# SHANGZHE DI

(86)15652587063 ⋄ shangzhe.di@gmail.com ⋄ dszdsz.cn

## RESEARCH INTEREST

Video Understanding, Multi-modal Learning

## EDUCATION

**Shanghai Jiao Tong University** — Apr 2023 - Present
*PhD student, supervised by Prof. Weidi Xie* — *Shanghai, China*

**Beihang University** — Sep 2020 - Jan 2023
*M.Eng. in Computer Science, supervised by Prof. Si Liu* — *Beijing, China*

**Technical University of Munich** — Apr 2019 - Aug 2019
*Exchange Program* — *Munich, Germany*

**Beihang University** — Sep 2016 - Jun 2020
*B.Eng. in Software Engineering* — *Beijing, China*

## INTERNSHIP

**ByteDance** — Nov 2024 - Present
*Research Topic: Multi-task Self-supervised Visual Learning* — *Shanghai, China*

**Alibaba** — Apr 2024 - Sep 2024
*Research Topic: Streaming Video Understanding* — *Hangzhou, China*

**SenseTime** — Aug 2022 - Jan 2023
*Research Topic: Vision Foundation Models* — *Beijing, China*

## PUBLICATIONS

$*$ stands for equal contribution.

[1] **Shangzhe Di**$^*$, Zeren Jiang$^*$, et al. "Video Background Music Generation with Controllable Music Transformer." In *ACM International Conference on Multimedia (ACM MM)*, 2021, Best Paper Award.

[2] Yulu Gao, Chonghao Sima, Shaoshuai Shi, **Shangzhe Di**, et al. "Sparse Dense Fusion for 3D Object Detection." In *International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[3] Zizheng Xun$^*$, **Shangzhe Di**$^*$, et al. "Linker: Learning Long Short-term Associations for Robust Visual Tracking." In *IEEE Transactions on Multimedia (TMM)*, 2023.

[4] **Shangzhe Di**, Weidi Xie. "Grounded Question-Answering in Long Egocentric Videos." In *IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.

[5] Qirui Chen, **Shangzhe Di**, Weidi Xie. "Grounded Multi-Hop VideoQA in Long-Form Egocentric Videos." In *The 39th AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[6] **Shangzhe Di**, et al. "Streaming Video Question-Answering with In-context Video KV-Cache Retrieval." In *The 13th International Conference on Learning Representations (ICLR)*, 2025.

[7] Yudi Shi, **Shangzhe Di**, Qirui Chen, Weidi Xie. "Enhancing Video-LLM Reasoning via Agent-of-Thoughts Distillation." In *IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[8] Yibin Yan*, Jilan Xu*, **Shangzhe Di**, Yikun Liu, Yudi Shi, Qirui Chen, Zeqian Li, Yifei Huang, Weidi Xie. "Learning Streaming Video Representation via Multitask Training." In *International Conference on Computer Vision (ICCV)*, 2025, Oral.

[9] Zeqian Li, **Shangzhe Di**, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, Weidi Xie. "Universal Video Temporal Grounding with Generative Multi-modal Large Language Models." In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

## HONORS AND AWARDS

**Best Paper Award**, ACM MM, 2021

Best Video Award, IJCAI, 2021

AntiUAV Challenge Runner-up ($1,000), ICCV, 2021

AOT Challenge Runner-up ($7,500), ICCV, 2021

First Prize Scholarship × 2 (Top 10%), Beihang University, 2019 & 2021

**Full Scholarship for Exchange Program**, China Scholarship Council, 2019

**Special Prize Scholarship** (Top 3%), Beihang University, 2018

## PROFESSIONAL ACTIVITIES

**Reviewer** of ICLR, CVPR, NeurIPS, and ICML.

## INVITED TALKS

**Learning to Generate Video Background Music** Sep 26, 2021

*JIG Graduate Academic Forum 2021*

**Towards Long-form Video Question-Answering** Dec 16, 2024

*NICE No.38*

## REFEREES

**Weidi Xie**, Associate Professor, Shanghai Jiao Tong University, weidi@sjtu.edu.cn

**Si Liu**, Professor, Beihang University, liusi@buaa.edu.cn