Light Up the Shadows: Enhance Long-Tailed Entity Grounding with Concept-Guided Vision-Language Models

Yikai Zhang[♠], Qianyu He[♠], Xintao Wang[♠],
Siyu Yuan[♡], Jiaqing Liang[♡] [*], Yanghua Xiao[♠][*]
[♠]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[♡]School of Data Science, Fudan University
{ykzhang22,qyhe21,xtwang21,syyuan21}@m.fudan.edu.cn,
{liangjiaqing, shawyh}@fudan.edu.cn

Abstract

Multi-Modal Knowledge Graphs (MMKGs) have proven valuable for various downstream tasks. However, scaling them up is challenging because building large-scale MMKGs often introduces mismatched images (*i.e.*, noise). Most entities in KGs belong to the long tail, meaning there are few images of them available online. This scarcity makes it difficult to determine whether a found image matches the entity. To address this, we draw on the Triangle of Reference Theory and suggest enhancing vision-language models with concept guidance. Specifically, we introduce COG, a two-stage framework with **CO**ncept-**G**uided vision-language models. The framework comprises a CONCEPT INTEGRATION module, which effectively identifies image-text pairs of long-tailed entities, and an EVIDENCE FUSION module, which offers explainability and enables human verification. To demonstrate the effectiveness of COG, we create a dataset of 25k image-text pairs of long-tailed entities. Our comprehensive experiments show that COG not only improves the accuracy of recognizing long-tailed image-text pairs compared to baselines but also offers flexibility and explainability.[1]

1 Introduction

Multi-Modal Knowledge Graphs (MMKGs) are knowledge graphs that integrate and align information from diverse modalities (*e.g.*, text and images) (Ferrada et al., 2017; Liu et al., 2019a; Wang et al., 2020). Due to the growing demand for multimodal intelligence and extensive knowledge in various applications (Hou et al., 2019; Marino et al., 2021), MMKGs have received increasing attention in recent years.

Although the number of images in current MMKGs has increased, their coverage and accuracy are limited (Oñoro-Rubio et al., 2017; Wang et al., 2020), especially for long-tailed entities (*i.e.*, less common entities). Figure 1 illustrates that the trends in an entity's click frequency and the number of matching images found are similar, both displaying a long-tailed pattern, indicating the scarcity of images of long-tailed entities.

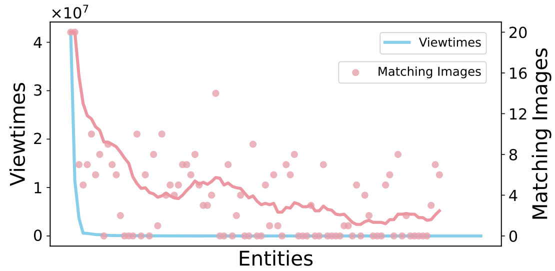


Figure 1: We randomly select 100 entities from the large-scale knowledge graph CN-DBpedia (Xu et al., 2017) and add human annotations. The blue line represents the changes in the entities' *viewtimes*, which indicates their click frequency. The red dots indicate the number of correctly matched images found in the top 20 search results for each entity, and the red line smooths out these data points.

Aligning long-tailed entities with appropriate images (*i.e.*, entity grounding) is crucial in constructing MMKGs. First, it expands the scope and completeness of MMKGs, which traditionally focus on common entities (Oñoro-Rubio et al., 2017; Wang et al., 2020). Second, adding visual content for long-tailed entities boosts efficiency in downstream tasks (Pezeshkpour et al., 2018; Chen et al., 2022b,a). Third, pairing images with long-tailed entities provide valuable training data for developing and refining vision-language models, particularly for rare or domain-specific entities.

Grounding long-tailed entities in MMKGs is challenging. Current methods (Oñoro-Rubio et al., 2017; Liu et al., 2019a; Wang et al., 2020) for entity grounding rely on web resources, especially search engines. They collect images by matching

[*] Corresponding authors.

[1] Resources of this paper can be found at https://github.com/ykzhang721/COG.

Findings of the Association for Computational Linguistics: ACL 2024, pages 13379–13389
August 11-16, 2024 ©2024 Association for Computational Linguistics

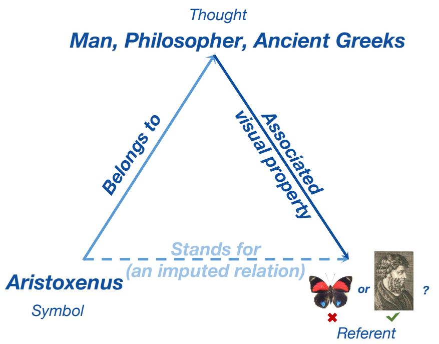


Figure 2: This figure shows that when searching for an entity named *Aristoxenus*, the search engine returns two images. By applying concepts, we can conclude that the target *Aristoxenus* refers to a person, not a butterfly.

image captions to entity names. While effective for well-known entities (Liu et al., 2019a; Wang et al., 2020), these approaches struggle with long-tailed entities due to several limitations: 1) Search engines used for text matching show suboptimal performance because entity grounding involves matching images and texts; 2) Although pre-trained vision-language models (PVLMs) have shown impressive performance in various cross-modal tasks, they encounter challenges in recognizing long-tailed entities due to their infrequent appearance during pre-training; 3) Current methods lack the ability to explain their image selection, which is important for ensuring quality.

