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

Anonymous ACL submission

Abstract

Multi-Modal Knowledge Graphs (MMKGs) are a type of Knowledge Graph (KG) that integrates information from various modalities and holds significant application value. However, the construction of MMKGs often introduces mismatched images (*i.e.*, noise). Due to the power-law distribution of images on the internet for entities, a large number of long-tail entities have very few images. Existing methods struggle to accurately identify images of long-tail entities. To address this issue, we draw inspiration from the Triangle of Reference theory and propose to enhance the pre-trained visual-language models with concepts. Specifically, we propose a two-stage framework containing two modules, i.e., Concept Integration and Evidence Fusion. The Concept Integration module aims to accurately recognize image-text pairs associated with long-tail entities, thereby improving MMKG quality. Additionally, our Evidence Fusion module can provide explainability regarding the results, which facilitates human verification, further enhancing long-tail entity grounding. Finally, we construct a dataset of 25k image-text paris of long-tail entities. Comprehensive experiments show our method outperforms the baseline, achieving an average increase of about 20% in Mean Reciprocal Rank (MRR) in the ranking task and approximately 85% in F1 in the classification task.

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 multi-modal intelligence and extensive knowledge in various applications, such as visual question answering (Marino et al., 2021) and image captioning (Hou et al., 2019), MMKGs have received increasing attention in recent years.

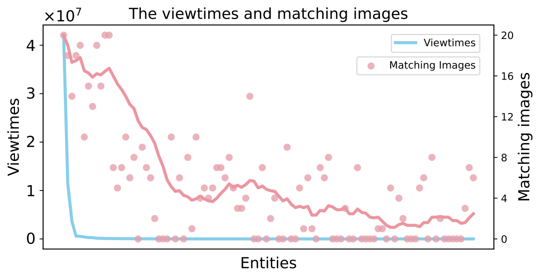

Figure 1: We randomly select 100 entities from a large-scale KG CN-DBpedia (Xu et al., 2017) and make human annotations. The blue line represents the change in the entity's *viewtimes*, which reflects the click frequency of the entity. The red points represents the number of correct matches for the entity in the top 20 results for a Google search for images of the entity and the red line is the result of smoothing the data points.

While the quantity of images within current MMKGs has steadily increased, these collections have restricted coverage and accuracy, particularly concerning less common entities (*i.e.*, long-tail entities). As shown in Figure 1, the change trends of the entity's *viewtimes* and the number of found matching images are roughly the same, and both show a long-tail distribution. It demonstrates the long-tail entities have rare images.

In constructing Multi-Modal Knowledge Graphs (MMKGs), aligning long-tail entities with proper images (*i.e.*, entity grounding (Zhu et al., 2022)) is important. First, incorporating images for long-tail entities significantly bolsters the completeness and breadth of knowledge graphs, ensuring a more comprehensive coverage across diverse subjects and domains. Second, the integration of visual content for long-tail entities in MMKGs enhances user engagement and efficiency in information retrieval, particularly beneficial in visually-driven learning and search contexts. Third, the pairing of images with long-tail entities can serve as valuable training data, aiding in the development and refinement of robust vision-language models, especially for less common or domain-specific entities.

However, grounding long-tail entities in MMKG is non-trivial. Existing methods for entity grounding (Wang et al., 2020; Oñoro-Rubio et al., 2017; Liu et al., 2019a) primarily rely on web resources, particularly search engines. These methods gather images by matching strings in image captions with entity names and then ranking them based on click frequency. While these methods prove effective for widely-recognized entities (Liu et al., 2019a; Wang et al., 2020), they face challenges with long-tail entities. Specifically, existing methods have several limitations: (1) Search engines are used for text matching, but entity grounding involves the matching of images and texts. (2) Although pre-trained vision-language models (PVLMs) like CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) have shown impressive performance in various cross-modal tasks, they encounter challenges in identifying long-tail entities due to their infrequent appearance during pre-training. (3) None of these methods can explain why one image is chosen over another, which is crucial for further human verification.

To tackle above challenges, we design a holistic and explainable two-stage framework aiming at enabling PVLMs to effectively leverage concepts for long-tail entity grounding.

