Yonatan Bitton, Curriculum Vitae, December 2024

Contact Information

yonatanbitton.github.io yonatanbitton1@gmail.com linkedin.com/in/yonatanbitton

Current Positions

Senior Research Scientist, Google Research April 2024-Present
Advancing multimodal consistency. Developing feedback models for text-to-image and text-to-video applications and enhance multimodal factuality to ensure the accuracy of text generated from visual sources.

Research Scientist, Google Research June 2023-April 2024
Vision-and-language. Recent works include image-text alignment, improving text-to-image models, and visual instruction tuning.

Education

PhD in Computer Science, The Hebrew University of Jerusalem 2020–2023
Advisors: Prof. Gabriel Stanovsky and Prof. Roy Schwartz
Thesis: Bridging Vision and Language with Data.

MSc in Computer Science, *magna cum laude*, Ben Gurion University 2019–2020
Advisors: Prof. Michael Elhadad and Prof. Eitan Bachmat
Thesis: Cross-lingual entity linking and visual question answering. GPA 97

BSc in Computer Science, Ben Gurion University, 2015-2019 2015–2019

Work Experience[†]

Research Intern, Google 2022-2023
Cerebra team: focusing on conversational AI, engaged with leading language models (LaMDA, PaLM, BARD); leveraged synthetic data for query generation, crafted personalized agents, and augmented LLM memory capabilities.

Applied Scientist, Amazon Lab126 2019-2022
Visual Fitness Halo Team - Developed a virtual fitness trainer, specializing in 2D/3D pose estimation, action recognition, error correction, on-device deployment and more.

Researcher, IBM Research 2017-2019
Developing machine-learning methods to detect frauds

Peer-Reviewed Publications

[*] indicates equal contribution. For abstracts and more information, see **Google Scholar**.

[1] **RefVNLI: Towards Scalable Evaluation of Subject-driven Text-to-image Generation**
Slobodkin. A, Taitelbaum. H, **Bitton. Y**, Gordon. B, Sokolik. M, Bitton-Guetta. N, Gueta. A, Rassin. R, Laish. I, Lischinski. D, Szpektor. I
April 2025 *arXiv preprint*

[2] **VideoPhy2: Challenging Action-Centric Physical Commonsense Evaluation of Video Generation**
Bansal. H[*], Peng. C[*], **Bitton. Y**[*], Goldenberg. R, Grover. A, Chang. K-W
March 2025 *arXiv preprint*

[3] **PaliGemma 2: A Family of Versatile VLMs for Transfer**
Steiner. A, Pinto. A. S, Tschannen. M, Keysers. D, Wang. X, **Bitton. Y**, Gritsenko. A, Minderer. M, Sherbondy. A, Long. S, Qin. S, Ingle. R, Bugliarello. E, Kazemzadeh. S, Mesnard.

[†] Parallel to studies.

T, Alabdulmohsin. I, Beyer. L, Zhai. X
December 2024 *arXiv preprint: 2412.03555*

[4] **Bridging the Visual Gap: Fine-Tuning Multimodal Models with Knowledge-Adapted Captions**
Yanuka. M, Ben Kish. A, **Bitton. Y**, Szpektor. I, Giryes. R
November 2024 *arXiv preprint: 2411.09018*

[5] **KITTEN: A Knowledge-Intensive Evaluation of Image Generation on Visual Entities**
Huang. H-P, Wang. X, **Bitton. Y**, Taitelbaum. H, Tomar. G. S, Chang. M-W, Jia. X, Chan. K. C. K, Hu. H, Su. Y-C, Yang. M-H
October 2024 *arXiv preprint: 2410.11824*

[6] **Visual Riddles: A Commonsense and World Knowledge Challenge for Large Vision and Language Models**
Bitton-Guetta. N, Slobodkin. A, Maimon. A, Habba. E, Rassin. R, **Bitton. Y**, Szpektor. I, Globerson. A, Elovici. Y
July 2024 *NeurIPS 2024, Datasets and Benchmarks Track*