To address these challenges, we develop COG, a two-stage framework that uses **CO**ncept-**G**uided vision-language models to ground long-tailed entities. Initially, as shown in Figure 2, the Triangle of Reference Theory (McElvenny, 2014) explains the connection between *Thought*, *Symbol*, and *Referent*. This demonstrates how humans link an entity to a real-world object through concepts. Inspired by this theory, we improve PVLMs to accurately identify images of long-tailed entities using concepts. Notably, our COG allows for the use of replaceable PVLMs, meaning it can integrate any module for matching images and texts, thus offering flexibility in application. Next, we investigate how the choice of concept selection strategy affects the performance of COG, considering the varying scope of concepts. Lastly, COG not only enhances recognition accuracy but also ensures explainability, providing a basis for further quality control (*i.e.*, human verification).

To sum up, our contributions are as follows:

- We introduce a flexible and explainable two-stage framework COG, using concept-guided PVLMs to recognize image-text pairs of long-tailed entities.
- We examine and discuss how choosing different concepts affects the experimental results.
- Through extensive experiments, we show that our method significantly enhances the accuracy of long-tailed entity grounding. It also allows for human verification, ensuring finer quality control.

2 Related Work

2.1 Multi-Modal Knowledge Graph Construction

A Multi-Modal Knowledge Graph (MMKG) is a unified information representation that integrates data from various modalities, such as text, images, and audio, into a single interconnected graph (Zhu et al., 2022). Existing methods for entity grounding in MMKGs can be categorized into two main groups: *1) Methods based on online encyclopedias* (Ferrada et al., 2017; Alberts et al., 2020): These methods link existing encyclopedic multimedia resources (*e.g.*, Wikimedia Commons, Wikipedia, ImageNet (Deng et al., 2009)) by associating texts with images to construct MMKGs; *2) Methods based on web search engines* (Oñoro-Rubio et al., 2017; Liu et al., 2019a; Wang et al., 2020): These methods directly search for images of entities using web search engines. This approach is more flexible than using online encyclopedic multimedia data, as it allows for expansion based on existing filtered and refined KGs. However, it tends to prioritize popular entities because entity images adhere to a power-law distribution. For long-tailed entities lacking web images, search engines, despite providing ranked results, can easily return incorrect or mismatched images, leading to noise. In this paper, we propose COG that leverages concepts to reduce this kind of noise for long-tailed entity grounding.

2.2 Pre-Trained Vision-Language Models

Pre-trained Vision-Language Models (PVLMs) are designed to process visual and textual data, aiming to align image-text data through extensive cross-modal pre-training. Many approaches conduct contrastive pre-training on large-scale datasets (Radford et al., 2021; Jia et al., 2021; Yuan et al., 2021; Li et al., 2022; Yu et al., 2022). CLIP (Radford et al., 2021) employs a self-supervised method with 400 million internet-sourced image-text pairs to enhance modality alignment. ALIGN (Jia et al., 2021) uses a dual-encoder and trains with over a billion pairs, while BLIP (Li et al., 2022) improves multimodal task performance by filtering out low-quality data. In COG, we design a two-stage framework to guide PVLMs with concepts. PVLMs serve as a module to generate similarity between texts and images in COG. This module can be replaced by other methods with similar functionality, making our framework more flexible and general.

2.3 Long-Tailed Classification

Some researchers in computer vision focus on the problem of long-tailed image classification. This issue primarily concerns classifying images that are infrequent in the training set. Various datasets (Liu et al., 2019b; Cui et al., 2019) are used to evaluate the ability to learn classification with limited samples. However, the problem we discuss in this paper is different from traditional image classification. Long-tailed entity grounding is not confined to known, fixed classes, which still include numerous images on the web. Identifying images of specific and unknown entities poses a significant challenge to PVLMs. To address this issue, we suggest creating connections between entities and images for PVLMs using concepts, aiming for accurate recognition.

3 COG

3.1 Problem Definition

A Multi-Modal Knowledge Graph (MMKG) is a knowledge graph where nodes are entities or images, and edges represent their relationships. In MMKG, triplets are defined as $(e, \textit{has image}, i)$, where e is the textual entity and i is its corresponding image, indicating a *has image* relationship. To match images with entities in MMKG (*i.e.*, entity grounding in MMKG), common methods typically follow a two-step process. First, they rank the collected images based on their relevance to the given entity, which can be modeled as a **Ranking** task. To formalize this, given a corrupted triplet $(e, \textit{has image}, ?)$ in MMKG, this sub-task aims to predict the removed image i. Then, they select the top-n images and classify whether the image is related to the given entity in order, which can be modeled as a binary **Classification** task. To formalize this, each triplet $(e, \textit{has image}, i)$ can be classified as $True$ if the image correctly matches the entity; otherwise, the triplet is classified as $False$. We design both tasks to thoroughly simulate the entity grounding process and conduct rich experiments on COG.

