QR-CLIP: Introducing Explicit Knowledge for Location and Time Reasoning

WEIMIN SHI, State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China
DEHONG GAO, School of Cybersecurity, Northwestern Polytechnical University, Xi'an, China
YUAN XIONG, State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China
ZHONG ZHOU, Zhongguancun Laboratory, Beijing, China and State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China

This article focuses on reasoning about the location and time behind images. Given that pre-trained vision-language models (VLMs) exhibit excellent image and text understanding capabilities, most existing methods leverage them to match visual cues with location and time-related descriptions. However, these methods cannot look beyond the actual content of an image, failing to produce satisfactory reasoning results, as such reasoning requires connecting visual details with rich external cues (e.g., relevant event contexts). To this end, we propose a novel reasoning method, *QR-CLIP*, that aims at enhancing the model's ability to reason about location and time through interaction with external explicit knowledge such as Wikipedia. Specifically, *QR-CLIP* consists of two modules: (1) The *Quantity* module abstracts the image into multiple distinct representations and uses them to search and gather external knowledge from different perspectives that are beneficial to model reasoning. (2) The *Relevance* module filters the visual features and the searched explicit knowledge and dynamically integrates them to form a comprehensive reasoning result. Extensive experiments demonstrate the effectiveness and generalizability of *QR-CLIP*. On the WikiTiLo dataset, *QR-CLIP* boosts the accuracy of location (country) and time reasoning by 7.03% and 2.22%, respectively, over previous SOTA methods. On the more challenging TARA dataset, it improves the accuracy for location and time reasoning by 3.05% and 2.45%, respectively. The source code is at https://github.com/Shi-Wm/QR-CLIP.

CCS Concepts: • **Computing methodologies** → **Knowledge representation and reasoning**; **Visual content-based indexing and retrieval**; • **Information systems** → **Information retrieval diversity**;

Additional Key Words and Phrases: Multimodal Learning, Visual Reasoning, CLIP, Distributed Cognition

This work was supported by the Natural Science Foundation of China under Grant No. 62272018.
Authors' Contact Information: Weimin Shi, State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China; e-mail: shiwm@buaa.edu.cn; Dehong Gao, School of Cybersecurity, Northwestern Polytechnical University, Xi'an, China; e-mail: dehong.gdh@nwpu.edu.cn; Yuan Xiong, State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China; e-mail: xiongyuanxy@buaa.edu; Zhong Zhou (corresponding author), Zhongguancun Laboratory, Beijing, China and State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China; e-mail: zz@buaa.edu.cn.

ACM 1551-6865/2024/11-ART358
https://doi.org/10.1145/3689638

ACM Reference format:
Weimin Shi, Dehong Gao, Yuan Xiong, and Zhong Zhou. 2024. QR-CLIP: Introducing Explicit Knowledge for Location and Time Reasoning. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 11, Article 358 (November 2024), 22 pages.
https://doi.org/10.1145/3689638

1 Introduction

Reasoning about implicit information behind images, such as location and time, enables models to analyze the social context and event details behind images more deeply, thereby enhancing the intelligence of multimedia applications like news analysis [1, 3], intelligent recommendation [4, 6], and social media management [7, 9]. However, this reasoning requires models to not only rely on visual evidence (e.g., geographic markers, architectural styles, and environmental characteristics) but also to connect with external cues (e.g., historical events and cultural attributes). Such a process challenges models' comprehensive reasoning and multidimensional data processing capabilities.

Previous research, such as **Cross-View Time (CVT)** model [10] and **Cross-View Feature Transport (CVFT)** technique [11], use pre-collected remote sensing data to predict the location and time of images via feature-matching methods [12]. However, these methods usually involve high costs for data acquisition and processing and have issues with long update cycles and limited spatiotemporal coverage [13, 15]. With the rapid development of artificial intelligence, **vision-language models (VLMs)** like **Contrastive Language-Image Pre-training (CLIP)** [16] and **Bootstrapped Language-Image Pre-training (BLIP)** [17] have emerged in an endless stream, which are trained on large-scale image-text pair datasets with extensive knowledge, including geographical information and historical context, showing excellent visual and textual understanding capabilities. For instance, VLMs can easily connect an image of the Eiffel Tower with related texts about Paris landmarks. Based on this, **Time and Place for Reasoning beyond the image (TARA)** method [18] proposes the CLIP+Seg model, which matches landmarks, buildings, people, etc., in images with the model's knowledge to predict the location and time related to the image. **Wiki**Common **Ti**mes and **Lo**cation (WikiTiLo) [19] introduces a two-stage reasoning task to uncover whether VLMs can recognize the location and time-relevant features and further reason about them. However, reasoning about the location and time behind images requires the model to look beyond the actual content of an image, and relying solely on image information often fails to produce satisfactory reasoning results. By contrast, human reasoning ability can be expanded and enhanced through the interaction between individuals, tools, and the environment, as demonstrated by Hutchins' distributed cognition [20, 22]. As shown in Figure 1, humans can analyze landmarks, text, and other content in images and use various tools (e.g., search engines and image recognition software) to acquire environmental knowledge (e.g., event background and social context), thus accurately reasoning the information contained in the images.

To enable VLMs to reason like humans, inspired by the above theory, this article proposes a novel reasoning method called *Knowledge Quantity and Relevance Optimization CLIP* (*QR-CLIP*). *QR-CLIP* searches for and effectively uses environmental knowledge from multiple perspectives to enhance VLMs' ability to reason about implicit information behind images. As shown in Figure 2, in the reasoning process, our method involves not only comprehending image details, such as Cristiano Ronaldo and the language on the sign but also searching for related knowledge in the environment, such as searching for the latest updates on Cristiano Ronaldo from Twitter or Wikipedia. By combining the above information, the model can infer that the photo was taken at Cristiano Ronaldo's unveiling ceremony with Al-Hilal Club in Riyadh. In this process, the acquisition and application of knowledge allow the model to look beyond the image for more accurate reasoning. To ensure the reliability and applicability of the reasoning process, we further introduce two types

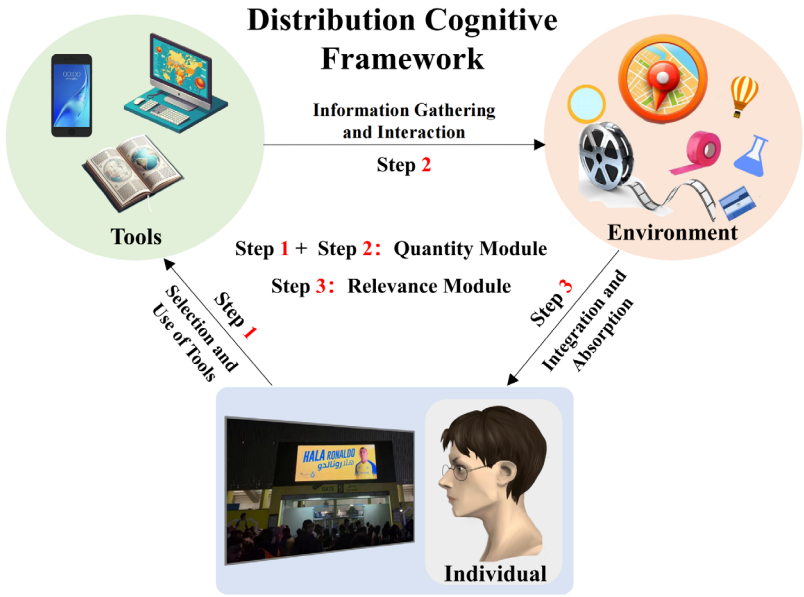

Fig. 1. The diagram of the principle of distributed cognition. When reasoning about implicit information behind images, individuals engage in interactive exchanges with the environment using various tools such as smartphones, books, and computers. Through this interactive process, the individual acquires a comprehensive and profound cognition of the image's information.

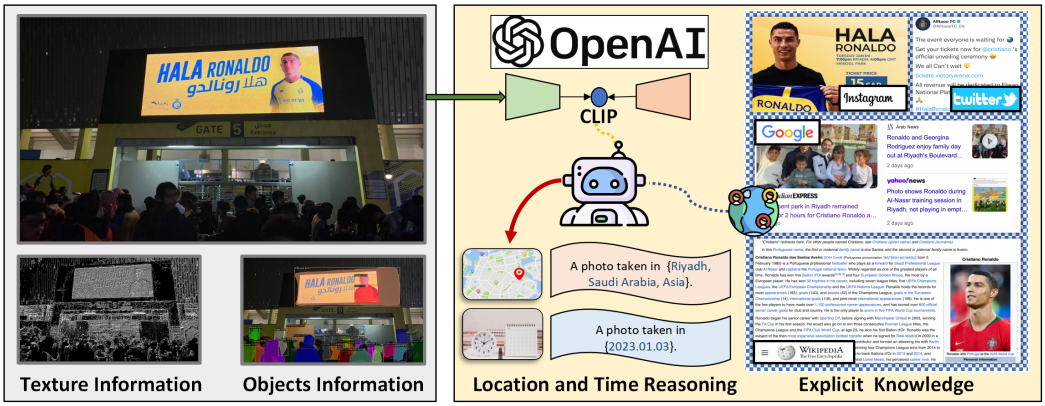

Fig. 2. Comparing traditional computer vision tasks (left) with location and time reasoning (right), it becomes clear that location and time reasoning requires more human experience and knowledge (a.k.a. explicit knowledge) rather than just simple image color, texture, and object information.

of environmental knowledge in this study. One is human-curated databases (e.g., Wikipedia entries on geographical locations and historical events), and the other is task-specific language models (e.g., language models trained on location and time-related corpora). We term the above knowledge as **Explicit Knowledge (EK)**, which is clearly defined and easily accessed in the environment.

