Qidong Huang

Researcher, Qwen-VL Team, Alibaba Group

Alibaba Chaoyang Technology Park C Area
Chaoyang, Beijing, China
(+86) 13085060686
hqd214215@gmail.com
https://shikiw.github.io/

Short Biography

I am currently a researcher working at Qwen-VL team, focusing on post-training techniques for Qwen-VL series. I received my PhD degree and Bachelor degree at University of Science and Technology of China (USTC). I have published more than 10 papers at top-tier conferences and journals. I am working closely with Shuai Bai, Dongdong Chen, Xiaoyi Dong, and Gang Hua.

Education

09/2020–06/2025 **PhD of Cyberspace Technology**, *University of Science and Technology of China*, Hefei, China, CAS Key Laboratory of Electromagnetic Space Information. Supervised by Prof. Weiming Zhang, Prof. Nenghai Yu.

09/2016–06/2020 **Bachelor of Electronic Information**, *University of Science and Technology of China*, Hefei, China, Supervised by Prof. Weiming Zhang.

Experience

05/2025–present **Researcher**, *Qwen-VL Team, Alibaba Group*, Beijing, China.
Key member of VL post-training, core contributor of Qwen3-VL and Qwen3-Omni series, mainly focusing on strong-to-weak distillation training (including text/VL distillation) and STEM reasoning in SFT/RFT.

08/2023–04/2025 **Research Intern**, *Shanghai AI Laboratory*, Shanghai, China.
Member of InternLM-XComposer Group, supervised by Xiaoyi Dong, Jiaqi Wang. Research in multi-modal LLMs, especially in inference-time scaling for image/video caption scaling, cross-modal alignment, efficient training/inference, and multi-modal hallucination.

05/2022–07/2022 **Research Intern**, *iFlyTek Research*, Hefei, China.
Member of Avatar Strip, supervised by Shan He. Research in Chinese text-to-image model based on conditional diffusion, focusing on large-scale image-text datasets such as Wukong.

Skills

⋆ **Expertise in VL Post-Training :** I am playing a core role in the VL post-training pipeline, focusing on :

- **1) Strong-to-Weak Distillation Training :** This is the 2nd stage in current Qwen-VL post-training, where I co-developed the Megatron/FSDP training framework for both text and VL distillation. I have rich distillation experience for both training skills and data collection strategies, successfully improving the text abilities and STEM reasoning for different sizes of Qwen-VL models (from large-scale MoE models like 235A22 on 512+ GPUs, to the smaller dense models like 2B, 4B, 8B, establishing distillation as the superior alternative to RL for these lightweight models). We also use the targeted VL distillation to resolve domain-specific performance bottlenecks appearing in post-training.
- **2) STEM Reasoning in SFT/RFT :** I am also responsible for improving STEM reasoning abilities in the first stage of VL post-training (as a good cold start for distillation training and RL). Currently, I am also designing the advanced thinking patterns (inspired by Gemini-3-Pro) to facilitate better RFT training in next-generation of Qwen-VL models.

- ⋆ **Expertise in Academic Areas :** My researches during PhD career mainly focus on multi-modal LLMs, including scalable image/video captioning, cross-modal alignment, efficient training/inference, and multi-modal hallucination.
  - — **1) Image/Video Caption Scaling :** I am the core authors of ScaleCap and CapRL. ScaleCap proposes a scalable captioning strategy that generates comprehensive and detailed image captions. CapRL presents an effective decoupled two-stage training scheme with verifiable caption reward to boost captioning model.
  - — **2) MIR&MoCa :** We propose an effective and reliable metric named MIR for quantifying MLLM pre-training, and a light-weight modality calibration module MoCa to facilitate cross-modal alignment.
  - — **3) MMRC :** We propose a multi-modal conversation benchmark MMRC for evaluating open-ended abilities of MLLMs.
  - — **4) PyramidDrop :** We propose an efficient training/inference framework for MLLMs through vision redundancy reduction, especially working for high-resolution MLLMs and video LLMs, achieving ∼50% acceleration for models like LLaVA series and Video-LLMs.
  - — **5) OPERA :** We delve into the underlying causes of multi-modal hallucinations and give an explanation based on information attenuation. Based on this, we propose a training-free decoding algorithm to mitigate the hallucination issue. This work has earned over 50,000 reads and 4,000 shares on social media, with nearly 400 citations within two years.

Publications (First Author)

