# Computer Vision and Image Understanding
# Special Issue Proposal on Trustworthy Cross-Modal Reasoning for Video-Language Understanding

**Guest Editors**

Dr. **Dan Guo**, Professor, Hefei University of Technology, China, guodan@hfut.edu.cn

Dr. **Zhun Zhong**, Assistant Professor, University of Trento, Italy, zhunzhong007@gmail.com

Dr. **Subhankar Roy**, Postdoctoral Researcher, Télécom Paris, France, subhankar.roy@telecom-paris.fr

Dr. **Linchao Zhu**, Professor, Zhejiang University, China, linchao.zhu@uts.edu.au

Dr. **Chuang Gan**, 1) Assistant Professor, UMass Amherst, and 2) Researcher, MIT-IBM Watson AI Lab, America, chuangg@cics.umass.edu

Dr. **Meng Wang**, Professor, Hefei University of Technology, China, wangmeng@hfut.edu.cn

## 1. Scope and Motivation

In past decades, we have witnessed the rapid development of smartphone cameras, device storage, and 5G networks, which facilitate user-generated video creation and daily content sharing on diverse topics, such as travel, sports, and music. Due to the explosive growth of user-generated video data on the internet together with the urgent requirement of a joint understanding of videos and languages, cross-modal analysis and reasoning has become an active research field and attracted a huge amount of research attention from the CV, NLP, and Multimedia communities in recent years. With the vast success of deep CNNs and transformers, the visual perception of image content has been significantly boosted, sometimes even surpassing humans. However, existing techniques of image-oriented cross-modal analysis cannot process the video-language understanding tasks well, e.g., dense video captioning, text-based video moment localization, and video question answering, due to the complex temporal characteristics of videos and the challenging video-language semantic alignment. Video-language understanding and reasoning are long-standing problems for the CV and Multimedia community. By endowing an AI machine with the cross-modality reasoning ability for video-language understanding, AI researchers expect the machine to “think” like a human and then make trustable decisions. That is the reason why cross-modal is so important and why it can attract world-wide research interest.

Although considerable improvements have been made in the research on video-oriented cross-modal reasoning, it is still in its early stages and requires deeper exploration by the community. Videos offer the promise of understanding not only what can be discerned from a single image (e.g., scenes, people, and objects) but also multi-frame event temporality, causality, and dynamics. The video-oriented cross-modal reasoning should reach a deeper understanding of complex (temporal, causal) events in the multimodal video-language context. Most existing efforts primarily aim to improve in-domain performance while overlooking how to truly capture the essence of cross-modal reasoning. A recent study [1] has pointed out that performance-driven learning modes are easily susceptible to spurious

correlations hidden in datasets and thus usually yield accurate but unreliable in-domain results. Despite the rich multimodal cues (e.g., appearance, motion, action, relations, audio, linguistics, and events) provided by the video-language context, it is still very hard to effectively perform cross-modal reasoning over multimodal cues for trustworthy and comprehensive video-language understanding. Especially the fundamental question in video-language understanding (What makes a video task uniquely suited for videos, beyond what can be understood from a single image?) is usually overlooked by researchers and has yet to be well answered.

Recently, trustworthy cross-modal reasoning arouses serval emerging/new research trends, mainly including: 1) **Defense methods against adversarial attacks for robust cross-modal reasoning** [2-5]. Cross-modal semantics-consistency is essential for video-language systems. To resist the noise within visual, textual, and audio data in the video-language context, defense methods adopt adversarial attack-defense mechanisms to pursue noise-agnostic cross-modal reasoning. Popular studies for adversarial learning include discovering and harnessing adversarial attacks, synthesizing adversarial noise, and instructive adversarial learning such as clean-label targeted attacks, etc. 2) **Causality-inspired domain generalization methods for fair video-language understanding** [6-10]. In the video-language field, fairness is always affected by various data biases, such as language, visual, and individual biases. The primary idea of domain generalization is to identify stable features or mechanisms that remain invariant across the different distributions. Many generalization approaches employ causal theories to describe the invariance since causality and invariance are inextricably intertwined at high-level semantic reasoning. Therefore, causality-aware domain generalization is employed to improve the generalizability of intelligent models via causal data augmentation, causal representation learning, and transferring causal mechanisms. 3**) Explainable cross-modal reasoning via knowledge-driven techniques** [11-15]. With the advancement of intelligent applications, more complex and difficult tasks have paid attention to explainable cross-modal reasoning. Some approaches are proposed to address cross-modal reasoning based on implicit and explicit knowledge acquisitions, multi-modal knowledge representation, multi-source knowledge fusion, and knowledge-data dual-driven inference, etc. Researchers still attempt to develop various novel knowledge-inspired deep networks, which are beneficial to improve the effectiveness and interpretability of video language reasoning. **4) Trustworthy language-to-video generation under privacy-preserving and security-controlled prerequisites [16-20].** Recently, the rapid development of big models has driven the research and application upsurge of AIGC (Artificial Intelligence Generated Content). Developing high-quality videos through novel language-to-video and video-to-video generation models has been significant for various video applications, such as video editing, virtual reality, and multimedia content creation. However, AIGC may bring privacy leakage and uncontrollable security issues, so it is necessary and valuable to explore trustworthy generation solutions for video-language application scenarios.

