former models (Kenton & Toutanova, 2019; Dosovitskiy et al., 2020).

To simplify the illustration and better represent our model in the following sections:

$$[\texttt{CLS}]_i^v \leftarrow \text{Enc}_v([\hat{\texttt{CLS}}]_i^v) \text{ and } [\texttt{CLS}]^t \leftarrow \text{Enc}_t([\hat{\texttt{CLS}}]^t), \quad (2)$$

by default, when talking about $[\texttt{CLS}]$ ($[\texttt{CLS}]_i^v$ and $[\texttt{CLS}]^t$), they are the CLIP's output embeddings rather than the input token ($[\hat{\texttt{CLS}}]$), the same as shown in Fig. 2.

However, the output embedding `[CLS]` cannot adequately represent an image as the single embedding provides limited cues to the location and time reasoning. Therefore, we consider enlarging the representations for an image. It is evident that in real life, *we would get different ideas about what an image means from different individuals*. Following in this vein, we propose a simple yet effective methods: in our technical implementation, we are inspired by MVR (Zhang et al., 2022) and introduce additional $[\texttt{CLS}]_i^v (i = 1, ..., n)$ to replace the original single `[CLS]` representation. Through the observation of ablations, we finally use 6 different $[\hat{\texttt{CLS}}]_i^v$ at the beginning of the image patch token embeddings ($\hat{I} = \hat{I}_1^{patch}, ..., \hat{I}_7^{patch}$), like ($[\hat{\texttt{CLS}}]_1^v ... [\hat{\texttt{CLS}}]_6^v\ \hat{I}$), and after going through the encoder Enc_v, we get a list of embeddings ($[\texttt{CLS}]_1^v ... [\texttt{CLS}]_6^v\ I$). Using this design, the pre-trained model can investigate an image from multiple perspectives and dimensions. It follows the Q-principle and increases the quantity of information from multiple perspectives.

Since the text contains explicit semantic information and most language inputs carry clear messages, we only use the original $[\texttt{CLS}]^t$ at the beginning of the text token embedding (T), like ($[\texttt{CLS}]^t\ T$). Hence, we search for corresponding information through the image from the CLIP model by conducting:

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

In the following fine-tuning or searching for open-world knowledge, each $[\texttt{CLS}]_i^v$ of the Enc_v calculates the similarity with $[\texttt{CLS}]^t$ of the candidate information by inner-product.

Location/Time Fine-tune. We first initialize and position-encode each $[\texttt{CLS}]_i^v$ individually, aiming to extend the distance between each $[\texttt{CLS}]_i^v$. Then, we fine-tune CLIP with local and global losses (He et al., 2020; Zhang et al., 2022) to ensure each $[\texttt{CLS}]_i^v$ is aligned with the location and time linguistic features $[\texttt{CLS}]^t$.

For the local loss, the correspondence between each $[\texttt{CLS}]_i^v$ and $[\texttt{CLS}]^t$ is achieved by a contrastive learning loss:

$$L_{local} = -\frac{1}{i+1} \sum_0^i \log \frac{e^{f_i(q_v, k_{t+})}}{\sum_1^n [e^{f_i(q_v, k_{t+})} + e^{f_i(q_v, k_{t-})}]}, \quad (4)$$

here, q_v denote the query image embedding ($[\texttt{CLS}]_i^v$); k_{t+} and k_{t-} are the positive and negative key text embeddings (a batch of $[\texttt{CLS}]^t$). We calculate the correlation score between them by inner product $f_i(x, y)$.

Then, the global loss further constrains the correspondence between image features and location/time features, and the calculation method is as follows:

$$L_{global} = -\log \frac{e^{f_{max}(q_v, k_{t+})}}{\sum_1^n [e^{f_{max}(q_v, k_{t+})} + e^{f_{max}(q_v, k_{t-})}]}, \quad (5)$$

where $f_{max}(q_v, k_t) = \max_i \{f_i(q_v, k_t)\}$, $\max_i \{\}$ represents the maximum value. The entire training loss is defined as a linear combination of two losses as $L_{total} = L_{local} + L_{global}$.

Open-World Knowledge Search. After fine-tuning, each $[\texttt{CLS}]_i^v$ output by CLIP-V can represent image location/time information from various perspectives. We use these different representations to retrieve more valuable open-world knowledge from the OWK dataset (Sec 4.1) to increase the quantity of knowledge.

Given an image I and the corresponding Open-World Knowledge ($O = T_1^{owk}, T_2^{owk}, ..., T_k^{owk}, k = 122,408$), the process of searching follows Eq. 3: each $[\texttt{CLS}]_i^v$ calculates the similarity with 122,408 candidate Wikipedia corpus (OWK). Here, we select the candidate Wikipedia with the top-1 similarity for each $[\texttt{CLS}]_i^v$, yielding a total of 6 OWKs. After that, the Quantity module (sec 3.2) finished its job by collecting a list of highly-related OWKs items that the next Relevance module (sec 3.3) would use as input for CLIP-T.

3.3. Relevance Module

Scoring Mechanism. The amount of useful information varies for each $[\texttt{CLS}]_i^v$ of an image and the corresponding embeddings of open-world knowledge. As a result, it is critical to weigh the importance of different features dynamically. Based on the above motivation, we propose a scoring mechanism to further highlight the relevant features.

We adopt two layers of `MLP` ($\text{MLP}_{2-layer}$) as our relevance scoring component and find it helpful:

$$W^x = \text{MLP}_{2-layer}([\texttt{CLS}]_i^x), \quad (6)$$

Here, $[\texttt{CLS}]_i^x$ is the input embedding, and W^x is the calculated weight. We use contrastive learning to optimize the model. To facilitate implementation, we directly adopt the loss functions from the first step of the Quantity module (sec 3.2). In this case, we keep the CLIP-T and CLIP-V frozen and only update the parameters of the relevance scoring component.

In the local loss, the information of two features is integrated tto optimize the scoring mechanism jointly:

$$f_i(q, k_+) = (W_i^{owk} \times [\texttt{CLS}]_i^{owk} + W_i^v \times [\texttt{CLS}]_i^v), \cdot F^{gt}, \quad (7)$$