To effectively utilize EK, *QR-CLIP* introduces two additional modules based on CLIP: the *Quantity* module and the *Relevance* module. Among them, the *Quantity* module helps the model search and gather knowledge that is beneficial to reasoning from the external environment. Unlike traditional transformer-based models like BERT [23] and Vision Transformer [24], which use a single `[CLS]`

token to represent the input. We develop additional [CLS] tokens as cognitive tools to simulate different human cognitive perspectives of the same image. Each [CLS] processes and interprets the input information in a unique way. When combined with knowledge from different perspectives, these tokens contribute to a more comprehensive understanding of the specific input. For example, different tokens correspond to information about the weather, landmarks, and objects in the image, which can be used to retrieve rich EK to assist reasoning. On the other hand, the *Relevance* module is used to integrate the retrieved EK to form a comprehensive reasoning result. In this module, we design a scoring mechanism to refine both the image features and EK features. Specifically, we introduce an adaptive weighting and feature filtering method to dynamically evaluate information relevant to location and time reasoning in different features, thereby reducing the impact of noisy features. This makes the model robust to changes in feature and modality quality and enhances the credibility of the final reasoning results. For example, based on the vision and EK embeddings extracted in Figure 2, the model adaptively increases the weights of visual features such as persons, stadiums, and billboards, as well as EK features of relevant descriptions such as Saudi Arabia and Al Nassr FC, while reducing the weights of irrelevant or noisy features.

The experiments have shown the effectiveness of our *QR-CLIP* model. On the more challenging TARA dataset, *QR-CLIP* achieves an accuracy (or Rank@1) of 19.51%, which is a 3.05% improvement compared to the previous **State of the Art (SOTA)** on location reasoning. Moreover, for time reasoning tasks, our model achieved an accuracy of 3.45%, representing a significant improvement of 2.45%. Additionally, the Rank@5 also improved from 5.53% to 10.97%. It should be noted that the accuracy of location and time reasoning was calculated under challenging conditions: location reasoning required precision at the district level, and time reasoning to specific dates, such as Dongcheng District, Beijing, China, and Asia in 2017-08-01. To provide a more comprehensive assessment of the model, we utilized the Example-F1 [18]. This metric evaluates the model's accuracy in multi-level label prediction tasks, for instance, assigning scores based on the model's accuracy in reasoning years or months. Specifically, our model achieved a score of 51.25% in location reasoning and 50.53% in time reasoning on Example-F1, exceeding the previous SOTA by 7.64%. Overall, our key contributions can be summarized as:

- —We propose a novel location and time reasoning method, *QR-CLIP*, which searches and utilizes EK in the environment to enhance the ability of VLMs to infer implicit information behind images.
- —We propose a *Quantity* module, which develops additional [CLS] tokens to help the model search and gather EK that is beneficial to reasoning from various perspectives.
- —We propose a *Relevance* module, which employs a scoring mechanism to refine and integrate the visual features and the retrieved EK features to form comprehensive reasoning results.

Comprehensive experiments on the TARA dataset demonstrate the effectiveness of our method. In particular, our method achieved an accuracy improvement of 3.05%/2.45% in location and time reasoning tasks compared to the previous SOTA method.

2 Related Work

2.1 Location and Time Reasoning

Location and time reasoning aims to extract spatial and temporal information from inputs. Some pioneering works propose to predict user locations from social media texts [25], extract temporal information from various texts [26], and deduce spatiotemporal information from news articles [27]. These methods demonstrate the potential to understand the spatiotemporal information of data across diverse textual sources.

However, as the scale of multimedia data grows, the limitations of extracting information simply from textual inputs have become increasingly apparent, and researchers have turned their attention to inferring location and time information from visual inputs. For example, CVT [10] supports tasks like image geolocation and fake news detection by dynamically mapping time/location metadata to visual attributes. CVFT [11] simulates human behavior in remembering the relative spatial positions of objects or buildings during navigation and introduces a domain transfer cross-view feature transmission method. This method determines the location of an image by matching it with aerial views in a database. Despite the success of these methods in specific scenarios, they often rely on extensive pre-collected Geographic Information System data, which limits the practicality of these methods for widespread deployment.

Recently, the emergence of pre-trained VLMs offers new solutions for location and time reasoning, as they associate images with a wide range of world knowledge (such as geographic, temporal, and event information, etc.). Based on this fact, TARA [18] proposes the CLIP+seg method. This method improves upon CLIP by identifying specific objects in images (e.g., landmarks, buildings, and people) to predict the location and time when the image was taken. INFOSEEK [28] conducts a large-scale **Visual Question Answering (VQA)** experiment, focusing on answering questions about the location, time, and object attributes in images to enhance the ability of VLMs to infer implicit information behind images. CogBench [29] performs a thorough evaluation of the cognitive abilities of VLMs, including location reasoning and special time reasoning tasks, on a cognitive assessment benchmark constructed for image reasoning and description. Further, WikiTiLo [19] builds a dataset consisting of images with a broader temporal span and unbiased location distribution, providing a more comprehensive and accurate benchmark for evaluating and improving the reasoning capabilities of these VLMs. Based on this benchmark, WikiTiLo introduces a two-stage reasoning task that enables VLMs to identify location and time-related features and perform further reasoning.

Due to the inability to effectively access knowledge beyond the images, reasoning based solely on the observed visual content makes these VLMs perform poorly in reasoning about location and time. In this context, this article proposes a novel reasoning model, *QR-CLIP*, which searches for and utilizes relevant knowledge from different perspectives, expanding the breadth and depth of the model's understanding of multimedia data, thereby improving the accuracy of location and time reasoning behind images.

2.2 VLMs

Pre-trained VLMs, which connect visual concepts with textual descriptions, have indicated remarkable performance across a variety of downstream tasks, such as image retrieval [30], dense prediction [31], and VQA [32]. As a milestone, CLIP [16], which adopts a contrastive learning method [33] on a vast collection of image-text pairs, exhibits excellent transferability over 30 classification datasets. Inspired by this work, numerous follow-ups have been proposed to improve the training strategy (e.g., Tip-adapter [34], A Large-scale ImaGe and Noisy-text embedding [35], **Self-supervision meets Language-Image Pre-training (SLIP)** [36], BLIP [17], and Pyramidclip [37]) or apply it to other fields (e.g., CLIP-Event [1] and CrowdCLIP [38]).

However, despite the impressive ability of existing VLMs to match visual cues with textual semantics, they fail to further uncover implicit information behind the images based on this matching. Instead, we propose the *Quantity* module and *Relevance* module. The *Quantity* module helps VLMs understand images from different perspectives and seek beneficial environmental knowledge. The *Relevance* module further enhances the reliability and accuracy of VLMs in the reasoning process. These two modules effectively expand the functionality of VLMs (e.g., CLIP) and enhance their reasoning capabilities.

3 Approach

3.1 Preliminary

Task Background. The Current AI methods are relatively weak in cognizing and reasoning the information concealed within an image. The goal of this article is to let the model reason the location and time based on image input [18]: given an image I, we need the model ($\mathrm{M}(I)$) to predict the location (Pred_l) and time (Pred_t).

VLMs. VLMs leverage visual-language pre-training method to learn both visual and language representations from large-scale image-text pairs. They generally consist of an image encoder (CLIP-V) Enc_v and a text encoder (CLIP-T) Enc_t, which are jointly trained to, respectively, map input images and texts into a unified representation space. Specifically, the image encoder uses ResNet [39] or ViT [24] with a global attention pooling layer to generate a class token $[\mathtt{CLS}]^v$ that represents the global feature of input image I, while the text encoder adopts a Transformer [40] to extract the embedding $[\mathtt{CLS}]^t$ of the input text T. For simplicity, we represent the above process as:

$$[\mathtt{CLS}]^v \leftarrow \mathrm{Enc}_v(I) \text{ and } [\mathtt{CLS}]^t \leftarrow \mathrm{Enc}_t(T), \tag{1}$$

Afterwards, contrastive learning [33, 41, 42] is employed as their training objective, with ground-truth image-text pairs treated as positive samples and mismatched image-text pairs constructed as negative samples. Using large-scale image-text pairs for model training, VLMs (e.g., CLIP [16], PyramidCLIP [37]) have powerful visual language understanding capabilities. In this work, we mainly verify the effectiveness of the proposed method based on CLIP.

Our Pipeline. To enhance the reasoning capability of VLMs (e.g., CLIP), we propose *QR-CLIP*. As shown in Figure 3, *QR-CLIP* consists of two modules: the *Quantity* module and the *Relevance* module. The *Quantity* module helps the model search and gather knowledge that is beneficial to reasoning from the external environment, which is crucial for expanding cognitive resources and enhancing reasoning abilities. The *Relevance* module integrates this knowledge with a scoring mechanism to form a comprehensive reasoning result. The two modules work together to further improve the time and location reasoning performance of VLMs.

