Shilong Liu Postdoctoral Research Fellow at Princetion AI Lab

sl8264@princeton.edu slongliu86@gmail.com

Homepage Google Scholar Linkedin

Highlights

- **Recent Research Topics:** LLM Agents, Multimodal Learning, and Physical Intelligence.
- Over **13,000** Google Scholar citations, with the majority as (co-)first or leading author.
- Over **30,000** GitHub stars, with the majority as (co-)first or leading author.
- **4** papers recognized among the "Top 15 Most Influential Papers" of their respective conferences by Paper Digest.
- Open-sourced models receive over **2 million downloads per month**. Grounding DINO is the most downloaded zero-shot object detection model on Hugging Face.
- (Co-)Supervised more than 5 students to publish paper on top-tier conference.

Employment History

Oct 2025 – Present **Princeton AI Lab, Princeton University.** Postdoc Research Fellow. *Princeton, NJ, USA*

Jun 2025 – Oct 2025 **Bytedance Seed.** Research Scientist. *Beijing, China*

Jul 2024 – Jan 2025 **NVIDIA Research.** Research internship. *Santa Clara, CA, USA*
Worked on vision-language models, contributed to Eagle2 [24].

Jan 2024 – Fri 2024 **Shengshu-Tech** Research internship. *Beijing, China*
Worked on Vidu [20], a top-tier video generation model.

May 2023 – Sep 2023 **Microsoft Research, Redmond.** Research internship. *Redmond, WA, USA*
Proposed LLaVA-Plus [3], enables vision language models with vision experts as tools.

Jul 2021 – May 2023 **IDEA Research** Research internship. *Shenzhen, China*
Worked on open-world visual recognitions. Proposed Grounding DINO [4], Grounded-SAM [34], DINO [7], DAB-DETR [8], etc.

Education

Sep 2020 – Jun 2025 **Ph.D., Department of Computer Scientist, Tsinghua University**
Topics: Computer Vision; Multimodal Learning; Agents
Outstanding thesis / Outstanding Graduate of Tsinghua University

Sep 2016 – Jun 2020 **B.Eng, Department of Industry Engineering, Tsinghua University**

Academic Awards

Apr 2024 **WAIC Yunfan Award – Rising Star** *(15 people a year.)*

Feb 2024 **KAUST AI Rising Star** *(15% accept ratio.)*

Oct 2023 **CCF-CV Academic Emerging Scholar Award** *(3 people a year. CCF stands for the China Computer Federation, the leading professional organization for computer science in China.)*

Oct 2024 **Innovation 84 Scholarship** *(Top-Tier Scholarship at the maximum amount at Tsinghua University.)*

Highlight Publications

1. H.-a. Gao, Z. Zhang, T. Luo, K. Yang, X. Juan, J. Qiu, T. Chen, B. He, H. Zhao, H. Zhou, **S. Liu**, and M. Wang, *Cubebench: Diagnosing interactive, long-horizon spatial reasoning under partial observations*, 2026. arXiv: 2512.23328 [cs.AI]. URL: https://arxiv.org/abs/2512.23328.
2. J. Feng, Y. Zhang, C. Zhang, Y. Lu, **S. Liu**, and M. Wang, *Web world models*, 2025. arXiv: 2512.23676 [cs.AI]. URL: https://arxiv.org/abs/2512.23676.
3. **S. Liu**, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu, L. Zhang, J. Gao, and C. Li, "Llava-plus: Learning to use tools for creating multimodal agents," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 126–142, ISBN: 978-3-031-72970-6.
4. **S. Liu**, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 38–55, ISBN: 978-3-031-72970-6.
5. J. Qiu, X. Qi, H. Wang, X. Juan, Y. Wang, Z. Zhao, J. Geng, J. Guo, P. Li, J. Shi, **S. Liu**, and M. Wang, *Alita-g: Self-evolving generative agent for agent generation*, 2025. arXiv: 2510.23601 [cs.AI]. URL: https://arxiv.org/abs/2510.23601.
6. J. Qiu, X. Qi, T. Zhang, X. Juan, J. Guo, Y. Lu, Y. Wang, Z. Yao, Q. Ren, X. Jiang, X. Zhou, D. Liu, L. Yang, Y. Wu, K. Huang, **S. Liu**, H. Wang, and M. Wang, *Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution*, 2025. arXiv: 2505.20286 [cs.AI]. URL: https://arxiv.org/abs/2505.20286.
7. H. Zhang, F. Li, **S. Liu**, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," in *ICLR 2023*, 2023.
8. **S. Liu**, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=oMI9PjOb9Jl.