3.2 Concept Selection

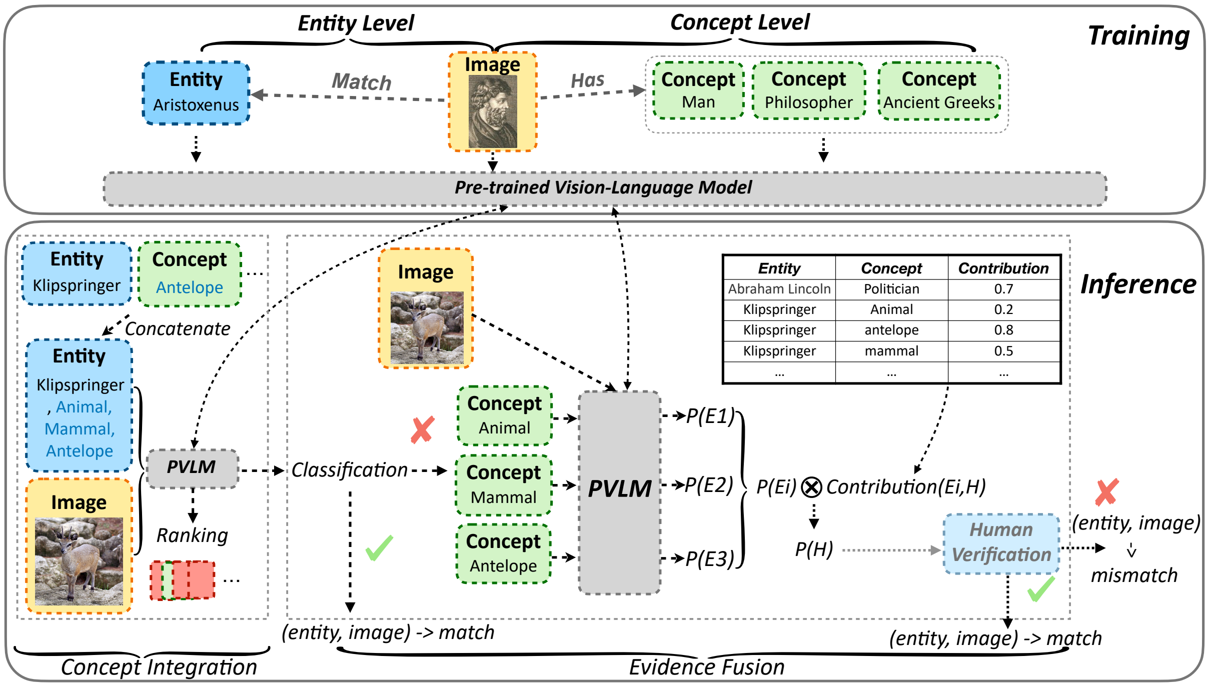
A concept typically refers to a group of entities sharing common characteristics. These concepts can be categorized based on the number of entities they encompass, indicating different levels of granularity. To identify which concepts are most effective for concept guidance, we explore the influence of utilizing various concepts. It is common for entities to embody multiple concepts, each displaying distinct levels of granularity. According to (Wang et al., 2015), humans mainly utilize Basic-level Categorization (BLC), a mid-level concept, for everyday thinking.

Motivated by this understanding, we evaluate the efficiency of BLC concepts compared to a broader range of concepts. We describe BLC concepts as those represented by a single word and examine their performance against that of all concepts. The experiments outlined in § 5.2 demonstrate how different strategies for selecting concepts influence performance.

3.3 Details of COG

Concept guidance initially necessitates that PVLMs possess the capability to identify concepts and images. Subsequently, they should use this ability to conduct thorough analyses for long-tailed entity grounding.

Figure 3 illustrates that, to train models for concept recognition, we use contrastive learning on both the entity and concept levels, as detailed in § 3.3.1. When performing inference with our finely tuned model, we introduce a two-stage framework consisting of two modules in § 3.3.2: CONCEPT INTEGRATION and EVIDENCE FUSION. CONCEPT INTEGRATION combines all available information to directly predict whether an image matches the corresponding text. Given that images of long-tailed entities are rare and valuable, we develop EVIDENCE FUSION to reassess the image candidates discarded by CONCEPT INTEGRATION. EVIDENCE FUSION offers clear evidence by breaking down various attributes of the entity, thus providing a basis for human annotation.



Entity	Concept	Contribution
Abraham Lincoln	Politician	0.7
Klipspringer	Animal	0.2
Klipspringer	antelope	0.8
Klipspringer	mammal	0.5
...	...	...

Figure 3: Overview of COG. COG uses contrastive learning on entity and concept levels for model training. At the inference stage, we utilize a two-stage framework with CONCEPT INTEGRATION and EVIDENCE FUSION modules. CONCEPT INTEGRATION aims for direct prediction of image-text matches using concept guidance, while the EVIDENCE FUSION module reassesses discarded image candidates from CONCEPT INTEGRATION, particularly valuable for rare, long-tailed entities.

