

Contact Information	<a href="https://github.com/yonatanbitton">yonatanbitton.github.io</a>	<a href="mailto:yonatanbitton1@gmail.com">yonatanbitton1@gmail.com</a>	<a href="https://linkedin.com/in/yonatanbitton">linkedin.com/in/yonatanbitton</a>
Current Positions	<div><div>Senior Research Scientist, Google Research</div><div>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.</div></div> <div><div>Research Scientist, Google Research</div><div>Vision-and-language. Recent works include <a href="#">image-text alignment</a>, improving <a href="#">text-to-image models</a>, and visual instruction tuning.</div></div>		
Education	<div><div>PhD in Computer Science, The Hebrew University of Jerusalem</div><div><i>Advisors: Prof. Gabriel Stanovsky and Prof. Roy Schwartz</i></div><div>Thesis: Bridging Vision and Language with Data.</div></div> <div><div>MSc in Computer Science, <i>magna cum laude</i>, Ben Gurion University</div><div><i>Advisors: Prof. Michael Elhadad and Prof. Eitan Bachmat</i></div><div>Thesis: Cross-lingual entity linking and visual question answering. GPA 97</div></div> <div><div>BSc in Computer Science, Ben Gurion University, 2015-2019</div></div>		
Work Experience <sup>†</sup>	<div><div>Research Intern, Google</div><div>Cerebra team: focusing on conversational AI, engaged with leading language models (LaMDA, PaLM, BARD); leveraged synthetic data for <a href="#">query generation</a>, crafted personalized agents, and augmented LLM memory capabilities.</div></div> <div><div>Applied Scientist, Amazon Lab126</div><div>Visual Fitness Halo Team - Developed a virtual fitness trainer, specializing in 2D/3D pose estimation, action recognition, error correction, on-device deployment and more.</div></div> <div><div>Researcher, IBM Research</div><div>Developing machine-learning methods to detect frauds</div></div>		
Peer-Reviewed Publications	<p>* indicates equal contribution. For abstracts and more information, see <a href="#">Google Scholar</a>.</p> <div><div>[1] <a href="#">Video-STaR: Self-Training Enables Video Instruction Tuning with Any Supervision</a></div><div>Zohar. O, Wang. X, <b>Bitton. Y</b>, Szpektor. I, Yeung-Levy. S</div><div>arXiv preprint arXiv:2407.06189</div></div> <div><div>[2] <a href="#">VideoPhy: Evaluating Physical Commonsense for Video Generation</a></div><div>Bansal. H, Lin. Z, Xie. T, Zong. Z, Yarom. M, <b>Bitton. Y</b>, Jiang. C, Sun. Y, Chang. K-W, Grover. A</div><div>arXiv preprint arXiv:2406.03520</div></div> <div><div>[3] <a href="#">TALC: Time-Aligned Captions for Multi-Scene Text-to-Video Generation</a></div><div>Bansal. H, <b>Bitton. Y</b>, Yarom. M, Szpektor. I, Grover. A, Chang. K-W</div><div>arXiv preprint arXiv:2405.04682</div></div> <div><div>[4] <a href="#">ImageInWords: Unlocking Hyper-Detailed Image Descriptions</a></div><div>Garg. R, Burns. A, Ayan. B, <b>Bitton. Y</b>, Montgomery. C, Onoe. Y, Bunner. A, Krishna. R,</div></div>		

---

<sup>†</sup> Parallel to studies.

Baldrige. J, Soricut. R  
arXiv preprint arXiv:2405.02793

- [5] **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, Baldrige. J  
The European Conference on Computer Vision (**ECCV 2024**)
- [6] **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**)
- [7] **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**)
- [8] **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**)
- [9] **Mismatch Quest: Visual and Textual Feedback for Image-Text Misalignment**  
Gordon. G\*, **Bitton. Y\***, Shafrir. Y, Garg. R, Chen. X, Lischinski. D, Cohen-Or D, Szpektor. I  
arXiv preprint The European Conference on Computer Vision (**ECCV 2024**)
- [10] **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**)
- [11] **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**)
- [12] **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**)
- [13] **Read, Look or Listen? What's Needed for Solving a Multimodal Dataset**  
Madvil. N, **Bitton. Y**, Schwartz. R  
arXiv preprint
- [14] **Transferring Visual Attributes from Natural Language to Verified Image Generation**  
Valerio. R, Bordalo. J, Yarom. M, **Bitton. Y**, Szpektor. I, Magalhaes. J  
arXiv preprint
- [15] **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**)
- [16] **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**
- [17] **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**)

- [18] **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/>
- [19] **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**
- [20] **WHOOOPS! 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
- [21] **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**
- [22] **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”)
- [23] **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**
- [24] **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**)
- [25] **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 ( <i>magna cum laude</i> )	2020
	Computer Science Department Research Excellence Award for journal publication	2020

## Advising

STUDENTS MENTORED AT THE HEBREW UNIVERSITY	Oren Sultan	2023
	Nitzan Bitton-Guetta	2022-2023
	Netta Madvil	2022-2023
	Ron Yosef	2022-2023

## Professional Activities

### CONFERENCE REVIEWER

Conference on Empirical Methods in Natural Language Processing (EMNLP), Industry Track 2023  
Annual Meeting of the Association of Computational Linguistics (ACL) 2023  
North American Chapter of the Association of Computational Linguistics (NAACL) 2022  
NeurIPS Datasets and Benchmarks 2022  
Annual Meeting of the Association of Computational Linguistics (ACL) 2021  
Conference on Empirical Methods in Natural Language Processing (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](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](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>