[7] **DataComp-LM: In search of the next generation of training sets for language models**
Li. J, Fang. A, Smyrnis. G, Ivgi. M, Jordan. M, Gadre. S, Bansal. H, Guha. E, Keh. S, Arora. K, Garg. S, Xin. R, Muennighoff. N, Heckel. R, Mercat. J, Chen. M, Gururangan. S, Wortsman. M, Albalak. A, **Bitton. Y**, Nezhurina. M, Abbas. A, Hsieh. C, Ghosh. D, Gardner. J, Kilian. M, Zhang. H, Shao. R, Pratt. S, Sanyal. S, Ilharco. G, Daras. G, Marathe. K, Gokaslan. A, Zhang. J, Chandu. K, Nguyen. T, Vasiljevic. I, Kakade. S, Song. S, Sanghavi. S, Faghri. F, Oh. S, Zettlemoyer. L, Lo. K, El-Nouby. A, Pouransari. H, Toshev. A, Wang. S, Groeneveld. D, Soldaini. L, Koh. P, Jitsev. J, Kollar. T, Dimakis. A, Carmon. Y, Dave. A, Schmidt. L, Shankar. V
June 2024 *Neural Information Processing Systems (NeurIPS 2024)*

[8] **Contrastive Sequential-Diffusion Learning: An approach to Multi-Scene Instructional Video Synthesis**
Ramos. V, **Bitton. Y**, Yarom. M, Szpektor. I, Magalhaes. J
July 2024 *IEEE Winter Conference on Applications of Computer Vision (WACV 2025)*

[9] **Beyond Thumbs Up/Down: Untangling Challenges of Fine-Grained Feedback for Text-to-Image Generation**
Collins. K. M, Kim. N, **Bitton. Y**, Rieser. V, Omidshafiei. S, Hu. Y, Chen. S, Dutta. S, Chang. M, Lee. K, Liang. Y, Evans. G, Singla. S, Li. G, Weller. A, He. J, Ramachandran. D, Dvijotham. K. D
June 2024 arXiv preprint:2406.16807

[10] **Video-STaR: Self-Training Enables Video Instruction Tuning with Any Supervision**
Zohar. O, Wang. X, **Bitton. Y**, Szpektor. I, Yeung-Levy. S
arXiv preprint arXiv:2407.06189

[11] **VideoPhy: Evaluating Physical Commonsense for Video Generation**
Bansal. H, Lin. Z, Xie. T, Zong. Z, Yarom. M, **Bitton. Y**, Jiang. C, Sun. Y, Chang. K-W, Grover. A
arXiv preprint arXiv:2406.03520

[12] **TALC: Time-Aligned Captions for Multi-Scene Text-to-Video Generation**
Bansal. H, **Bitton. Y**, Yarom. M, Szpektor. I, Grover. A, Chang. K-W
arXiv preprint arXiv:2405.04682

[13] **ImageInWords: Unlocking Hyper-Detailed Image Descriptions**
Garg. R, Burns. A, Ayan. B, **Bitton. Y**, Montgomery. C, Onoe. Y, Bunner. A, Krishna. R, Baldridge. J, Soricut. R
arXiv preprint arXiv:2405.02793

[14] **DOCCI: Descriptions of Connected and Contrasting Images**
Onoe. Y, Rane. S, Berger. Z, **Bitton. Y**, Cho. J, Garg. R, Ku. A, Parekh. Z, Pont-Tuset. J, Tanzer. G, Wang. Su, Baldridge. J
The European Conference on Computer Vision **(ECCV 2024)**

[15] **A Chain-of-Thought Is as Strong as Its Weakest Link: A Benchmark for Verifiers of Reasoning Chains**
Jacovi. A, **Bitton. Y**, Bohnet. B, Herzig. J, Honovich. O, Tseng. M, Collins. M, Aharoni. R, Geva. M
Annual Meeting of the Association of Computational Linguistics (**ACL 2024**)

[16] **ParallelPARC: A Scalable Pipeline for Generating Natural-Language Analogies**
Sultan. O*, **Bitton. Y***, Yosef. R, Shahaf. D
North American Chapter of the Association of Computational Linguistics (**NAACL 2024**)

[17] **Generating Coherent Sequences of Visual Illustrations for Real-World Manual Tasks**
Bordalo. J, Ramos. V, Valério. R, Glória-Silva. D, **Bitton. Y**, Yarom. M, Szpektor. I, Magalhaes. J
Annual Meeting of the Association of Computational Linguistics (**ACL 2024**)