3.2 Quantity Module

The *Quantity* Module aims to expand the cognitive resources of the model. To achieve this, we propose first encoding multiple distinct $[\mathtt{CLS}]_i^v$ tokens to represent the image, allowing the model to seek EK from various perspectives. Next, we fine-tune CLIP on our location and time reasoning task to further enhance its performance. Finally, we use the fine-tuned model for EK search.

Introducing additional [CLS]. Vanilla CLIP utilizes a single class token $[\mathtt{CLS}]$ to summarize the global features of an image. However, single $[\mathtt{CLS}]$ is inadequate in representing an image comprehensively, as it provides limited location and time reasoning cues. Therefore, we propose to expand the image representations. It is evident that in real life, individuals can achieve a more comprehensive and accurate understanding of images by integrating the information and functionality of various tools. In this vein, we propose to introduce additional $[\mathtt{CLS}]_i^v$ tokens to describe images from multiple perspectives, which can be expressed as:

$$[\mathtt{CLS}]_i^v \leftarrow \mathrm{Enc}_v(I), \tag{2}$$

where i represents the count of $[\mathtt{CLS}]$ tokens in a given image, ranging from 1 to n. By default, we set $n = 6$. After passing through the encoder Enc_v, we get a list of embeddings ($[\mathtt{CLS}]_1^v ... [\mathtt{CLS}]_6^v\, I$). Using this design, each $[\mathtt{CLS}]_i^v$ token is treated as a separate cognitive tool, simulating the framework of distributed cognition. This approach enables the pre-trained model to incorporate multiple perspectives, enhancing the richness of the captured EK.

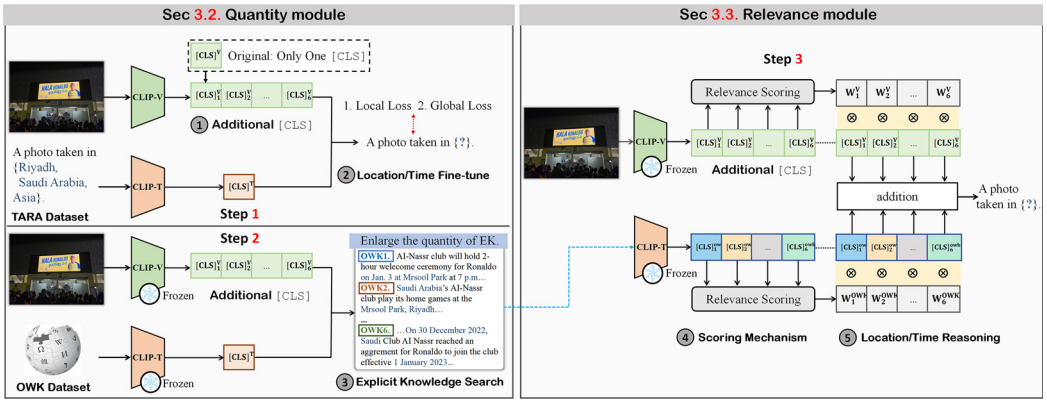

Fig. 3. The *QR-CLIP* pipeline consists of two modules: the *Quantity* Module (Section 3.2) and the *Relevance* Module (Section 3.3). Each step described below corresponds to Step 1, Step 2, and Step 3 in Figure 1. In Step 1, we add additional [CLS] tokens to simulate the use of different cognitive tools by individuals. We then design local and global loss functions to guide location/time fine-tuning. Then, we freeze the fine-tuned CLIP-V and CLIP-T models and utilize them to search for EK from our EK dataset (Section 4.1). In the *Relevance* Module, we use a scoring mechanism to weigh the most valuable information from CLIP-T and CLIP-V. After multiplying scoring weights for vision and language, we add them for the final similarity calculation.

Regarding the encoding, since the input text contains explicit semantic information and most language inputs convey clear messages, we directly utilize the original $[\texttt{CLS}]^t$ as the input feature embedding. Afterwards, we utilize image features $[\texttt{CLS}]_i^v$ to retrieve relevant textual information:

$$([\texttt{CLS}]^t) \cdot ([\texttt{CLS}]_i^v), \tag{3}$$

here, $\cdot$ denotes the inner product operation. In the fine-tuning or EK search process, each $[\texttt{CLS}]_i^v$ from Enc_v calculates its similarity with the $[\texttt{CLS}]^t$ of the candidate information.

Location/Time Fine-tune. We further fine-tune CLIP with local and **global losses (GLs)** [43, 44] to ensure that each $[\texttt{CLS}]_i^v$ is aligned with the linguistic features of location and time $[\texttt{CLS}]^t$. The **local loss (LL)** is utilized to construct multiple different $[\texttt{CLS}]_i^v$, while ensuring that they encode the visual features of the image from diverse perspectives. This loss function consists of **multi-view contrastive learning (MVC)** and multi-view regularization. Among them, the alignment between each $[\texttt{CLS}]_i^v$ and $[\texttt{CLS}]^t$ is achieved through the MVC:

$$L_{MVC} = -\log \frac{e^{f(q_v^i,\, k_{t+})}}{e^{f(q_v^i,\, k_{t+})} + e^{f(q_v^i,\, k_{t-})}}, \tag{4}$$

here, q_v^i denotes the query image embedding ($[\texttt{CLS}]_i^v$), while k_{t+} and k_{t-} represent the positive and negative key text embeddings, in a batch of $[\texttt{CLS}]^t$. $f(x, y)$ denotes inner product function to calculate the similarity between x and y.

Since multiple $[\texttt{CLS}]_i^v$ correspond to one $[\texttt{CLS}]^t$, L_{MVC} tends to cluster $[\texttt{CLS}]_i^v$ together. To overcome this issue, we add a regularization term to separate the distance between each $[\texttt{CLS}]_i^v$, promoting them to learn and represent information in the image independently from diverse perspectives:

$$L_{MVR} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{f\left(q_v^i,\, q_v^j\right)}{||q_v^i|| \cdot ||q_v^j||}, \tag{5}$$

where n represents the number or $[\texttt{CLS}]_i^v$ vectors, i ranges from 1 to $n-1$, and j ranges from $i+1$ to n. This implies that the calculation of L_{MVR} takes into account all possible pairs of $[\texttt{CLS}]_i^v$

vectors. The numerator of the formula uses the inner product function, $f(qv^i,\ q_v^j)$, to gauge the similarity between each pair of $[\texttt{CLS}]_i^v$ vectors. The denominator normalizes the distance between the vectors by the product of their magnitudes $||qv^i|| \cdot ||q_v^j||$, which encourages the model to learn discriminative features that are independent of vector length.

During the optimization phase, the model strives to increase the distances between pairs of $[\texttt{CLS}]_i^v$, thereby enhancing its ability to discern image features from diverse perspectives. However, each $[\texttt{CLS}]_i^v$ vector faces distinct learning challenges, resulting in uneven rates of training progress. To counteract this, we use a dynamically balanced learning strategy:

$$w^i = \text{softmax}(1 - acc^i), \tag{6}$$

where the variable w^i denotes the dynamic learning weight for each $[\texttt{CLS}]_i^v$. This weight is calculated by applying a softmax function over the $(1 - acc^i)$, where acc^i represents the prediction accuracy achieved by the model using each $[\texttt{CLS}]_i^v$ during the training process. In essence, this approach dynamically adjusts the learning priorities, offering more attention to instances of $[\texttt{CLS}]_i^v$ that exhibit slower progress or present greater learning challenges.

Based on the above-mentioned discussions, the LL can be defined as:

$$L_{local} = \sum_{i=1}^{n} w^i L_{MVC} + \lambda L_{MVR}, \tag{7}$$

which aims to minimize the distance between each $[\texttt{CLS}]_i^v$ and its corresponding sentence embedding ($[\texttt{CLS}]_t$), while simultaneously maximizing the distance between different $[\texttt{CLS}]_i^v$.

Besides the LL, we further introduce a GL to constrain the correspondence between image features and location/time features. The calculation for this constraint is as follows:

$$L_{global} = -\log \frac{e^{f_{mean}(q_v,\, k_{t+})}}{e^{f_{mean}(q_v,\, k_{t+})} + e^{f_{mean}(q_v,\, k_{t-})}}, \tag{8}$$

here, we have the function $f_{mean}(q_v, k_t) = \frac{1}{n} \sum_{i=1}^{n} f(q_v^i, k_t)$. L_{global} aims to enhance the learning of global correspondence by integrating the mean correlation score across various perspectives.

Finally, the total training objective can be formulated as:

$$L_{total} = L_{local} + L_{global}, \tag{9}$$

which not only encourages the model to learn robust correspondences between diverse visual perspectives and the text but also enables it to capture the overall alignment of the image features with respect to location and time attributes mentioned in the text.

EK Search. After fine-tuning, each $[\texttt{CLS}]_i^v$ outputted by CLIP-V is capable of representing image location and time information from various perspectives. Thus, we use them to search more valuable EK from the EK dataset (Section 4.1), facilitating interaction between the model and the environment. Specifically, given an image I and its corresponding EK ($O = T_1^{EK}, T_2^{EK}, ..., T_k^{EK}, k = 122{,}408$), the search process follows Equation (3): each $[\texttt{CLS}]_i^v$ calculates the similarity with 122,408 candidate Wikipedia corpus (EK). Here, we select the Wikipedia candidate with the highest similarity for each $[\texttt{CLS}]_i^v$, yielding a total of n EK entries.

