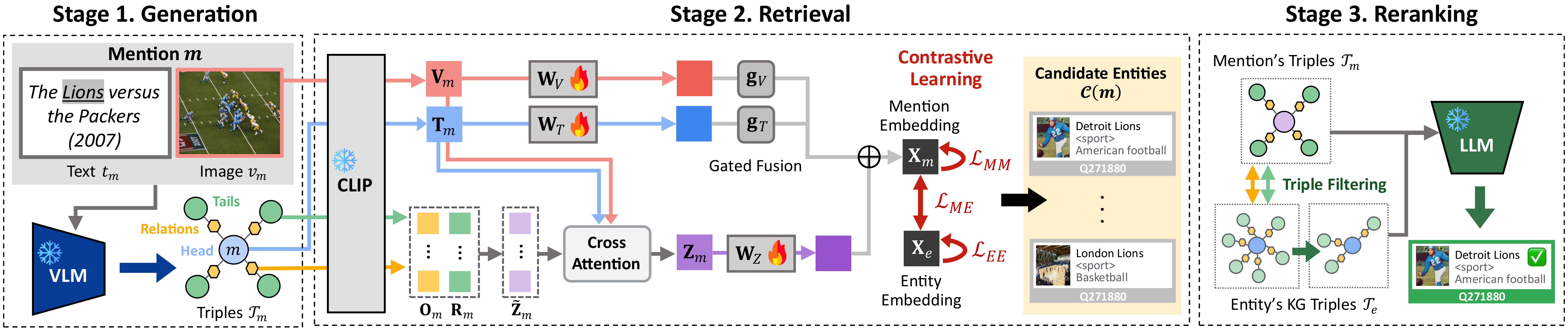
KGMEL: Knowledge Graph-Enhanced Multimodal Entity Linking

Juyeon Kim — KAIST — juyeonkim@kaist.ac.kr
Geon Lee — KAIST — geonlee0325@kaist.ac.kr
Taeuk Kim — Hanyang University — kimtaeuk@hanyang.ac.kr
Kijung Shin — KAIST — kijungs@kaist.ac.kr

