

# CVPR 2024 速報

---

森江梨花, Qiu Yue, Yanjun Sun, 山田亮佑, 大谷豪, 松尾雄斗, 児玉憲武,  
中原龍一, 柳凜太郎, 中村凌, 篠原崇之, Hao Guoqing, 和田唯我, 柴田優斗,  
大久保蓮, 篠田理沙, 柴田直生, 木林佑太, 今井悠人, 上田樹, 阿部純, 舘野将寿, 八木  
拓真, 和田唯臣, 石井湧太, 石川竜之介, 桐生一輝, 鳥見晃平, 西村和也, 元田智大, 園  
山昌司, 古川毅流, 金子知紘, 井手康允,  
大塚大地, 速水亮, 井口悠司, 前伸一, 片岡裕雄

## CVPR 2024 の動向・気付き

---

- 今回どんな研究が流行っていた？
- 海外の研究者は何をしている？
- 「動向」や「気付き」をまとめました

## CVPRの大規模化(1/3)

- 投稿数の増加(9,155→11,532)に伴ってArea Chairや査読者が増加
  - Area Chair: 400+ → 477
  - 査読者: 6,625 → 9,872

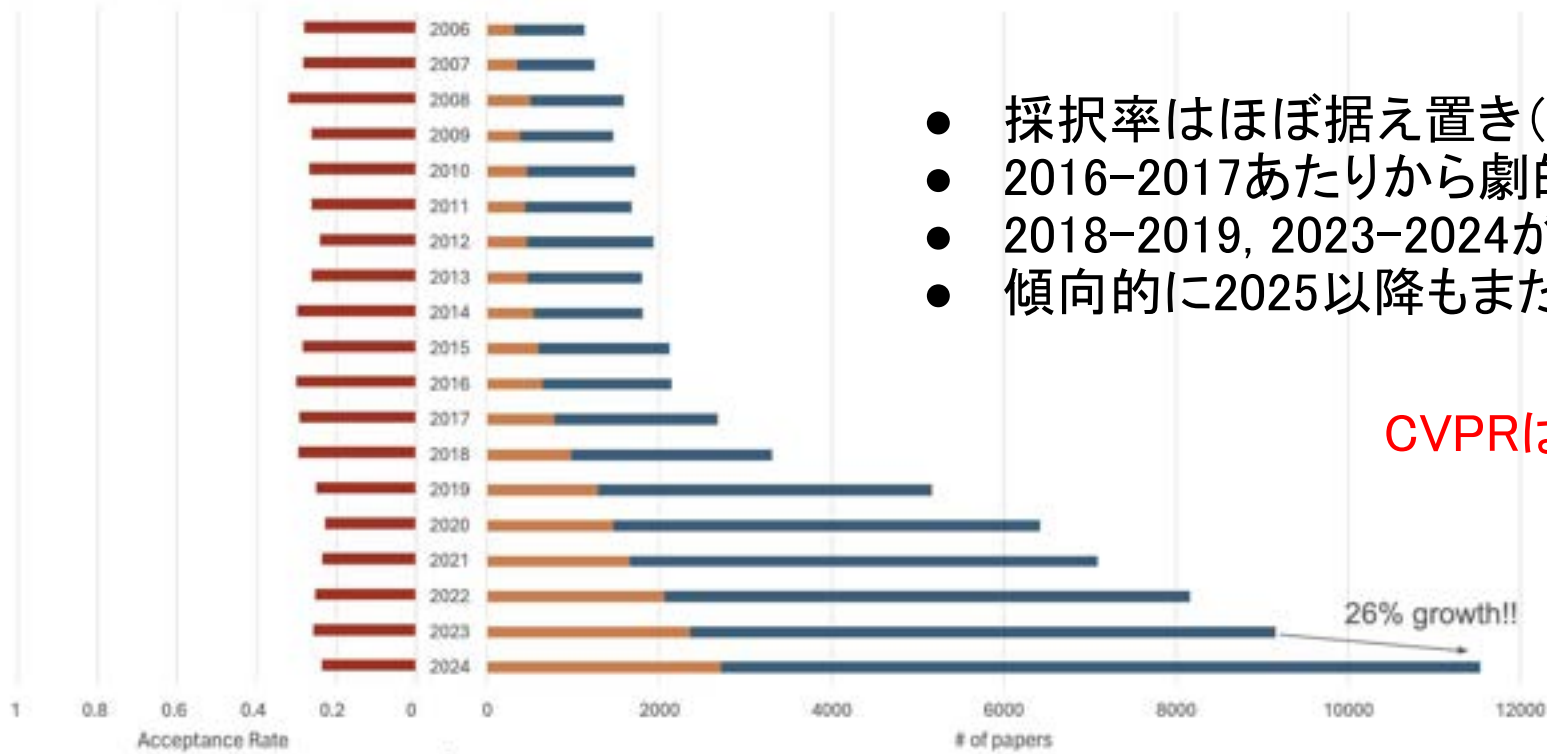
→ Area Chairや査読者の確保・査読の質の担保が今後の課題
- Workshopも123件採択
  - 似たようなテーマのワークショップが存在
  - 投稿先が割れたのか、proceedingsが1件というWorkshopも...
  - LIMITもICCV'23の約1/3程度の投稿数に → 招待ポスターを追加
  - 著名な研究者が何件も発表を抱えている・講演内容の重複
    - 主催者はプログラムの考案で頭を抱えそう...
    - 投稿者は日程や内容によって投稿するWorkshopを選べるという点ではプラス?
    - Workshopの開催判定に労力を割く余裕があまりないのかも..?

# CVPR 2024 の動向・気付き(2/116)

## CVPRの大規模化(2/3)

- 投稿数の増加(9,155 → 11,532)に伴ってArea Chairや査読者が増加

**New records: 11,532 submissions, 2,719 accepted papers!**



- 採択率はほぼ据え置き(か微減)につき, 論文数単純増加
- 2016-2017あたりから劇的に論文増加傾向
- 2018-2019, 2023-2024が特に投稿数の伸びを記録
- 傾向的に2025以降もまだ頭打ちにはならない

CVPRはどこまで行くのか？

## CVPRの大規模化 (3/3)

- 投稿数の増加に伴ってAwardsも増加 (5件 → 10件)
  - 次点まで発表していたのが昨年との違い
    - 投稿数が多いからか甲乙付け難かった？
  - Best Paperの候補論文数も倍になっている (12件 → 24件)
    - 増やしすぎるとAwardsのプレミア感は薄れそう...?