Research Publications

Check my Google Scholar page for a full list.

1. H.-a. Gao, Z. Zhang, T. Luo, K. Yang, X. Juan, J. Qiu, T. Chen, B. He, H. Zhao, H. Zhou, **S. Liu**, and M. Wang, *Cubebench: Diagnosing interactive, long-horizon spatial reasoning under partial observations*, 2026. arXiv: 2512.23328 [cs.AI]. URL: https://arxiv.org/abs/2512.23328.
2. J. Feng, Y. Zhang, C. Zhang, Y. Lu, **S. Liu**, and M. Wang, *Web world models*, 2025. arXiv: 2512.23676 [cs.AI]. URL: https://arxiv.org/abs/2512.23676.
3. H.-a. Gao, J. Geng, W. Hua, M. Hu, X. Juan, H. Liu, **S. Liu**, J. Qiu, X. Qi, Y. Wu, H. Wang, H. Xiao, Y. Zhou, S. Zhang, J. Zhang, J. Xiang, Y. Fang, Q. Zhao, D. Liu, Q. Ren, C. Qian, Z. Wang, M. Hu, H. Wang, Q. Wu, H. Ji, and M. Wang, *A survey of self-evolving agents: On path to artificial super intelligence*, 2025. arXiv: 2507.21046 [cs.AI]. URL: https://arxiv.org/abs/2507.21046.
4. Q. Jiang, F. Li, Z. Zeng, T. Ren, **S. Liu**, and L. Zhang, "T-rex2: Towards generic object detection via text-visual prompt synergy," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 38–57, ISBN: 978-3-031-73414-4.
5. Z. Jiang, S. Xie, W. Li, W. Zu, P. Li, J. Qiu, S. Pei, L. Ma, T. Huang, M. Wang, and **S. Liu**, *Zoom in, click out: Unlocking and evaluating the potential of zooming for gui grounding*, 2025. arXiv: 2512.05941 [cs.CV]. URL: https://arxiv.org/abs/2512.05941.

6 A. Y. Li, B. Yu, D. Lei, T. Ren, and **S. Liu**, *Chain-of-ground: Improving gui grounding via iterative reasoning and reference feedback*, 2025. arXiv: 2512.01979 [cs.AI]. URL: https://arxiv.org/abs/2512.01979.

7 F. Li, H. Zhang, P. Sun, X. Zou, **S. Liu**, C. Li, J. Yang, L. Zhang, and J. Gao, "Segment and recognize anything at any granularity," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 467–484, ISBN: 978-3-031-73195-2.

8 H. Li, H. Zhang, **S. Liu**, Z. Zeng, T. Ren, F. Li, and L. Zhang, "Taptr: Tracking any point with transformers as detection," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 57–75, ISBN: 978-3-031-72640-8.

9 Z. Li, G. Chen, **S. Liu**, S. Wang, V. VS, Y. Ji, S. Lan, H. Zhang, Y. Zhao, S. Radhakrishnan, N. Chang, K. Sapra, A. S. Deshmukh, T. Rintamaki, M. Le, I. Karmanov, L. Voegtle, P. Fischer, D.-A. Huang, T. Roman, T. Lu, J. M. Alvarez, B. Catanzaro, J. Kautz, A. Tao, G. Liu, and Z. Yu, *Eagle 2: Building post-training data strategies from scratch for frontier vision-language models*, 2025. arXiv: 2501.14818 [cs.CV]. URL: https://arxiv.org/abs/2501.14818.

10 **S. Liu**, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu, L. Zhang, J. Gao, and C. Li, "Llava-plus: Learning to use tools for creating multimodal agents," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 126–142, ISBN: 978-3-031-72970-6.

11 **S. Liu**, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 38–55, ISBN: 978-3-031-72970-6.

12 J. Qiu, X. Qi, H. Wang, X. Juan, Y. Wang, Z. Zhao, J. Geng, J. Guo, P. Li, J. Shi, **S. Liu**, and M. Wang, *Alita-g: Self-evolving generative agent for agent generation*, 2025. arXiv: 2510.23601 [cs.AI]. URL: https://arxiv.org/abs/2510.23601.

13 J. Qiu, X. Qi, T. Zhang, X. Juan, J. Guo, Y. Lu, Y. Wang, Z. Yao, Q. Ren, X. Jiang, X. Zhou, D. Liu, L. Yang, Y. Wu, K. Huang, **S. Liu**, H. Wang, and M. Wang, *Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution*, 2025. arXiv: 2505.20286 [cs.AI]. URL: https://arxiv.org/abs/2505.20286.