Therefore, a special issue on “Trustworthy Cross-modal Reasoning for Video-Language Understanding” is urgently required to track the continual growth of research, primarily related to the robustness, fairness, explainability, and security of video-oriented cross-modal

reasoning. This special challenge aims to bring together researchers interested in new and innovative solutions that will advance research on trustworthy video-language understanding and domain-specific applications. If approved, we believe that this special issue will largely impact researchers and practitioners working in related areas across academia and industry.

**Reference:**


[1] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. CVPR , pp. 2917-2927. 2022.

[2] Madeline Schiappa, Shruti Vyas, Hamid Palangi, Yogesh Rawat, Vibhav Vineet. Robustness Analysis of Video-Language Models Against Visual and Language Perturbations. NeurIPS, 2022.

[3] Jianping Zhang, Yizhan Huang, Weibin Wu, Michael Lyu. Transferable Adversarial Attacks on Vision Transformers with Token Gradient Regularization. CVPR, 2023.

[4] Jiaming Zhang, Qi Yi, Jitao Sang. Towards Adversarial Attack on Vision-Language Pre-training Models. ACM MM 2022.

[5] Yawen Zeng, Da Cao, Shaofei Lu, Hanling Zhang, Jiao Xu, Zheng Qin. Moment is Important: Language-Based Video Moment Retrieval via Adversarial Learning. ACM Transactions on Multimedia Computing, Communications, and Applications, 2022.

[6] Wei Wang, Junyu Gao, Changsheng Xu. Weakly-Supervised Video Object Grounding via Causal Intervention. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[7] Xin Wang, Xiaohan Lan, Wenwu Zhu. Video Grounding and Its Generalization, ACM MM, 2022.

[8] Wuyang Li, Xinyu Liu, Xiwen Yao, Yixuan Yuan. SCAN: Cross Domain Object Detection with Semantic Conditioned Adaptation. AAAI, 2022.

[9] Jiangmeng Li, Yanan Zhang, Wenwen Qiang, Lingyu Si, Chengbo Jiao, Xiaohui Hu, Changwen Zheng, Fuchun Sun. Object Detection from A Conditional Causal Perspective. AAAI, 2023.

[10] Pan Deng, Yu Zhao, Junting liu, Jia Xiaofeng, Mulan Wang. Spatio-temporal Neural Structural Causal Models for Bike Flow Prediction. AAAI, 2023.

[11] Zhenwei Shao, Zhou Yu, Meng Wang, Jun Yu. Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering. CVPR, 2023.

[12] Zihui Xue, Sucheng Ren, Zhengqi Gao, Hang Zhao. Multimodal Knowledge Expansion. ICCV, 2021.

[13] Hantao Yao, Rui Zhang, Changsheng Xu. Visual-Language Prompt Tuning with Knowledge-guided Context Optimization. CVPR, 2023.

[14] Jianguo Mao, Wenbin Jiang, Hong Liu, Xiangdong Wang, Yajuan Lyu. Inferential Knowledge-Enhanced Integrated Reasoning for Video Question Answering. AAAI, 2023.

[15] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, Qi Wu. MuKEA: Multimodal Knowledge Extraction and Accumulation for Knowledge-Based Visual Question Answering. CVPR 2022.