## 日本人とCVPR (1/3)

- 日本のプレゼンスが低下し続けている
  - 日本からの絶対的な採択数は増加(2023: 60 → 2024: 66)
  - 一方で日本人著者比率(2023: 0.92% → 2024: 0.85%)や日本人著者が絡む論文比率(2023: 2.55% → 2024: 2.43%)は減少傾向
  - つまり、世界の成長速度に日本が若干フォローできていないということになる(もちろん、日本だけではなくが勢いを出したい)

次2ページに図表を引用

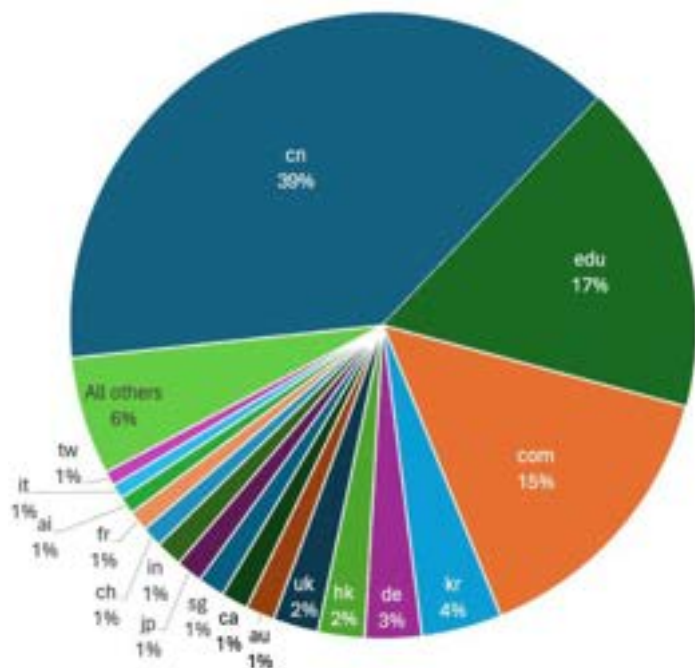
# CVPR 2024 の動向・気付き (5/116)

## 日本人とCVPR (2/3)

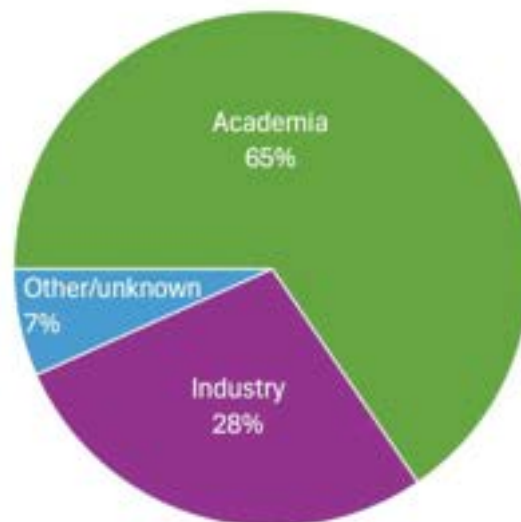
**35,691 registered authors!**

- Opening Remarksにおける“Authors by top-level email domain”で確認すると(恐らく)jp という部分が日本からの著者比率になっていて、1%と表示されている(次ページの0.85%という数字が対応?)
- データからは Academia + Student からの投稿が多い
- Student(53%) vs. Non-student(47%) はほぼ同等

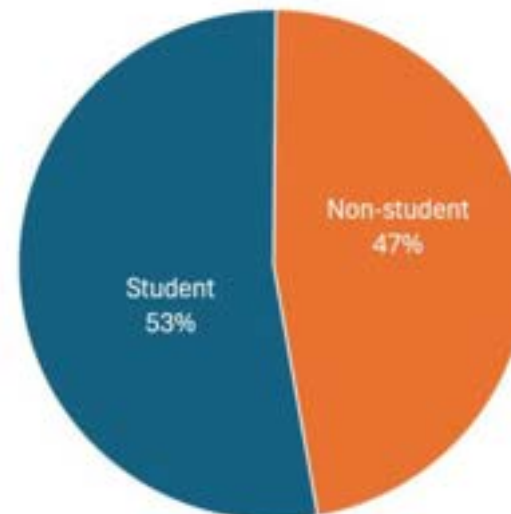
AUTHORS BY TOP-LEVEL EMAIL DOMAIN



AUTHORS BY INSTITUTION TYPE



AUTHORS BY STUDENT STATUS



引用: <https://media.eventhosts.cc/Conferences/CVPR2024/OpeningRemarkSlides.pdf>

# CVPR 2024 の動向・気付き(6/116)

## 日本人とCVPR (3/3)

開催年	論文数	著者数	平均著者数	日本人著者数	日本人著者比率	日本人著者を含む論文	日本人著者が絡む論文比率
2014	540	1,881	3.48	45	2.39%	26	4.81%
2015	602	2,207	3.67	33	1.50%	20	3.32%
2016	643	2,487	3.87	45	1.81%	21	3.27%
2017	783	3,185	4.07	61	1.92%	29	3.70%
2018	979	4,214	4.30	93	2.21%	38	3.88%
2019	1,294	5,863	4.53	86	1.47%	40	3.09%
2020	1,466	6,970	4.75	65	0.93%	38	2.59%
2021	1,660	8,087	4.87	72	0.89%	42	2.53%
2022	2,063	10,874	5.27	108	0.99%	52	2.52%
2023	2,357	12,722	5.40	117	0.92%	60	2.55%
2024	2,716	15,288	5.63	130	0.85%	66	2.43%

10年間で日本人の著者比率は3分の1近くに、論文比率は半分近くにまで減少

表2 CVPR投稿論文全体の著者数に占める日本人比率の推移

引用: <https://research-p.com/column/1889>



# CVPR 2024 の動向・気付き(7/116)

## プログラムから見るCVPR

- Oral, Keynote, Panelと「口頭発表」重視のプログラムになっている(2023はポスターによるインタラクティブ重視)

