Goals and Contributions

Given the challenges associated with training large multimodal models from scratch, we identify the opportunity to utilize the vast amount of existing pretrained unimodal models. By extracting and combining knowledge from these models, we can generate multimodal models, for the proof-of-concept developed in this work vision-language models, that are *smaller* in size and significantly *cheaper* to train. Importantly, the approach is not limited to vision and language, meaning it can be adapted to other combinations of modalities wherever pretrained unimodal models are available.

Pretrained unimodal models, such as those for computer vision and natural language processing, have already learned rich representations within their respective domains through extensive (pre-)training. By aligning and integrating these representations, we can create a multimodal model without the need for large-scale training from scratch. This methodology can be extended to other modalities, such as audio, video, by leveraging existing pretrained models in those domains.

It is particularly interesting to use an end-to-end self-supervised approach because it has the advantage that it avoids the scalability issues associated with supervised learning, such as the need for large labeled datasets. Self-supervised learning allows us to leverage the raw data available across modalities, making the training process more scalable and cost-effective. Additionally, using self-supervised methods throughout, including the pretrained parts, ensures that the entire end-to-end process benefits from scalable learning techniques, regardless of the specific modalities involved.

The goal is to develop a multimodal model that performs well enough to serve as a successful proof-of-concept in the vision-language domain, demonstrating that this approach is viable. While it may not match the performance of large enterprise models—a goal that is generally unrealistic for individual researchers without access to extensive computational resources—achieving competitive results with significantly less computational expensive would proof the concept. Moreover, success in the vision-language domain suggests that this approach can be applied to other modality combinations, further broadening its impact.

This strategy would make multimodal learning more accessible to smaller research groups and individual researchers, increasing the diversity of research in the field across various modalities.

Based on this, our contributions are as follows:
We develop an end-to-end self-supervised learning approach for generating multimodal models from unimodal components, which is significantly **cheaper** to train and **smaller** in size compared to existing multimodal (vision-language) models. We show that while our approach does not reach the performance of state-of-the-art multimodal models, it is **competitive** with other vision-language models in some benchmarks.

As part of our research, we also find a promising new approach to generate smaller unimodal models that are to some extent competitive with their larger counterpart.