Required Background Knowledge

This work assumes that the reader has a thorough understanding of machine learning, deep learning, and neural networks. In particular, familiarity with concepts such as loss functions, activation functions (e.g. softmax), backpropagation, and optimization algorithms in the context of neural networks is essential. Since this work focuses on vision-language models, a basic understanding of computer vision, and most importantly, natural language processing (NLP) is required. The latter includes knowledge about tokenization, word embeddings, and while RNNs are not used in this work, familiarity with them and the **self-attention mechanism** is recommended.

Because the work involves utilizing pretrained models, the reader should be aware of concepts like transfer learning, finetuning pretrained models, and using pretrained models as feature extractors.

A deep understanding of multimodal models and how they are trained is not required, as this work provides a detailed explanation of the methodology used.