

Xichen Pan

xichenpan.com

E-mail : xcpan.mail@gmail.com

Mobile : +86 186 535 00448

EDUCATION

- **Shanghai Jiao Tong University** Shanghai, China
B.Eng. in Computer Science (**Outstanding Graduate**), advised by Prof. Zhouhan Lin *Sept. 2018 – June 2022*
Overall: 88.42/100, Major: 91.29/100

RESEARCH INTERSECTS

- **Multimodal Deep Learning:** Multimodal deep learning (including audio-visual and vision-language), especially multimodal representation learning and multimodal self-supervised pre-training

PUBLICATIONS & MANUSCRIPTS

- Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhui Chen. **Synthesizing Coherent Story with Auto-Regressive Latent Diffusion Models**, *CVPR 2023 Under Review* [pdf]
- Xichen Pan, Zekai Li, Yichen Gong, Xinbing Wang, and Zhouhan Lin. **Towards Diverse Lip Reading Representations**, *ICASSP 2023 Under Review*
- Xichen Pan. **Multimodal Audio-Visual Speech Recognition System Based On Pre-trained Models**, *Bachelor Thesis at Shanghai Jiao Tong University (Best Thesis Award, 1st/150)* [news]
- Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. **Leveraging Unimodal Self-Supervised Learning for Multimodal Audio-Visual Speech Recognition**, *ACL 2022 Main Conference* [pdf]

EXPERIENCE

- **Microsoft Research** Beijing, China
Multimodal GPT *Nov. 2022 – Present*
StarBridge Research Assistant, mentored by Li Dong
 - Participated in multimodal GPT, a big text and image decoder model with visual, language and audio conditions.
- **Alibaba Group** Beijing, China
Synthesizing Coherent Story with Auto-Regressive Latent Diffusion Models *Sept. – Nov. 2022*
Research Intern, mentored by Pengda Qin
 - Proposed a history-aware auto-regressive conditioned latent diffusion model named AR-LDM, which first successfully leverages diffusion models for story visualization and continuation with relative FID score improvements of 70% and 20% over previous SoTA, respectively.
 - Introduced the VIST dataset and showed AR-LDM is capable of real-world story synthesis.
 - Proposed a simple but efficient adaptation method, allowing AR-LDM to generalize to unseen characters.
- **Horizon Robotics** Beijing, China
Towards Diverse Lip Reading Representations *Apr. 2021 – July 2022*
Research Intern, mentored by Yichen Gong
 - Improved the diversity of lip reading representations by using an attention mask to maintain and incorporate contextual information. Alleviated the over-smoothing problem of Transformer in word-level lip reading. The proposed method achieved new SoTA audio-visual speech recognition performance on Lip Reading in the Wild.
- **John Hopcroft Center for Computer Science, Shanghai Jiao Tong University** Shanghai, China
Leveraging Unimodal Self-Supervised Learning for Multimodal AVSR *Apr. – Sept. 2021*
Research Intern, advised by Prof. Zhouhan Lin
 - Successfully leveraged unimodal self-supervised pre-training for multimodal audio-visual speech recognition for the first time, achieved a word error rate (WER) of 2.6% on Lip Reading Sentences 2 (LRS2), raising the SoTA performances with a relative improvement of 30%. The proposed audio-only and visual-only models also reached a WER of 2.7% and 43.2%, respectively.
 - Largely improved models' noise robustness, as well as reduced the need of labeled aligned data through the extra self-supervised pre-training.

SKILLS

- **Programming Languages:** C/C++, Python
- **Packages:** PyTorch, Lightning, Transformers, Diffusers, fairseq, WandB, Hydra, OpenCV, h5py, NumPy, PyQt5