Broader Impact

This work focused on reducing the computational cost of training vision-language models by leveraging pretrained unimodal models. Even though our approach does not reach state-of-the-art performance, which is not to be expected given that the models we compare to have been developed by organizations such as OpenAI, Meta, and Microsoft, it clearly demonstrates the potential of leveraging existing components to generate new model paradigms, such as multimodal models.

We hope that this proof-of-concept will inspire other researchers to explore new efficient ways of training multimodal models, and any models in general, to make the technology more accessible. While the current trend in deep learning is