Muhammad Maaz

+971 (52) 532-6156
Abu Dhabi, UAE
mmaaz60@gmail.com

Google Scholar | Homepage | LinkedIN | GitHub

Citations: 7000+

Research Interests

Multimodal Large Language Models (MLLMs) for in-the-wild image and video understanding, multimodal reasoning, grounding, and long-video understanding.

Education

2023 – 2026 **Ph.D. Computer Vision** Mohamed bin Zayed University of Artificial Intelligence.
Thesis Title: Towards Video Understanding and Reasoning via Multimodal Models.
Advisor: Prof. Salman Khan

2021 – 2023 **M.Sc. Computer Vision** Mohamed bin Zayed University of Artificial Intelligence.
Thesis Title: Class-agnostic Object Detection with Multi-modal Transformer.
Advisor: Prof. Salman Khan

2014 – 2018 **B.Sc. Electrical Engineering** University of Engineering and Technology.
Thesis Title: Intelligent Discipline Maintenance System with Client-Server Topology.
Advisor: Prof. Kashif Javed

Experience

Jun. 2026 – Continue **Research Scientist** @ Meta – Abu Dhabi, UAE (Remote)
I am working with SAM team under Fundamental AI Research (FAIR) pillar.

Mar. 2025 – Aug. 2025 **Research Scientist Intern (Internship # 2)** @ Meta – San Francisco, USA
Conducted research on incorporating promptable image segmentation into Multimodal Large Language Models (MLLMs) to improve fine-grained visual reasoning.

Jun. 2024 – Dec. 2025 **Research Scientist Intern (Internship # 1)** @ Meta – San Francisco, USA
Developed Perception Language Model (PLM), a family of open and fully reproducible multimodal large language models (MLLMs). PLM matches or exceeds the performance of recent state-of-the-art MLLMs such as InternVL3 and Qwen2.5VL, using fully open data.

Jul. 2020 – Dec. 2020 **Computer Vision Engineer**, Hazen.ai – Lahore, PAK
Developed a traffic light phase detection solution for road safety applications. Trained a neural network using triplet loss to learn robust embeddings for different traffic light phases.

Jun. 2018 – Jul. 2020 **Software Engineer**, Confiz Limited – Lahore, PAK
Developed a person detection and tracking system to identify engaged and ignored areas in retail stores, and a face recognition system to generate visitor demographics and visit frequencies.

Selected Research

arXiv 2026 **Video-R2: Reinforcing Consistent and Grounded Reasoning in Multimodal Language Models**
Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Khan
| Paper | Code | Hugging Face |

arXiv 2026 **Video-CoM: Interactive Video Reasoning via Chain of Manipulations**
Hanoona Rasheed, Mohammed Zumri, **Muhammad Maaz**, Ming-Hsuan Yang, Fahad Khan, Salman Khan
| Paper | Code | Hugging Face |

Selected Research (continued)

NeurIPS 2025
Spotlight
2300+ ☆
70+ Citations

PerceptionLM: Open-Access Data and Models for Detailed Visual Understanding
Muhammad Maaz*, Jang Hyun Cho*, Andrea Madotto*, Effrosyni Mavroudi*, Triantafyllos Afouras*, Tushar Nagarajan*, Yale Song*, Tengyu Ma*, Shuming Hu*, Suyog Jain, Miguel Martin, Huiyu Wang, Hanoona Rasheed, Peize Sun, Po-Yao Huang, Daniel Bolya, Nikhila Ravi, Shashank Jain, Tammy Stark, Shane Moon, Babak Damavandi, Vivian Lee, Andrew Westbury, Salman Khan, Philipp Krähenbühl, Piotr Dollár, Lorenzo Torresani, Kristen Grauman, Christoph Feichtenhofer
| Paper | Code | Hugging Face |

ACL 2024
1500+ ☆
1700+ Citations

Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models
Muhammad Maaz*, Hanoona Rasheed*, Salman Khan, Fahad Khan
| Paper | Code | Hugging Face | Benchmark |

arXiv Preprint
290+ ☆
110+ Citations

VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding
Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Khan
| Paper | Code | Hugging Face |

CVPR 2024
950+ ☆
600+ Citations

GLaMM: Pixel Grounding Large Multimodal Model
Hanoona Rasheed*, **Muhammad Maaz***, Sahal Shaji Mullappilly, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Erix Xing, Ming-Hsuan Yang, Fahad Khan
| Paper | Code | Hugging Face | Dataset | Website |