First, inspired by the Triangle of Reference theory shown in Figure 2, we use concepts to guide PVLMs in accurately identifying images of long-tail entities. Second, we analyze the impact of the selection of different concepts on results. Third, our two-stage framework contain an Evidence Fusion module that can provide envidences for results. When introducing human verification, these evidences can significantly improve labeling accuracy.

To sum up, the contributions of this paper are as follows:

- We introduce concept guidance to enhance PVLMs' ability to recognize images of long-tail entities and develop an effective two-stage framework for incorporating concepts.
- We compare and analyze the impact of selecting different concepts on experimental results
- Our extensive experiments show that our method can effectively improve the accuracy of long-tail entity grounding and also offers explanation, which is benefical for human verification to further improve performance.

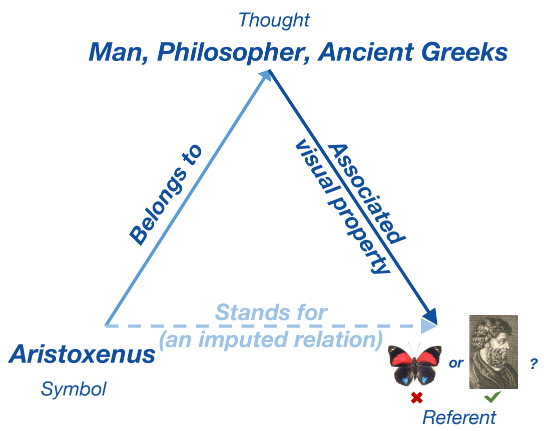

Figure 2: The Triangle of Reference theory (McElvenny, 2014) use the triangle's three vertices represent *symbol*, *thought* and *referent*. *Thought* build an bridge between *symbol* and *referent*. This figure illustrates that for an entity named *Aristoxenus*, the search engine retrieves two images. Deciding which one is correct based solely on the entity name is challenging because we don't know *Aristoxenus*. However, by utilizing concepts, we can determine that the target *Aristoxenus* refers to a person, not a butterfly.

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 sources, such as text, images, and audio, into a single interconnected graph. Existing methods for entity grounding in MMKGs fall into two main categories: *1) Methods based on online encyclopedias* (Ferrada et al., 2017; Alberts et al., 2020): which link existing encyclopedic multimedia resources (Wikimedia Commons, Wikipedia, ImageNet (Deng et al., 2009)) associating texts with images to construct MMKGs. *2) Methods based on web search engines* (Oñoro-Rubio et al., 2017; Wang et al., 2020; Liu et al., 2019a): which directly search for images of entities using web search engines. This method is more flexible than using online encyclopedic multimedia data, and it allows for expansion based on existing filtered and refined KGs. However, due to the constraint that entities and the associated images follow a power-law distribution shown in Figure 1, these works often focus on popular entities. Since long-tail entities may not have images on the web and the search engine always give ranked results although all the images are mismatched, it is easy to return wrong images, thus introducing noise. Considering that certain long-tail entities may have few available images, we propose a framework that leverages concepts for reducing the noise and provides explanation for further huamn verification.

2.2 Pre-trained Vision-Language Models (PVLMs)

Pre-trained Vision-Language Models (PVLMs) are models pre-trained on cross-modal data and can process visual and textual information simultaneously. PVLMs aim to align image and text data through large-scale cross-modal pre-training. CLIP (Radford et al., 2021) employs a self-supervised approach, leveraging a dataset of 400 million image-text pairs collected from the internet. This vast dataset significantly enhances alignment across diverse modalities. ALIGN (Jia et al., 2021), on the other hand, adopts a dual-encoder structure and a notably larger dataset, consisting of over a billion image-text pairs. In contrast, BLIP (Li et al., 2022) takes a different approach by filtering out poor-quality data to further optimize the performance of multi-modal tasks.

2.3 Long-Tail Classification

Some researchers in the field of computer vision focus on the long-tailed image classification problem. Various datasets (Liu et al., 2019b; Cui et al., 2019) are employed to assess the capability of learning classification with limited samples. However, our objective diverges from the traditional image classification. Rather than determining image labels, we aim to determine whether an image match a specific entity.