Through the above process, the *Quantity* module collects EK that is highly relevant to the location and time reasoning task, which is crucial to improving the performance of the method.

3.3 Relevance Module

As the images and the retrieved EK inevitably contain redundant information that is irrelevant to location and time reasoning, this causes unnecessary interference with the final reasoning process. To utilize them more reasonably to obtain more accurate reasoning results, we further propose the *Relevance* Module, which adopts a scoring mechanism to emphasize and highlight relevant features and suppress irrelevant features.

Specifically, we first adopt two-layers MLP ($\text{MLP}_{2-layer}$) as relevance scoring component:

$$W^x = \text{MLP}_{2-layer}([\texttt{CLS}]_i^x), \tag{10}$$

to evaluate the significance of different features. Here, $[\texttt{CLS}]_i^x$ is the input embedding, W^x is the calculated weight. Then, we perform adaptive fusion of different features based on the predicted weights to form a comprehensive feature representation:

$$[\texttt{CLS}]^{fused} = \sum_{1}^{n}(W_i^{EK} \times [\texttt{CLS}]_i^{EK} + W_i^v \times [\texttt{CLS}]_i^v), \tag{11}$$

here, W_i^{EK} and W_i^v are the weights of the $[\texttt{CLS}]_i^{EK}$ and $[\texttt{CLS}]_i^v$. Finally, we calculate the similarity between $[\texttt{CLS}]^{fused}$ and the embeddings of candidate locations/times to perform the final reasoning.

Moreover, to train the *Relevance* module, we adopt the same loss functions as the *Quantity* Module (Section 3.2), i.e., local and GLs. In particular, we maintain the fine-tuned CLIP-T and CLIP-V frozen, and solely update the parameters of the relevance scoring component.

By utilizing the CLIP (which is pre-trained 400 M explicit corpus) and subsequently fine-tuning it by adding additional [CLS] with location-and-time-specific data, the model can reason about meta information more effectively. Further, we enhance its performance by retrieving valuable EK and utilizing it as auxiliary cues. Finally, we filter the vision and EK embeddings with a scoring mechanism, enabling the model to achieve more effective reasoning.

4 Experiments

4.1 Experimental Settings

Dataset. In this study, we evaluate our *QR-CLIP* method using the TARA dataset [18], which comprises 15,429 samples of news pictures and their location and time descriptions. We train on 12,306 instances and test on 1,644 instances to assess the effectiveness of location and time reasoning. Further, we evaluate the method's generalization performance across four different datasets. Among them, TARA-Dev [18] contains 1,552 images different from the TARA test set. TARA-Interest [18] comprises 30 images related to news events occurring after January 2021, which is the cut-off date for the CLIP model. The **Commonsense and Factual Reasoning (COFAR)** dataset [45] includes landmark images with descriptions to verify the model's understanding of location-related events. WikiTiLo dataset [19] consists of 6,296 images annotated with specific times and locations, spanning 30 countries across four continents to minimize distribution bias. Additionally, The EK for our method is derived from the Wikipedia-based Image Text dataset [46]. We selected 122,408 texts from the 37.5 million English Wikipedia that correspond to the specific countries and years as our EK.

Evaluation Metrics. For a fair comparison, we first follow the same evaluation metrics as outlined in the TARA benchmark [18]: Accuracy (Rank@1) and Example-F1. Accuracy is calculated as the proportion of correctly predicted samples to the total number of samples. It measures whether the model's reasoning results accurately include all the information in the location and time labels. For example, the model needs to accurately predict all the information in {"*Dongcheng District, Beijing, China, Asia, 2017-08-01*"} to be considered accurate. Example-F1 is calculated by comparing reasoning results with hierarchical labels:

$$\text{Example-F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2\,|\text{GT}_i \cap \text{Pred}_i|}{|\text{GT}_i| + |\text{Pred}_i|}, \tag{12}$$

where GT_i represents the hierarchical label, and Pred_i represents the hierarchical reason. If the entire label is {"*Zurich, Switzerland, Europe*"}, the progressive hierarchical labels consist of three combinations of true labels: {"*Zurich, Switzerland, Europe*"}, {"*Switzerland, Europe*"} and {"*Europe*"}.

Table 1. Summary of the Performance for Different Baselines on the TARA Dataset [18]

ID	Method	Training Mode	Accuracy (Rank@1)	Rank@5	Example-F1	F1-Score
			Location Reasoning			
1	ResNet-50 [39]	Fine-tune	3.18%	9.82%	22.19%	2.27%
2	Swin-T [47]	Fine-tune	6.70%	17.07%	33.56%	5.02%
3	CLIP [16]	Zero-Shot	11.11%	27.85%	44.96%	9.74%
4	BLIP [17]	Zero-Shot	4.07%	12.26%	36.01%	2.92%
5	SLIP [36]	Zero-Shot	3.17%	11.52%	32.85%	2.33%
6	PyramidCLIP [37]	Zero-Shot	4.66%	13.75%	34.64%	2.75%
7	CLIP† [18]	Fine-tune	15.72%	37.13%	49.74%	13.82%
8	PyramidCLIP† [37]	Fine-tune	7.13%	32.47%	48.23%	5.86%
9	CLIP+Seg [18]	Fine-tune	16.46%	37.48%	50.52%	14.63%
10	*QR-CLIP* (Ours)	Fine-tune	**19.51%**	**38.48%**	**51.25%**	**17.65%**
			Time Reasoning			
11	ResNet-50 [39]	Fine-tune	0.84%	5.14%	39.99%	0.46%
12	Swin-T [47]	Fine-tune	0.97%	5.53%	43.95%	0.72%
13	CLIP [16]	Zero-Shot	0.46%	2.42%	39.90%	0.25%
14	BLIP [17]	Zero-Shot	1.69%	3.99%	43.27%	0.20%
15	SLIP [36]	Zero-Shot	0.32%	2.15%	32.89%	0.71%
16	PyramidCLIP [37]	Zero-Shot	1.15%	3.61%	41.51%	0.33%
17	CLIP† [18]	Fine-tune	1.00%	2.99%	43.09%	0.54%
18	PyramidCLIP† [37]	Fine-tune	1.73%	4.32%	43.77%	1.41%
19	CLIP+Seg [18]	Fine-tune	0.92%	3.15%	42.89%	0.71%
20	*QR-CLIP* (Ours)	Fine-tune	**3.45%**	**10.97%**	**50.53%**	**1.49%**

Here, CLIP and PyramidCLIP use the ViT-B/32 model. SLIP uses ViT-B/16 model. BLIP uses the 129M model. The symbol † denotes that the model was fine-tuned. The best results are in bold.

In addition, the performance of methods is further evaluated using Rank@5 and F1-Score. Rank@5 measures the model's accuracy in the Top 5 reasoning results. F1-Score is calculated as the harmonic mean of the model's precision and recall, evaluating whether the model accurately reasons.

Implementation Details. *QR-CLIP* is based on CLIP+VIT-B/32 model with an input size of 224×224. It is implemented on the PyTorch 1.10.1 platform with the Adam optimizer to update the neural network's weights and biases. The training batch size is 32, and the initial learning rate is $1e-6$. Our model utilizes a pre-trained model and fine-tuned process on an NVIDIA RTX 3090 GPU running CUDA 11.7.1.

4.2 Comparison with SOTA Methods

In this section, we compare *QR-CLIP* with the current state-of-the-art location and time reasoning methods. By analyzing and comparing these methods on multiple key metrics, we indicate the performance advantages of our method.

(1) *Location Reasoning*. We compare the results of *QR-CLIP* with other methods for location reasoning in Table 1. In this experiment, both *ResNet-50* and *Swin-T* models were initialized with ImageNet [48] pre-trained weights and subsequently fine-tuned for location and time reasoning tasks using the TARA dataset with an additional classification head. Our *QR-CLIP* model achieves an accuracy of 19.51% (Rank@1). Additionally, it attains an Example-F1 score of 51.25% for the hierarchical labels. All the results collectively show that our method outperforms other methods.

Compared with ResNet-50 [39] and Swin-T [47], vanilla CLIP achieves an improvement of 7.93% and 4.41% in location reason accuracy *(IDs: 1, 2, 3)*. It is evident that in comparison to the vision model only trained on ImageNet [48], CLIP already possesses a certain level of knowledge for reasoning. Furthermore, compared to other VLMs, CLIP shows advantages in location reasoning in both zero-shot and fine-tuned settings. For instance, compared to BLIP [17] and PyramidCLIP [37],

CLIP improves location reasoning accuracy by 7.04% and 6.45%, respectively. The improvement is more significant for CLIP†, reaching 11.65% and 11.06%, respectively *(IDs: 3–8)*. This is mainly because CLIP is trained on a larger dataset of image-text pairs, which not only provides the model with richer world knowledge but also a powerful visual encoder. In contrast, by integrating CLIP into our method, *QR-CLIP* further improves location reasoning accuracy by 8.40%, demonstrating its effectiveness in improving the reasoning ability of VLMs *(IDs: 3 vs 10)*.