	Wednesday	Thursday	Friday
8:30 AM	Opening		
9:00 AM	<b>Orals</b> 1A Low-level vision 1B Vision and Graphics 1C Humans: Face, body, pose, gesture, movement	<b>Orals</b> 3A 3D from single view 3B Vision, Language, and Reasoning 3C Medical and Physics-based vision	<b>Orals</b> 5A Datasets and evaluation 5B 3D from multiview and sensors 5C Low-shot, self-supervised, semi-supervised learning
10:30 AM	Posters	Posters	Posters
12:00 PM	Lunch	Lunch	Lunch
1:00 PM	<b>Orals</b> 2A Image & Video Synthesis 2B Deep learning architectures & techniques 2C 3D from multiview and sensors	<b>Orals</b> 4A Autonomous navigation & egocentric vision 4B 3D Vision 4C Action and motion	<b>Orals</b> 6A Low-level vision and remote sensing 6B Image & Video Synthesis 6C Multi-modal learning
2:45 PM	<b>Keynote</b> Joshua Bongard	<b>Keynote</b> David Baker	<b>Keynote</b> Sofia Crespo
4:00 PM	<b>Panel</b> Societal opportunities and challenges of AI	<b>PAMI-TC Meeting</b>	<b>Panel</b> CVPR: Past, Present, and Future
5:15 PM	Posters	Posters	Posters
7:00 PM		Reception and Music	

- 参加者レベルでは「ポスター回るの無理」という声→ 90分で400件のポスターは1件あたり約10秒(ムリ)
- 会場都合でポスターが別の建物、移動に10分前後要する

# CVPR 2024 の動向・気付き(8/116)

## Paper Decisions (1/4)

### □ 選定プロセス

- 各論文3つ以上の査読, ACによりAccept/Rejectを推薦, PCにより最終確認
- 上位90件(3.3%)がOral, 上位324件(11.9%)がHighlight poster

### Paper decisions

- Each paper received 3 reviews and a meta-review from an Area Chair.
- Decisions made within triplets of ACs.
- We decided to bring back orals, but also keep the CVPR 2023 innovation of “highlighted” posters
- Orals and highlight candidates collected from the ACs
- Overall acceptance rate: 23.6%
  - 90 (3.3%) of papers are Orals+posters
  - 324 (11.9%) of papers are “highlights” posters, with special annotation in the program
  - 2,305 additional posters

#### PROGRAM GUIDE

325 SchurVNS: Schur Complement-Based Liq  
Navigation System, Yanfei Fan, Tianyu Zh  
326 READ: Retrieval-Enhanced Asymmetric Di  
Planning, Takeru Oba, Matthew Walter, Ni  
327 Retrieval-Augmented Embodied Agents, I  
Xiaofeng Mou, Jian Tang  
328 Collaborative Semantic Occupancy Predi  
Feature Fusion in Connected Automated V  
Chenwei Liang, Hu Cao, Zhiran Yan, Walte  
Andreas Festag, Alois Knoll  
329 Diffusion-EDFs: Bi-equivariant Denoising I  
SE(3) for Visual Robotic Manipulation, Hy  
Hyunseok An, Junwoo Chang, Joochan Se  
Kim, Chaewon Hwang, Jongeun Choi, Robi  
330 Adaptive VIO: Deep Visual-Inertial Odome  
Continual Learning, Youqi Pan, Wugen Zh  
Hongbin Zha  
331 F3Loc: Fusion and Filtering for Floorplan I  
Changan Chen, Rui Wang, Christoph Vogt  
332 Gaussian Splatting SLAM,  
Hidenobu Matsuki, Riku Murai, Paul H.J. Ke  
333 SUGAR: Pre-training 3D Visual Representati  
Shizhe Chen, Ricardo Garcia, Ivan Laptev, Ct  
334 ManipLLM: Embodied Multimodal Large L  
Object-Centric Robotic Manipulation, Xia  
Yiran Geng, Haoran Geng, Yuxing Long, Ya  
Jiaming Liu, Hao Dong  
335 Open-Vocabulary Object 6D Pose Estimati  
Davide Boscaini, Changjie Oh, Andrea Ca  
336 Hierarchical Diffusion Policy for Kinematic  
Robotic Manipulation, Xiao Ma, Sumit Pat  
Stephen James  
337 Smart Help: Strategic Opponent Modeling  
Adaptive Robot Assistance in Households  
Wang, Siwen Xie, Anji Liu, Lifeng Fan  
338 Generalizing 6-DoF Grasp Detection via D  
Knowledge, Hualiang Ma, Modi Shi, Boye  
339 A Simple and Effective Point-based Netw  
6-DOFs Pose Relocalization, Hongwei Ren  
Zhou, Heohan Fu, Yulong Huang, Bokun Ch  
340 Neural Visibility Field for Uncertainty-Driv  
Mapping, Shangjie Xue, Jesse DLK Franay  
Panagiotis Tsoltra, Danfei Xu  
341 SPIN: Simultaneous Perception Interaction I  
Lippal, Ananya Agarwal, Haoyu Xiong, Kinn

## Paper Decisions (2/4)

### □ Area Chairs

□ やはりたくさんいる...