[16] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, Jie Tang. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. ICLR, 2023.
[17] Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, Elisa Ricci. Playable Video Generation. CVPR, 2021.
[18] Ivan Skorokhodov, Sergey Tulyakov, Mohamed Elhoseiny. StyleGAN-V: A Continuous Video Generator with the Price, Image Quality and Perks of StyleGAN2. CVPR, 2022.
[19] Minsoo Kang, Doyup Lee, Jiseob Kim, Saehoon Kim, Bohyung Han. Variational Distribution Learning for Unsupervised Text-to-Image Generation. CVPR, 2023.
[20] Yaosi Hu, Chong Luo, Zhenzhong Chen. Make It Move: Controllable Image-to-Video Generation with Text Descriptions. CVPR, 2022.

**2. Topics of the Special Issue**

The purpose of this special issue is to solicit high-quality, high-impact, and original papers on current developments in cross-modal reasoning for video-language understanding. We are interested in submissions covering topics of particular interest that include but are not limited to the following:

- New datasets for trustworthy video-language understanding
- Adversarial learning for robust multimodal representation
- New methods for robust video summarization
- Cross-modal semantics-consistent representation learning
- Domain generalization in video-language understanding
- Causal learning for trustworthy multimodal reasoning
- Unfair bias measurement and mitigation in video-language understanding
- Explainable multimodal data fusion and interaction
- Brain-inspired networks for explainable cross-modal reasoning
- Trustworthy reasoning algorithm in video dialog
- Knowledge-driven explainable cross-modal reasoning
- Text-guided visual-textual reasoning and generation
- Privacy protection and security control in cross-modal AIGC
- Applications of trustworthy video-language understanding

**3. Rationales**

**1) How is the special issue related to the Computer Vision and Image Understanding journal?** The central focus of Computer Vision and Image Understanding (CVIU) journal is the computer analysis of pictorial information. CVIU publishes papers covering all aspects of image analysis from the low-level, iconic processes of early vision to the high-level, symbolic processes of recognition and interpretation. A wide range of topics in the image understanding area is covered, including papers offering insights that differ from predominant views. Trustworthy cross-modal reasoning for video-language understanding involves different forms of multimedia data, such as vision, language, and audio. The special issue specifically focuses on the trustworthy representation, analysis, and interaction of these data, which is an area of increasing importance and relevance in the field of image

understanding. The topics of our special issue fit very well with the scope and aim of the CVIU journal.

**2) Why is the topic of the special issue important?** Due to the rapid growth of deep learning technologies, considerable improvement has been made in video-language understanding. More recently, there has been an emerging research trend toward studying cross-modal reasoning, which aims to improve the fine-grained representation of heterogeneous data. It is critically important for a multimodal system, especially in some specific domains, such as healthcare services, fintech, and self-driving cars, therefore gaining increasing research interest from multiple communities. Therefore, the topics of our proposed special issue are highly important.

**3) Why the special issue may attract a significant number of submissions?** Multimodal reasoning is an emerging research area and has gained intensive attention from researchers in both academia and industry. Recently, a large body of work has been proposed to study video-language understanding and reasoning in top conferences (*e.g.*, MM, CVPR, ICCV, ECCV, NeurIPS, ICLR). With this special issue, we would like to showcase the significant progress that has been made within the video-language community at large over the past years.

**4. Strategy for the Paper Recruitment**

1) **(Academic Partners)** We have a diverse team of Guest Editors from China, Singapore, France, and Italy. Prior to writing the proposal, we contacted researchers from a wide range of institutes, and they all expressed interest in this special issue. These institutes include but are not limited to: Tsinghua University (China), National University of Singapore (Singapore), University of Technology Sydney (Australia), Hefei University of Technology (China), Zhengzhou University (China), Peking University (China), Monash University (Australia), University of Science and Technology of China (China), Xidian University (China), and University of North Carolina at Chapel Hill (United States).

2) (**Social Media**) Moreover, we will publicize this special issue through various venues, such as Twitter, LinkedIn, Reddit, and working connections, to attract submissions from the related research communities.

3) (**Conference**) We also have contacted the organizing committees of several international multimedia/computer vision conferences to invite the top-ranked and topic-related conference papers (extension version) for submission in our proposed SI. The conferences include but are not limited to: International Conference on Multimedia and Expo (ICME 2023, Brisbane, Australia), IEEE Multimedia Big Data (BigMM 2023, Laguna Hills, Canada),

Chinese Conference on Pattern Recognition and Computer Vision (PRCV 2023, Xiamen, China), ACM Multimedia (MM 2023, Ottawa, Canada)

Due to the reputation of the editor team and diverse strategies of paper recruitment, we are confident of attracting a batch of high-quality submissions (e.g., greater than 30) and will accept no more than 10 manuscripts for publication. We have contacted several reputed researchers in the computer vision and multimedia community for contributing an excellent survey on the SI's topic.