3.3.1 Contrastive Learning on Two Levels

When training PVLMs, we designate a text as t and an image as i. First, we input both t and i into the PVLM. The model generates a $prediction$ that indicates the degree of alignment between t and i, as demonstrated here:

$$logit = PVLM(t, i) \tag{1}$$

$$Sigmoid(logit) = \frac{1}{1 + e^{-logit}} \tag{2}$$

$$prediction = Sigmoid(logit) \tag{3}$$

In this equation, both t and i symbolize the text and image inputs respectively. The $prediction$ value signifies the model's prediction of the similarity between the image and the text. If the $prediction$ surpasses a certain threshold, we deem it as a match; otherwise, it is considered a mismatch.

Next, we train the model using contrastive learning, which includes in-batch negative samples. Each batch contains n samples, and n represents the batch size. A sample consists of a pair (t, i), which stands for a text and an image. As shown in Figure 3, we create contrastive samples on both the entity and concept levels. We define t_i as the combination of the i-th entity and its concepts: $t_i = e_i, c_1, c_2, \ldots, c_m$, where e_i is the i-th entity and c_j is the j-th concept of the entity.

On the entity level, we use p_{t_a, i_b} to represent the $prediction$ of the concatenation of the a-th concatenated text and the b-th image, and $l_{a,b}$ to represent the label whether it matches. Then, we obtain L_{entity} in a batch:

$$L_{entity} = -\sum_{a=1}^{n}\sum_{b=1}^{n} BCE(l_{a,b}, p_{t_a, i_b}) \tag{4}$$

where BCE is the binary cross entropy function.

Similarly, we first obtain concepts related to a-th entity e_a using $C(e_a)$. Assuming there are m concepts of e_a, p_{c_k, i_b} represents the $prediction$ of the k-th concept and the b-th image and $l_{n \times m \times n}$ represents a matrix where $l_{a,k,b}$ is 1 if the b-th entity has the k-th concept of the a-th entity; otherwise, $l_{a,k,b}$ is 0. The concept loss $L_{concept}$ is calculated as:

$$L_{concept} = \tag{5}$$

$$-\sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{k=1}^{len(C(e_a))} BCE(l_{a,k,b}, p_{c_k, i_b})$$

Finally, we update the model parameters by the loss *L* below:

$$L = L_{entity} + L_{concept} \tag{6}$$

3.3.2 Concept-Guided Image-Text Recognition

CONCEPT INTEGRATION In CONCEPT INTEGRATION, we directly concatenate all concepts c related to the entity e as t and input the concatenated text t and image i into the PVLM. For example, take the entity *Jay Chou* associated with concepts like *singer*, *actor*, and *director*. The concatenated text would be *Jay Chou, singer, actor, director.* The PVLM generates a $prediction$ for input t, i and we set an threshold for judging whether an image-text pair matches.

While CONCEPT INTEGRATION improves performance in experiments, it acts as a black-box model lacking explanatory capability. Additionally, images of long-tailed entities are scarce. The black-box approach's prediction lacks credibility, potentially causing errors or the loss of correct images. Therefore, we introduce EVIDENCE FUSION to re-judge the samples discarded in CONCEPT INTEGRATION.

EVIDENCE FUSION For a more comprehensive understanding of EVIDENCE FUSION, we first define:

Definition $P()$ represents the probability of occurrence. E and H represent the evidence events and the ultimate conclusion, respectively. $P(E)$ and $P(H)$ are utilized to express the probability of E and H. Additionally, $Con(E, H)$ is the contribution of evidence E on conclusion H.

In our task, the evidence *E* refers to the image matching the concept of the entity, while the conclusion *H* is that the image matches the entity.

In EVIDENCE FUSION, essentially, we transform the task of matching an entity and an image into a comprehensive analysis of the matching between the concepts of the entity and the image. In Figure 3, E_i represents an image matching a concept. For example, we define evidence E_1 as *The object in the image is an animal* and evidence E_2 as *The object in the image is an antelope*. Correspondingly, H can be *The object in the image is Klipspringer*. As a result, we directly utilize the prediction of the image and the concept as $P(E)$, where each E_i corresponds to a $P(E_i)$.

The contribution of each evidence E on the conclusion H is different. For example, *The object in the image is an antelope* provides more information than *The object in the image is an animal* for judging the image matching *Klipspringer* due to its narrower scope. To measure this contribution, we define $Con(E_i, H)$ for each E_i as follows:

$$Con(E_i, H) = \begin{cases} \frac{\frac{1}{log(num)} - \frac{1}{ents}}{1 - \frac{1}{ents}} & \text{if } num \geq n \\ 1 & \text{if } num < n \end{cases} \tag{7}$$

where num denotes the number of entities that contain this concept, $ents$ denotes the number of all the entities, and n is the base of log for scaling. (We use 10 in this paper.) Notably, the contribution is based on the distribution of entities and concepts in the test set, independent of the training process.