3 Problem Definition

Multi-Modal Knowledge Graph(MMKG) is a type of knowledge graph in which nodes can be entities or images and edges represent the relationships between them. The triplets in MMKG can be defined as $(e, \textit{has image}, i)$, where e denotes the textual entity, i denotes its matching image, thus their relationship can be represented as *has image*.

To match images for entities in MMKG (i.e. entity grounding in MMKG), existing 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 images i. Then, they select the top-n images and classify whether the image is related to the given entity, which can be modeled as a **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*.

4 Concept Selection

To figure out what concepts are suitable for concept guidance, we explore the influence of employing various concepts in this section. An entity often contains multiple concepts, and these concepts have different granularities. As suggested by (Wang et al., 2015), humans comprehend the world by Basic-level Categorization (BLC), which refers to a mid-level concept that people tend to use in daily cognition.

Motivated by this, we compare the performance under BLC concepts and all concepts. Specifically we treat concepts consisting of only one word as Basic-level Categorization (BLC) concepts and compare the performance of using BLC concepts and using all concepts. The experiments in Sec 6 demonstrate how different concept selection strategies impact the results.

5 Concept-guided Method

Incorporating concepts is not easy, in order to ensure both effectiveness and explainability, we design an two-stage framework, as illustrated in Figure 3.

During training, we calculate contrastive losses at both entity and concept levels. Then we fine-tune the model through this loss, and the fine-tuned model is used in the inference stage. When inferencing, we use the trained model as a part of our two-stage framework to predict. The framework contains two modules Concept Integration and Evidence Fusion. Concept Integration directly concatenates entities and concepts to enhance PVLM. The prediction from Concept Integration can be used to ranking and classification. Evidence Fusion mainly processes those pairs that the predictions is not true, because images of long-tail entities are rare and valuable. Evidence Fusion can provide evidence by separately predicting whether each concept matches the image and the evidence is useful for human verification.

5.1 Contrastive Learning on Two Levels

During the training of the Pre-trained Vision-Language Model (PVLM), we designate a text as *t* and an image as *i*. We first input both *t* and *i* into the PVLM. The model then produces a *prediction*, indicating the degree of alignment between *t* and *i*, as shown below:

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

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

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

In this equation, *t* and *i* represent the text and image inputs, respectively. The *prediction* value, ranging between [0, 1], indicates the model's prediction of the match between the image and the text. If the *prediction* exceeds 0.5, we consider it a match; otherwise, it is considered a mismatch.

Next, we train the model using contrastive learning with in-batch negative samples. In each batch, there are *n* samples, where *n* denotes the batch size. Each sample is a pair (*t*, *i*), representing a text and an image. As illustrated in Figure 3, we formulate contrastive samples at both entity and concept levels. We let t_i be the concatenation of the i-th entity and all concepts of the entity as:

$$t_i = e_i, c_1, c_2 \ldots \tag{4}$$

where *e* represents an entity, and *c* represents a concept.

At the entity level, we use p_{t_a,i_b} to represent the *prediction* of the concatenation of the i-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:

$$L_{entity} = -\sum_{a=1}^{n}\sum_{b=1}^{n} BCE(l_{a,b}, p_{t_a,i_b}) \tag{5}$$

where BCE is 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_{akb} is 0. The concept loss $L_{concept}$ loss is calculated as:

$$L_{concept} = \tag{6}$$

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

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

$$L = L_{entity} + L_{concept} \tag{7}$$

5.2 Concept-Guided Image-Text Cognition

As illustrated in Figure 3, we design a two-stage framework for incorporating concepts.

5.2.1 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 resulting concatenated text would be *"Jay Chou,singer,actor,director"*.

While Concept Integration improves performance in experiments, it acts as a black-box model lacking explanatory capability. Additionally, images of long-tail entities are scarce. The black-box approach's prediction lack credibility, potentially causing errors or the loss of correct images. Therefore, we introduce Evidence Fusion to re-judge the samples whose prediction is less than 0.5 in Concept Integration.