GitHub: https://github.com/juyeonnn/KGMEL

SIGIR 2025

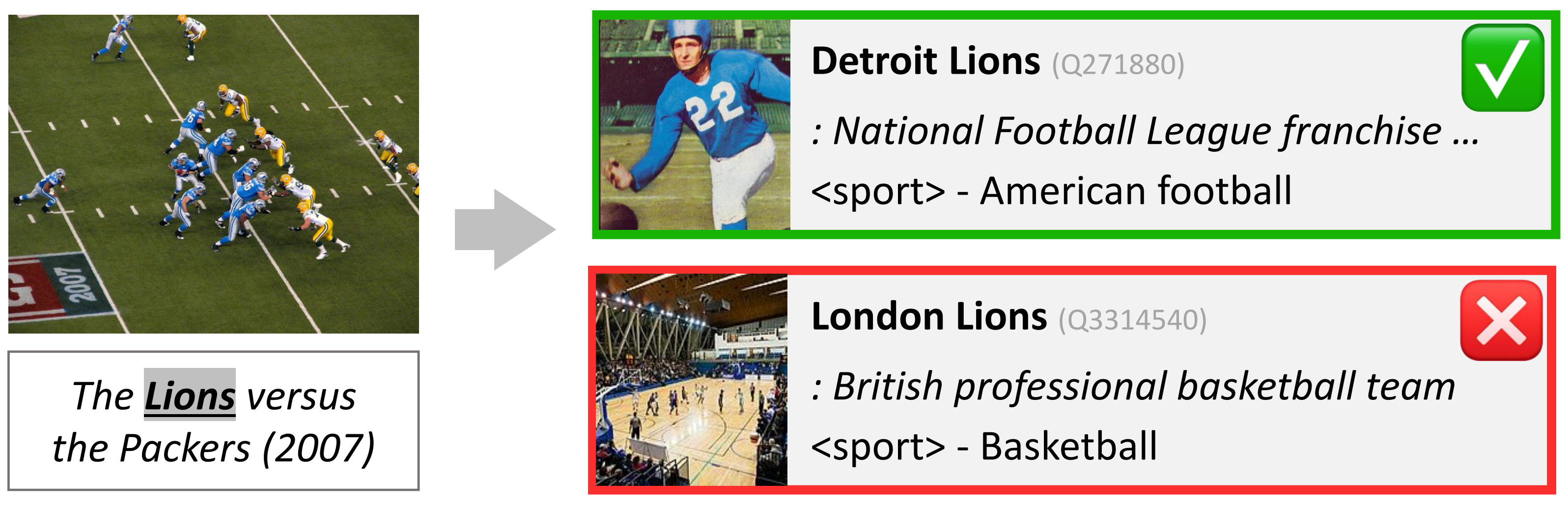


Summary

- **MEL** aims at aligning **mentions** with their corresponding **entities** in a knowledge base using **both visual and textual information** from mentions.
- **Key Observations: KG triples show significant potential for MEL.**
  - **O1. Abundance of KG triples:** Entities are usually associated with **lots of KG triples**, often **richer** than their text descriptions.
  - **O2. Triples as semantic bridges:** Triples help **disambiguate** between entities, that are otherwise **indistinguishable from text and image alone**.