Fianlly, the $P(H)$ is calculated as:

$$P(H) = \frac{1}{n} \sum_{i=1}^{n} P(E_i) \cdot Con(E_i, H) \tag{8}$$

In this equation, n denotes the number of concepts of the entity. $P(H)$ represents the probability of the conclusion H, and we utilize the threshold to determine whether the conclusion H is classified as $True$ or $False$.

3.4 Human Verification

Because images of long-tailed entities are rare and valuable, the method also supports human verification to preserve more correct images. The scarcity of visual representations for long-tailed entities makes it challenging for annotators to assess the relevance of images directly. However, evidence from EVIDENCE FUSION helps overcome this challenge. COG uses EVIDENCE FUSION to review images discarded in CONCEPT INTEGRATION and offer explanations as evidence. For example, as shown in Figure 3, it may be difficult for humans to determine if an image depicts a *Klipspringer*. However, providing evidence such as *The image matches a mammal* and *The image matches an antelope* greatly assists humans in making more accurate recognition.

4 Experimental Setup

4.1 Data Collection

To address the absence of a suitable dataset for long-tailed image-text recognition, we employ a rule-based method that uses entity linking to accurately identify images of long-tailed entities. Using

Statistic	Number
Total Entities	25,166
BLC Concepts	1,278
Total Concepts	10,702
Average BLC Concepts per Entity	2.78
Average Concepts per Entity	4.45

Table 1: Statistics of the dataset.

Method	Precision	Recall	F1
Entity Linking	98	62	75

Table 2: The results of the entity linking method for finding matching images.

this dataset, we evaluate our method on various downstream tasks and offer a detailed analysis.

Although some long-tailed image classification datasets exist (Liu et al., 2019b; Cui et al., 2019), they are not suitable for our tasks. The long tail of these datasets is designed for model training rather than representing genuine scarcity. To tackle this, we select long-tailed entities from an actual KG CN-DBpedia (Xu et al., 2017). Using entity linking (Chen et al., 2018), we collect pertinent images for these entities. As a result, we obtain a dataset with 25,166 image-text pairs of long-tailed entities and convert these entity names into English.

Selection of Long-Tailed Entities To identify long-tailed entities, we analyze the distribution of entities in CN-DBpedia, focusing on a property called *viewtimes* that reflects their click frequency. We randomly select 100 entities from the knowledge graph and further examine their *viewtimes*, as shown in Figure 1. Our observations reveal a positive correlation between an entity's *viewtimes* and the number of its images. Thus, we classify entities with *viewtimes* below 100,000 as long-tailed entities, which typically have few or no images available online.

Grounding Long-Tailed Entities through Entity Linking In response to the inaccurate recognition of PVLMs for long-tailed entities, we use the entity linking approach to find correct images, as depicted in Figure 4. Initially, we search for entity names using a search engine. Following that, we employ short text entity linking (Chen et al., 2018) on the caption of the top search result image to pair it with the target entity. If the entity name is included in the linking results, the image is considered a match. We pick 100 entities with fewer than 100,000 *viewtimes*, search for images through Google, and manually annotate whether the first image corresponds to the entity. As shown in Table 2, our strategy yields high precision, allowing us to accurately generate a dataset of image-text pairs of long-tailed entities.

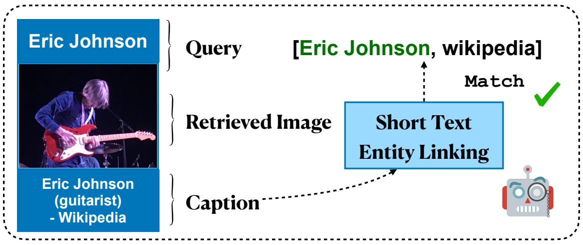

Eric Johnson
Query
[Eric Johnson, wikipedia]
Match
Retrieved Image
Short Text Entity Linking
Eric Johnson (guitarist) - Wikipedia
Caption

Figure 4: The process of obtaining correct images through short text entity linking.

Statistics We use CN-Probase (Chen et al., 2019), a comprehensive Chinese concept graph, to gather concepts related to entities from CN-DBpedia. Table 1 shows that our dataset contains approximately 10k concepts for 25k entities, averaging 4.45 concepts per entity. Although BLC concepts represent a small portion, each entity has about 3 BLC concepts on average, demonstrating the abundance and prevalence of BLC concepts.

4.2 Settings

Metrics For classification, we evaluate the performance of COG using accuracy, precision, recall, and F1 score. For ranking, we use metrics like Mean Reciprocal Rank (MRR), which is the average of the reciprocal ranks of the first correct answer, Mean Rank (MR), the average rank of the correct answer, and Hit@k, the percentage of correct candidates in the top-k predictions of the model in the test set. We set k to 1, 5, and 10 in our experiments.

Model Choice We conduct COG on three PVLMs, including CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and BLIP (Li et al., 2022).