□ 各人が25本前後の論文を担当している(投稿数をほぼ当分割: 25 x 477)

### Area Chairs (n=477)

Aayush Bansal, Abby Stylianou, Abhinav Shrivastava, Abhishek Gupta, Adam Czejka, Adam Kortylewski, Adriana Kovashka, Aishwarya Agrawal, Akihiro Sugimoto, Alejandro F. Frangi, Alexander C. Berg, Alexander T Tsoh, Alexandre Alahi, Almoor Reza, Alreza Fathi, Aijosa Osep, Ameesh Makadia, Amir Zamir, Andre Araujo, Andrea Tagliasacchi, Andrew Owens, Angel X Chang, Angela Yao, Anh Tuan Tran, Animesh Gang, Anna Rohrbach, Anpei Chen, Anurag Aramb, Ansha Nagrani, Arun Malley, Asako Kanezaki, Ashok Veeraraghavan, Aswin C. Sankaranarayanan, Baoyuan Wu, Bastian Leibe, Ben Mildenhall, Bin Fan, Bing Su, Binh-Son Hua, Björn Ommer, Bo Chen, Bolei Zhou, Boqing Gong, Boxin Shi, Boyi Li, Brendan Tran Morris, Bruce Allen Maxwell, Bryan A. Plummer, Bryan Morse, Burns Ham, Carl Doersch, Carl Vondrick, Chang Xu, Chao Ma, Chao-Yuan Wu, Chaowei Xiao, Charles Fowlkes, Chen Change Loy, Chen Sun, Chen Wang, Chengjiang Long, Chenliang Xu, Chenyou Fan, Chetan Arora, Chong Luo, Christian Rupprecht, Christopher Clark, Christopher Funk, Chuhan Gao, Chuhan Zou, Chunyu Wang, Chang Xie, Cornelia Fermüller, Cristian Canton Ferrer, Culling Lan, Cuong V. Nguyen, Dahun Kim, Dan Xu, Danna Gurari, David B. Lindell, David Fouhey, David Novotny, Deng-Ping Fan, Devis Tuia, Di He, Di Huang, Diane Larlus, Dim Papadopoulos, Dima Damen, Dimosthenis Karatzas, Dinesh Jayaraman, Dingwen Zhang, Dong Chen, Dong Xu, Dongfang Liu, Eddy Ilg, Elisa Ricci, Emily Morgan Hand, Enrique Dunn, Eshed Ohn-Bar, Evan Shelhamer, Evangelos Kalogerakis, Fahad Khan, Farshad Faghri, Fatma Güney, Federico Tombari, Feng Lu, Francesco Moreno-Noguer, Gang Yu, Gedas Bertasius, Georgia Gkioxari, Georgios Pavlakos, Georgios Tzimiropoulos, Gim Hee Lee, Giovanni Maria Farinella, Gordon Wetstein, Guanbin Li, Guansong Pang, Gül Varol, Guosheng Lin, Guoyu Lu, Hadi Pouransari, Hae-Gon Jeon, Hajime Nagahara, Hamid Rezaeifathi, Han-Jia Ye, Hanbyul Joo, Hao Chen, Hao Dong, Henghui Ding, Hengshuang Zhao, Hideki Nakayama, Hilde Kuehne, Hirokatsu Kataoka, Hiroshi Kawasaki, Hong Chang, Hong Xuan, Hongyang Li, Hossein Rahmani, Hsin-Ying Lee, Hu Han, Hualun Jiang, Hui Ji, Huijuan Xu, Huwen Chang, Hung-Yu Tseng, Hyun Soo Park, Hyunung Shim, Hyunwoo J. Kim, Iacopo Masi, Ioannis Gkioulekas, Ishan Misra, Ishani Chakraborty, Jaesik Park, James Hays, Jason J Corso, Jasper Uijlings, Jean-François Lalonde, Jeany Son, Jiahuan Zhou, Jiajun Wu, Jian Wang, Jianfei Cai, Jiankang Deng, Jianwen Xie, Jimel Yang, Jingya Wang, Jinwei Gu, Joao P Barreto, Jonathan T. Barron, Jong Chul Ye, Jonghyun Choi, Jordi Pont-Tuset, Joseph Tighe, Judy Hoffman, Jufeng Yang, Jun Liu, Jun-Yan Zhu, Jungseok Joo, Junseok Kwon, Junsong Yuan, Kai Han, Kaiyang Zhou, Kannappan Palaniappan, Karteek Alahari, Katerina Fragkiadaki, Keji Yanai, Ken Sakurada, Kevin J Liang, Khosrui, Kiara Ehsani, Kihyuk Sohn, Kris M. Kitani, Kristin Dana, Kwang In Kim, Kwang Moo Yi, Lam M. Nguyen, Lamberto Ballan, Laura Leal-Taixá, Laurent Kneip, Le Lu, Lei Wang, Lef Kobbelt, Li Eran Li, Li Fuxin, Li Yi, Li Zhang, Liang Lin, Liang Zheng, Liang-Chieh Chen, Limin Wang, Lingjie Liu, Lingxi Xie, Linjie Yang, Liwei Wang, Long Chen, Lourdes Agapito, Lu Jang, Lu Sheng, Lu Yuan, Luca Weihs, M. Salman Asif, Mahdi S. Hosseini, Mang Ye, Manohar Paluri, Manolis Savva, Marcello Pellillo, Marco Lorenzi, Marcus Rohrbach, Margret Keuper, Maria Vakalopoulou, Mark Yatskar, Massimiliano Mancini, Matej