- ⋆ Long Xing*, **Qidong Huang***, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jinsong Li, Shuangrui Ding, Weiming Zhang, Nenghai Yu, Jiaqi Wang, Feng Wu, Dahua Lin. ScaleCap : Scalable Image Captioning via Dual-Modality Debiasing. Arxiv preprint 2506.19848, 2025. (*Long Xing and Qidong Huang contribute equally.)
- ⋆ **Qidong Huang**, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Dahua Lin, Weiming Zhang, Nenghai Yu. Deciphering Cross-Modal Alignment in Large Vision-Language Models with Modality Integration Rate. *IEEE/CVF International Conference on Computer Vision (**ICCV**), 2025.*
- ⋆ **Qidong Huang**, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, Nenghai Yu. OPERA : Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2024. (Highlight, 2.8% of submissions)*
- ⋆ **Qidong Huang**, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, Kui Zhang, Gang Hua, Nenghai Yu. PointCAT : Contrastive Adversarial Training for Robust Point Cloud Recognition. *IEEE Transactions on Image Processing (**TIP**), 2024.*
- ⋆ **Qidong Huang**, Xiaoyi Dong, Dongdong Chen, Yinpeng Chen, Lu Yuan, Gang Hua, Weiming Zhang, Nenghai Yu. Improving Adversarial Robustness of Masked Autoencoders via Test-time Frequency-domain Prompting. *IEEE/CVF International Conference on Computer Vision (**ICCV**), 2023.*
- ⋆ **Qidong Huang**, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, Nenghai Yu. Diversity-Aware Meta Visual Prompting. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2023.*
- ⋆ **Qidong Huang**, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, Nenghai Yu. Shape-invariant 3D Adversarial Point Clouds. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2022.*
- ⋆ **Qidong Huang***, Jie Zhang*, Wenbo Zhou, Weiming Zhang, Nenghai Yu. Initiative Defense against Facial Manipulation. *AAAI Conference on Artificial Intelligence (**AAAI**), 2021.* (*Qidong Huang and Jie Zhang contribute equally.)

Publications (Collaborate)

- ⋆ Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, **Qidong Huang**, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, Ke Zhu. Qwen3-VL Technical Report. ArXiv preprint 2511.21631, 2025.
- ⋆ Qwen Team, Alibaba Group. Qwen3-Omni Technical Report. ArXiv preprint 2509.17765, 2025.
- ⋆ Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, **Qidong Huang**, Jiaqi Wang, Feng Wu, Dahua Lin. CapRL : Stimulating Dense Image Caption Capabilities via Reinforcement Learning. Arxiv preprint 2509.22647 (**Under Review**), 2025.
- ⋆ Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, **Qidong Huang**, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, Yutong Xie, Imran Razzak, Zongyuan Ge, Jionglong Su, Junjun He, Yu Qiao. MMRC : A Large-Scale Benchmark for Understanding Multimodal Large Language Model in Real-World Conversation. The 63rd Annual Meeting of the Association for Computational Linguistics (**ACL**), 2025.
- ⋆ Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, **Qidong Huang**, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Anyi Rao, Jiaqi Wang, Li Niu. Light-A-Video : Training-free Video Relighting via Progressive Light Fusion. *IEEE/CVF International Conference on Computer Vision (**ICCV**), 2025.*
- ⋆ Long Xing, **Qidong Huang**, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, Dahua Lin. PyramidDrop : Accelerating Your Large Vision-Language Models via Pyramid Visual Redundancy Reduction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2025.*
- ⋆ Likai Liang, **Qidong Huang**, Weiming Zhang, Wenying Zhang. RDPI : Defending against Multi-Turn Jailbreak Attacks via Response-Based Dynamic Prompt Inference. (**Under Review**), 2024.
- ⋆ Feifei Wang, Zhentao Tan, Tianyi Wei, Yue Wu, **Qidong Huang**†. SimAC : A Simple Anti-Customization Method against Text-to-Image Synthesis of Diffusion Models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2024. († Corresponding author)*
- ⋆ Kui Zhang, Hang Zhou, Jie Zhang, **Qidong Huang**, Weiming Zhang, Nenghai Yu. Ada3Diff : Defending against 3D Adversarial Point Clouds via Adaptive Diffusion. *ACM International Conference on Multimedia (**MM**), 2023*
- ⋆ Jie Zhang, Dongdong Chen, **Qidong Huang**, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, Nenghai Yu. Poison ink : Robust and invisible backdoor attack. *IEEE Transactions on Image Processing (**TIP**), 2022.*
- ⋆ Han Fang, Dongdong Chen, **Qidong Huang**, Jie Zhang, Zehua Ma, Weiming Zhang and Nenghai Yu. Deep Template-based Watermarking. *IEEE Transactions on Circuits and Systems for Video Technology (**TCSVT**), 2020.*

Services

- ⋆ Reviewer for CVPR 2022-2026
- ⋆ Reviewer for ICCV 2023-2025
- ⋆ Reviewer for ECCV 2022-2024

- ⋆ Reviewer for ACL 2025
- ⋆ Reviewer for ICML 2025
- ⋆ Reviewer for ICLR 2024-2025
- ⋆ Reviewer for NeurIPS 2024-2025
- ⋆ Reviewer for AAAI 2025
- ⋆ Reviewer for IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
- ⋆ Reviewer for IEEE Transactions on Neural Networks and Learning Systems (TNNLS)
- ⋆ Reviewer for IEEE Transactions on Image Processing (TIP)
- ⋆ Reviewer for Pattern Recognition (PR)

Talk

2025 Toward Efficient & Effective Multi-Modal LLMs. Shanghai Innovation Institute.

2024 Exploring MLLM's Hallucination from A Causal Attention Perspective. AI SPOT, OpenMMLab.

Awards & Honors

2025 Chinese Academy of Sciences (CAS) President Award

2025 USTC Outstanding PhD Thesis Nomination Award

2025 USTC Outstanding Graduates

2024 China National Scholarship

2021 China National Scholarship

2023 "Internet +" Innovation and Entrepreneurship Competition, Provincial Bronze Award

2023 Anheng Information Scholarship

2015 National High School Mathematics League Provincial First Prize

2014 National High School Mathematics League Provincial First Prize