

# SHIZHE CHEN

Email: shizhe.chen@inria.fr

Website: <https://cshizhe.github.io>

## EMPLOYMENT

---

<b>INRIA, Researcher</b> Vision-and-language, embodied AI	<i>Oct. 2023 - Present</i>
<b>INRIA, Postdoc</b> Vision-and-language navigation	<i>Mar. 2021 - Sep. 2023</i> Advisor: Ivan Laptev and Cordelia Schmid
<b>Carnegie Mellon University, Researcher</b> Zero-shot action recognition	<i>Jun. 2020 - Dec. 2020</i> Advisor: Dong Huang
<b>University of Adelaide, Visiting Scholar</b> Controllable image caption generation	<i>Jul. 2019 - Oct. 2019</i> Advisor: Qi Wu
<b>Microsoft Research Asia, Research Intern</b> Neural storyboard creation	<i>Dec. 2018 - Jun. 2019</i> Mentors: Jianlong Fu and Ruihua Song
<b>Carnegie Mellon University, Visiting Scholar</b> Video caption generation	<i>Oct. 2017 - Oct. 2018</i> Advisor: Alexander Hauptmann

## EDUCATION

---

<b>Renmin University of China</b> Ph.D in Computer Science	<i>Sep. 2015 - Jun. 2020</i> Advisor: Qin Jin
<b>Renmin University of China</b> B.S in Computer Science	<i>Sep. 2011 - Jun. 2015</i> Advisor: Qin Jin

## HONORS

---

★ Baidu Academic: Top Global Young Chinese Female Scholars in AI.	2023
★ AI TIME: AI 2000 Most Influential Scholar Honorable Mention in Multimedia.	2022 - 2024
★ Outstanding Graduate Award (Highest honor for graduate set by the government of Beijing).	2020
★ JingDong Scholarship (10 students at Renmin University).	2019
★ ACM Multimedia Student Travel Grant.	2019
★ ICMR Best Paper Runner up.	2018
★ Baidu Scholarship (10 Ph.D students worldwide).	2017
★ National Scholarship for Ph.D Students.	2016
★ ACM Multimedia Student Travel Grant.	2016
★ National Scholarship for Undergraduate Students.	2013

## AWARDS

---

★ Ranked 2nd in CSIG REVERIE VLN Challenge.	2022
★ Ranked 1st in ICCV HIRV Workshop REVERIE & SOON VLN Challenges.	2021
★ Ranked 1st in CVPR ActivityNet Entities Object Localization Challenge.	2021
★ Ranked 1st in CVPR ActivityNet Dense Video Captioning Challenge.	2018 - 2020
★ Ranked 1st in NIST Trecvid Video to Text Challenge.	2017 - 2020
★ Ranked 2nd in CVPR Video Understanding Pentathlon Challenge.	2020
★ First Prize in Zhijiang Cup Global AI Competition Video Captioning Challenge.	2019

- ★ Ranked 2nd and won Outstanding Method Prize in ICCV VATEX Video Captioning Challenge. 2019
- ★ Ranked 1st in ACM Multimedia AVEC Emotion Recognition Challenge. 2017 - 2019
- ★ Ranked 1st in ACM Multimedia Video to Language Grand Challenge. 2016 - 2017
- ★ Ranked 2nd in ACM Multimedia AVEC Emotion Recognition Challenge. 2016
- ★ Ranked 1st in MediaEval Emotion Impact of Movies Task. 2016
- ★ Ranked 2nd in CCPR Multimodal Emotion Recognition Challenge. 2016
- ★ Second Prize in Chinese Big Data Contest P2P Task. 2015
- ★ Second Prize in IBM Bleumix Cognitive Computation Development Contest. 2015
- ★ First Prize in National College Student Information Security Contest. 2014
- ★ Second Prize in Chinese Big Data Contest Baidu IERMu Task. 2014
- ★ Meritorious Winner in American Mathematical Contest in Modeling. 2014
- ★ National Second Prize in China Undergraduate Mathematical Contest in Modeling. 2013

## PROFESSIONAL ACTIVITIES

---

### Area chair

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2023, 2024, 2025.
- European Conference on Computer Vision (ECCV) 2024.
- International Conference on Computer Vision (ICCV) 2023.
- Conference on Neural Information Processing Systems (NeurIPS) 2023, 2024, 2025.
- International Conference on Learning Representations (ICLR) 2023, 2024, 2025.
- International Conference on Machine Learning (ICML) 2025.
- ACM Multimedia (ACM MM) 2022, 2023.
- Asian Conference on Computer Vision (ACCV) 2024.