Besides, compared to CLIP† and the previous SOTA method CLIP+Seg [18], *QR-CLIP* exhibits a significant accuracy improvement of 3.79% and 3.05%, respectively, along with a corresponding increase in F1-Score by 3.83% and 3.02%, respectively *(IDs: 7, 9, 10)*. Other evaluation metrics (e.g., Rank@5 and Example-F1) also improved. These results show that *QR-CLIP* can effectively utilize EK to enable the model to understand and distinguish more detailed visual attributes, creating stronger connections between image and location information. However, we have observed that the improvement in Example-F1 is not as apparent. We argue that this is because of the mechanism of Example-F1. To illustrate, consider Figure 2, which contains many elements of Arabia, such as turbans and Arabic writing. It is not difficult for many models to recognize that this image was captured in the Middle East and to predict its hierarchical label as {"*Asia*"}. However, they failed when asked to predict the entire label *Riyadh, Saudi Arabia, Asia*. Therefore, the discrepancy in other metrics may be more noticeable.

(2) *Time Reasoning*. Table 1 also presents the performance of our method and existing techniques for time reasoning. The Accuracy (Rank@1) of *QR-CLIP* is 3.45%, and Example-F1 is 50.53%; compared to the CLIP model, the two metrics have been improved by 2.99% and 10.63%, respectively *(IDs: 13, 20)*. Compared with CLIP† and CLIP+Seg, which are also based on fine-tuned CLIP, our method achieves improvements of 2.45% and 2.53% in time reasoning accuracy, respectively. Compared with traditional image classification methods, *QR-CLIP* exhibits advantages in all metrics *(IDs: 17, 19, 20)*. In addition, we find that in the time reasoning task, existing VLM-based methods struggle with achieving effective time reasoning because images often lack features that directly indicate specific dates.

It is not surprising that even for humans, determining the time a photo was taken may be difficult, as illustrated by the sample image in Figure 2. For instance, if one is unfamiliar with Cristiano Ronald or lacks specific knowledge, they may not recognize that the time stamp on the image, {"*03-01-2023*"} indicates the date the photo was taken. Nevertheless, our method is effective, achieving an improvement of +2.45% in predicting time compared to CLIP†.

4.3 Method Generalization Validation

In this section, we analyze the generalizability of *QR-CLIP*, by showing its performance on different datasets, its results when using different VLMs, and its outcomes using various types of EK.

(1) *Testing on Different Datasets*. We first evaluate the zero-shot performance of QR-CLIP on three datasets (e.g., TARA-Dev [18], TARA-Interest [18] and COFAR dataset [45]) without any extra training. Next, we conduct a comprehensive evaluation on the WikiTiLo dataset [19], which has a longer time span and a more even geographic distribution to verify the reasoning capability of *QR-CLIP* in larger scenarios.

As illustrated in Table 2, the experimental results show the efficacy of our *QR-CLIP* in comparison to other methods. On the TARA-Dev dataset, *QR-CLIP* achieved an accuracy of 20.35% in location reasoning and 6.53% in time reasoning, surpassing both the CLIP and CLIP+Seg models *(IDs: 21–26)*. These results validate the effectiveness of the model in handling diverse and previously unseen images. A similar trend was observed in the TARA-Interest dataset, where *QR-CLIP* attained an accuracy of 58.62% *(ID: 29)* in location reasoning and 20.69% *(ID: 32)* in time reasoning. This not only

Table 2. Performance Comparison of Location and Time Reasoning Tasks across Different Datasets

ID	Method	Accuracy (Rank@1)	Rank@5	Example-F1
	TARA-Dev [18]			
	Location Reasoning			
21	CLIP [16]	10.99%	29.72%	45.90%
22	CLIP+Seg [18]	15.88%	39.15%	51.83%
23	*QR-CLIP* (Ours)	**20.35%**	**40.23%**	**51.60%**
	Time Reasoning			
24	CLIP [16]	0.53%	1.82%	42.14%
25	CLIP+Seg [18]	0.53%	2.50%	43.55%
26	*QR-CLIP* (Ours)	**6.53%**	**18.89%**	**53.26%**
	TARA-Interest [18]			
	Location Reasoning			
27	CLIP [16]	13.33%	27.85%	56.44%
28	CLIP+Seg [18]	23.33%	37.48%	63.11%
29	*QR-CLIP* (Ours)	**58.62%**	**86.20%**	**80.46%**
	Time Reasoning			
30	CLIP [16]	0.00%	1.85%	24.56%
31	CLIP+Seg [18]	3.33%	9.48%	24.43%
32	*QR-CLIP* (Ours)	**20.69%**	**41.38%**	**60.34%**
	COFAR [45]			
	Location Reasoning			
33	CLIP [16]	70.96%	84.29%	81.97%
34	CLIP+Seg [18]	70.00%	83.33%	80.05%
35	*QR-CLIP* (Ours)	**71.42%**	**85.71%**	**85.14%**

Here, we show the results on TARA-Dev [18], TARA-Interest [18] and COFAR [45]. Notably, our method was tested without any additional training. The best results are in bold.

Table 3. The Performance Comparison of Different Methods on the WikiTiLo Dataset [19]

			Times			Country			Region		
ID	Method	Training Mode	Accuracy	Precision	F1-score	Accuracy	Precision	F1-score	Accuracy	Precision	F1-score
36	CLIP [16]	Zero-Shot	78.57%	70.66%	70.66%	44.28%	43.11%	40.19%	63.65%	67.34%	64.42%
37	BLIP [17]	Zero-Shot	30.95%	46.81%	46.14%	35.23%	35.30%	30.07%	46.51%	57.02%	49.05%
38	CLIP†[16]	Fine-tune	89.37%	85.67%	85.04%	57.83%	55.42%	54.15%	79.37%	79.36%	79.11%
39	BLIP†[17]	Fine-tune	86.51%	81.36%	80.51%	47.77%	45.39%	41.25%	75.08%	75.10%	75.26%
40	*QR-CLIP* (Ours)	Fine-tune	**91.59%**	**89.25%**	**88.23%**	**64.86%**	**62.86%**	**61.80%**	**83.77%**	**83.48%**	**83.17%**

Here, CLIP uses the ViT-B/32 model. BLIP uses the 129M model. The symbol † denotes that the model was fine-tuned. The best results are in bold.

shows the generalization capability of *QR-CLIP* but also suggests its potential uses in evolving real-world scenarios. Furthermore, on the COFAR dataset, *QR-CLIP* once again outperforms other models with a location reasoning accuracy of 71.42% *(ID: 35)*. Given that landmark descriptions contain more context information, our model has achieved significant growth in all metrics concerning location reasoning. This result corroborates the robustness and adaptability of our model to different data types and tasks.

The experimental results on the WikiTiLo dataset [19] are shown in Table 3. The results indicate that the CLIP variants outperform the BLIP model overall *(IDs: 36–39)*. This is because CLIP uses a larger pre-training dataset, which allows it to capture and understand visual features more

Table 4. Performance Results of Adding the Quantity Module and Relevance Module to PyramidCLIP on the TARA Dataset [18]

ID	Method	Training Mode	Accuracy (Rank@1)	Rank@5	Example-F1
Location Reasoning					
41	PyramidCLIP [37]	Zero-Shot	4.66%	13.75%	34.64%
42	PyramidCLIP† [37]	Fine-tune	7.13%	32.47%	48.23%
43	QR-PyramidCLIP	Fine-tune	**13.65%**	**35.40%**	**50.73%**
Time Reasoning					
44	PyramidCLIP [37]	Zero-Shot	1.15%	3.61%	41.51%
45	PyramidCLIP† [37]	Fine-tune	1.73%	4.32%	43.77%
46	QR-PyramidCLIP	Fine-tune	**3.15%**	**11.03%**	**48.29%**

Here, the symbol † denotes that the model was fine-tuned. QR-PyramidCLIP represents PyramidCLIP with the Quantity and Relevance Modules. The best results are in bold.

effectively and includes more implicit knowledge of location and time. Moreover, both CLIP and BLIP show improved performance when fine-tuned, as fine-tuning enhances the models' adaptability to downstream tasks *(ID: 36 vs 38, 37 vs 39)*. However, reasoning about location and time is a fine-grained task that requires models to distinguish more detailed visual cues at the knowledge level, such as understanding different geographical and cultural elements. Relying solely on extracted visual features, the models still do not achieve satisfactory accuracy. Further, compared to CLIP† method, *QR-CLIP* shows notable performance improvements, increasing the accuracy of location (Region) reasoning by 4.40% and time reasoning by 2.22% *(IDs: 38 vs 40)*. This further suggests that leveraging EK from the environment can more accurately distinguish visual cues, thereby effectively enhancing the model's reasoning performance. Additionally, the improved performance shows the potential of our model to reason about location and time in a wider range of application scenarios.

(2) *Performance Analysis Based on Different VLMs.* To furthur validate the generalizability of our method, we add *Quantity* module and *Relevance* module to PyramidCLIP [37], naming it QR-PyramidCLIP. As shown in Table 4, the experimental results indicate that QR-PyramidCLIP improves accuracy by 6.52% and 1.42% over PyramidCLIP† in location and time reasoning tasks, respectively *(IDs: 42 vs 43, 45 vs 46)*. These results indicate that QR-PyramidCLIP effectively enhances the reasoning performance of the original PyramidCLIP, validating the rationality of designed *Quantity* and *Relevance* modules. Additionally, these results confirm that introducing EK is an effective solution for improving the reasoning capabilities of VLMs.

