Zhexiao Xiong

+1-314-319-2407 | x.zhexiao@wustl.edu | Linkedin | Github | Personal-Webpage

Biography

I am a fourth-year CS Ph.D. candidate at Washington University in St. Louis(WashU), advised by **Prof. Nathan Jacobs**. My research interests lie in computer vision and multi-modal learning, with a focus on generative models, vision-language models (VLMs) and AIGC-related topics. In particular: (1) Unifying vision understanding and generation, world models; (2) Controllable & personalized image/video generation and editing; (3) Integration of VLMs with generative models; (4) Generative models for 3D vision, including neural rendering, cross-view synthesis, and novel view synthesis.

Education

- **Washington University in St. Louis** — *2022.08 – 2027.05(Expected)*
  Ph.D. Candidate in Computer Science — St. Louis, MO, USA
  Advisor: Prof. **Nathan Jacobs**
- **Tianjin University** — *2018.09 – 2022.06*
  B.S. in Electrical and Information Engineering — Tianjin, China

Work Experience

- **Bosch Research** [🌐] — *2025.06 – 2025.09*
  Research Intern — Sunnyvale, CA, USA
  - Researched on Unified Visual Understanding, Planning and Generation Models for autonomous driving.
- **OPPO US Research Center** [🌐] — *2024.05 – 2024.08*
  Research Intern — Palo Alto, CA, USA
  - Researched on text-guided 3D Scene Generation, use Large-language model(LLM)-based dreaming and video generation models to generate both geometric and semantic consistent 3D scene.
- **OPPO Research Institute** [🌐] — *2022.02 – 2022.05*
  Research Intern — Beijing, China
  - Researched on image matting, proposed a framework to use human pose as guidance to achieve whole body matting.
- **Institute of Automation, Chinese Academy of Sciences(CASIA)** [🌐] — *2021.01 – 2022.01*
  Research Intern — Beijing, China
  - Researched on model compression and network pruning, especially the application on Vision Transformers.

Selected Publications

C=Conference, J=Journal, P=Pre-print

[P.1] **Zhexiao Xiong**, Wei Xiong, Jing Shi, He Zhang, Yizhi Song, Nathan Jacobs. **GroundingBooth: Grounding Text-to-Image Customization**. *Arxiv Pre-print*.

[P.2] **Zhexiao Xiong**, Xin Ye, Burhan Yaman, Yiren Lu, Feng Qiao, Liu Ren, Nathan Jacobs. **UniDrive: Unified Understanding, Planning and Generation World Model For Autonomous Driving**. *In Submission*.

[P.2] **Zhexiao Xiong**, Wei Xiong, Yizhi Song, Liu He, Yu Yuan, Nathan Jacobs. **Physics-Align A Video: Physics-Coherent Image-to-Video Generation through Feature and 3D Representation Alignment**. *In Submission*.

[C.1] **Zhexiao Xiong**, Zhang Chen, Zhong Li, Yi Xu, Nathan Jacobs. **PanoDreamer: Consistent Text to 360-Degree Scene Generation**. In *CVPR Workshops (CV4Metaverse), 2025*.

[C.2] Wanzhou Liu*, **Zhexiao Xiong***, Xinyu Li, Nathan Jacobs. **DeclutterNeRF: Generative-Free 3D Scene Recovery for Occlusion Removal**. In *CVPR Workshops (CV4Metaverse), 2025*.

[C.3] Feng Qiao, **Zhexiao Xiong**, Eric Xing, Nathan Jacobs. **GenStereo: Towards Open-World Generation of Stereo Images and Unsupervised Matching**. *International Conference on Computer Vision(ICCV), 2025*.

[C.4] Feng Qiao, **Zhexiao Xiong**, Xinge Zhu, Yuexin Ma, Qiumeng He, Nathan Jacobs. **MCPDepth: Omnidirectional Depth Estimation via Stereo Matching from Multi-Cylindrical Panoramas**. *In IEEE/CVF Winter Conference on Applications of Computer Vision(WACV), 2026*.

[C.5] **Zhexiao Xiong**, Feng Qiao, Yu Zhang, Nathan Jacobs. **StereoFlowGAN: Co-training for Stereo and Flow with Unsupervised Domain Adaptation**. In *British Machine Vision Conference(BMVC), 2023*.

[C.6] Xin Xing, **Zhexiao Xiong**, Abby Stylianou, Srikumar Sastry, Liyu Gong, Nathan Jacobs. **Vision-Language Pseudo-Labels for Single-Positive Multi-Label Learning**. In *CVPR Workshops(CVPRW), 2024*.

