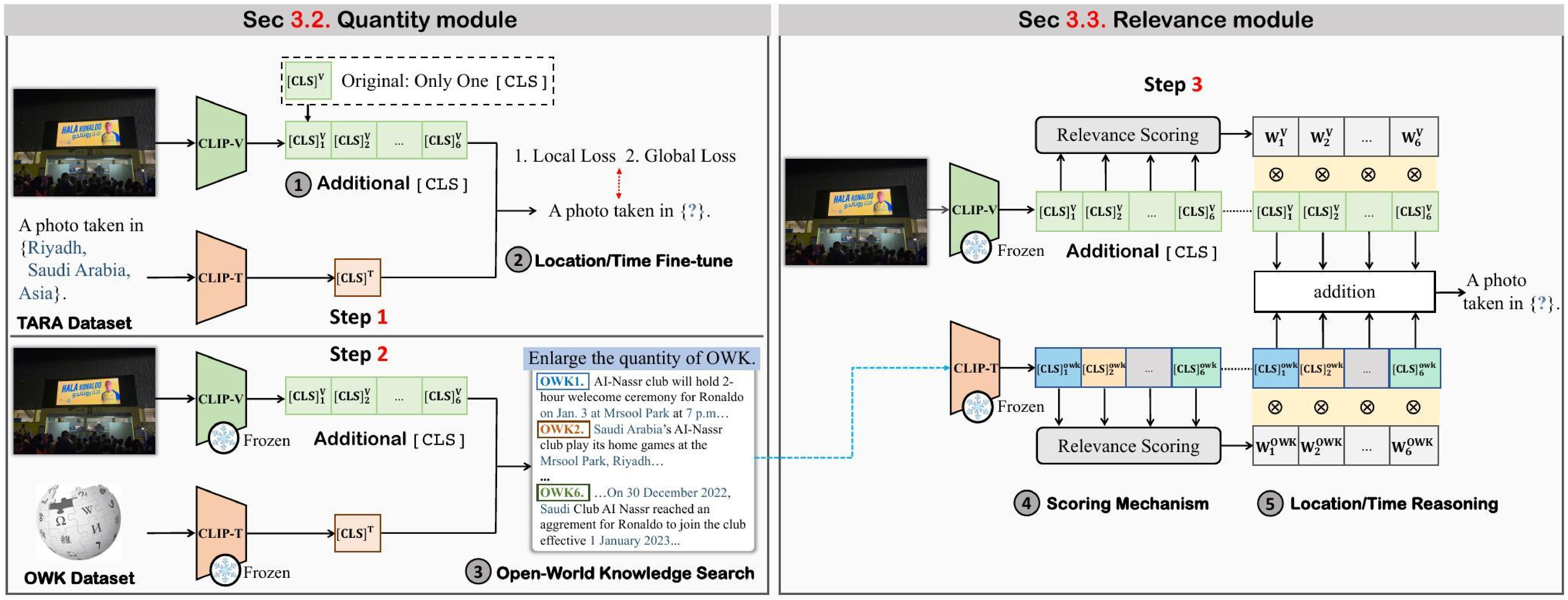
Sec 3.2. Quantity module
[CLS]^V Original: Only One [CLS]
CLIP-V
[CLS]$^V_1$ [CLS]$^V_2$ ... [CLS]$^V_6$
1 Additional [CLS]
1. Local Loss 2. Global Loss
A photo taken in {?}.
A photo taken in {Riyadh, Saudi Arabia, Asia}.
CLIP-T
[CLS]^T
2 Location/Time Fine-tune
TARA Dataset
Step 1
Step 2
Frozen
Additional [CLS]
Enlarge the quantity of OWK.
OWK Dataset
3 Open-World Knowledge Search
Sec 3.3. Relevance module
Step 3
Relevance Scoring
W$^V_1$ W$^V_2$ ... W$^V_6$
⊗ ⊗ ⊗ ⊗
[CLS]$^{owk}_1$ [CLS]$^{owk}_2$ ... [CLS]$^{owk}_6$
addition
A photo taken in {?}.
W$^{OWK}_1$ W$^{OWK}_2$ ... W$^{OWK}_6$
4 Scoring Mechanism
5 Location/Time Reasoning

OWK1. AI-Nassr club will hold 2-hour welecome ceremony for Ronaldo on Jan. 3 at Mrsool Park at 7 p.m…

OWK2. Saudi Arabia's AI-Nassr club play its home games at the Mrsool Park, Riyadh…

...