**5. Review Process**

The review process will comply with the standard review process of the Computer Vision and Image Understanding journal. Each paper will receive at least three reviews from experts in the field.

**6. Important Dates**

- Submission deadline: December 15, 2023
- First-round decision notification: March 15, 2024
- Revised manuscript due: May 15, 2024
- Final decision notification: July 30, 2024
- Camera-ready version: November 30, 2024

**7. Relationship to Related Special Issues**

In recent years, cross-modal analysis and reasoning has become an active research field. Many special issues are proposed to seek original contributions towards multimedia content understanding [1-4]. To address rapidly growing interest in artificial intelligence (AI) for multimedia processing, some special issues focused on making AI models transparent, interpretable, and accountable [1-3].


[1] Special Issue on Trustworthy Multimedia Computing and Applications in Urban Scenes, ACM TOMM, 2022. (https://dl.acm.org/pb-assets/static_journal_pages/tomm/pdf/TOMM_cfp_SI_TrustworthyMultimedia-1643357525513.pdf)

[2] Special Issue on Trustworthy Multimedia Big Data Computing for Next-Generation Multimedia Systems, Computing and Information Technology, 2023. (http://cit.fer.hr/index.php/CIT/announcement/view/18)

[3] Special Series on AI in Signal & Data Science - Toward Explainable, Reliable, and Sustainable Machine Learning, IEEE JSTSP, 2023. (https://signalprocessingsociety.org/blog/ieee-jstsp-special-series-ai-signal-data-science-toward-explainable-reliable-and-sustainable)

[4] Special Issue on Pre-trained Models for Multi-modality Understanding, IEEE TMM, 2023.

(https://signalprocessingsociety.org/blog/ieee-tmm-special-issue-pre-trained-models-multi-modality-understanding）

The above special issues reflect the emerging interest of the community in trustworthy multimedia computing, explainable machine learning, and pre-trained models for multi-modality understanding. They focused on technical advances or applications in some specific domains, such as trustworthy algorithms for city-scale human and vehicle analysis, privacy-enhanced computation for big data, and energy-efficient machine learning models. There has not been a special issue that specifically addresses the challenge of trustworthy cross-modal reasoning for video-language understanding, which requires novel techniques for understanding, thinking, and reasoning the complex relationships between different modalities like humans, and ensuring that the reasoning process is transparent and interpretable. The proposed special issue aims to fill this gap by bringing together researchers to explore new approaches for cross-modal reasoning that can improve the robustness, fairness, explainability, and security of video-language understanding systems. Compared to the related special issues, we emphasize the importance of trustworthy reasoning and its implications for real-world applications such as video question answering, video dialog, and video summarization. Overall, this proposed special issue is new in its focus on trustworthy cross-modal reasoning and its potential impact on advancing video-language understanding research.

**Biography of Guest Editors**

**Dan Guo** (https://vut-hfut.github.io/ ) is a Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, China. She received the BE degree in computer science and technology from Yangtze University, China, in 2004 and the PhD degree in system analysis and integration from the Huazhong University of Science and Technology, China, in 2010. She is a professor at the School of Computer and Information, Hefei University of Technology. Her research interests include computer vision, machine learning, and intelligent multimedia content analysis.

[1] **Dan Guo**, Hui Wang, Meng Wang. "Context-aware graph inference with knowledge distillation for visual dialog." *IEEE TPAMI*, 2021.
[2] Jinxing Zhou, **Dan Guo**, Meng Wang. "Contrastive positive sample propagation along the audio-visual event line." *IEEE TPAMI*, 2022.
[3] **Dan Guo**, Hui Wang, Hanwang Zhang, Zheng-Jun Zha, Meng Wang. "Iterative context-aware graph inference for visual dialog". *CVPR*, 2020.
[4] **Dan Guo,** Wengang Zhou, Anyang Li, Houqiang Li, Meng Wang. "Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation." *IEEE TIP*, 2019.
[5] Kun Li, **Dan Guo**, Meng Wang. "Proposal-free video grounding with contextual pyramid network." *AAAI*, 2021.