Datasets for Downstream Tasks Firstly, the image-text pairs gathered are separated into training, validation, and test sets using an 8:1:1 ratio. This results in 20,132 training, 2,517 validation, and 2,517 test samples. Each piece of training data follows the $(entity, image, label)$ structure, with all labels being $True$.

In terms of ranking, the validation and test sets include samples with one entity and 50 candidate

Models	MR ↓	MRR ↑	Hit@1 ↑	Hit@5 ↑	Hit@10 ↑
CLIP	13.22	27.10	15.45	27.25	52.36
w/ *Stage1*	**5.51**	**50.14**	**33.65**	**58.72**	**84.51**
ALIGN	13.04	27.72	15.97	28.29	52.88
w/ *Stage1*	**5.47**	**49.81**	**33.73**	**57.37**	**84.74**
BLIP	14.21	21.09	8.34	21.37	49.30
w/ *Stage1*	**7.04**	**38.00**	**19.39**	**46.60**	**77.91**

Table 3: Results for the ranking task. *Stage1* represents CONCEPT INTEGRATION in our framework.

Models	Accuracy	Precision	Recall	F1
CLIP	67.44	62.37	88.37	73.13
w/ *Stage1*	83.63	81.67	87.10	84.30
w/ *Stage1+2*	**83.87**	80.92	88.64	**84.60**
ALIGN	68.12	63.12	89.38	73.99
w/ *Stage1*	**83.19**	77.82	92.84	84.67
w/ *Stage1+2*	83.13	77.84	92.67	**84.68**
BLIP	68.55	61.58	91.30	71.30
w/ *Stage1*	79.41	76.61	84.70	80.45
w/ *Stage1+2*	**79.42**	76.42	85.10	**80.53**

Table 4: Results for the classification task. *Stage1* and *Stage2* repersents CONCEPT INTEGRATION and EVIDENCE FUSION in our framework respectively.

images. To conserve computational resources, we choose 49 negative samples randomly, following the methods of (Teru et al., 2020; Zha et al., 2022); only one image is correct.

For classification, we add an equal number of negative samples to the validation and test sets by replacing images from different entities. As a result, the classification dataset contains 20,132 training samples, and each of the validation and test sets contains 5,034 samples.

Implementation Details We perform our experiments with a single RTX3090 GPU. The batch sizes are set to 64 for CLIP (Radford et al., 2021), 4 for ALIGN (Jia et al., 2021), and 16 for BLIP (Li et al., 2022). We use the AdamW optimizer with a learning rate of 1e-5.

For the classification task, we apply a threshold of 0.5 to decide if an image corresponds to a text, since the $prediction$ varies from 0 to 1.

5 Results and Analysis

5.1 Main Results

Table 3 compares the performance with and without concept guidance in the ranking task. Incorporating CONCEPT INTEGRATION (*Stage1*) significantly enhances the performance of all PVLMs, highlighting the importance of integrating concepts. Table 4 shows performance in the classification task under three conditions: without concepts, with only CONCEPT INTEGRATION, and with both CONCEPT INTEGRATION and EVIDENCE FUSION. CONCEPT INTEGRATION considerably improves recognition accuracy through the appropriate integration of concepts. Additionally, EVIDENCE FUSION, aimed at providing explainability, further enhances performance.

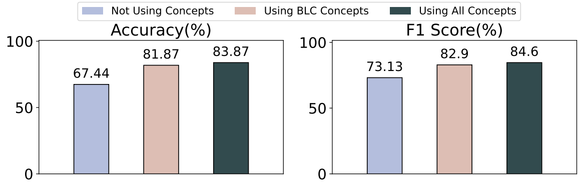


Figure 5: Comparison of using different concepts in our framework. *Not Using Concepts* represents using only entity names. *Using BLC Concepts* and *Using All Concepts* represents using BLC and all concepts respectively.

Experimental results show that integrating concepts allows PVLMs to more effectively align image and text modalities. PVLMs associate images with various concepts related to entities, not just the names of entities, during pre-training. CONCEPT INTEGRATION improves the recall of knowledge gained in pre-training. However, relying solely on this method is inadequate due to its black-box nature. As a result, we introduce the EVIDENCE FUSION module, which breaks down the recognition process into multiple pieces of evidence for the conclusion. This decomposition maintains effective performance while also making it more convincing and explicit.

Notably, Tables 3 and 4 demonstrate the superior performance of our method across different models, showcasing its flexibility. All models can determine whether a piece of text and an image match are suitable for integration into our method. Therefore, our method is model-pluggable and supports the replacement of different PVLMs or other methods used for image-text matching, significantly enhancing its transferability.

5.2 Analysis

How do different concept selections affect performance? To explore how concept selection affects results, we employ the same method but with different concepts. Figure 5 shows the influence of using various concepts on the classification task.

Methods	Accuracy	Precision	Recall	F1
Search Engine	61.11	61.11	100	75.77
CLIP	73.33	73.13	89.09	80.33
w/ *Stage1+2*	**81.11**	82.76	87.27	**84.96**

Table 5: *Search Engine* indicates entity grounding via search engine. CLIP represents not using concepts. *Stage1* and *Stage2* repersents CONCEPT INTEGRATION and EVIDENCE FUSION in our framework respectively.