5.2.2 Evidence Fusion

In Concept Integration, we leverage the generalization capability of PVLM, enabling the model to effectively recognize a subset of long-tail entities. During Concept Integration, PVLM produces a prediction value *prediction*. If *prediction* exceeds 0.5, we regard the text and image matching. If *prediction* is below 0.5, we proceed to Evidence Fusion, where we apply our Evidence Fusion method for re-judgement.

For a more comprehensive understanding of Evidence Fusion, we 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, *P(E, H)* is defined to evaluate the influence of evidence *E* on conclusion *H*.

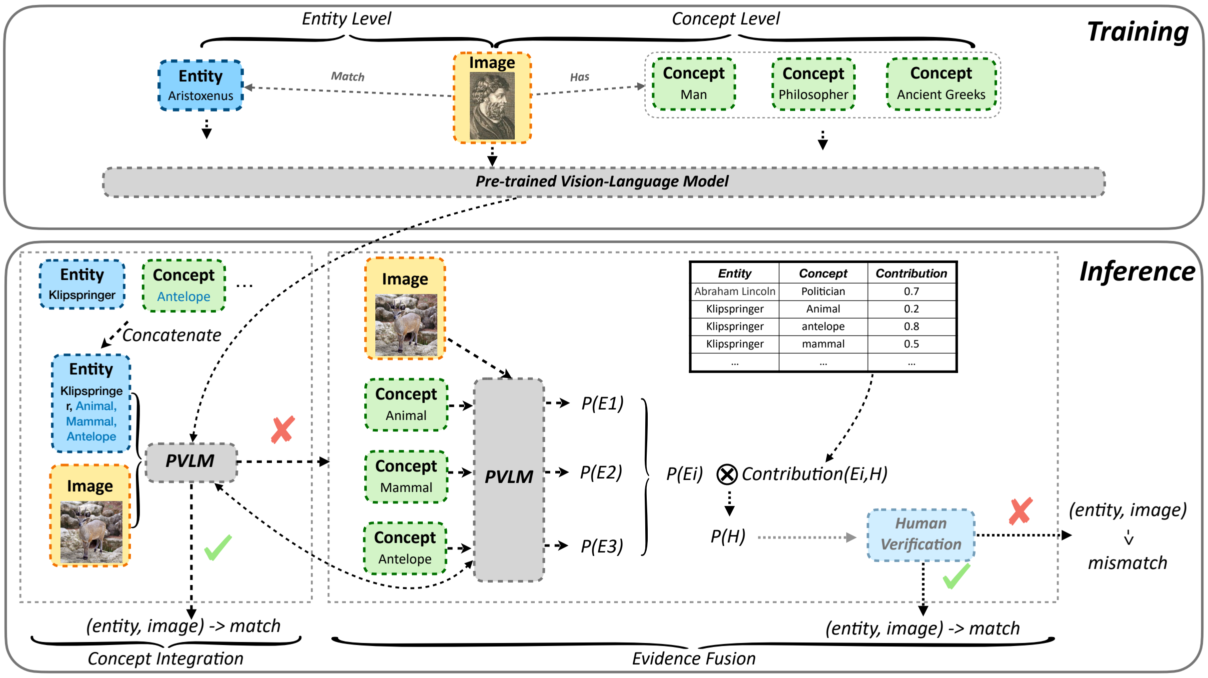

Figure 3: Our framework. During training, we follow the way of contrastive learning to generate samples. In inference, we initially concatenate entities and concepts and then input the concatenated text and images directly into the PVLM in Concept Integration. If the prediction is *False*, proceed to Evidence Fusion. In Evidence Fusion, we calculate the weighted average of predictions and contribution for each concept and image.

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

In Evidence Fusion, essentially, we fundamentally transform the task of matching an entity and an image into a comprehensive analysis of the matching between concepts and the image. In Figure 3, E_i represents an image matching a concept. For example, we can 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 influence of each evidence E on the conclusion H is different. For example, *"Be an antelope"* provides more information than *"Be an animal"* for judging the image matching *"Klipspringer"* due to its narrower scope. To measure this influence, we define $CF(E_i, H)$ for each E_i as follows:

$$P(E_i,H) = \begin{cases} \frac{\frac{1}{log(num)} - \frac{1}{ents}}{1 - \frac{1}{ents}} & \text{if } num \geq 10 \\ 1 & \text{if } num < 10 \end{cases} \quad (8)$$

where *num* denotes the number of entities which contain this concept, *ents* denotes the number of all the entities and the base of *log* for scaling is 10.