**Zhun Zhong** (https://zhunzhong.site ) received his Ph.D. degree in 2019 from Xiamen University. He is now an assistant professor at the University of Trento and was a postdoc at

the same place. He is committed to designing robust and scalable visual recognition systems for real-world applications. He has published more than 30 peer-reviewed papers in top conferences and journals. He was an area chair or a senior program committee in several top conferences, e.g., ACM MM, AAAI, and IJCAI. He received the Outstanding Reviewer Award at CVPR 2020 and NeurIPS 2021. He was a guest editor of the International Journal of Computer Vision and the International Journal of Applied Earth Observation and Geoinformation and Electronics. He is now an Associate Editor of Image and Visual Computing. He was selected as the AI 2000 Most Influential Scholar Honorable Mention in AAAI/IJCAI in 2021 and 2022.

[1] Nan Pu, **Zhun Zhong**, Nicu Sebe. "Dynamic Conceptional Contrastive Learning for Generalized Category Discovery."*CVPR*, 2023.
[2] **Zhun Zhong**, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, Nicu Sebe. "Openmix: Reviving Known Knowledge for Discovering Novel Visual Categories in an Open World." *CVPR*, 2021.
[3] Yuyang Zhao, **Zhun Zhong,** Na Zhao, Nicu Sebe, Gim Hee Lee. "Style-Hallucinated Dual Consistency Learning for Domain Generalized Semantic Segmentation." *ECCV*. 2022.
[4] Wei Wang, **Zhun Zhong**, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, Nicu Sebe. "Dynamically Instance-Guided Adaptation: A Backward-Free Approach for Test-Time Domain Adaptive Semantic Segmentation." CVPR, 2023.
[5] **Zhun Zhong**, Yuyang Zhao, Gim Hee Lee, Nicu Sebe. "Adversarial Style Augmentation for Domain Generalized Urban-Scene Segmentation." *NeurIPS*, 2022.

**Subhankar Roy** (https://roysubhankar.github.io/ ) is a postdoctoral researcher at Telecom Paris, France. He received his Ph.D in the year 2022 from the University of Trento, Italy, where he was supervised by Prof. Elisa Ricci and Prof. Nicu Sebe. His primary areas of research involve adapting neural networks to data distribution and semantic shifts, especially while working with limited, weak or no supervision. In particular, he works in domain adaptation, model adaptation, continual learning and open-world recognition applied to image classification, action recognition and semantic segmentation. He frequently publishes in top-tier computer vision conferences and is an active member of the program committees of such conferences.

[1] Giacomo Zara, **Subhankar Roy**, Paolo Rota, Elisa Ricci. "AutoLabel: CLIP-based framework for Open-set Video Domain Adaptation." *CVPR*, 2023.
[2] Tianyu Li, **Subhankar Roy**, Huayi Zhou, Hongtao Lu, Stéphane Lathuilière. "Contrast, Stylize and Adapt: Unsupervised Contrastive Learning Framework for Domain Adaptive Semantic Segmentation."*CVPR*, 2023.
[3] **Subhankar Roy**, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, Arno Solin. "Uncertainty-guided source-free domain adaptation." *ECCV*, 2022.
[4] **Subhankar Roy**, Evgeny Krivosheev, Zhun Zhong, Nicu Sebe, Elisa Ricci. "Curriculum graph co-teaching for multi-target domain adaptation." *CVPR*, 2021.
[5] Yasser Benigmim, **Subhankar Roy**, Slim Essid, Vicky Kalogeiton, Stéphane Lathuilière. "One-shot Unsupervised Domain Adaptation with Personalized Diffusion Models."*CVPR*. 2023.

**Linchao Zhu** (http://ffmpbgrnn.github.io/) is a Professor with the College of Computer Science at Zhejiang University, China. He received the BE degree from Zhejiang University, China, in 2015, and the Ph.D. degree in computer science from the University of Technology Sydney, Australia, in 2019. His research focus includes video representation learning, few-shot learning, transfer learning and self-supervised learning. So far, he has published more than 60 papers in top-tier scientific conferences such as AAAI, IJCAI, NIPS, CVPR, ICCV, and IEEE journals such as IEEE TPAMI, IEEE TIP, IEEE TMM.

