ID	Method	Accuracy (∼ Rank@1)	Rank@5	Example-F1
	Location Reasoning (Only QM)			
27	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*LL*	16.77%	36.89%	49.63%
28	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*GL*	17.13%	37.75%	49.87%
29	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*LL*+*GL*	17.25%	37.80%	49.98%
	Time Reasoning (Only QM)			
30	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*LL*	1.23%	5.33%	43.59%
31	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*GL*	1.92%	5.60%	44.62%
32	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*LL*+*GL*	2.00%	5.37%	45.60%
	Location Reasoning (QR-CLIP: QM+RM)			
33	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*LL*	19.11%	37.74%	50.51%
34	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*GL*	18.36%	37.87%	50.03%
35	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*LL*+*GL*	19.31%	38.78%	50.96%
	Time Reasoning (QR-CLIP: QM+RM)			
36	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*LL*	3.22%	11.57%	47.80%
37	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*GL*	2.98%	9.60%	46.32%
38	CLIP+$[\texttt{CLS}]_i^v$(n=6)+*LL*+*GL*	3.53%	10.90%	47.89%

Table 3. The impact of various loss functions and components on performance. *LL*, *GL* indicate the local loss and global loss, respectively. QR-CLIP means the model contains entirely Quantity module (QM: Sec. 3.2) and Relevance module (RM: Sec. 3.3).

ID	Candidate OWKs	Accuracy (∼ Rank@1)	Rank@5	Example-F1
	Location Reasoning			
39	29,243	17.75%	37.75%	50.32%
40	52,159	18.90%	37.94%	50.84%
41	122,408	19.31%	38.78%	50.96%
	Time Reasoning			
42	29,243	1.59%	5.87%	45.64%
43	52,159	2.96%	10.65%	47.77%
44	122,408	3.53%	10.90%	47.89%

Table 4. The results of the effect of increasing the candidate numbers of Open-World Knowledge (OWK).

ID	Method	Accuracy (∼ Rank@1)	Rank@5	Example-F1
	Location Reasoning			
45	Score$_v$	16.43%	36.74%	49.97%
46	Score$_t$	18.38%	37.53%	50.19%
47	Proposed	19.31%	38.78%	50.96%
	Time Reasoning			
48	Score$_v$	2.76%	10.59%	47.53%
49	Score$_t$	2.92%	10.37%	47.60%
50	Proposed	3.53%	10.90%	47.89%

Table 5. The effect of different scoring mechanisms on network performance, where Score$_v$ indicates that only images are scored and Score$_t$ means scoring open-world knowledge only.

increases image cues by constructing multiple perspectives and has a promising benefit.

Effectiveness of Losses and Modules. We further analyze the impact of losses, the Quantity Module (Sec 3.2) and Relevance Module (Sec 3.3). As shown in Tab. 3, adding both the local and global losses will increase model performance, which first indicates the effectiveness of these two losses (ID: 27-38). When we compare the Quantity module to the entire QR-CLIP, we can see that the Relevance module significantly improved the reasoning abilities (ID: 29,32,35,38), which verifies that the whole designs of the two modules are reasonable.

Impact of Open-World Knowledge. To validate the performance of different amounts of open-world knowledge, we conduct an experiment to vilify whether increasing the number of OWKs is beneficial. As shown in Tab. 4, when 122,408 OWKs were added, compared to the method without open-world knowledge, the network was able to make more accurate predictions (absolute lift by 2.06% and 1.53%) about location and time (ID: 18,25,41,44). These results show that our method can effectively use open-world knowledge to improve the model's accuracy for image location and time. Besides, the performance gradually improves as the number of OWKs increases (ID: 39-44). It also shows that our method can further explore a more extensive range of open-world knowledge. Nonetheless, comparing each $[\texttt{CLS}]_i^v$ with $122,408$ OWKs is already time-consuming and limits our ability to increase the amount; in the future, we will find a more efficient way to overcome this challenge.

Performance of Scoring Mechanism. This part analyzes the performance of different scoring mechanisms in the Relevance module (Sec 3.3), and the experimental results are shown in Tab. 5. When we used the Score$_v$, some image features were even weakened, and the time and location prediction accuracy decreased after fusing open-world knowledge (ID: 45,47,48,50). When use the scoring mechanism on text (Score$_t$)—only the open world knowledge was weighted during the fusion process—the accuracy of location and time prediction was improved by 1.13% and 0.92%, respectively (ID: 45,46,48,49). This indicates that the weights have a significant influence on the final predictions. When both image and open-world knowledge embeddings are scored, the accuracy of location and time predictions increases by 2.06% and 1.50%, respectively (ID: 45,47,48,50). This implies that providing only the information required for the final prediction helps our QR-CLIP better understand abstract information and caters to the idea of the QR rule.

4.4. Visualization

We provide a visual demonstration for QR-CLIP in Fig. 3. Taking the fourth picture as an example, when we use vanilla CLIP (Radford et al., 2021) as the baseline, as can be seen, it performs worse in this case, achieving lower Example-F1 scores (22.22%). After using additional $[\texttt{CLS}]$ and fine-tuned them using global and local losses, our QR-CLIP detects an image from different perspectives and get higher scores (28.57%). After that, QR-model retrieves six OWKs used as language input; the six OWKs all describe the abstract information expressed in the image content: an election meeting. In addition, each piece of knowledge contains much information about the time associated with the meeting. The scoring mechanism then assigns different weights to each OWK, with the OWK that lacks valuable time information receiving a lower weight, guiding the model to pay attention to the correct time information.