(3) *Search Knowledge From Language Model.* Existing language models have gathered a vast amount of knowledge through training on extensive text data. In this experiment, we utilize task-specific language models as explicit EK. As shown in Figure 4, to familiarize the model with the text distribution relevant to the task, we used 122,408 candidate Wikipedia data *(ID: 77)*, updating the parameters of the GPT-2 language model in an unsupervised method [49]. The experiment employs Magic [50] decoding methods. By integrating the similarity between the token and the image generated at each step by the language model into the decoding score, we execute a knowledge search from the language model.

As shown in Table 5, the experimental results demonstrate that the incorporation of task-specific language models as EK enhances the model's capabilities in location and time reasoning tasks *(IDs: 50, 54)*. The application of EK, derived from both human-curated databases (Wikipedia) and language model (GPT-2), shows an improvement in the performance of both location and time reasoning tasks compared to methods without any EK *(IDs: 47–50, 51–54)*. This highlights the

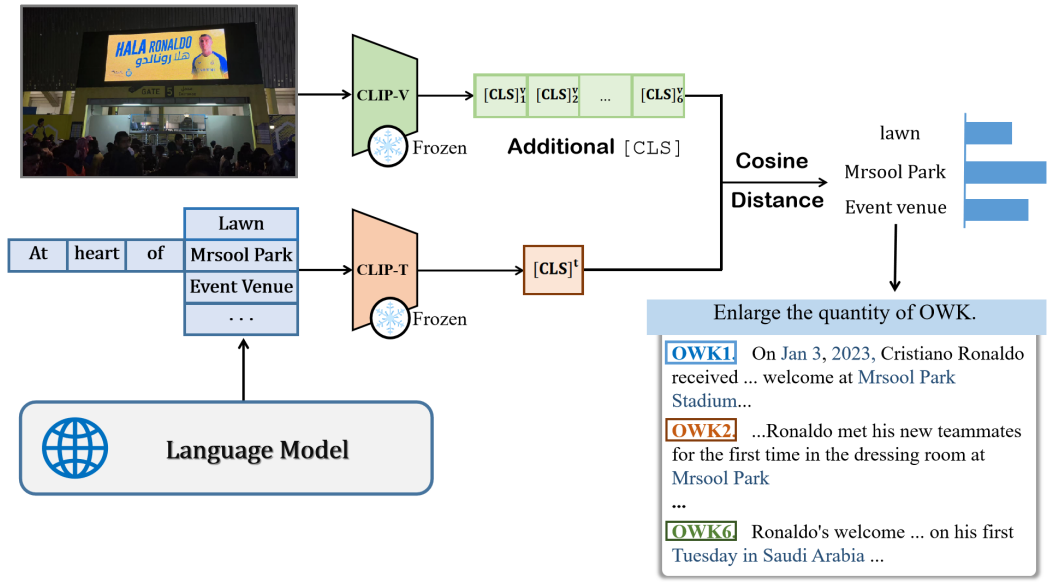

Fig. 4. The illustration of searching EK from language models. The CLIP-V outputs diverse representations of image location and time as $[\mathrm{CLS}]_i^v$, which enables semantic alignment between generated results and input images. This alignment facilitates the selection of the most suitable token for EK search based on visual information. Each $[\mathrm{CLS}]_i^v$ yields a corresponding textual knowledge, consistent with step 2 in Figure 3.

Table 5. The Impact Results of Different EK Sources on Location and Time Reasoning Tasks

ID	EK Source	Method	Accuracy (Rank@1)	Rank@5	Example-F1
Location Reasoning					
47	*N/A*	CLIP	15.72%	37.13%	49.74%
48	*N/A*	CLIP+Seg	16.46%	37.48%	50.52%
49	*Wikipedia*	QR-CLIP	19.51%	**38.48%**	**51.25%**
50	*GPT-2*	QR-CLIP	**19.73%**	38.25%	51.19%
Time Reasoning					
51	*N/A*	CLIP	1.0%	2.99%	43.09%
52	*N/A*	CLIP+Seg	0.92%	3.15%	42.89%
53	*Wikipedia*	QR-CLIP	**3.45%**	**10.97%**	**50.53%**
54	*GPT-2*	QR-CLIP	3.37%	9.51%	46.33%

The EK sources utilized include None (*N/A*), Wikipedia, and a language model (*GPT-2*). The best results are in bold.

significance of EK and showcases the versatility of our approach. We effectively leverage various types of EK in the environment to enhance the model's reasoning abilities.

In summary, our model successfully motivates visual language models to perform higher-level reasoning while maintaining the potent generalization capabilities of large-scale pre-training models. As a result, this research not only provides substantial support for further studies and application of location and time reasoning tasks but also carries positive implications for enhancing the generalization performance of large-scale visual language models in specific tasks.

Table 6. The Impact of Various Loss Functions and Components on Performance

ID Method	Accuracy (Rank@1)	Rank@5	Example-F1
Location Reasoning (Only QM)			
55 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}LL$	16.56%	37.08%	49.85%
56 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}GL$	17.04%	37.26%	49.93%
57 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}LL{+}GL$	**17.47%**	**38.00%**	**50.10%**
Time Reasoning (Only QM)			
58 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}LL$	1.31%	5.56%	44.83%
59 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}GL$	1.84%	5.77%	44.53%
60 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}LL{+}GL$	**2.03%**	**6.33%**	**45.72%**
Location Reasoning (QR-CLIP: QM+RM)			
61 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}LL$	19.19%	37.12%	50.63%
62 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}GL$	18.61%	37.49%	50.91%
63 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}LL{+}GL$	**19.51%**	**38.48%**	**51.25%**
Time Reasoning (QR-CLIP: QM+RM)			
64 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}LL$	3.12%	**11.49%**	48.85%
65 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}GL$	2.78%	9.86%	47.11%
66 CLIP+$[\texttt{CLS}]_i^v(n{=}6){+}LL{+}GL$	**3.45%**	10.97%	**50.53%**

LL, *GL* indicate the local loss and global loss, respectively. **QR-CLIP** means the model contains entirely Quantity Module (QM: Section 3.2) and Relevance Module (RM: Section 3.3). The best results are in bold.

4.4 Ablation Study

We conduct ablation studies to analyze the effectiveness of each component in our method. Specifically, we (1) first analyze the performance of the proposed local and GL functions under differ module settings. (2) Then, we analyze the impact of different numbers of [CLS] tokens on method performance. (3) Next, we evaluate the impact of varying amounts of EK in the environment. (4) Afterward, we test the performance differences of distinct scoring mechanisms. (5) Finally, we analyze the impact of multi-view representation images.

(1) *Effectiveness of Losses and Modules.* We analyze the impact of different loss functions, i.e., LL and GL, as well as the contributions of the *Quantity* (Section 3.2) and *Relevance* Modules (Section 3.3) to the performance of the model. As shown in Table 6, compared to training with only one loss function, both modules achieved performance improvements when combining the two loss functions, demonstrating their combined potential. We attribute the performance improvement to the fact that these two loss functions help the model integrate image information from multiple perspectives. Additionally, by adding the *Relevance* module, we significantly improve the reasoning abilities of the model. These experimental results validate the rationality of our model design: first, acquiring extensive knowledge through the *Quantity* module, and then effectively integrating this information using the *Relevance* module to boost the performance.

(2) *Impact of Additional* [CLS]*' Number.* Following the model design process, we first analyze the impact of different numbers of [CLS] tokens on model performance in the *Quantity* module. As shown in Table 7, in location reasoning, compared to CLIP† with a single [CLS]*(ID: 7)*, the model's accuracy improves by 1.81% and 1.75% when the number of [CLS] tokens is $n = 4$ and $n = 6$, respectively *(IDs: 7 vs 68, 7 vs 69)*. The results indicate that the additional [CLS] effectively increases image cues by constructing multiple perspectives, which has promising benefits. Therefore, we choose $n = 6$ for subsequent experiments to acquire as much EK as possible from different

Table 7. Performance of Additional [CLS] in **QR-CLIP** with Different Numbers

ID	Method	Accuracy (Rank@1)	Rank@5	Example-F1
Location Reasoning				
67	CLIP+$[\texttt{CLS}]_i^v(n{=}2)$	16.91%	37.91%	49.47%
68	CLIP+$[\texttt{CLS}]_i^v(n{=}4)$	**17.53%**	**38.10%**	50.03%
69	CLIP+$[\texttt{CLS}]_i^v(n{=}6)$	17.47%	38.06%	**50.10%**
70	CLIP+$[\texttt{CLS}]_i^v(n{=}8)$	16.78%	37.40%	48.71%
Time Reasoning				
71	CLIP+$[\texttt{CLS}]_i^v(n{=}2)$	1.90%	5.25%	45.62%
72	CLIP+$[\texttt{CLS}]_i^v(n{=}4)$	1.99%	**5.38%**	45.68%
73	CLIP+$[\texttt{CLS}]_i^v(n{=}6)$	**2.03%**	5.33%	**45.72%**
74	CLIP+$[\texttt{CLS}]_i^v(n{=}8)$	1.66%	5.16%	45.27%

Here, n represents the number of [CLS] tokens. The best results are in bold.