[1] Yaowei Li, Ruijie Quan, **Linchao Zhu**, Yi Yang. "Efficient Multimodal Fusion via Interactive Prompting." *CVPR*, 2023.
[2] Shannan Guan, Haiyan Lu, **Linchao Zhu**, Gengfa Fang." PoseGU: 3D human pose estimation with novel human pose generator and unbiased learning." *CVIU*, 2023.
[3] Yang Jin, **Linchao Zhu**, Yadong Mu. "Complex Video Action Reasoning via Learnable Markov Logic Network." *CVPR*, 2022.
[4] Juncheng Li, Siliang Tang, **Linchao Zhu**, Wenqiao Zhang, Yi Yang, Tat-Seng Chua, Fei Wu. "Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding." *IEEE TPAMI*, 2023.
[5] Difei Gao, Luowei Zhou, Lei Ji, **Linchao Zhu**, Yi Yang, Mike Zheng Shou. "MIST: Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering." *CVPR*, 2023.
[6] Juncheng Li, Xin He, Longhui Wei, Long Qian, **Linchao Zhu**, Lingxi Xie, Yueting Zhuang, Qi Tian, Siliang Tang. “Fine-Grained Semantically Aligned Vision-Language Pre-Training”. NeurIPS 2022

**Chuang Gan** (https://people.csail.mit.edu/ganchuang/) is an Assistant Professor at UMass Amherst and a researcher at MIT-IBM Watson AI Lab, American. He received the Ph.D. at Tsinghua University. His research lies at the intersection of computer vision, AI, cognitive science, and robotics. The overarching goal of his research is to build a human-like common sense machine that is capable of sensing, reasoning, and acting in the physical world. His works have been recognized by Microsoft Fellowship, Baidu Fellowship, and media coverage from CNN, BBC, The New York Times, WIRED, Forbes, and MIT Tech Review. He has served as an Area Chair for ICLR 2023, CVPR 2023, NeurIPS 2023, ICML 2023, ICCV 2023, and ECCV 2022.

[1] Kexin Yi, Jiajun Wu, **Chuang Gan**, Antonio Torralba, Pushmeet Kohli, Joshua B Tenenbaum. "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding." NeurIPS, 2018.
[2] Kun Su, Kaizhi Qian, Eli Shlizerman, Antonio Torralba, **Chuang Gan**. "Physics-Driven Diffusion Models for Impact Sound Synthesis from Videos." CVPR, 2023.
[3] Aisha Urooj, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Bousselham, **Chuang Gan**, Niels Lobo, Mubarak Shah. "Learning Situation Hyper-Graphs for Video Question Answering." CVPR, 2023.

[4] Peihao Chen, Dongyu Ji, Kunyang Lin, Runhao Zeng, Thomas Li, Mingkui Tan, **Chuang Gan**. "Weakly-supervised multi-granularity map learning for vision-and-language navigation." NeurIPS, 2022.
[5] Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H Li, Mingkui Tan, **Chuang Gan**. "Masked Motion Encoding for Self-Supervised Video Representation Learning." CVPR, 2023.

**Meng Wang** (https://sites.google.com/view/meng-wang/home) is a Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, China. He is a Fellow of IEEE and IAPR. He received his B.E. degree and Ph.D. degree in the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science from the University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2008, respectively. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He has authored more than 200 book chapters, journal and conference papers in these areas. He is the recipient of the ACM SIGMM Rising Star Award 2014. He has extensive editorial experience, including serving as an associate editor of IEEE TKDE, IEEE TCSVT, IEEE TMM, and IEEE TNNLS. He was the General Co-Chair of ICMR 2021, PCM 2018 and MMM 2013, and the Program Co-Chair of ICIMCS 2013.

[1] Tianyu Chang, Xun Yang, Tianzhu Zhang, **Meng Wang**. "Domain Generalized Stereo Matching via Hierarchical Visual Transformation." CVPR, 2023.
[2] Xun Yang, Fuli Feng, Wei Ji, **Meng Wang**, Tat-Seng Chua. "Deconfounded video moment retrieval with causal intervention." SIGIR, 2021.
[3] Jinxing Zhou, Dan Guo, **Meng Wang**. “Contrastive Positive Sample Propagation along the Audio-Visual Event Line”. IEEE TPAMI 2022
[4] Lianli Gao, Yu Lei, Pengpeng Zeng, Jingkuan Song, **Meng Wang**, Heng Tao Shen. "Hierarchical representation network with auxiliary tasks for video captioning and video question answering." IEEE TIP, 2021.
[5] Zhenguang Liu, Pengxiang Su, Shuang Wu, Xuanjing Shen, Haipeng Chen, Yanbin Hao, **Meng Wang**. "Motion Prediction Using Trajectory Cues". ICCV, 2021.