Kristan, Matt Feiszli, Matthew Brown, Matthew O'Toole, Matthieu Cord, Mehrdad Farajtabar, Miaomiao Liu, Michael Maire, Michael Wray, Mike Zheng Shou, Min H. Kim, Min Sun, Mingkai Tan, Mingsheng Long, Mingze Xu, Minh Hoai, Mingjie Cai, Minsu Cho, Mohamed Elhoseiny, Mohit Gupta, Mohsen Ali, Muhammad Haris Khan, Negar Rostamzadeh, Neill D. F. Campbell, Niaryi Li, Nicy Mitra, Ning Yu, Noah Snavely, Norimichi Uchita, Nuno Vasconcelos, Oana-Maria Camburu, Oisin Mac Aodha, Oliver Wang, Oncel Tuzel, Pablo Arbeláez, Paolo Favaro, Pascal Mettes, Paul Hongsuck Seo, Pedro Morgado, PENG WANG, Peter Hedman, Peter Vincent Gehler, Peter Wonka, Peyman Milanfar, Phlippos Mordochai, Ping Hu, Ping Luo, Piotr Bojanowski, Piotr Koniusz, Pratul P. Srivastava, Praveen Krishnan, Qi Dai, Qi Wu, Qian Yu, Qifeng Chen, Qilong Wang, Qixing Huang, Rafal Mantiuk, Raffay Hamid, Rahaf Aljundi, Ran He, Rana Hanocka, Ranjay Krishna, Ravi Ramamoorthi, Raymond A. Yeh, Rei Kawakami, Richard Zhang, Robby T. Tan, Rohit Girdhar, Roman Pfugfelder, Ryoma Bise, Sai-Kit Yeung, Salman Khan, Sanjeev Jagannatha Koppel, Sara Beery, Sarah Adel Bargal, Sarah Rowlands, Sathya Narayanan N. Aakur, Satoshi Tsutsui, Saurabh Gupta, Saurabh Singh, Scott McCloskey, Scott Workman, Seong Joon Oh, Ser-Nam Lim, Sergey Tulyakov, Seunghoon Hong, Seungyong Kim, Seungyong Lee, Shalini De Mello, Shang-Hong Lai, Shanshan Zhang, Shaohui Lin, Sharon X Huang, Shayok Chakraborty, Shenghua Gao, Shentong Wang, Shijian Lu, Shin'ichi Satoh, Shizhe Chen, Shohei Nobuhara, Shubham Tuliani, Shuicheng YAN, Siavash Arjomand Bigdeli, Siyu Tang, Song Bai, Srinath Sridhar, Stefan Lee, Stefano Mattoccia, Stella X. Yu, Stephan Alaniz, Stephan Richter, Stephen James, Stratis Gavves, Subhanshi Venugopalan, Suha Kwak, Tae Hyun Kim, Tae-Hyun Oh, Tae-Kyun Kim, Taesung Park, Taesup Moon, Takayuki Okatani, Tanmay Gupta, Tat-Jun Chin, Tat-Jun Chin, Tatsuya Harada, Thibaut Durand, Thomas Kipf, Tian Han, Tianfan Xue, Ting Yao, Ting Zhang, Tinne Tuytelaars, Todd Zickler, Tolga Birdal, Tony Tung, Toshihiko Yamaseki, Tsung-Yi Lin, Tushar Nagarajan, Ulugbek S. Kamilov, Unnat Jain, Varun Jampani, Venkatesh Babu Radhakrishnan, Viktoriia Sharmanska, Vincent Lepetit, Vincent Sitzmann, Vishal M. Patel, Vitor Albani, Vittorio Murino, Vlad I. Morariu, Wangmeng Zuo, Waqas Sultan, Wei Liu, Wei Yang, Wei-Chih Hung, Wei-Chiu Ma, Weihong Deng, Weisong Shi, Wen Li, Wenguan Wang, Wenzhi Ren, William Robson Schwartz, Xi Peng, Xiangyu Xu, Xiangyu Yue, Xiaoguang Han, Xiaojuan Qi, Xiaoming Liu, Xiaowei Zhou, Xiaoyu Wang, Xihui Liu, Xin Wang, Xin Yang, Xin Yu, Xingang Pan, Xinggang Wang, Xinlei Chen, Xu-Shen Wei, Xueting Li, Xuming He, Xun Cao, Yale Song, Yan Zhang, Yan Zhang, Yanhao Yang, Yang Song, Yang Wang, Yang You, Yannis Kalantidis, Yaoyao Liu, Yasushi Makihara, Yebin Liu, Yedid Hoshen, Yen-Yu Lin, Yezhou Yang, Yi Wu, Yi-Hsuan Tsai, Yi-Ting Chen, Yibing Song, Yifei Huang, Yijun Li, Yin Li, Ying Fu, Ying Wu, Yinghuan Shi, Yingli Tian, Yingqiang Zheng, Yisen Wang, Yiyi Liao, Yonatan Bisk, Yong Jae Lee, Yossi Keller, Yu Cheng, Yu Kong, Yu Li, Yu Wu, Yu-Chuan Su, Yu-Xiong Wang, Yuandong Tian, Yuchao Dai, Yuki M Asano, Yulun Zhang, Yumin Suh, Yunchao Wei, Yung-Yu Chuang, Yunzhu Li, Yusuke Sugano, Yuyin Zhou, Zaid Harouthi, Zhangyang Wang, Zhaowei Cai, Zhaoxiang Zhang, Zhe Gan, Zheng Zhang, Zhengming Ding, Zhengqi Li, Zhirong Wu, Zhizhong Han, Zhun Zhong, Ziwel Liu, Zongwei Zhou, Zoltan Kira, Zuxuan Wu, Zuzana Kukelova