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

$$P(H) = \frac{1}{n} \sum_{i=1}^{n} P(E_i) \cdot P(E_i, H) \quad (9)$$

In this equation, n denotes the number of concepts of the entity. *P(H)* represents the probability of the conclusion H and we utilize a threshold of 0.5 to determine whether the conclusion H is classified as *True* or *False*.

5.3 Human Verification

Because images of long-tail entities are very valuable, we introduce human verification to further improve the recall rate. In our method, Evidence Fusion repredicts images discarded in Concept Integration and generates evidence as explanations. Due to the rarity of visual representations of long-tail entities on the Internet, it is challenging for annotators to directly determine if an image matches a long-tail entity. However, the evidence generated in Evidence Fusion effectively compensates for this limitation. So we provide the evidence to aid human verification and the experiments highlights the importance of evidence.

6 Experiments

Because of the lack of suitable long-tail entity image-text pair datasets, we construct a new dataset containing 25k long-tail entities. Based on this dataset, we conduct two different downstream tasks to prove the effectiveness of our framework. We also show that human verification with evidence can further improve the accuracy of entity grounding.

6.1 Data Collection

Although there are some long-tail image classification datasets (Liu et al., 2019b; Cui et al., 2019), all of them have limitations. Because these datasets are often assembled from the web and the image resources are usually rich, the *long-tail* is for model training rather than real scarcity. However, for our research, we need entities with extremely scarce images. To meet this requirement, we choose long-tail entities from an actual Knowledge Graph (KG). Because in a real KG, many long-tail entities face difficulty finding corresponding images on the web, creating a scenario where PVLMs have not encountered such data during pre-training.

Consequently, we first collect long-tail entities from CN-DBpedia (Xu et al., 2017), a large-scale structured knowledge graph with millions of entities. To ground these entities, we then use entity linking (Chen et al., 2018) to collect relevant images from the internet. Finally, we obtain a dataset with 25,166 image-text pairs of long-tail entities and translate them to english.

6.1.1 Selection of Long-Tail Entities

For obtaining long-tail entities, we analyze the distribution of entity images and we find that entities in CN-DBpedia have a property called *viewtimes*, indicating their click frequency on the web.

To further investigate, we randomly select 100 entities from the knowledge graph and analyze their *viewtimes*, as shown in Figure 1. We find that there's a positive correlation between an entity's *viewtimes* and the quantity of its images on the internet. Therefore, we choose entities with *viewtimes* under 100,000 as long-tail entities.

6.1.2 Grounding Long-tail Entities through Entity Linking

To address the lack of images for long-tail entities, we use the entity linking method to find appropriate images, as depicted in Figure 4. First, we search for entity names from CN-DBpedia using

$Precision(\%)$	$Recall(\%)$	$F1(\%)$
98	62	75

Table 1: The results of using the entity linking method to determine 100 long-tail entity image-text pairs.

a search engine. Then, we apply short text entity linking (Chen et al., 2018) to the caption of the first search result image to link it with the relevant entity. If the entity name is in the linking results, we consider the image to be a match for the entity.

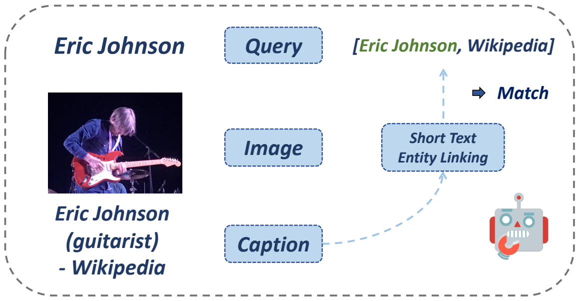

Figure 4: The process of obtaining an accurate image through short text entity linking (Chen et al., 2018). The entity linking method can establish a connection between a piece of text and the entity within CN-DBpedia. If the linking result includes the queried entity, the image is matching.