14 J. Qiu, F. Xiao, Y. Wang, Y. Mao, Y. Chen, X. Juan, S. Zhang, S. Wang, X. Qi, T. Zhang, Z. Yao, J. Guo, Y. Lu, C. Argon, J. Cui, D. Chen, J. Zhou, S. Zhou, Z. Zhou, L. Yang, **S. Liu**, H. Wang, K. Huang, X. Jiang, Y. Cao, Y. Chen, Y. Chen, Z. Chen, R. Dai, M. Deng, J. Fu, Y. Gu, Z. Guan, Z. Huang, X. Ji, Y. Jiang, D. Kong, H. Li, J. Li, R. Li, T. Li, Z. Li, H. Lian, M. Lin, X. Liu, J. Lu, J. Lu, W. Luo, Z. Luo, Z. Pu, Z. Qiao, R. Ren, L. Wan, R. Wang, T. Wang, Y. Wang, Z. Wang, Z. Wang, Y. Wu, Z. Wu, H. Xin, W. Xing, R. Xiong, W. Xu, Y. Shu, Y. Xiao, X. Yang, Y. Yang, N. Yi, J. Yu, Y. Yu, H. Zeng, D. Zhang, Y. Zhang, Z. Zhang, Z. Zhang, X. Zheng, P. Zhou, L. Zhong, X. Zong, Y. Zhao, Z. Chen, L. Ding, X. Gao, B. Gong, Y. Li, Y. Liao, G. Ma, T. Ma, X. Sun, T. Wang, H. Xia, R. Xian, G. Ye, T. Yu, W. Zhang, Y. Wang, X. Gao, and M. Wang, *On path to multimodal historical reasoning: Histbench and histagent*, 2025. arXiv: 2505.20246 [cs.AI]. URL: https://arxiv.org/abs/2505.20246.

15 J. Yang, A. Zeng, T. Ren, **S. Liu**, F. Li, R. Zhang, and L. Zhang, "Ed-pose++: Enhanced explicit box detection for conventional and interactive multi-object keypoint detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 7, pp. 5636–5654, 2025. DOI: 10.1109/TPAMI.2025.3555527.

16 H. Zhang, H. Li, F. Li, T. Ren, X. Zou, **S. Liu**, S. Huang, J. Gao, Leizhang, C. Li, and J. Yang, "Llava-grounding: Grounded visual chat with large multimodal models," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 19–35, ISBN: 978-3-031-72775-7.

17 F. Bao, C. Xiang, G. Yue, G. He, H. Zhu, K. Zheng, M. Zhao, **S. Liu**, Y. Wang, and J. Zhu, *Vidu: A highly consistent, dynamic and skilled text-to-video generator with diffusion models*, 2024. arXiv: 2405.04233 [cs.CV]. URL: https://arxiv.org/abs/2405.04233.

18 F. Bao, C. Xiang, G. Yue, G. He, H. Zhu, K. Zheng, M. Zhao, **S. Liu**, Y. Wang, and J. Zhu, "Vidu: A highly consistent, dynamic and skilled text-to-video generator with diffusion models," *arXiv preprint arXiv:2405.04233*, 2024.

19 C. Chen, Y. Guo, F. Tian, **S. Liu**, W. Yang, Z. Wang, J. Wu, H. Su, H. Pfister, and S. Liu, "A unified interactive model evaluation for classification, object detection, and instance segmentation in computer vision," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 1, pp. 76–86, 2024. DOI: 10.1109/TVCG.2023.3326588.

20 F. Li, Q. Jiang, H. Zhang, T. Ren, **S. Liu**, X. Zou, H. Xu, H. Li, J. Yang, C. Li, L. Zhang, and J. Gao, "Visual in-context prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 12 861–12 871.

21 J. Li, **S. Liu**, Z. Liu, Y. Wang, K. Zheng, J. Xu, J. Li, and J. Zhu, "Instructpix2nerf: Instructed 3d portrait editing from a single image," *ICLR*, 2024.

22 Y. Li, F. Wei, C. Zhang, and H. Zhang, *Eagle-2: Faster inference of language models with dynamic draft trees*, 2024. arXiv: 2406.16858 [cs.CL]. URL: https://arxiv.org/abs/2406.16858.

23 T. Ren, Q. Jiang, **S. Liu**, Z. Zeng, W. Liu, H. Gao, H. Huang, Z. Ma, X. Jiang, Y. Chen, Y. Xiong, H. Zhang, F. Li, P. Tang, K. Yu, and L. Zhang, *Grounding dino 1.5: Advance the "edge" of open-set object detection*, 2024. arXiv: 2405.10300 [cs.CV]. URL: https://arxiv.org/abs/2405.10300.

