David Wan

Email: davidwan@cs.unc.edu
Website: meetdavidwan.github.io

EDUCATION

University of North Carolina, Chapel Hill | *Ph.D. in Computer Science* **2021 - Present**
- Advisor: Mohit Bansal
- Google PhD Fellowship in NLP for 2024 and 2025

Columbia University, New York | *M.S. in Computer Science* **2020 - 2021**
- Advisor: Kathleen McKeown
- Thesis: Methods for Cross-Language Search and Summarization for Low-Resource Languages

Columbia University, New York | *B.A. in Computer Science* **2016 - 2020**
- Concentration in Linguistics

RESEARCH & INDUSTRY EXPERIENCE

Google Research **New York, NY**
Research Intern | *Hosts:* Sebastien Baur and Gaurav Singh Tomar *May 2025 - Aug 2025*

Salesforce AI Research **Palo Alto, CA**
Research Intern | *Hosts:* Shafiq Joty and Jesse Vig *May 2024 - Oct 2024*

FAIR Labs at Meta **Seattle, WA**
Research Intern | *Hosts:* Ramakanth Pasunuru and Asli Celikyilmaz *May 2023 - Dec 2023*

Alexa AI at Amazon **Seattle, WA**
Research Intern | *Hosts:* Mengwen Liu and Markus Dreyer *May 2022 - Oct 2022*

Columbia University **New York, NY**
Research Assistant | *Advisor:* Michael Collins *Nov 2018 - May 2019*

RESEARCH INTEREST

Interests Natural Language Generation (Summarization, Machine Translation), Factuality, Multimodal

RESEARCH PUBLICATIONS

CLaMR: Contextualized Late-Interaction for Multimodal Content Retrieval
David Wan, Han Wang, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal.
Preprint on ArXiv. [Paper] [Code]

GenerationPrograms: Fine-grained Attribution with Executable Programs
David Wan, Eran Hirsch, Elias Stengel-Eskin, Ido Dagan, and Mohit Bansal.
COLM 2025. [Paper] [Code]

QAPyramid: Fine-grained Evaluation of Content Selection for Text Summarization
Shiyue Zhang, **David Wan**, Arie Cattan, Ayal Klein, Ido Dagan, and Mohit Bansal.
COLM 2025. [Paper] [Code]

MAMM-Refine: A Recipe for Improving Faithfulness in Generation with Multi-Agent Collaboration
David Wan, Justin Chen, Elias Stengel-Eskin, and Mohit Bansal.
NAACL 2025. [Paper] [Code]

On Positional Bias of Faithfulness for Long-form Summarization
David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty.
NAACL 2025. [Paper] [Code]

Localizing Factual Inconsistencies in Attributable Text Generation
Arie Cattan, Paul Roit, Shiyue Zhang, **David Wan**, Roee Aharoni, Idan Szpektor, Mohit Bansal, and Ido Dagan.
Preprint on ArXiv. [Paper] [Code]

Contrastive Region Guidance: Improving Grounding in Vision-Language Models Without Training
David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal.
ECCV 2024. [Paper] [Code]

ACUEval: Fine-grained Hallucination Evaluation and Correction for Abstractive Summarization
David Wan, Koustuv Sinha, Srini Iyer, Asli Celikyilmaz, Mohit Bansal, and Ramakanth Pasunuru.
ACL 2024. [Paper] [Code]

HistAlign: Improving Context Dependency in Language Generation by Aligning with History
David Wan, Shiyue Zhang, and Mohit Bansal.
EMNLP 2023. [Paper] [Code]

Extractive is not Faithful: An Investigation of Broad Unfaithfulness Problems in Extractive Summarization
Shiyue Zhang, **David Wan**, and Mohit Bansal.
ACL 2023. [Paper] [Code]

Faithfulness-Aware Decoding Strategies for Abstractive Summarization
David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal.
EACL 2023. [Paper] [Code]

Evaluating and Improving Factuality in Multimodal Abstractive Summarization
David Wan and Mohit Bansal.
EMNLP 2022. [Paper] [Code]

Constrained Regeneration for Cross-Lingual Query-Focused Extractive Summarization
Elsbeth Turcan, **David Wan**, Faisal Ladhak, Petra Galuscakova, Sukanta Sen, Svetlana Tchistiakova, Weijia Xu, Marine Carpuat, Kenneth Heafield, Douglas Oard, and Kathleen McKeown.
ACL 2022. [Paper]

FactPEGASUS: Factuality-Aware Pre-training and Fine-tuning for Abstractive Summarization
David Wan and Mohit Bansal.
NAACL 2022. [Paper] [Code]

Segmenting Subtitles for Correcting ASR Segmentation Errors
David Wan, Chris Kedzie, Faisal Ladhak, Elsbeth Turcan, Petra Galuscakova, Elena Zotkina, Zhengping Jiang, Peter Bell, and Kathleen McKeown.
EACL 2021. [Paper]

Incorporating Terminology Constraints in Automatic Post-Editing
David Wan, Chris Kedzie, Faisal Ladhak, Marine Carpuat, and Kathleen McKeown.
WMT 2020. [Paper] [Code]

Subtitles to Segmentation: Improving Low-Resource Speech-to-Text Translation Pipelines
David Wan, Zhengping Jiang, Chris Kedzie, Elsbeth Turcan, Peter Bell, and Kathy McKeown.
CLSSTS 2020. [Paper]

ACHIEVEMENTS AND AWARDS

Google PhD Fellowship in Natural Language Processing, Google. 2024

One of twelve students globally to receive full funding for two years

Theodore R. Bashkow Award, Dept. of Computer Science, Columbia University. 2020

One of three undergrads awarded for excelling in independent projects

PROFESSIONAL SERVICES

Reviewer

- EMNLP NewSumm Workshop 2025
- NeurIPS 2025
- ACL Rolling Review, May 2025
- ACL Rolling Review, December 2024
- ACL Rolling Review, October 2024
- ACL Rolling Review, June 2024
- ACL Rolling Review, April 2024
- ACL Rolling Review, February 2024
- ACL Rolling Review, December 2023
- EMNLP 2023
- ACL 2023
- EMNLP 2022

INVITED TALKS

Research Trend AI **Summer 2025**

- CLaMR: Contextualized Late-Interaction for Multimodal Content Retrieval

TEACHING EXPERIENCE

Columbia University **New York, NY**
Teaching Assistant for Natural Language Processing *Jan - May 2019, Jan - May 2020*
Columbia University **New York, NY**
Tutor for Natural Language Processing *Sep 2019 - Dec 2019*