Table 8. The Results of the Effect of Increasing the Candidate Number of EK

ID	Candidate EK	Accuracy (Rank@1)	Rank@5	Example-F1
Location Reasoning				
75	29,243	18.26%	37.97%	50.29%
76	52,159	18.83%	38.29%	50.45%
77	122,408	**19.51%**	**38.48%**	**51.25%**
Time Reasoning				
78	29,243	2.26%	6.77%	47.85%
79	52,159	2.88%	10.67%	48.52%
80	122,408	**3.45%**	**10.97%**	**50.53%**

The best results are in bold.

perspectives. However, when n increases to 8, the model's performance slightly declines *(IDs: 68–70, 72–74)*. This may be due to the lack of unique features in an image to support too many [CLS] tokens, resulting in information redundancy.

(3) *Impact of EK's Number.* To validate the impact of varying amounts of EK in the environment, we conducted an experiment to determine whether increasing the number of EK is beneficial. As shown in Table 8, the addition of 122,408 EK resulted in more accurate reasons by the network (lift by 2.04% and 1.42%) for location and time *(IDs: 69 vs 83, 73 vs 86)*, compared to the method without EK. These results show that our method effectively utilizes EK to enhance the accuracy of the model for image location and time. Besides, the performance gradually improves as the number of EK increases *(IDs: 75–77, 78–80)*. It also shows that our method has the capability to explore a wider range of EK. However, comparing each $[\texttt{CLS}]_i^v$ with 122,408 EK is already time-consuming and limits the ability to increase the amount. In the future, we will strive to find a more efficient approach to overcome this challenge.

(4) *Analysis of Scoring Mechanism.* We evaluate the performance of different scoring mechanisms in the *Relevance* Module (Section 3.3), and the results are shown in Table 9. When utilizing Score_v, certain image features may be weakened, and the accuracy of time and location reasoning may decrease after fusing EK. When using the scoring mechanism on text (Score_t), only EK was considered during the fusion process. As a result, the accuracy of location and time reasoning improved by 3.11% and 0.11%, respectively *(IDs: 81 vs 82, 84 vs 85)*. This suggests that the weights exert a

Table 9. The Effect of Different Scoring Mechanisms on Network Performance

ID	Method	Accuracy (Rank@1)	Rank@5	Example-F1
Location Reasoning				
81	Score_v	15.45%	35.63%	47.87%
82	Score_t	18.56%	37.55%	50.94%
83	Proposed	**19.51%**	**38.48%**	**51.25%**
Time Reasoning				
84	Score_v	2.72%	10.61%	50.38%
85	Score_t	2.83%	10.43%	50.42%
86	Proposed	**3.45%**	**10.97%**	**50.53%**

Where score_v indicates that only images are scored and score_t means scoring EK only. The best results are in bold.

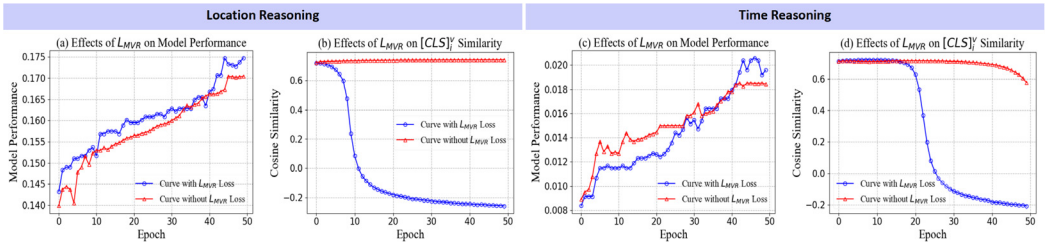

Fig. 5. This figure illustrates the impact of diversity on the performance of the *Quantity* Module. Among them, graph (a) compares the model's performance in location reasoning tasks with the blue curve representing the model's performance with L_{MVR} loss and the red curve representing the model's performance without L_{MVR} loss. Graph (b) tracks the changes in cosine similarity among $[\texttt{CLS}]_i^v$ during training. Similarly, graphs (c) and (d) present the results of experiments in time reasoning tasks.

significant influence on the final reasoning. When both image and EK embeddings are scored, the accuracy of location and time reasoning increases by 4.06% and 0.73%, respectively *(IDs: 81 vs 83, 84 vs 86)*. It is evident that by scoring both image and EK embeddings simultaneously, the model can more effectively optimize the connection between visual cues and EK, thereby improving the model's reasoning performance.

(5) *Impact of Multi-view representation images.* In our method, we enhance the model's ability to understand implicit information and search for valuable knowledge in the environment by encouraging it to represent images from multiple perspectives. To verify the effectiveness of this design, we evaluate the impact of the diversity of $[\texttt{CLS}]_i^v$ in the *Quantity* module and the duplication rates in the EK on the performance of method.

As shown in Figure 5, We first use the L_{MVR} (Equation (5)), which increases the distance between different $[\texttt{CLS}]_i^v$ during training, as the experimental variable to analyze the impact of the diversity of $[\texttt{CLS}]_i^v$ in the *Quantity* module on model performance. The incorporation of L_{MVR} leads to a peak accuracy of 17.47% (graph(a): blue curve), surpassing the model without L_{MVR}, which achieves an accuracy of 17.04% (graph(a): red curve). In addition, the inclusion of L_{MVR} results in a decrease in the similarity between each $[\texttt{CLS}]_i^v$ (graph(b): blue curve). Similar trends are also observed in the experimental results of time reasoning tasks (graph(c), graph(d)). These results show that the introduction of L_{MVR} enhances the dissimilarity between each $[\texttt{CLS}]_i^v$, allowing the model to effectively differentiate and capture multiple perspectives, thereby improving model performance.

Table 10. The Impact of Diversity in EK on Relevance Module Performance

ID	Method	Duplication Rate	Accuracy (Rank@1)	Rank@5	Example-F1
Location Reasoning					
87	Uniform Search	100%	19.17%	38.11%	50.97%
88	Distinct Search	0%	18.88%	37.97%	50.43%
89	Proposed	67.62%	**19.51%**	**38.48%**	**51.25%**
Time Reasoning					
90	Uniform Search	100%	3.16%	10.63%	50.11%
91	Distinct Search	0%	3.03%	10.74%	49.60%
92	Proposed	67.10%	**3.45%**	**10.97%**	**50.53%**

Where "Uniform Search" represents the approach of using the highest-scoring wiki entry for all searches, and "Distinct Search" involves searching through diverse wiki entries with a zero duplication rate. The best results are in bold.

Then, we evaluated the effect of diversity on performance using search-derived EK. As shown in Table 10, the use of the Distinct Search method results in a decrease in location reasoning accuracy compared to proposed method, due to the lack of diversity in the obtained EK *(IDs: 87 vs 89)*. Compared to other methods, despite obtaining six entirely different EKs, the model using Distinct Search had the lowest reasoning accuracy *(IDs: 87–89)*. A similar pattern was observed in Time Reasoning *(IDs: 90–92)*. The proposed method assigns unique scores to each $[\texttt{CLS}]_i^v$ and their corresponding EK, leading to superior performance in both tasks. It achieves a desirable equilibrium between diversity and reasoning capability, even in the presence of about 67% duplication in EK. Conversely, Uniform and Distinct Searches, focusing on consistency and diversity, resulted in lower performance.

4.5 Visualization

We present several visual demonstrations of *QR-CLIP* in Figure 6. The first figure shows the model's performance in a location reasoning task. Our *QR-CLIP* demonstrates significant improvements compared to vanilla CLIP [16], which served as a baseline and achieved lower Example-F1 scores (0%). *QR-CLIP* utilizes image search to obtain EKs containing location information related to visual content, such as Thailand. The scoring mechanism assigns weights to each EK, favoring those rich in valuable location details, thus guiding the model to emphasize the most relevant location information.

In the third picture, we explore the application of *QR-CLIP* to time reasoning. Here, vanilla CLIP serves as the baseline and achieves lower Example-F1 scores (50.00%). However, after using additional [CLS] and fine-tuning them using global and LLs, our *QR-CLIP* detects an image from different perspectives and gets higher scores (66.67%). Subsequently, *QR-CLIP* retrieves six EK used as language input, all of which describe the information expressed in the image content: a public participation activity. In addition, each piece of knowledge contains a wealth of information regarding the time associated with the activity. The scoring mechanism assigns varying weights to each EK, with the EK lacking valuable time information receiving a lower weight. This guides the model to focus on the correct time information.