[18] **Mismatch Quest: Visual and Textual Feedback for Image-Text Misalignment**
Gordon. G*, **Bitton. Y***, Shafir. Y, Garg. R, Chen. X, Lischinski. D, Cohen-Or D, Szpektor. I
arXiv preprint The European Conference on Computer Vision **(ECCV 2024)**

[19] **VideoCon: Robust Video-Language Alignment via Contrast Captions**
Bansal. H, **Bitton. Y**, Szpektor. I, Kai-Wei. C, Grover. A
Conference on Computer Vision and Pattern Recognition (**CVPR 2024**)

[20] **VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by Real-World Use**
Bitton. Y*, Bansal. H*, Hessel. J*, Shao. R, Zhu. W, Awadalla. A, Gardner. J, Taori. R, Schimdt. L
Neural Information Processing Systems Datasets and Benchmarks Track (**NeurIPS 2023**)

[21] **VisIT-Bench: A Benchmark for Vision-Language Instruction Following Inspired by Real-World Use**
Bitton. Y*, Bansal. H*, Hessel. J*, Shao. R, Zhu. W, Awadalla. A, Gardner. J, Taori. R, Schimdt. L
Neural Information Processing Systems Datasets and Benchmarks Track (**NeurIPS 2023**)

[22] **Read, Look or Listen? What's Needed for Solving a Multimodal Dataset**
Madvil. N, **Bitton. Y**, Schwartz. R
arXiv preprint

[23] **Transferring Visual Attributes from Natural Language to Verified Image Generation**
Valerio. R, Bordalo. J, Yarom. M, **Bitton. Y**, Szpektor. I, Magalhaes. J
arXiv preprint

[24] **What You See is What You Read? Improving Text-Image Alignment Evaluation**
Bitton. Y*, Yarom. M*, Changpinyo. S, Aharoni. R, Herzig. J, Lang. O, Ofek. E, Szpektor. I
Neural Information Processing Systems (**NeurIPS 2023**)

[25] **q2d: Turning Question into Dialogs to Teach Models How to Search**
Bitton. Y, Cohen. S, Hakimi. I, Lewenberg. Y, Aharoni. R, Weinreb. E,
Conference on Empirical Methods in Natural Language Processing: **EMNLP 2023**

[26] **DataComp: In search of the next generation of multimodal datasets via data scaling**
Yitzhak. S, Ilharco. G, Fang. A, Hayase. J, Smyrnis. G, Nguyen. T, Marten. R, Wortsman. M, Ghosh. D, Zhang. J, Orgad. E, Entezari. R, Daras. G, Pratt. S, Ramanujan. V, **Bitton. Y**, Mussmann. S, Vencu. R, Cherti. M, Krishna. R, Wei. P, Saukh. O, Ratner. A, Song. S, Hajishirzi. H, Farhadi. A, Beaumont. R, Oh. S, Dimakis. A, Jitsev. J, Carmon. Y, Shankar. V, Schmidt. L
Neural Information Processing Systems Datasets and Benchmarks Track (**NeurIPS 2023**)

[27] **OpenFlamingo: An open-source framework for training vision-language models with in-context learning**
Awadalla. A, Gao. I, Gardner. J, Hessel. J, Hafany. Y, Zhu. W, Gedre. S, **Bitton. Y**, Kalyani. M, Kornblith. S, Koh. P, Ilharco. G, Wortsman. M, Schmidt. L
Blog release: https://laion.ai/blog/open-flamingo/

[28] **IRFL: Image Recognition of Figurative Language**
Yosef. R, **Bitton. Y**, Shahaf. D
Findings of the Conference on Empirical Methods in Natural Language Processing: **EMNLP 2023**

[29] **WHOOPS! A Vision-and-Language Commonsense Benchmark of Heterogeneous Objects and Situations**
Guetta. N*, **Bitton. Y***, Hessel. J, Schmidt. L, Elovici. Y, Stanovsky. G, Schwartz. R,
International Conference on Computer Vision (**ICCV 2023**)
Neural Information Processing Systems Creative AI Track (**NeurIPS 2023**) - Gallery