Methods	Accuracy	Precision	Recall	F1
COG	80.00	76.78	86.00	81.13
Human Verification	75.00	68.38	93.00	78.81
+ Evidence	**83.00**	77.50	93.00	**84.54**

Table 6: COG shows the performance of our two-stage framework. *Human Verification* shows results from COG with human verification, and *Evidence* denotes human verification with extra evidence from EVIDENCE FUSION.

It indicates that using Basic-level Categorization (BLC) concepts improves performance, but incorporating all concepts leads to optimal outcomes. This suggests that BLC concepts are effective for identifying unfamiliar entities, and fine-grained concepts are also crucial as PVLMs can utilize detailed concept knowledge. Thus, we decide to include all concepts in our main experiments. However, Table 1 and Figure 5 also demonstrate that BLC concepts, being more common, still offer competitive performance. Considering that fine-grained concepts may be rare and difficult to collect, as highlighted by some studies (Li et al., 2021; Yuan et al., 2022, 2023), BLC concepts present a practical alternative.

How does the performance of the method measure up against traditional approaches? To show the comparison, we select 100 long-tailed entities and annotate whether the first searched image and the entity match. We then employ a trained CLIP model, using only entity names and our concept-guided framework COG. As shown in Figure 5, conventional methods that depend on direct search engine may attain a 100% recall rate, but suffer from low precision, resulting in unsatisfactory F1 scores. On the other hand, using CLIP demonstrates enhancements, stressing the importance of PVLMs in entity grounding. Our COG further amplifies the performance of PVLMs by incorporating concepts, underlining the method's efficiency and the significant advantage gained from integrating concepts.

How does EVIDENCE FUSION support human verification? Due to the scarcity of images of less common entities, we choose not to discard images labeled as incorrect by CONCEPT INTEGRATION, but rather to use EVIDENCE FUSION for re-judging. In Figure 6, two entities are named *Alexander Hamilton*. When the goal is to find an image of the musician *Alexander Hamilton* but an image of the politician with the same name is retrieved instead, EVIDENCE FUSION helps clarify this error. It shows that the evidence *The person in the image is a man* is true, but *The person in the image is a musician* and *The person in the image is an English actor* are false. This evidence explains why the image does not match the musician *Alexander Hamilton*, aiding in the identification of mismatches and supporting human annotators in their verification work, especially for less common entities where direct judgment is challenging.

To further investigate the role of evidence in human annotation, we conduct a test with 200 image-text pairs. These are evenly divided into positive and negative samples, and a two-stage classification method is applied. For samples identified as mismatches, we engage five annotators to review these mismatches. The accuracy and F1 scores are recalculated after this annotation. As indicated in Table 6, our findings demonstrate that explainability significantly enhances the verification process, highlighting the value of evidence in recognizing unfamiliar entities.

Is our method general and robust? We randomly select 100 common entities (with *viewtimes* over 1,000,000, in contrast to long-tailed entities with *viewtimes* under 100,000), using the best checkpoint trained in Table 4. As illustrated in Figure 7, COG is beneficial for both long-tailed and common entities. This proves that concept

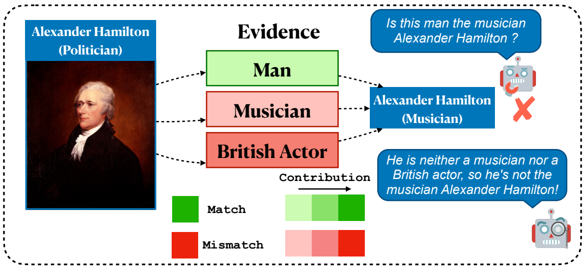

Alexander Hamilton (Politician)
Evidence
Man
Musician
British Actor
Alexander Hamilton (Musician)
Is this man the musician Alexander Hamilton ?
He is neither a musician nor a British actor, so he's not the musician Alexander Hamilton!
Contribution
Match
Mismatch

Figure 6: The process of recognizing a long-tailed entity with evidence.

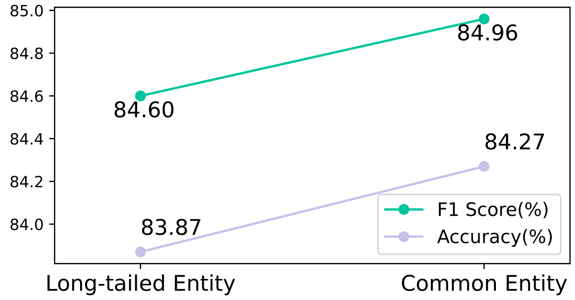


Figure 7: Comparison of our method on different entities.

guidance brings a generally effective enhancement, aligning with human cognition.