- **Challenges**
  - **C1. Structured KG data is provided for entities, but not for mentions.**
  - **C2. Too many irrelevant KG triples:** Only a few are useful for linking.
- **Proposed Method: KGMEL**
  - A novel three-stage **generate-retrieve-rerank** framework.
  - Effectively leverages **KG triples** to enhance MEL.
- **Experiments**: KGMEL achieves **SOTA performance** across all benchmarks, outperforming the best competitor by up to **19.13%** in terms of H@1.

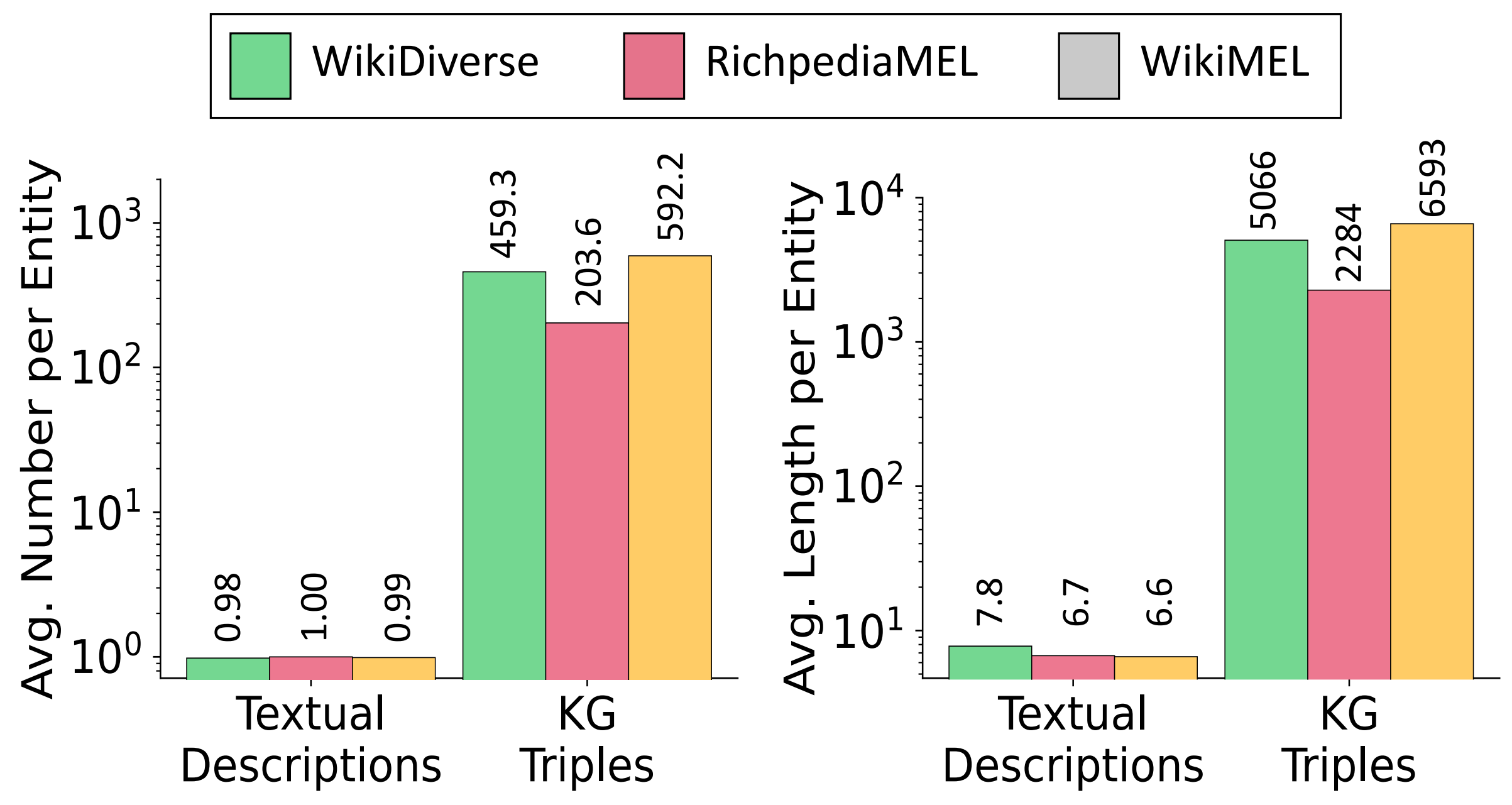
Background: MEL (Multimodal Entity Linking)