## Paper Decisions (3/4)

- Reviewers
  - ACよりたくさんいる...(多すぎ)
  - 査読者のおかげでCVPRコミュニティは回っている



## Paper Decisions (4/4)

- ❑ Outstanding Reviewers (TOP2%の査読者)
  - ❑ 査読者も評価してもらえて、ひとつのAwardとして機能する

### Outstanding Reviewers

Aadarsh Sahoo, Aaron Walsman, Abhinav Shukla, Adnan Qayyum, Adrian Lopez, Adrian Penate-Sanchez, Ahmed Taha, Alessio Xompero, Alexander Hermans, Alexandros Haliassos, Alvaro Budria, Amine Bourki, Ana C Murillo, Anastasia Antsiferova, Andrea Conti, Andrew Brown, Anh-Quan Cao, Anil Usumezbas, Anna Kukleva, Arjun Karpur, Armin Hadzic, Ashish Ramayee Asokan, Assia Benbihi, Ayush Jain, Ayush Saraf, Benjamin Busam, Bicheng Xu, Bingliang Jiao, Bumsoo Kim, Carlo Masone, Carlos Rodriguez-Pardo, Changqing Zhang, Chenglin Yang, Chengze Miaomiao Li, Cheston Tan, Chiao An Yang, Chi Xu, Chungyun Park, Cristian Rodriguez-Opazo, Dahyun Kang, Daqi Liu, Davide Cozzolino, David Fan, Da-Wei Zhou, Despoina Paschalidou, Di Hu, Dominique Ginjac, Donald G. Dansereau, Dongbo Min, Dong-Jun Han, Effrosyni Mavroudi, Elisavet Konstantina, Stathopoulou, Evangelos Kazakos, Fabio Tosi, Faisal Z Qureshi, Forrester Cole, Francis Engelmann, Francois Rameau, Fuyong Xing, Ganesh Subramanian Iyer, Gilles Puy, Ginger Delmas, Goutam Bhat, Guoli Jia, Gurusurthy Swaminathan, Hakan Bilen, Hanchen Xie, Hannah Kerner, Haomin Liu, Hao Zhu, Hawook Jeong, Hazel Doughty, Hermann Blum, Hongwei Bran Li, Huan Wang, Hui Liang, Inhwan Bae, Ioannis Papoutsis, Iro Laina, Ishan Rajendrakumar Dave, Jaeyoo Park, Jan Eric Lenssen, Jangwon Lee, Jaskirat Singh, Jens Behley, Jiangpeng He, Jianyi Wang, Jieji Ren, Jihyong Oh, Jingkang Yang, Jingyuan Yang, Joe Mathai, Joerg Stueckler, Jonghee Kim, Juan Camilo Perez, Julian Straub, Junhwa Hur, Jun Xiao, Junyao Hu, Kevin Qinghong Lin, Khang Truong Giang, Kiru Park, Konda Reddy Mopuri, Kongming Liang, Konstantinos Batsos, Kumar Ashutosh, Lin Geng Foo, Lin Zhu, Luisa Verdoliva, M. Saquib Sarfraz, M. Usman Rafique, Mahmoud Afifi, Mannat Singh, Martin R. Oswald, Matteo Poggi, Matthew Joseph Leotta, Mattia Savardi, Mayu Otani, Michaël Gharbi, Michael S. Brown, Michael Weinmann, Michel Antunes, Ming Jiang, Mohamed E. Hussein, Mohammad Reza, Hosseinzadeh Taher, Monica Hernandez, Mo Zhou, Nan Pu, Nicholas Kolkin, Paola Cascante-Bonilla, Patrick Ebel, Paul-Edouard Sarlin, Philipp Lindenberger, Qiang Xu, Qian Zheng, Qize Yang, Renato Martins, Reuben Tan, Riccardo Volpi, Richard Newcombe, Ruicong Liu, Ruihan Gao, Saihui Hou, Samaneh Azadi, Satwik Kottur, Satya Krishna Gorti, Shuhong Zheng, Shu Zhang, Shyamal Buch, Siamul Karim Khan, Sima Behpour, Simon Stepputtis, Siyan Dong, Sourav Garg, Spencer Whitehead, Stefan Leutenegger, Stefano Berretti, Sunghoon Im, Swathikiran Sudhakaran, Terrance E. Boult, Thomas Hummel, Toby Perrett, Torsten Sattler, Touqeer Ahmad, Valerio Cambareni, Vishaal Udandarao, Volker Rodehorst, Williemi, Wojciech Zielonka, Wonwoong Cho, Xiang An, Xiang Li, Xiangyu He, Xiaohan Zhang, Xingyi Zhou, Xiongwei Wu, Yanbin Liu, Yankai Jiang, Yannick Hold-Geoffroy, Yi Dong, Yihua Zhang, Yinan Zhao, Yongzhen Huang, Yuanhao Cai, Yuesong Wang, Yu-Lun Liu, Yunqi Miao, Yunze Man, Yurui Zhu, Yusuke Matsui, Yu-Wei Chao, Yuyu Guo, Zan Gojcic, Zerong Zheng, Zheheng Jiang, Zhekun Luo, Zheng Chen, Zhibin Liao, Zhicheng Zhang, Zhihe Lu, Zhou Yu

## Best Paper Awards (1/6)

### □ 選定プロセス

- Area Chairが賞候補論文を推薦
- Program Chairが24 → 10本の論文を選考

### Best Paper Awards

- 24 best paper award candidates were nominated by the Area Chairs, marked in the program
- From the candidates, 10 papers were selected for an award by the Program Chairs
  - 4 Best Student Paper Runners-Up
  - 2 Best Student Papers
  - 2 Best Paper Runners-Up
  - 2 Best Papers
- We'll present certificates to the award winners at the PAMI-TC meeting on Thursday

#### PROGRAM GUIDE

3	EgoGen: An Egocentric
4	Learning to Segment 4
5	Producing and Leverage
5	Prediction, Xunjiang C
	Pavone, Boris Ivanovic
13:00 - 14:30 Orals 4B: 2	
1	SceneFun3D: Fine-Grained
	Understanding in 3D Sc
	Federico Tombari, Robe
2	SpiderMatch: 3D Shap
2	Geometric Consistent
3	PaSCo: Urban 3D Plan
4	Awareness, An/Quar
4	PlatoNeRF: 3D Reconst
5	Bounce Lidar, Zofei King
	Yuchen Fan, Christian F
5	A Subspace-Constrain
	Structure from Motion
13:00 - 14:30 Orals 4C: 2	
1	Modeling Multimodal
	Baselines with Dense
	Bolin Lai, Fiona Ryan, J
2	An N-Point Linear Sol
	Event Cameras, Ling
	Scaramuzza, Laurent A
3	RoHM: Robust Human
	Siwei Zhang, Bharat Li
	Petr Kadlecek, Siyu Ta
4	Temporally Consistent
5	Unsupervised Action 2

## Best Paper Awards (2/6)

### Best Student Paper Runners-Up (1/2)

#### SpiderMatch: 3D Shape Matching with Global Optimality and Geometric Consistency

Paul Roetzer<sup>1</sup> Florian Bernard<sup>1</sup>  
University of Bonn<sup>1</sup>

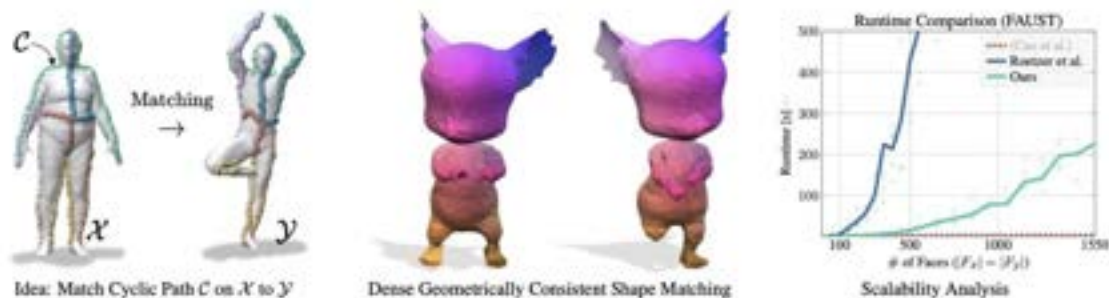


Figure 1. (Left) Schematic illustration of our method: we extract a cyclic path  $C$  on the surface of 3D shape  $X$ , and match  $C$  to the 3D target shape  $Y$ . (Middle) If  $C$  covers the whole source shape  $X$  (not shown for clarity, see Sec. 4.1 for details), our method can be used to compute globally optimal results for dense non-rigid and geometrically consistent 3D shape matching. (Right) We show runtimes w.r.t. shape resolution and compare to Roetzer et al. [55] (which is the only existing method that achieves geometric consistency while having a global flavour and being initialization-free) and Cao et al. [14] (which does not provide any geometric consistency guarantees).

