4.2. Comparative Results

Location Reasoning. We compare the results of our QR-CLIP with other methods for location reasoning in Tab. 1. QR-CLIP, achieves accuracy of 19.31% Accuracy (R@1). Meanwhile, its Example-F1 score for the hierarchical labels is 50.96%. All the results clearly show that our method outperforms other methods.

(1) Compared with ResNet-50 (He et al., 2016) and Swin-T (Liu et al., 2021), vanilla CLIP achieves 7.93% and 4.41% absolute improvement in location prediction accuracy (ID: 1,2,3). It indicates that compared with the vision model only trained on ImageNet (Deng et al., 2009), CLIP already possesses some knowledge for reasoning. Meanwhile, our QR-CLIP achieves a more significant advantage with 16.13% and 12.61% absolute improvements in terms of accuracy (ID: 1,2,6). These results show that traditional image classification methods cannot accomplish inference of the abstract information behind the images. While the CLIP model trained on large-scale internet data have the ability to identify locations based on image data, and QR-CLIP significantly enhances this capability.

(2) Besides, compared to CLIP† and the state-of-the-art method CLIP+Seg, QR-CLIP improves the accuracy by 3.59% and 2.85% absolute improvement, the F1-Score has increased by 3.88% and 3.07%, respectively (ID: 4,5,6). Other evaluation metrics also improved. The results show that QR-CLIP can effectively utilize open-world knowledge to establish a closer connection between image and location information through fine-tuning CLIP. However, we also find that the improvement in Example-F1 is not as obvious. We argue that this is because the mechanism of Example-F1: take the image of Fig. 1 as an example—the picture show many Arabia elements (turban and Arabic). 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.

Time Reasoning. Tab. 1 also presents the performance of our method and existing techniques for time reasoning. The Accuracy (R@1) of QR-CLIP is 3.53%, and Example-F1 is 47.89%; compared to the CLIP model, the two metrics have been absolutely improved by 3.07% and 7.99%, respectively (ID: 9,12). Compared with CLIP† and CLIP+Seg, which are also based on CLIP fine-tuning, our method obtains 2.53% and 2.61% improvement in the accuracy of prediction time, respectively. Compared with traditional image classification methods, QR-CLIP has achieved absolute advantages in all metrics (ID: 7,8,12). In addition, due to the lack of time-related information in the image, the prediction accuracy of fine-tuning CLIP methods for image time can only reach about 1%, which is significantly lower than the accuracy of location prediction (ID: 10,11). This is not surprising, also take the image on Fig. 1 as sample: even for humans, it is difficult to determine that {'*03-01-2023*'} is the time when this photograph was taken, if they are unfamiliar with Cristiano Ronaldo or some specific knowledge. Nevertheless, the method proposed in this paper is still effective (that our model achieves +253.00% relative lift) for predicting time and significantly closes the gap with location prediction.

ID	Method	Accuracy (∼ Rank@1)	Rank@5	Example-F1
		Location Reasoning		
13	CLIP+$[\texttt{CLS}^*]^v_i$(n=2)	9.69%	27.17%	44.37%
14	CLIP+$[\texttt{CLS}^*]^v_i$(n=4)	9.53%	26.25%	43.23%
15	CLIP+$[\texttt{CLS}^*]^v_i$(n=6)	9.21%	27.05%	43.69%
16	CLIP+$[\texttt{CLS}]^v_i$(n=2)	16.84%	37.47%	49.22%
17	CLIP+$[\texttt{CLS}]^v_i$(n=4)	17.11%	37.60%	49.51%
18	CLIP+$[\texttt{CLS}]^v_i$(n=6)	17.25%	37.80%	49.98%
19	CLIP+$[\texttt{CLS}]^v_i$(n=8)	17.03%	37.62%	48.93%
		Time Reasoning		
20	CLIP+$[\texttt{CLS}^*]^v_i$(n=2)	0.98%	3.03%	42.18%
21	CLIP+$[\texttt{CLS}^*]^v_i$(n=4)	1.03%	2.99%	43.98%
22	CLIP+$[\texttt{CLS}^*]^v_i$(n=6)	1.08%	3.15%	43.62%
23	CLIP+$[\texttt{CLS}]^v_i$(n=2)	1.84%	5.14%	45.57%
24	CLIP+$[\texttt{CLS}]^v_i$(n=4)	1.92%	5.21%	45.63%
25	CLIP+$[\texttt{CLS}]^v_i$(n=6)	2.00%	5.37%	45.60%
26	CLIP+$[\texttt{CLS}]^v_i$(n=8)	1.53%	5.06%	45.15%

Table 2. Performance of additional [CLS] in QR-CLIP with different number and prediction methods. Whereas $[\texttt{CLS}^*]^v_i$ refers to fusing all additional [CLS] by MLPs and then calculating the similarity with location and time labels, $[\texttt{CLS}]^v_i$ refers to calculating the similarity between each additional [CLS] with labels separately, and then using the ($[\texttt{CLS}]^v_i$-label) pair with the greatest similarity as the prediction. n represents the number of [CLS].

4.3. Ablation Study

Analysis on Additional [CLS]. Following the network design process, all experiments of this part were conducted on the setting with only the step 1 in Quantity module (Sec 3.2).

As shown in Tab. 2, both of different [CLS] aggregation methods and different numbers of [CLS] can affect network performance. Comparing $[\texttt{CLS}^*_i]$ and $[\texttt{CLS}]^v_i$ with the same number (*i.e.,* $n = 2$) of [CLS], the latter has 7.15% and 0.86% higher location and time prediction accuracy (ID: 13,16,20,23). Besides, the performance of $[\texttt{CLS}^*_i]$ is not significantly affected by the number of [CLS] (ID: 13-15,20-22). We argue that using MLP to aggregate the embeddings may destroy CLIP's original representation. It is better to separately calculate the similarities across each $[\texttt{CLS}]^v_i$ with the location and time labels and then select the one with the most significant value as the prediction. Then we analyze how different numbers of [CLS] affect the model performance. When n was increased to 8, no significant performance difference was observed, so we finally chose $n = 6$ in the following experiments (ID: 17-19, 24-26). The results indicate that the additional [CLS] effectively