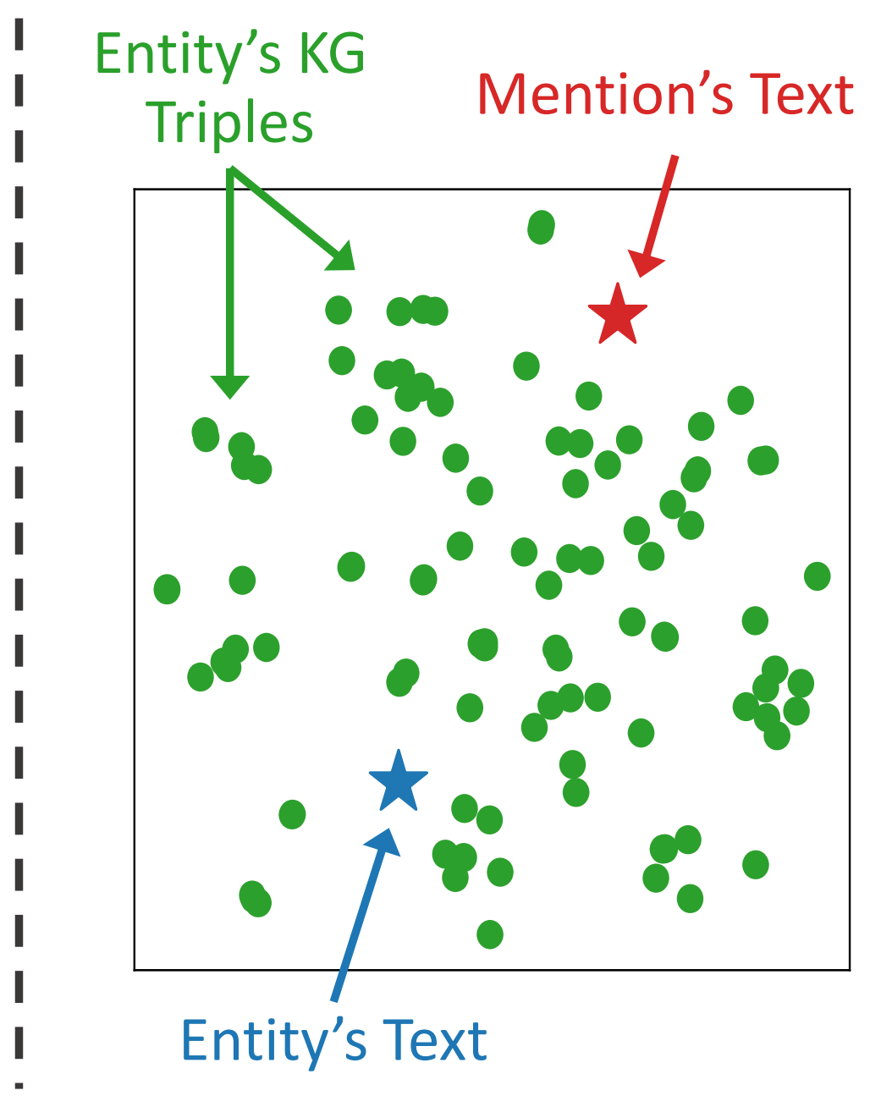
- **Task**: Identify the **ground-truth entity** $e_m \in \mathcal{E}$ that best matches **mention** m.
  - **Entity:** Given a set of entities $\mathcal{E}$, each **entity** $e \in \mathcal{E}$ is represented as **textual information** t_e, **visual information** v_e, and **KG triples** $\mathcal{T}_e$.
  - **Mention:** A **mention** m is represented as **textual information** t_m and **visual information** v_m.