[C.7] Subash Khanal, Eric Xing, Srikumar Sastry, Aayush Dhakal, **Zhexiao Xiong**, Adeel Ahmad, Nathan Jacobs. **PSM: Learning Probabilistic Embeddings for Multi-scale Zero-Shot Soundscape Mapping**. In *ACM Multimedia(ACM MM), 2024*.

[J.1] **Zhexiao Xiong**, Xin Xing, Scott Workman, Subash Khanal, Nathan Jacobs. **Mixed-View Panorama Synthesis using Geospatially Guided Diffusion**. *Transactions on Machine Learning Research(TMLR), 2025*.

[J.2] Nanfei Jiang, **Zhexiao Xiong**, Hui Tian, Xu Zhao, Xiaojie Du, Chaoyang Zhao, Jinqiao Wang. **PruneFaceDet: Pruning lightweight face detection network by sparsity training**. *Cognitive Computation and Systems, 2022*.

Projects

- **Unified Understanding, Planning and Generation model for Autonomous Driving** *2025.06 – present*
  Research Project during internship at Bosch Research
  - Developed a **world-model**-based framework that unifies trajectory planning and autoregressive future image generation, enhanced with Chain-of-Thought reasoning within a single **vision-language model (VLM)**.
  - Enabled thinking visually before planning, leading to more accurate and robust decision-making, and demonstrated significant gains on vision-language planning(VLP) benchmarks.
- **Physically Coherent Video Generation** *2025.02 – present*
  - Proposed a framework that leverages **vision-language model(VLM)**'s physics understanding to enable video generation with physically consistent motion and accurate 3D dynamics.
  - Achieved physically plausible video generation by combining relational alignment with foundation video understanding models, physics-aware feature encoding, and 3D geometry alignment.
- **Grounded text-to-image Customization** *2024.01 – 2024.09*
  Collabration with Adobe Research [🌐]
  - Proposed a framework that achieved zero-shot instance-level spatial grounding on both foreground subjects and background objects in the text-to-image customization task, enabling the customization of multiple subjects.
  - Our work is the first work to achieve a joint grounding on both subject-driven foreground generation and text-driven background generation.
  - Results show the effectiveness of our model in text-image alignment, identity preservation, and layout alignment.
- **Text to 360-Degree Scene Generation** *2024.05 – 2024.11*
  Research Project during internship at OPPO US Research Center
  - Proposed a holistic text to 360-degree scene generation pipeline, which achieved consistent text-to-360-degree scene generation with customized trajectory-guided scene extension.
  - Introduced semantically guided novel view synthesis into the refinement of 3D-GS optimization, reducing artifacts and improving geometric consistency.
- **Mixed-View Panorama Synthesis Using Geospatially-Guided Diffusion** *2023.05 – 2023.11*
  [🌐]
  - Introduced the task of mixed-view panorama synthesis, where the goal is to synthesize a novel panorama given a small set of input panoramas and a satellite image of the area.
  - Introduced an approach that utilizes diffusion-based modeling and an attention-based architecture for extracting information from all available input imagery.
- **Open-World Generation of Stereo Images and Unsupervised Matching** *2024.09 – 2025.03*
  [🌐]
  - Proposed GenStereo, a novel diffusion-based framework for open-world stereo image generation with applications in unsupervised stereo matching.
- **Co-training for Stereo and Flow with Unsupervised Domain Adaptation** *2023.01 – 2023.05*
  [🌐]
  - Built an end-to-end joint learning framework to combine unsupervised domain translation with optical flow estimation and stereo matching in the absence of real ground truth optical flow and disparity.
  - Applied novel constraints on the cycle domain translation process to achieve cross-domain translation with global and local consistency.
  - Employed task-specific multi-scale feature warping loss and iterative feature warping loss during the training phase to regulate the training process in both spatial and temporal dimensions.
- **Vision-Language Pseudo-Labels for Single-Positive Multi-Label Learning** *2022.11 – 2023.05*
  [🌐]
  - Proposed a novel approach called Vision-Language Pseudo-Labeling (VLPL), which uses a vision-language model to suggest strong positive and negative pseudo-labels, and outperforms the current SOTA methods by 5.5% on Pascal VOC, 18.4% on MS-COCO, 15.2% on NUS-WIDE, and 8.4% on CUB-Birds.

Services

- **Reviewer:**CVPR(2025), ECCV(2024), NeurIPS(2024,2025), ICML(2025), ICLR(2025), ICCV(2025)
- **Teaching Services (WashU):**CSE 559A Computer Vision **(Teaching Assistant/Grader)**

Technical Skills

Programming: Python, C/C++, Java, Matlab
Deep Learning Frameworks: Pytorch, Tensorflow
Research Frameworks: Diffusion models, VLMs,Transformer, GAN, 3DGS, NeRF, CNN, CLIP
Languages: English, Chinese