OWK6. …On 30 December 2022, Saudi Club AI Nassr reached an aggrement for Ronaldo to join the club effective 1 January 2023...

Figure 2. The pipeline of QR-CLIP. It has two modules: Quantity module (Sec 3.2) and Relevance module (Sec 3.3). In step 1, we add additional [CLS] to mimic different perspectives of different individuals and design the local and global loss to guide the location/time fine-tuning. Then, we freeze the fine-tuned CLIP-V and CLIP-T and use them to search for open-world knowledge from our OWK dataset (Sec 4.1). In the Relevance module, we use a scoring mechanism to weights the most valuable information from CLIP-T and CLIP-V. After multiplying the scoring weights for vision and language separately, we add them for final similarity calculation.

2.2. Location and Time Reasoning

Existing language models achieved significant success on a wide range of tasks that require an understanding of language (Kenton & Toutanova, 2019; Lewis et al., 2020). Also, the vision models (He et al., 2016; Dosovitskiy et al., 2020; Liu et al., 2021) can predict the correct class label of an image from thousands of options. But they still struggle with many reasoning tasks, such as discerning the abstractive meanings (*e.g.,* time, location, event) of images (Yang et al., 2020; Tahmasebzadeh et al., 2021) or performing mathematical calculations (Lewkowycz et al., 2022) and science deduction (Degrave et al., 2022). However, humans can do these things well because real brains are more powerful than artificial neural networks in many ways and actively learn to conduct abstract reasoning (Schmidhuber, 2015). We focus on location and time reasoning (Fu et al., 2022), which needs a model to think and reason beyond the actual content of an image. Compared to other reasoning tasks, it differs in that most of the time there aren't enough visual cues to make inferences, so auxiliary knowledge is required.

3. Approach

3.1. Preliminary

Task Background. Current AI methods are comparably weak in deducing the abstract information hidden behind an image. The goal of this paper is to let the model reason the location and time based on image input (Fu et al., 2022): Given an image I, we need the model ($\mathrm{M}(I)$) to predict the location (Pred_l) and time (Pred_t).

Horn's QR Theory. In cognitive research (Allott, 2013), it is believed that human reasoning is the process of obtaining the best correlations. The QR theory (Horn, 1984) builds a more clear definition of the above correlations, where Q stands for the quantity principle, which requires the maximization of valuable content, and R means the relevance principle that requires the minimization of form and extracts most related information:

$$\text{Correlation} \uparrow \longleftrightarrow Q \uparrow \times R \uparrow . \tag{1}$$

Therefore, we should simultaneously consider the quantity and relevance of information to achieve a higher correlation.

Our Pipeline. Following the human reasoning process, this paper proposes QR-CLIP. It is based on the contrastive language-image pretraining (CLIP) model (Radford et al., 2021), which was pre-trained with 400M internet data. The QR-CLIP is made up of the quantity module (Sec 3.2) and the relevance module (Sec 3.3). The quantity module aims to add diversity to the model's outputs. Here, we introduce the additional [CLS] method, which aims to generate multiple distinct $\texttt{[CLS]}_i$ to imitate different perspectives of a single image. The relevance module uses a scoring mechanism to weigh the retrieved open-world knowledge and image features to ensure the most relevant information and combine them for final prediction.

3.2. Quantity Module

Additional [CLS]. We employ CLIP (Radford et al., 2021) as our basic architecture. In vanilla CLIP, [C$\hat{\text{L}}$S] is used to distinguish the input token that represents the whole input features which is a common practice in other trans-