### Workshop co-organizer

- CVPR 2025 Generalization in Robotics Manipulation Workshop and Challenges
- ICMR 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding
- CVPR 2020 Workshop on Language & Vision with applications to Video Understanding

### Conference & journal reviewer

- Conferences: CVPR, ICCV, ECCV, AACL, ACM MM, ACL, EMNLP, COLM, IROS, COLM
- Journals: TPAMI, TMM, TOMM, T-RO, RA-L

### Supervision

- PhD students: Francois Porcher (co-supervise), Paul Pacaud (co-supervise), Sara Pieri (co-supervise), Francois Garderes (co-supervise), Zeeshan Khan (co-supervise), Thomas Chabal (co-advise), Riccardo Garcia (co-supervised), Zerui Chen (co-supervised), Pierre-Louis Guhur (co-advised), Anwen Hu (co-advised)
- Master students: Kejun Lin (co-advise), Qingrong He (co-advised), Yuqing Song (co-advised), Yida Zhao (co-advised)

### Teaching

- Course project advisor for Object Recognition and Computer Vision, 2021-2024
- Guest lecturer at Spoken Language Processing, 2019
- Teaching assistant for Data Structure and Algorithm, 2015
- Teaching assistant for Programming Practice with C Language, 2015

## INVITED PRESENTATIONS

---

- ★ Talk at CVPR 2025 Workshop on Computer Vision in the Wild. 06/2025
- ★ Talk at GDR IASIS Workshop on Deformable Object Modeling Trends: from Perception to Applications. 04/2025
- ★ Talk at Imaging in Paris Seminar. 06/2024

★ Talk at Gaoling School of Artificial Intelligence, Renmin University of China.	02/2024
★ Talk at BAAI Young Researcher Seminar.	08/2023
★ Talk at CAAI Tutorial on Embodied AI.	07/2023
★ Talk at IMAGINE research group at Ecole des Ponts ParisTech.	12/2022
★ Presentation at WILLOW/SIERRA retreat, Saint-Raphaël.	10/2022
★ Presentation at Stanford Vision and Learning Lab iGibson and BEHAVIOR team.	09/2022
★ Talk at Renmin University of China: Multimodal Perception and Action.	07/2022
★ Talk at AI TIME: Embodied Vision-and-Language Navigation in 3D Environments.	04/2022
★ Talk at Microsoft Research Asia: Recent Advances in Vision-and-Language Navigation.	03/2022
★ Presentation at WILLOW/SIERRA retreat, Avignon.	10/2021
★ Talk at Tencent: Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning.	06/2020

## PUBLICATION

---

I have published over 40 scientific papers most of which appeared in international journals and major peer-reviewed conferences. The leading conferences in computer vision (ICCV, ECCV, CVPR), machine learning (NeurIPS, ICLR) and robotics (CoRL, ICRA, IROS). Overall, my publications have **over 3,800 citations** and my **h-index is 27** (both obtained from Google Scholar).

1. Shiyao Li, Antoine Guédon, Clémentin Boittiaux, **Shizhe Chen**, and Vincent Lepetit. Nextbest-path: Efficient 3d mapping of unseen environments. In *ICLR*, 2025
2. Thomas Chabal, **Shizhe Chen**, Jean Ponce, and Cordelia Schmid. Online 3d scene reconstruction using neural object priors. In *3DV*, 2025
3. Ricardo Garcia, **Shizhe Chen**, and Cordelia Schmid. Towards generalizable vision-language robotic manipulation: A benchmark and llm-guided 3d policy. In *ICRA*, 2025
4. Zerui Chen, **Shizhe Chen**, Etienne Arlaud, Ivan Laptev, and Cordelia Schmid. Vividex: Learning vision-based dexterous manipulation from human videos. In *ICRA*, 2025
5. **Shizhe Chen**, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. SUGAR: Pre-training 3d visual representations for robotics. In *CVPR*, 2024
6. **Shizhe Chen**, Ricardo Garcia, Cordelia Schmid, and Ivan Laptev. Polarnet: 3d point clouds for language-guided robotic manipulation. In *CoRL*, 2023
7. **Shizhe Chen**, Thomas Chabal, Ivan Laptev, and Cordelia Schmid. Object goal navigation with recursive implicit maps. In *IROS*, 2023
8. Ricardo Garcia, Robin Strudel, **Shizhe Chen**, Etienne Arlaud, Ivan Laptev, and Cordelia Schmid. Robust visual sim-to-real transfer for robotic manipulation. In *IROS*, 2023
9. Anwen Hu, **Shizhe Chen**, Liang Zhang, and Qin Jin. Explore and tell: Embodied visual captioning in 3d environments. In *ICCV*, 2023
10. Xu Gu, Yuchong Sun, Feiyue Ni, **Shizhe Chen**, Xihua Wang, Ruihua Song, Boyuan Li, and Xiang Cao. Tevis: Translating text synopses to video storyboards. In *ACM MM*, 2023
11. Zerui Chen, **Shizhe Chen**, Cordelia Schmid, and Ivan Laptev. gsdF: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *CVPR*, 2023
12. Anwen Hu, **Shizhe Chen**, Liang Zhang, and Qin Jin. Infometric: An informative metric for reference-free image caption evaluation. In *ACL*, 2023
13. Pierre-Louis Guhur, **Shizhe Chen**, Ricardo Garcia, Makarand Tapaswi, Ivan Laptev, and Cordelia Schmid. Instruction-driven history-aware policies for robotic manipulations. In *CoRL*, 2022