#### Image Processing GNN: Breaking Rigidity in Super-Resolution

Yuchuan Tian<sup>1</sup>, Hanting Chen<sup>2</sup>, Chao Xu<sup>1</sup>, Yunhe Wang<sup>2\*</sup>

<sup>1</sup> National Key Lab of General AI, School of Intelligence Science and Technology, Peking University. <sup>2</sup> Huawei Noah's Ark Lab.  
tianyc@stu.pku.edu.cn, yunhe.wang@huawei.com

#### Abstract

Super-Resolution (SR) reconstructs high-resolution images from low-resolution ones. CNNs and window-attention methods are two major categories of canonical SR models. However, these measures are rigid: in both operations, each pixel gathers the same number of neighboring pixels, hindering their effectiveness in SR tasks. Alternatively, we leverage the flexibility of graphs and propose the Image Processing GNN (IPG) model to break the rigidity that dominates previous SR methods. Firstly, SR is unbalanced in that most reconstruction efforts are concentrated to a small proportion of detail-rich image parts. Hence, we leverage degree flexibility by assigning higher node degrees to detail-rich image nodes. Then in order to construct graphs for SR-effective aggregation, we treat images as pixel node sets rather than patch nodes. Lastly, we

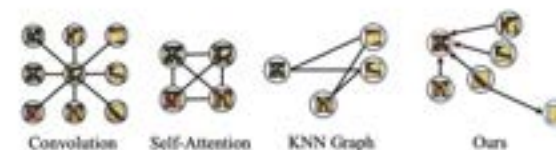


Figure 1. Convolution (left), Self-Attention (middle-left), KNN Graph Aggregation (middle-right), and Graph Aggregation in IPG (right). Compared with other methods, IPG graph aggregation considers the unbalanced nature of SR: detail-rich, high-frequency image nodes (red) have higher node degrees; while flat, low-frequency image nodes (blue) require fewer aggregations.

mote sensing [15], et cetera.

Despite various measures, it is taken for granted that mainstream SR models treat all pixels in a fairly rigid manner. For instance, as shown in Figure 1, in convolution layers of the CNN-based SR model VDSR [16], the

## Best Paper Awards (3/6)

### □ Best Student Paper Runners-Up (2/2)

#### Objects as volumes: A stochastic geometry view of opaque solids

Bailey Miller, Hanyu Chen, Alice Lai, Ioannis Gkioulekas  
Carnegie Mellon University

##### Abstract

We develop a theory for the representation of opaque solids as volumes. Starting from a stochastic representation of opaque solids as random indicator functions, we prove the conditions under which such solids can be modeled using exponential volumetric transport. We also derive expressions for the volumetric attenuation coefficient as a functional of the probability distributions of the underlying indicator functions. We generalize our theory to account for isotropic and anisotropic scattering at different parts of the solid, and for representations of opaque solids as stochastic implicit surfaces. We derive our volumetric representation from first principles, which ensures that it satisfies physical constraints such as reciprocity and reversibility. We use our theory to explain, compare, and correct previous volumetric representations, as well as propose meaningful extensions that lead to improved performance in 3D reconstruction tasks.

##### 1. Introduction

swers to queries such as “are two points mutually visible” (a visibility query) and “what is the distance to first intersection along a ray” (a ray-casting query), when the geometry occluding visibility and terminating rays is stochastic.

Volumetric representations for translucent and participating media are stochastic abstractions of their microscopic structure: Such media comprise numerous microscopic particles that reflect and occlude light rays. Modeling explicit microparticle configurations, and rendering light transport through them, is prohibitively expensive. As a compromise for efficiency, volumetric representations allow to simulate light transport in such media on average [6]. These representations replace explicit with statistical descriptions of microparticle configurations (e.g., average particle location, size, shape, and orientation), analogously to statistical BRDF models for surface microgeometry [13, 14, 24, 46, 60]. Computer graphics has developed volumetric representations for microparticle media that account for details such as varying particle size and material [16, 32]. shape and orienta-

#### Comparing the Decision-Making Mechanisms by Transformers and CNNs via Explanation Methods

Mingqi Jiang, Saeed Khorram, Li Fuxin  
Collaborative Robotics and Intelligent Systems (CoRIS) Institute  
Oregon State University

{jiangmi, khorrans, lif}@oregonstate.edu

##### Abstract

In order to gain insights about the decision-making of different visual recognition backbones, we propose two methodologies, sub-explanation counting and cross-testing, that systematically applies deep explanation algorithms on a dataset-wide basis, and compares the statistics generated from the amount and nature of the explanations. These methodologies reveal the difference among networks in terms of two properties called compositionality and disjunctivism. Transformers and ComNeXt are found to be more compositional, in the sense that they jointly consider multiple parts of the image in building their decisions, whereas traditional CNNs and distilled transformers are less compositional and more disjunctive, which means that they use

tant contributing factor is not the attention itself but those design principles? If so, which specific design principles particularly affect the decision-making of networks? Better answers to those questions would help us to gain more insights into those deep and complicated black-box networks.

