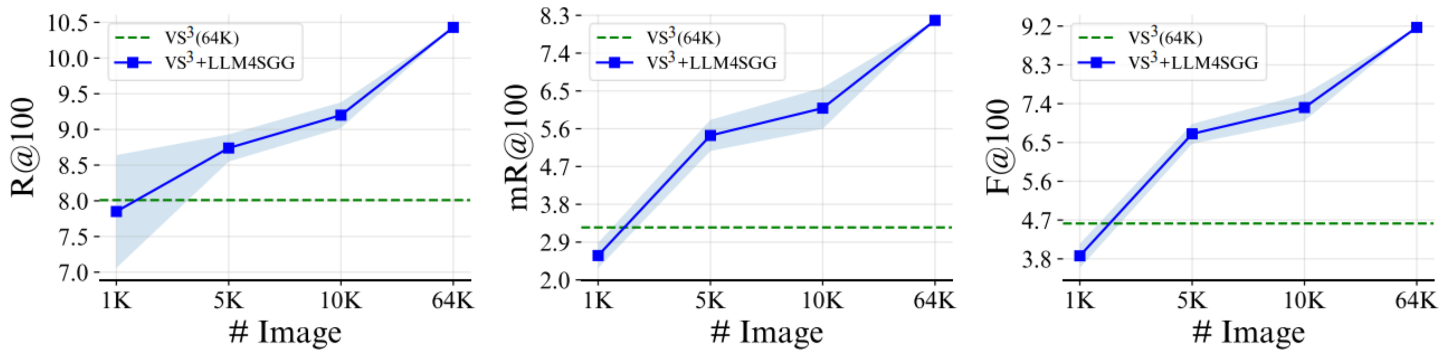
[1] Learning to Generate Scene Graph from Natural Language Supervision. Zhong et al. ICCV'21
[2] Learning to Generate Language-Supervised and Open-Vocabulary Scene Graph using Pre-trained Visual-Semantic Space. Zhang et al. CVPR'23

EXPERIMENT: DATA-EFFICIENCY

- **Question: Would LLM4SGG be effective despite having limited training data?**

- Total number of images: 64K

- Experiment: Performance is averaged by randomly extracting each of the following images five times: 1K (1.5%), 5K (7.8%), 10K (15.6%), 64K (100%)
  - Observation : Surpassing the performance of the baseline (VS^3) even with only 5K (7.8%) → Demonstrating Data-Efficiency
  - Another observation: Further performance increasement as the training data gradually increases to 10K and 64K



Performance over various number of images – Data efficiency

CONCLUSION

- **Existing Weakly Supervised SGG studies have mainly focused on grounding unlocalized triplets and image regions.**

- **However, we identify two issues within the triplet formation process: Semantic Over-simplification (Step 2) and Low-Density Scene Graph (Step 3).**

- **To alleviate them, we introduce LLM to the WSSGG task in Step 2 and Step 3.**

- **We observe that LLM4SGG significantly increases performance in terms of R@K and mR@K on both Visual Genome and GQA datasets.**
  - Demonstration of effectively alleviating the Semantic Over-simplification and Low-Density Scene Graph issues.

For more details of experiments, please refer to main paper.

THANK YOU

- Paper (Arxiv): https://arxiv.org/pdf/2310.10404
- Code: https://github.com/rlqja1107/torch-LLM4SGG

Paper

Code