We select 100 entities with viewtimes under 100,000, search for image using the Google search engine, and manually annotate whether the first image matches the entity. These images are used to assess the entity linking method, with the results shown in Table 1. The results indicate that our method achieves a high accuracy rate of 98%, enabling us to create a dataset of image-text pairs for long-tail entities with great precision.

We split the dataset into training, validation, and test sets in an 8:1:1 ratio, yielding 20,132 training, 2,517 validation, and 2,517 test samples. Each training sample follows the (entity, image, label) format, with all labels set to 1. For the ranking task, both validation and test sets contain samples with an entity and 50 candidate images, only one of which is correct. For the classification task, we expand the validation and test sets with an equal number of negative samples by replacing the image in a sample with one from a different entity. As a result, our classification dataset includes 20,132 training samples, 5,034 testing samples, and 5,034 validation samples.

Statistics	Quantity
Entities	25166
BLC Concepts	1278
Concepts	10702
BLC Concepts per entity (Avg)	2.78
Concepts per entity (Avg)	4.45

Table 2: Number of entities, number of concepts, and average number of concepts per entity in the dataset.

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: Stage1 repersents Concept Integration in our framework. We compared the effects of three PVLMs on ranking tasks, and the results show the advantages of our method.

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 separately.

6.1.3 Statistical Analysis

We use CN-Probase (Chen et al., 2019) to obtain the concepts related to the entities. CN-Probase is a comprehensive Chinese concept graph with about 17 million entities, 270,000 concepts, and 33 million isa relations.

Then, we conduct a statistical analysis of the entity concepts, with the results shown in Table 2. There are about 10k concepts for 25k entities and each entity owns 4.45 concepts. BLC concepts accounts for only a small portion of concepts, but on average there are nearly 3 BLC concepts per entity.

6.2 Experiment Setup

We conduct our experiments using a single RTX3090 GPU and set the batch size 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 AdamW optimizer and set the learning rate as 1e-5.

Metrics For classification, we evaluate the performance of our model using accuracy, precision, recall and F1 score. For ranking, we use various metrics, including Mean Reciprocal Rank (MRR), Mean Rank (MR), and Hit@k metrics.

Models We conduct our framework on 3 PVLMs, including CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and BLIP (Li et al., 2022).

6.3 Results

Concept Selection Table 5 shows the performance of using different concepts in the classification task, the results show that using all concepts is 1.4% higher than using BLC concepts on f1, and using BLC concepts is 9.77% higher than using no concepts on f1, indicating that (1) BLC concepts are helpful for recognizing unfamiliar entities. (2) Some fine-grained concepts are equally important because PVLMs can capture the knowledge of fine-grained concepts. Richer concepts have a better performance on enhancing PVLMs so that we use all the concepts for other experiments.

Ranking Table 3 displays the performance of various models, including CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and BLIP (Li et al., 2022). We first compare results using PVLM to evaluate only entity names and images, without concepts. Then, we compare these with outcomes from the concept-guided approach (using only Concept Integration). Our method shows significant improvements in all evaluation metrics, notably a 20.68% average increase in Mean Reciprocal Rank (MRR). This highlights the effectiveness of our concept-guided method in accurately ranking the correct images.

Classification Table 4 reports performance across three settings: without using concepts, using only Concept Integration, and employing both Concept Integration and Evidence Fusion. The results show that incorporating concepts significantly boosts effectiveness, leading to an average accuracy rate increase of around 14% and an average F1 score increase of about 10%.

We observe that integrating concepts directly aids PVLMs in aligning image and text modalities.

Concepts	Accuracy	Precision	Recall	F1
Not using concepts	67.44	62.37	88.37	73.13
BLC concepts	81.87	78.20	88.20	82.90
All concepts	**83.87**	**80.92**	**88.64**	**84.60**

Table 5: *Not using concepts* represents using only entity names. Both *BLC concepts* and *All concepts* use

Methods	Accuracy	Precision	Recall	F1
ours	80.00	76.78	86.00	81.13
w/o *Evidence*	75.00	68.38	93.00	78.81
w/ *Evidence*	**83.00**	**77.50**	**93.00**	**84.54**

Table 6: In this table, *ours* represents the results from our method. *w/o Evidence* and *w/ Evidence* respectively represent the results after human verification without evidence and with evidence.