The Lions versus the Packers (2007)
Detroit Lions (Q271880)
: National Football League franchise …
<sport> - American football
London Lions (Q3314540)
: British professional basketball team
<sport> - Basketball

Data Analysis: Triples in Knowledge Base

- **[Observation 1] Abundance of KG triples**
  - **Entities** are typically associated with **numerous structured KG triples** that are often **semantically richer** than concise textual descriptions.
- **[Observation 2] Triples as semantic bridges**
  - KG triples provide **contextual information** that helps **disambiguate** between entities that look **similar in text or image.**





Proposed Method: KGMEL

- **[Stage 1] Triple Generation of Mentions**
  - Use VLMs to **generate structured triples** from **mention's text and image.**
- **[Stage 2] Candidate Entity Retrieval**
  - **Encoding**: Obtain embeddings of **text**, **image**, and **triples**.
  - **Fusion:** Combine embeddings using **gated fusion** mechanism.
  - **Learning Objective:** Train with three **contrastive learning losses**.
- **[Stage 3] Entity Reranking**
  - **Triple Filtering:** Filter **entity triples** to retain only those related to **mention triples**, allowing LLMs to focus on the **most relevant information**.
  - **Zero-Shot Reranking:** Use LLMs to identify best matching entity.

Experimental Results

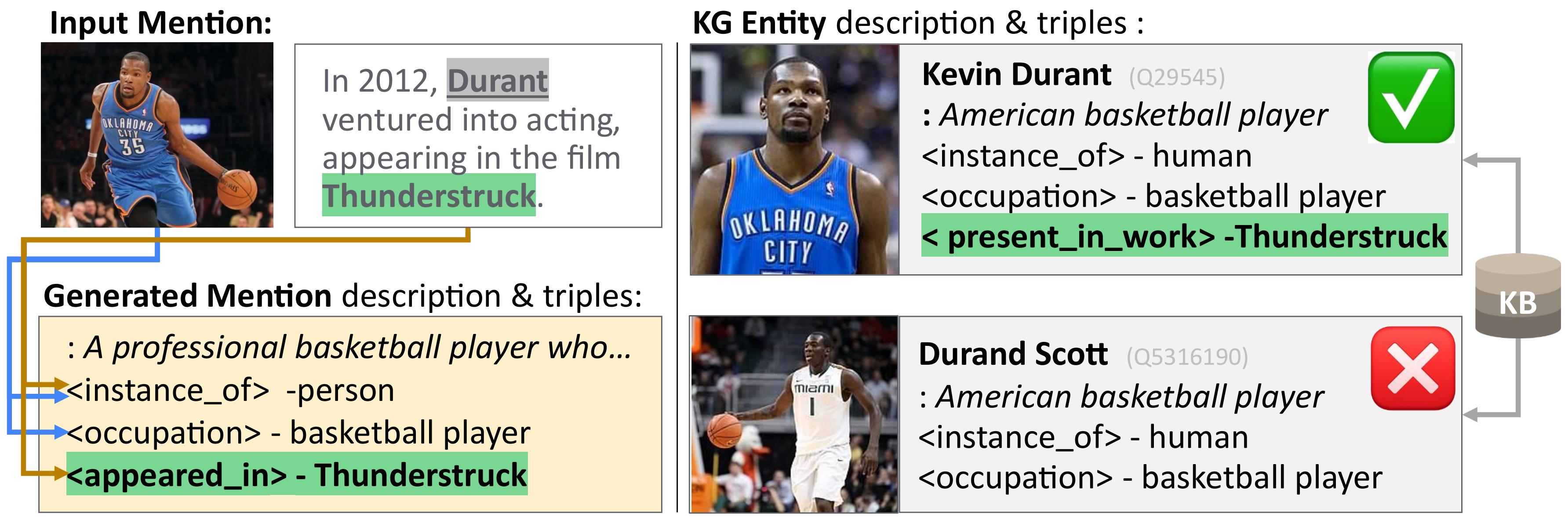
- **[RQ1] Accuracy**
  - **KGMEL** achieves **SOTA performance** across all three benchmarks.
  - **KGMEL** outperforms the best competitor by up to **19.13%** in terms of H@1.

	WikiDiverse	RichpediaMEL	WikiMEL
CLIP [23]	61.21	67.78	83.23
ViLT [9]	34.39	45.85	72.64
ALBEF [10]	60.59	65.17	78.64
METER [7]	53.14	63.96	72.46
DZMNED [20]	56.90	68.16	78.82
JMEL [1]	37.38	48.82	64.65
VELML [34]	54.56	67.71	76.62
GHMFC [28]	60.27	72.92	76.55
MIMIC [17]	63.51	81.02	87.98
OT-MEL [33]	66.07	83.30	<u>88.97</u>
MELOV [26]	67.32	84.14	88.91
M^{3}EL [8]	74.06	82.82	88.84
IIER [19]	69.47	<u>84.63</u>	88.93
GPT-3.5-turbo [21]	-	-	73.80
LLaVA-13B [13]	-	-	76.10
GEMEL [24]	-	-	82.60
GELR [15]	-	-	84.80
KGMEL (retrieval)	<u>82.12</u> ± 0.21	76.40 ± 0.30	87.29 ± 0.08
KGMEL (+ rerank)	**88.23** ± 0.29	**85.21** ± 0.24	**90.58** ± 0.25

- **[RQ2] Effectiveness**
  - Ablation study shows that **all KGMEL components** are **effective**.

	WikiDiverse	RichpediaMEL	WikiMEL	ΔAvg
KGMEL (retrieval)	**82.12**	**76.40**	**87.29**	-
w/o Image **V**	81.02	67.19	80.99	-5.54
w/o Triple **Z**	81.61	73.40	<u>85.95</u>	-1.62
w/o GateLayer $\mathbf{g}_*$	<u>81.73</u>	<u>74.38</u>	85.84	-1.29

- **[RQ3] Case Study**
  - **KGMEL** successfully **extracts** and **leverages** relevant triples from **textual** and **visual** information to identify the correct entity.


Input Mention:
In 2012, Durant ventured into acting, appearing in the film Thunderstruck.
Generated Mention description & triples:
: A professional basketball player who…
<instance_of> -person
<occupation> - basketball player
<appeared_in> - Thunderstruck
KG Entity description & triples :
Kevin Durant (Q29545)
: American basketball player
<instance_of> - human
<occupation> - basketball player
< present_in_work> -Thunderstruck
Durand Scott (Q5316190)
: American basketball player
<instance_of> - human
<occupation> - basketball player
KB