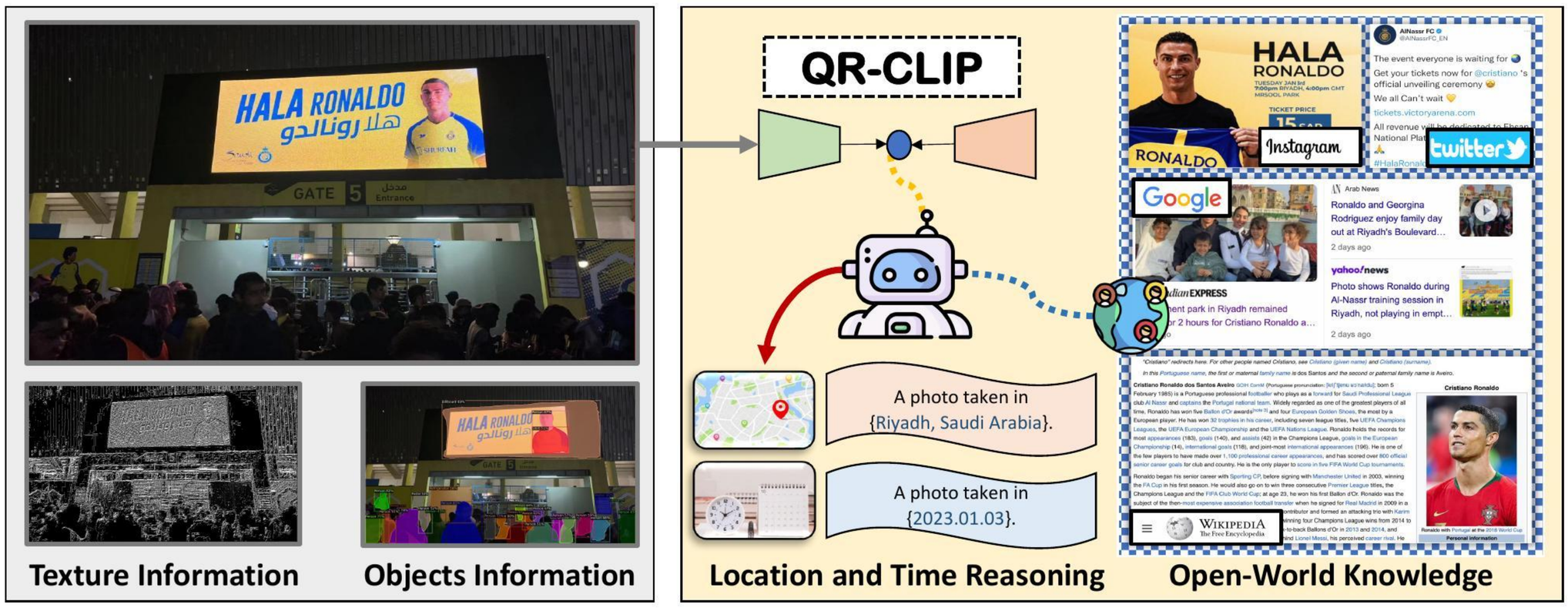

Figure 1. Comparisons of traditional computer vision tasks (*left*) with location and time reasoning (*right*). It is clear that—instead of simple image color, texture, and object information—location and time reasoning requires more human experiences and knowledge (*a.k.a.* open-world knowledge).

Regarding the Q-principle, we design the **Quantity** module: we utilize two techniques to improve the model to provide as much information as possible. For traditional transformer-based models (Kenton & Toutanova, 2019; Dosovitskiy et al., 2020), they use a single `[CLS]` to represent the input. Initially, we design additional `[CLS]` tokens that mimic different human perspectives on the same image. Since everyone has unique knowledge and experience, it is possible to gain a more comprehensive understanding of a given item by combining the knowledge of different individuals. This also inspires us to use each $\texttt{[CLS]}_i$ to retrieve various useful open-world knowledge to aid predictions. Furthermore, motivated by current contrastive learning methods (Zhang et al., 2022), we design the local and global loss for fine-tuning the CLIP model, ensuring our QR-CLIP is suitable for the tasks.

For the **Relevance** module, we design a scoring mechanism that weights the fusion of image and open-world knowledge embeddings. Like ordinary society, not all of the knowledge from individuals are correct. Thus our scoring mechanism is like an error correction tool to help the model select the most valuable information. It adaptively balances the different information and encourages the model to provide pertinent ones for location and time reasoning. Furthermore, the scoring mechanism balances vision and language knowledge, which means that when the quality of explicit OWK is poor, our scoring mechanism can focus more on original image features, greatly improving robustness.

The experiments indicate the strong abilities of our QR-CLIP model. Considering the accuracy (or Rank@1) achieves 19.31% (17.3% relative improvement *compared to previous SOTA*) on location reasoning, and 3.53% (253% relative improvement) on time reasoning.

Overall, our contributions can be categorized as:

- We design the **QR-CLIP**, which first investigates utilizing explicit open-world knowledge to help location and time reasoning.
- Our method achieves an average of 17.3% / 253% relative lifts on location and time reasoning tasks compared to the previous SOTA.
- In addition, the comprehensive experimental record will inspire the related field.

2. Related Work

2.1. Foundation Models

The emergence of foundation models is a relatively recent phenomenon (Bommasani et al., 2021), and has fundamentally altered the game rules of AI communities. They are commonly trained on a massive amount of unlabeled data at scale (typically via self-supervised learning (Radford et al., 2021)), making them adaptable to various downstream applications. Among these are popular models include GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022) as large language model, CLIP (Radford et al., 2021) and Flamingo (Alayrac et al., 2022) as vision-language model, Dall-E (Ramesh et al., 2021) and Stable Diffusion (Rombach et al., 2022) for text-to-image generation, GaTo (Reed et al.) as a generalist model, and ChatGPT (Ouyang et al., 2022) as human-like conversation agent, *etc*. This paper uses CLIP pre-trained with 400 million image-text pairs as baseline architecture. It learns excellent open-world knowledge and multi-modal representation by learning with such a large-scale corpus, making it ideal as the basic solution. Based on it, we made QR-CLIP to fit the location and time reasoning task.