24 Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, **S. Liu**, Y. Guo, and L. Zhang, "Recognize anything: A strong image tagging model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2024, pp. 1724–1732.

25 X. Zou, L. Li, J. Wang, J. Yang, M. Ding, Z. Yang, F. Li, H. Zhang, **S. Liu**, A. Aravinthan, Y. J. Lee†, and L. Wang†, *Interfacing foundation models' embeddings*, 2024.

26 F. Li, A. Zeng, **S. Liu**, H. Zhang, H. Li, L. Zhang, and L. M. Ni, "Lite detr: An interleaved multi-scale encoder for efficient detr," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 18 558–18 567.

27 F. Li, H. Zhang, H. Xu, **S. Liu**, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 3041–3050.

28 H. Li, H. Zhang, Z. Zeng, **S. Liu**, F. Li, T. Ren, and L. Zhang, "Dfa3d: 3d deformable attention for 2d-to-3d feature lifting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 6684–6693.

29 **S. Liu**, S. Huang, F. Li, H. Zhang, Y. Liang, H. Su, J. Zhu, and L. Zhang, "Dq-detr: Dual query detection transformer for phrase extraction and grounding," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 1728–1736, Jun. 2023. DOI: 10.1609/aaai.v37i2.25261.

30 **S. Liu**, T. Ren, J. Chen, Z. Zeng, H. Zhang, F. Li, H. Li, J. Huang, H. Su, J. Zhu, and L. Zhang, "Detection transformer with stable matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 6491–6500.

31 T. Ren, **S. Liu**, F. Li, H. Zhang, A. Zeng, J. Yang, X. Liao, D. Jia, H. Li, H. Cao, J. Wang, Z. Zeng, X. Qi, Y. Yuan, J. Yang, and L. Zhang, *Detrex: Benchmarking detection transformers*, 2023. arXiv: 2306.07265 [cs.CV]. URL: https://arxiv.org/abs/2306.07265.

[32] T. Ren, **S. Liu**, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded sam: Assembling open-world models for diverse visual tasks," in *ICCV 2023 Demo*, 2023. arXiv: 2401.14159 [cs.CV].

[33] Y. Shi, J. Wang, H. Cao, B. Tang, X. Qi, T. Yang, Y. Huang, **S. Liu**, L. Zhang, and H.-Y. Shum, "Toss: High-quality text-guided novel view synthesis from a single image," *ICLR*, 2023.

[34] J. Yang, A. Zeng, F. Li, **S. Liu**, R. Zhang, and L. Zhang, "Neural interactive keypoint detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 15 122–15 132.

[35] J. Yang, A. Zeng, **S. Liu**, F. Li, R. Zhang, and L. Zhang, "Explicit box detection unifies end-to-end multi-person pose estimation," in *International Conference on Learning Representations*, 2023. URL: https://openreview.net/forum?id=s4WVupnJjmX.

[36] H. Zhang, F. Li, **S. Liu**, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," in *ICLR 2023*, 2023.

[37] H. Zhang, F. Li, H. Xu, S. Huang, **S. Liu**, L. M. Ni, and L. Zhang, "Mp-former: Mask-piloted transformer for image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 18 074–18 083.

[38] H. Zhang, F. Li, X. Zou, **S. Liu**, C. Li, J. Yang, and L. Zhang, "A simple framework for open-vocabulary segmentation and detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 1020–1031.

[39] F. Li, H. Zhang, **S. Liu**, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 619–13 627.

[40] F. Li, H. Zhang, Y.-F. Zhang, **S. Liu**, J. Guo, L. M. Ni, P. Zhang, and L. Zhang, *Vision-language intelligence: Tasks, representation learning, and large models*, 2022. arXiv: 2203.01922 [cs.CV]. URL: https://arxiv.org/abs/2203.01922.

[41] **S. Liu**, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=oMI9PjOb9Jl.

[42] X. Yang, **S. Liu**, Y. Dong, H. Su, L. Zhang, and J. Zhu, "Towards generalizable detection of face forgery via self-guided model-agnostic learning," *Pattern Recognition Letters*, vol. 160, pp. 98–104, 2022, ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2022.06.007.

[43] **S. Liu**, L. Zhang, X. Yang, H. Su, and J. Zhu, *Query2label: A simple transformer way to multi-label classification*, 2021. arXiv: 2107.10834 [cs.CV]. URL: https://arxiv.org/abs/2107.10834.

[44] **S. Liu**, L. Zhang, X. Yang, H. Su, and J. Zhu, "Unsupervised part segmentation through disentangling appearance and shape," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 8355–8364.