Our framework gains robustness through concept aggregation. Incorporating concepts is beneficial but may introduce noise, particularly when PVLMs struggle with fine-grained concepts. The more detailed a concept is, the closer it is to the entity, which complicates recognition. We mitigate noise by aggregating multiple concepts. During CONCEPT INTEGRATION, we concatenate all the concepts, and in EVIDENCE FUSION, we synthesize all evidence to draw conclusions. In both modules, we find that fine-grained concepts offer improvements over using only BLC concepts, indicating that aggregation effectively reduces noise.

6 Conclusion

We propose a two-stage framework COG, using PVLMs with concept guidance to ground long-tailed entities in MMKGs effectively. Our experimental results demonstrate that COG greatly improves the ability of PVLMs to recognize image-text pairs of long-tailed entities. Furthermore, recognizing unfamiliar entities through concepts is convincing and provides clear evidence for human verification, suggesting a future direction for better handling long-tailed entity grounding.

Limitation

Throughout our method, we utilize concepts from CN-Probase, which contains noise. Both the quantity and quality of these concepts play a crucial role in determining the performance of our method. Exploring alternative concept generation methods can serve as a potential research question for future research. The improvement of concepts in the future is expected to contribute to the enhancement of our methods for more accurate long-tailed entity grounding.

Ethical Considerations

We hereby acknowledge that all authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct.

Use of Human Annotations All raters have been paid above the local minimum wage and consented to use the evaluation dataset for research purposes in our paper. Human annotations are only utilized in the methodological research stages to assess the proposed solution's feasibility. To guarantee the security of all annotators throughout the annotation process, they are justly remunerated according to local standards.

Risks The datasets used in this paper are obtained from public sources and anonymized to protect against any offensive information.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work is supported by National Natural Science Foundation of China (No. 62102095), and Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902). The computations in this research are performed using the CFFF platform of Fudan University.

References

Houda Alberts, Teresa Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2020. Visualsem: a high-quality knowledge graph for vision and language. *arXiv preprint arXiv:2008.09150*.

Jindong Chen, Ao Wang, Jiangjie Chen, Yanghua Xiao, Zhendong Chu, Jingping Liu, Jiaqing Liang, and Wei Wang. 2019. Cn-probase: a data-driven approach for large-scale chinese taxonomy construction. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1706–1709. IEEE.

Lihan Chen, Jiaqing Liang, Chenhao Xie, and Yanghua Xiao. 2018. Short text entity linking with fine-grained topics. In *Proceedings of the 27th ACM International conference on Information and Knowledge Management*, pages 457–466.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022a. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. 2022b. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Sebastián Ferrada, Benjamin Bustos, and Aidan Hogan. 2017. Imgpedia: a linked dataset with content-based analysis of wikimedia images. In *The Semantic Web–ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part II 16*, pages 84–93. Springer.

Jingyi Hou, Xinxiao Wu, Yayun Qi, Wentian Zhao, Jiebo Luo, and Yunde Jia. 2019. Relational reasoning using prior knowledge for visual captioning. *arXiv preprint arXiv:1906.01290*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Chenguang Li, Jiaqing Liang, Yanghua Xiao, and Haiyun Jiang. 2021. Towards fine-grained concept generation. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):986–997.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. 2019a. Mmkg: multi-modal knowledge graphs. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, pages 459–474. Springer.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. 2019b. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121.

James McElvenny. 2014. Ogden and richards' the meaning of meaning and early analytic philosophy. *Language Sciences*, 41:212–221.

Daniel Oñoro-Rubio, Mathias Niepert, Alberto García-Durán, Roberto González, and Roberto J López-Sastre. 2017. Answering visual-relational queries in web-extracted knowledge graphs. *arXiv preprint arXiv:1709.02314*.

Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding multimodal relational data for knowledge base completion. *arXiv preprint arXiv:1809.01341*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Komal Teru, Etienne Denis, and Will Hamilton. 2020. Inductive relation prediction by subgraph reasoning. In *International Conference on Machine Learning*, pages 9448–9457. PMLR.

Meng Wang, Haofen Wang, Guilin Qi, and Qiushuo Zheng. 2020. Richpedia: a large-scale, comprehensive multi-modal knowledge graph. *Big Data Research*, 22:100159.

Zhongyuan Wang, Haixun Wang, Ji-Rong Wen, and Yanghua Xiao. 2015. An inference approach to basic level of categorization. In *Proceedings of the 24th acm international on conference on information and knowledge management*, pages 653–662.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cn-dbpedia: A never-ending chinese knowledge extraction system. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 428–438. Springer.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Siyu Yuan, Deqing Yang, Jiaqing Liang, Zhixu Li, Jinxi Liu, Jingyue Huang, and Yanghua Xiao. 2022. Generative entity typing with curriculum learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3061–3073, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siyu Yuan, Deqing Yang, Jinxi Liu, Shuyu Tian, Jiaqing Liang, Yanghua Xiao, and Rui Xie. 2023. Causality-aware concept extraction based on knowledge-guided prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9255–9272, Toronto, Canada. Association for Computational Linguistics.

Hanwen Zha, Zhiyu Chen, and Xifeng Yan. 2022. Inductive relation prediction by bert. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5923–5931.

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*.