This thesis investigates the development of vision–language models, from image captioning to modern Multimodal Large Language Models.

It presents methods to improve the alignment between visual understanding and language generation, examines how external knowledge can be incorporated into multimodal models, and proposes approaches for more structured and reliable reasoning.

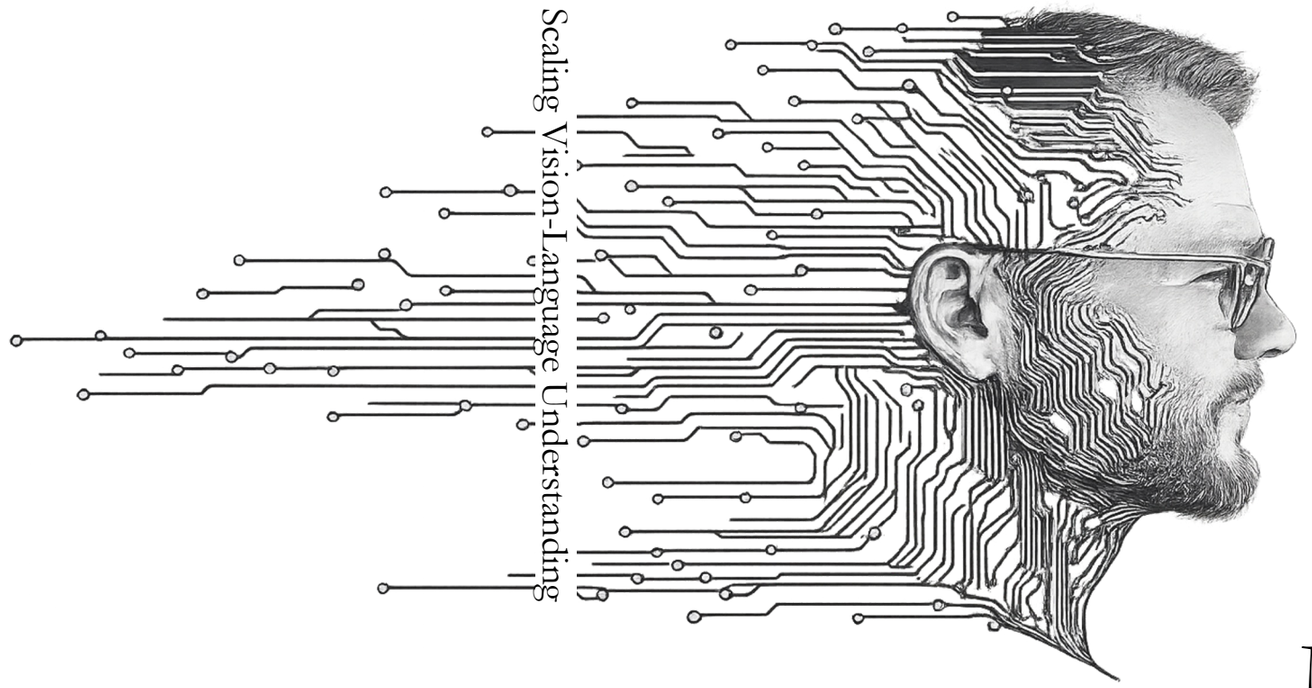
The results contribute to models that not only describe visual content, but also interpret and reason about it in a more meaningful way.

NICHOLAS MORATELLI

Scaling Vision-Language Understanding

Scaling Vision-Language Understanding

From Image Captioning to Knowledge-Grounded Multimodal Large Language Models



NICHOLAS MORATELLI