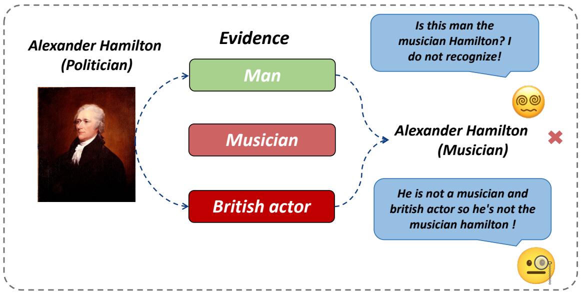

Figure 5: The light green color signifies that the evidence is true, indicating a match between the image and *"man"*. However, the contribution of this evidence is relatively low. On the other hand, the red color indicates that the images do not correspond to *"Musician"* and *"British Actor"*. These instances possess a higher discriminatory power and thus appear darker. By aggregating and comprehensively analyzing the aforementioned evidence, we can infer that the image is mismatching.

In the pre-training phase, PVLMs often associate images with a range of concepts related to entities, extending beyond entity names alone. Concept Integration improves the recall of knowledge acquired during pre-training. However, relying solely on this black box method is inadequate. Therefore, we introduced an Evidence Fusion module, utilizing concepts as evidence. This explicit imitation of the cognitive process maintains performance similar to the black-box method while crucially generating evidence, essential for human verification.

6.4 Explainability

Since images of long-tail entities are extremely rare, we do not readily discard images deemed incorrect by Concept Integration. Instead, we use Evidence Fusion to provide explanations. These explanations, consisting of evidence, significantly aid human judgment, as shown in Figure 5.

As shown in Figure 5, two entities share the name *"Alexander Hamilton"*. When aiming to ground an image for the musician *"Alexander Hamilton"* but accidentally retrieving an image of the politician with the same name, evidence fusion clarifies the mismatch. It indicates that while *"The person in the image is a man"* is true, *"The person in the image is a musician"* and *"The person in the image is an English actor"* are false. The evidence explains why the retrieved image does not match the musician *"Alexander Hamilton"*.

6.5 Human Verification

Evidence not only makes predictions more credible but also assists human annotators in verification. Since it's challenging for labelers to directly judge image-text pairs of long-tail entities, we provide explanations to assisit labelers.

Specifically, we select 200 image-text pairs with a 1:1 ratio of positive and negative samples. First, we use our two-stage method with fine-tuned CLIP to classify samples and calculate f1 score. Concurrently, Evidence Fusion outputs evidence for samples judged as mismatching. To prevent discarding potentially correct images, we hire annotators to relabel samples deemed mismatched. Following this, we recalculate the accuracy and F1 score and then compare the performance after annotation, both with and without evidence.

As Table 6 indicates, we engage five students as annotators and report the average score. The results demonstrate that the explainability provided by our method is necessary. For long-tail entity grounding tasks, human verification can be introduced when necessary to ensure the recall rate.

7 Conclusion

To ground long-tail entities effectively in a multimodal knowledge graph (MMKG), we propose a solution utilizing PVLMs with concept guidance. In order to ensure both effectiveness and explainability, we introduce a two-stage framework. We define two tasks that simulate the real-world entity grounding process, showcasing that our approach enhances results and provides explainability.

Limitation

Throughout our method, we utilize concepts from CN-Probase. 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 reaserch question for future research. The improvement of concepts in the future is expected to contribute to the enhancement of our methods for more accurate entity grounding.

Ethical Considerations

We provide details of our work to address potential ethical considerations.

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 early stages of methodological research to assess the feasibility of the proposed solution. To guarantee the security of all annotators throughout the annotation process, they are justly remunerated according to local standards. Human annotations are not employed during the evaluation of our method.

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

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.

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.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training 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*.

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.

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.

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*.