ICLR 2026

VideoMathQA: Benchmarking Mathematical Reasoning via Multimodal Understanding in Videos
Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, **Muhammad Maaz**, Ming-Hsuan Yang, Salman Khan, Fahad Khan
| Paper | Website | Hugging Face |

arXiv Preprint

Mobile-VideoGPT: Fast & Accurate Video Understanding Language Model
Abdelrahman Shaker, **Muhammad Maaz**, Chenhui Gou, Hamid Rezatofighi, Salman Khan, Fahad Khan
| Paper | Code | Hugging Face |

WACV 2025
Oral

PALO: A Polyglot Large Multimodal Model for 5B People
Hanoona Rasheed, **Muhammad Maaz**, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, Fahad Khan
| Paper | Code | Dataset |

CVPR 2023
300+ ☆
360+ Citations

Fine-tuned CLIP Models are Efficient Video Learners
Hanoona Rasheed, Muhammad Uzair Khattak, **Muhammad Maaz**, Salman Khan, Fahad Khan
| Paper | Code | Website |

CVPR 2023
820+ ☆
1800+ Citations

MaPLe: Multi-modal Prompt Learning
Muhammad Uzair Khattak, Hanoona Rasheed, **Muhammad Maaz**, Salman Khan, Fahad Khan
| Paper | Code | Website | Patent (**US12493741B2**) |

ECCV 2022
310+ ☆
175+ Citations

Class-agnostic Object Detection with Multi-modal Transformer
Muhammad Maaz*, Hanoona Rasheed*, Salman Khan, Fahad Khan, Rao M. Anwer, Ming-Hsuan Yang
| Paper | Code | Presentation |

NeuRIPS 2022
290+ ☆
240+ Citations

Bridging the Gap between Object and Image-level Representations for Open-Vocabulary Detection
Hanoona Rasheed*, **Muhammad Maaz***, Muhammad Uzair Khattak, Salman Khan, Fahad Khan
| Paper | Code | Website | Patent (**US12288372B2**) |

ECCVW 2022
Oral, 415+ ☆
540+ Citations

EdgeNeXt: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications
Muhammad Maaz*, Abdelrahman Shaker*, Hisham Cholakkal, Salman Khan, Syed Waqas Zamir, Rao M. Anwer, Fahad Khan
| Paper | Code | Website | Patent (**US12373672B2**) |

Selected Research (continued)

ICCV 2023
310+ ☆
390+ Citations

SwiftFormer: Efficient Additive Attention for Transformer-based Real-time Mobile Vision Applications
Abdelrahman Shaker, **Muhammad Maaz**, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, Fahad Khan
| Paper | Code |

IEEE TMI 2024
520+ ☆
630+ Citations

UNETR++: Efficient and Accurate 3D Medical Image Segmentation
Abdelrahman Shaker, **Muhammad Maaz**, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, Fahad Khan
| Paper | Code | Website | Patent (**US12373956B2**) |

Awards

AI Rising Star 2026	AI Rising Star Award at KAUST AI Rising Star Symposium 2026. Recognized for emerging research impact and contributions to the field of AI.
Graduate Research Excellence Award 2026	Recipient of the Graduate Research Excellence Award at MBZUAI's 5th Anniversary celebration, awarded as the only student honoree.
Google PhD Fellowship Award 2025	First in the MENA region and the first student in the UAE to receive the Google PhD Fellowship (2025 - 2026), selected as one of 255 fellows worldwide from 35 countries, with the fellowship awarded for research in Machine Perception.

Invited Talks

KAUST AI Rising Star Symposium 2026	Invited for a Research Talk at the 5th edition of AI Rising Star Symposium at King Abdullah University of Science and Technology (KAUST).
IndabaX Sudan 2025	Invited to deliver a research talk at IndabaX Sudan 2025, the annual Machine Learning and AI conference and one of the largest gatherings in the African AI community.

Community Contributions

2022 – Cont. Serving as Reviewer
Completed 40+ peer reviews across venues including Transformer Models in Vision, Special Issue in IEEE TPAMI, ECCVW 2022, CVPR 2023, ICCV 2023, ACCV 2024, ECCV 2024, ICCV 2025, NeurIPS 2025, ICLR 2026, CVPR 2026, ECCV 2026 and NeurIPS 2026.