14. **Shizhe Chen**, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *NeurIPS*, 2022
15. **Shizhe Chen**, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Learning from unlabeled 3d environments for vision-and-language navigation. In *ECCV*, 2022
16. Sipeng Zheng, **Shizhe Chen**, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *ECCV*, pages 297–313. Springer, 2022
17. **Shizhe Chen**, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, 2022
18. Sipeng Zheng, **Shizhe Chen**, and Qin Jin. Vrdformer: End-to-end video visual relation detection with transformers. In *CVPR*, pages 18836–18846, 2022
19. **Shizhe Chen**, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *NeurIPS*, 2021
20. Pierre-Louis Guhur, Makarand Tapaswi, **Shizhe Chen**, Ivan Laptev, and Cordelia Schmid. Airbert: In-domain pretraining for vision-and-language navigation. In *ICCV*, 2021
21. **Shizhe Chen** and Dong Huang. Elaborative rehearsal for zero-shot action recognition. *ICCV*, 2021
22. Anwen Hu, **Shizhe Chen**, and Qin Jin. Question-controlled text-aware image captioning. *ACM MM*, 2021
23. Yuqing Song, **Shizhe Chen**, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. Product-oriented machine translation with cross-modal cross-lingual pre-training. In *ACM MM*, pages 2843–2852, 2021
24. Yuqing Song, **Shizhe Chen**, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. Enhancing neural machine translation with dual-side multimodal awareness. *IEEE Transactions on Multimedia*, 2021
25. Chaorui Deng, **Shizhe Chen**, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In *CVPR*, pages 234–243, 2021
26. Yuqing Song, **Shizhe Chen**, and Qin Jin. Towards diverse paragraph captioning for untrimmed videos. In *CVPR*, pages 11245–11254, 2021
27. **Shizhe Chen**, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *CVPR*, pages 9962–9971, 2020
28. **Shizhe Chen**, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, pages 10638–10647, 2020
29. Anwen Hu, **Shizhe Chen**, and Qin Jin. Icecap: Information concentrated entity-aware image captioning. In *ACM Multimedia*, pages 4217–4225, 2020
30. Sipeng Zheng, **Shizhe Chen**, and Qin Jin. Skeleton-based interactive graph network for human object interaction detection. In *ICME*, pages 1–6, 2020
31. **Shizhe Chen**, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin, Xiaoyu Qi, Chunting Wang, and Jin Zhou. Neural storyboard artist: Visualizing stories with coherent image sequences. In *ACM Multimedia*, pages 2236–2244, 2019
32. Yuqing Song, **Shizhe Chen**, Yida Zhao, and Qin Jin. Unpaired cross-lingual image caption generation with self-supervised rewards. In *ACM Multimedia*, pages 784–792, 2019
33. Sipeng Zheng, **Shizhe Chen**, and Qin Jin. Visual relation detection with multi-level attention. In *ACM Multimedia*, pages 121–129, 2019

34. **Shizhe Chen**, Qin Jin, and Jianlong Fu. From words to sentences: A progressive learning approach for zero-resource machine translation with visual pivots. In *IJCAI*, pages 4932–4938, 2019
35. **Shizhe Chen**, Qin Jin, and Alexander G. Hauptmann. Unsupervised bilingual lexicon induction from mono-lingual multimodal data. In *AAAI*, pages 8207–8214, 2019
36. Weiyang Wang, Yongcheng Wang, **Shizhe Chen**, and Qin Jin. Youmakeup: A large-scale domain-specific multimodal dataset for fine-grained semantic comprehension. In *EMNLP*, pages 5136–5146, 2019
37. **Shizhe Chen**, Qin Jin, Jia Chen, and Alexander G Hauptmann. Generating video descriptions with latent topic guidance. *IEEE Trans. Multimedia*, 21(9):2407–2418, 2019
38. **Shizhe Chen**, Jia Chen, Qin Jin, and Alexander Hauptmann. Class-aware self-attention for audio event recognition. In *ICMR*, pages 28–36, 2018
39. **Shizhe Chen**, Jia Chen, Qin Jin, and Alexander Hauptmann. Video captioning with guidance of multimodal latent topics. In *ACM Multimedia*, pages 1838–1846, 2017
40. Qin Jin, Jia Chen, **Shizhe Chen**, Yifan Xiong, and Alexander Hauptmann. Describing videos using multi-modal fusion. In *ACM Multimedia*, pages 1087–1091, 2016
41. **Shizhe Chen** and Qin Jin. Multi-modal conditional attention fusion for dimensional emotion prediction. In *ACM Multimedia*, pages 571–575, 2016