5 Conclusion and Future Work

In our study, we developed *QR-CLIP*, a model inspired by Hutchins's distributed cognition theory, for image-based location and time reasoning tasks. It contains two Modules. The *Quantity* Module enhances cognitive abilities by providing a suite of cognitive tools aimed at aggregating a maximal

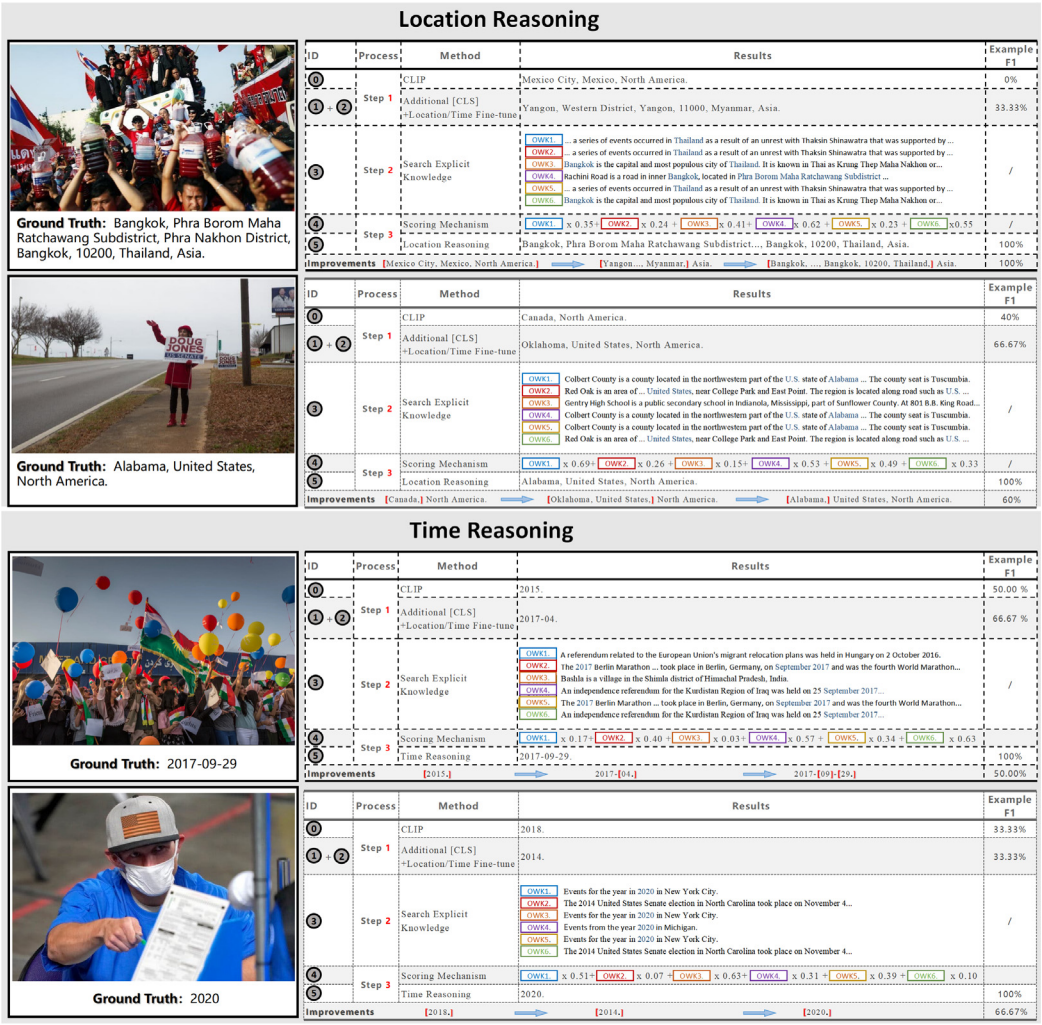

Fig. 6. We show visualizations of five procedures in *QR-CLIP*. For each process, readers can refer to Figure 3.

amount of EK from the surrounding environment, thereby broadening the scope of cognitive resources. The *Relevance* Module integrates relevant information from various cognitive tools to produce a comprehensive cognitive output. This synergy aligns with the distributed cognition theory, which posits that cognition is distributed among individuals, tools, and environments. Through this conceptual alignment, our *QR-CLIP* outperforms previous SOTA methods, achieving an improvement of 3.05% in location reasoning accuracy. In the challenging task of time reasoning, which demands reasoning the exact day in the benchmark, our model shows a significant improvement in accuracy, with a 2.45% increase at Rank@1 and a remarkable improvement from 5.53% to 10.97% at Rank@5. However, the model's performance is contingent on the quality and quantity of available knowledge. Insufficient knowledge may also impede the reasoning results. Future work will focus on enhancing reasoning capabilities by refining its architecture and algorithms to handle more complex tasks. Furthermore, incorporating more EK will augment the accuracy of reasoning.

Acknowledgments

The authors would like to thank Mingchen Zhuge from King Abdullah University of Science and Technology, Saudi Arabia, for his helpful discussions related to this work and assistance with writing. We are also grateful to Professors Ming-Ming Cheng and Deng-Ping Fan from Nankai University, Tianjin, China, for their valuable suggestions during the revision of the paper. Additionally, we appreciate the anonymous reviewers and the editor for their constructive feedback and support in improving the manuscript.

References

[1] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision And Pattern Recognition*, 16420–16429.

[2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[3] Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 3 (2021), 1–23.

[4] Lu Zhang, Jialie Shen, Jian Zhang, Jingsong Xu, Zhibin Li, Yazhou Yao, and Litao Yu. 2021. Multimodal marketing intent analysis for effective targeted advertising. *IEEE Transactions on Multimedia* 24 (2021), 1830–1843.

[5] Nengjun Zhu, Jian Cao, Kunwei Shen, Xiaosong Chen, and Siji Zhu. 2020. A decision support system with intelligent recommendation for multi-disciplinary medical treatment. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 1s (2020), 1–23.

[6] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. 2022. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2022), 4425–4445.

[7] Lu Zhang, Jingsong Xu, Yongshun Gong, Litao Yu, Jian Zhang, and Jialie Shen. 2021. Unsupervised image and text fusion for travel information enhancement. *IEEE Transactions on Multimedia* 24 (2021), 1415–1425.

[8] Jie Wen, Nan Jiang, Lang Li, Jie Zhou, Yanpei Li, Hualin Zhan, Guang Kou, Weihao Gu, and Jiahui Zhao. 2024. TA-detector: A GNN-based anomaly detector via trust relationship. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024). Retrieved from https://dl.acm.org/doi/pdf/10.1145/3672401

[9] Jialie Shen and Neil Robertson. 2021. BBAS: Towards large scale effective ensemble adversarial attacks against deep neural network learning. *Information Sciences* 569 (2021), 469–478.

[10] Tawfiq Salem, Scott Workman, and Nathan Jacobs. 2020. Learning a dynamic map of visual appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12435–12444.

[11] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. 2020. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 11990–11997.

[12] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4938–4947.

[13] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. Geollm: Extracting geospatial knowledge from large language models. arXiv:2310.06213. Retrieved from https://arxiv.org/abs/2310.06213

[14] Gabriele Berton, Carlo Masone, and Barbara Caputo. 2022. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4878–4888.

[15] Royston Rodrigues and Masahiro Tani. 2022. Global assists local: Effective aerial representations for field of view constrained image geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3871–3879.

[16] 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*. PMLR, 8748–8763.

[17] 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*. PMLR, 12888–12900.

[18] Xingyu Fu, Ben Zhou, Ishaan Chandratreya, Carl Vondrick, and Dan Roth. 2022. There's a time and place for reasoning beyond the image. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, Long Papers, 1138–1149.

[19] Gengyuan Zhang, Yurui Zhang, Kerui Zhang, and Volker Tresp. 2024. Can vision-language models be a good guesser? Exploring VLMs for times and location reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 636–645.
[20] Edwin Hutchins. 1991. *The Social Organization of Distributed Cognition.* American Psychological Association.
[21] E. Hutchins. 1995. *Cognition in the Wild*. MIT Press.
[22] Edwin Hutchins. 2000. *Distributed Cognition.* Elsevier Science.
[23] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics,* 4171–4186.
[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929.
[25] Lisi Chen and Shuo Shang. 2019. Region-based message exploration over spatio-temporal data streams. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 873–880.
[26] Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2022. Temporal common sense acquisition with minimal supervision. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics,* 7579–7589.
[27] Qiang Ning, Ben Zhou, Hao Wu, Haoruo Peng, Chuchu Fan, and Matt Gardner. 2022. A meta-framework for spatiotemporal quantity extraction from text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics,* Vol. 1, Long Papers, 2736–2749.
[28] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? arXiv:2302.11713. Retrieved from https://arxiv.org/abs/2302.11713
[29] Xiujie Song, Mengyue Wu, Kenny Q. Zhu, Chunhao Zhang, and Yanyi Chen. 2024. A cognitive evaluation benchmark of image reasoning and description for large vision language models. arXiv:2402.18409. Retrieved from https://arxiv.org/abs/2402.18409
[30] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21466–21474.
[31] Hyeongjun Kwon, Taeyong Song, Somi Jeong, Jin Kim, Jinhyun Jang, and Kwanghoon Sohn. 2023. Probabilistic prompt learning for dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6768–6777.
[32] Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. 2023. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5606–5611.
[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*. PMLR, 1597–1607.
[34] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*. Springer, 493–510.
[35] 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*. PMLR, 4904–4916.
[36] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*. Springer, 529–544.
[37] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. 2022. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 35959–35970.
[38] Dingkang Liang, Jiahao Xie, Zhikang Zou, Xiaoqing Ye, Wei Xu, and Xiang Bai. 2023. Crowdclip: Unsupervised crowd counting via vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2893–2903.
[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
[40] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* 568–578.

[41] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 6827–6839.

[42] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

[43] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

[44] Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Multi-view document representation learning for open-domain dense retrieval. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 5990–6000.

[45] Prajwal Gatti, Abhirama Subramanyam Penamakuri, Revant Teotia, Anand Mishra, Shubhashis Sengupta, and Roshni Ramnani. 2022. COFAR: Commonsense and factual reasoning in image search. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, Vol. 1, Long Papers, 1185–1199.

[46] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2443–2449.

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

[48] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[49] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS '22)*, 21548–21561.

[50] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022. Language models can see: Plugging visual controls in text generation. arXiv:2205.02655. Retrieved from https://arxiv.org/abs/2205.02655

Received 15 December 2023; revised 29 June 2024; accepted 15 August 2024