[30] **VASR: Visual Analogies of Situation Recognition**
Bitton. Y, Yosef. R, Strugo. E, Shahaf D, Schwartz. R, Stanovsky. G
Association for the Advancement of Artificial Intelligence (**AAAI 2023**)
Selected as an **Oral Presentation**

[31] **WinoGAViL: Gamified Association Benchmark to Challenge Vision-and-Language Models**
Bitton. Y*, Guetta. N*, Yosef. R, Bansal. M, Stanovsky. G, Schwartz. R,
Neural Information Processing Systems Datasets and Benchmarks Track (**NeurIPS 2022**)
Selected as a **Featured Presentation** (Updated version of "Oral Presentation")

[32] **Data Efficient Masked Language Modeling For Vision and Language**
Bitton. Y, Stanovsky. G, Elhadad. M, Schwartz. R,
Findings of the Conference on Empirical Methods in Natural Language Processing: **EMNLP 2021**

[33] **Automatic Generation of Contrast Sets from Scene Graphs: Probing the Compositional Consistency of GQA**
Bitton. Y, Stanovsky. G, Schwartz. R, Elhadad. M,
North American Chapter of the Association of Computational Linguistics (**NAACL 2021**)

[34] **Cross-lingual Unified Medical Language System entity linking in online health communities**
Bitton. Y, Cohen. R, Schifter. T, Bachmat. E, Elhadad. M, Elhadad. N
Journal of the American Medical Informatics Association (**JAMIA 2020**)

Selected Awards and Scholarships

PhD Awards

KLA Scholarship for Outstanding Graduate Students 2022

MSc Awards

Dean's Award for Excellence 2020
Graduated with honors (*magna cum laude*) 2020
Computer Science Department Research Excellence Award for journal publication 2020

Professional Activities

Organizer

The 3rd Workshop on Computer Vision in the Wild @ CVPR 2024

Area Chair

ACL ARR May 2025
ACL ARR February 2025
WACV 2025
The 3rd Workshop on Computer Vision in the Wild @ CVPR 2024

Conference Reviewer

ICCV	2025
ICLR	2025
NAACL	2025
NeurIPS Creative AI	2024
NeurIPS Datasets and Benchmarks	2024
EMNLP Industry Track	2023
ACL	2023
NAACL	2022
NeurIPS Datasets and Benchmarks	2022
ACL	2021
EMNLP	2021

Invited Talks

Bridging Vision and Language with Data: From Perception to Understanding — April-June 2023
Hebrew University of Jerusalem, NLP-IL Reading Group, Microsoft Israel (MSAI-HIVE team), Meta AI Research Tel-Aviv, Technion, Ben Gurion University, Google Tel-Aviv, Bar-Ilan University, IBM Research (Israel NLP team), Tel Aviv University
Talk record is available in **YouTube**

Commonsense Benchmarks for Vision and Language — November 2022
NLP Seminar at Cornell Tech, Google Research Israel, the Hebrew University of Jerusalem

q2d: Turning Questions into Dialogs to Teach Models How to Search — September 2022
Conversational applications with LLMs - Summit in Google Zurich

WinoGAViL: Gamified Association Benchmark to Challenge Vision-and-Language Models — June 2022
IBM Research Israel

VASR: Visual Analogies of Situation Recognition — May 2022
Computer Vision Seminar at the Hebrew University of Jerusalem

Open Source

Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images
Project website: https://whoops-benchmark.github.io/
Huggingface dataset: https://huggingface.co/datasets/nlphuji/whoops

WinoGAViL: Gamified Association Benchmark To Challenge Vision-And-Language Models
Project website: https://winogavil.github.io/
Software: https://github.com/WinoGAViL/WinoGAViL-experiments

VASR: Visual Analogies of Situation Recognition
Project website: https://vasr-dataset.github.io/
Software: https://github.com/vasr-dataset/vasr

Data Efficient Masked Language Modeling for Vision and Language
Software: https://github.com/yonatanbitton/data_efficient_masked_language_modeling_for_vision_and_language

Automatic Generation of Contrast Sets from Scene Graphs
Software: https://github.com/yonatanbitton/automatic_generation_of_contrast_sets_from_scene_graphs

Cross-lingual unified medical language system entity linking in online health communities
Software: https://github.com/yonatanbitton/mdtel