In this paper, we propose a novel methodology to examine these questions through deep explanation algorithms. Explanation algorithms have improved significantly in recent years and can generate accurate explanations that can be verified through intervention experiments on images [21, 26]. Recent search-based explanation algorithms can find a comprehensive set of minimally sufficient explanations (MSEs) [30], defined as the minimal set of patches that, if shown to the network, lead to predictions that are almost as confident as predictions from the full image. The



## Best Paper Awards (4/6)

### □ Best Student Paper

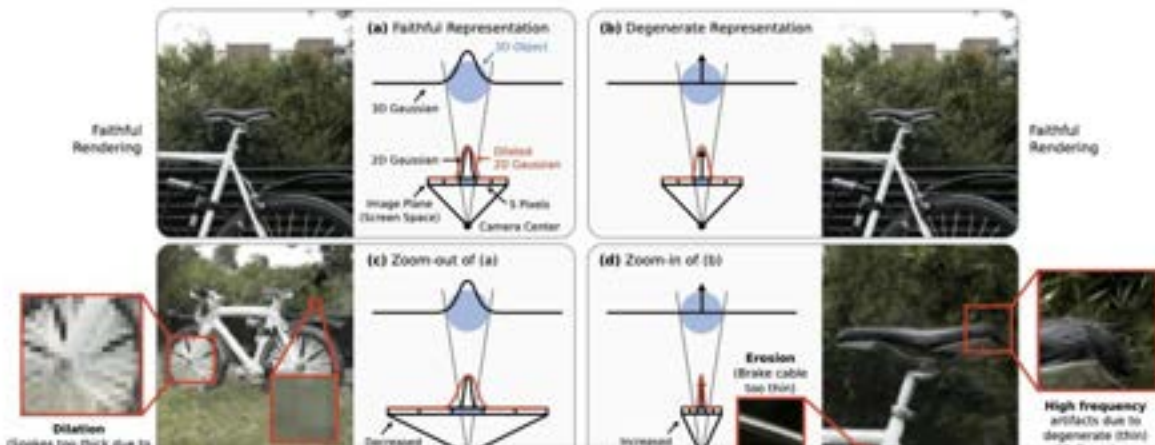
#### Mip-Splatting: Alias-free 3D Gaussian Splatting

Zehao Yu<sup>1,2</sup> Anpei Chen<sup>1,1,2</sup> Binbin Huang<sup>3</sup> Torsten Sattler<sup>4</sup> Andreas Geiger<sup>1,2</sup>

<sup>1</sup>University of Tübingen <sup>2</sup>Tübingen AI Center <sup>3</sup>ShanghaiTech University

<sup>4</sup>Czech Technical University in Prague

<https://niujinshuchong.github.io/mip-splatting>



#### BIOCLIP: A Vision Foundation Model for the Tree of Life

Samuel Stevens<sup>1\*</sup>, Jiaman Wu<sup>1\*</sup>, Matthew J Thompson<sup>1</sup>, Elizabeth G Campolongo<sup>1</sup>, Chan Hee Song<sup>1</sup>, David Edward Carlyn<sup>1</sup>, Li Dong<sup>2</sup>, Wasila M Dahdul<sup>3</sup>, Charles Stewart<sup>4</sup>, Tanya Berger-Wolf<sup>1</sup>, Wei-Lun Chao<sup>1</sup>, and Yu Su<sup>1†</sup>

<sup>1</sup>The Ohio State University, <sup>2</sup>Microsoft Research, <sup>3</sup>University of California, Irvine, <sup>4</sup>Rensselaer Polytechnic Institute

#### Abstract

*Images of the natural world, collected by a variety of cameras, from drones to individual phones, are increasingly abundant sources of biological information. There is an explosion of computational methods and tools, particularly computer vision, for extracting biologically relevant information from images for science and conservation. Yet most of these are bespoke approaches designed for a specific task and are not easily adaptable or extendable to new questions, contexts, and datasets. A vision model for general organismal biology questions on images is of timely need. To*

*research this, we create and release TreeOfLife-10M [79–81, 87, 88] into actionable information (e.g., species classification, individual identification, and trait detection) has accelerated and enabled new advances in tasks such as species delineation [32], understanding mechanisms of adaptation [23, 39], abundance and population structure estimation [3, 40, 58, 82], and biodiversity monitoring and conservation [83].*

*However, applying computer vision to answer a biological question is still a laborious task requiring substantial machine learning expertise and effort—biologists must manually label sufficient data for the specific taxa and task of interest, and find and train a suitable model for the task. Meanwhile, foundation models [33] such as CLIP [50] and*

## Best Paper Awards (5/6)

### Best Paper Runners-Up

#### EventPS: Real-Time Photometric Stereo Using an Event Camera

Bohan Yu<sup>1,2</sup> Jiej Ren<sup>3</sup> Jin Han<sup>4,5</sup> Feishi Wang<sup>1,2</sup> Jinxiu Liang<sup>1,2</sup> Boxin Shi<sup>1,2\*</sup>

<sup>1</sup> National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University

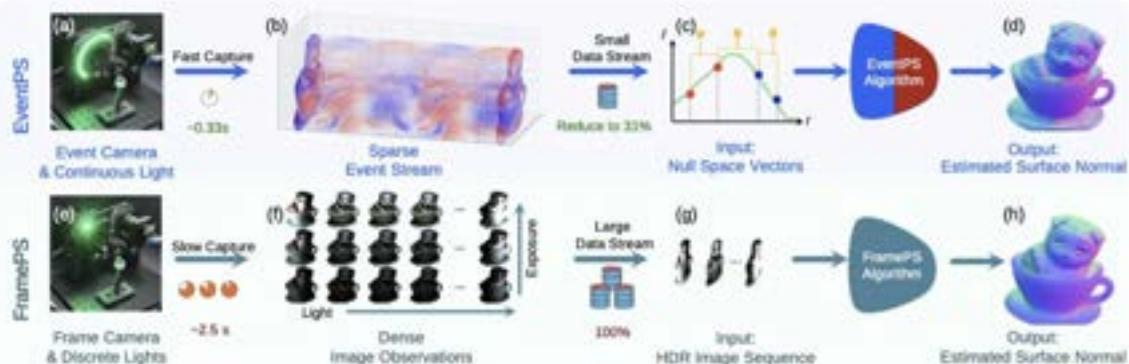
<sup>2</sup> National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

<sup>3</sup> School of Mechanical Engineering, Shanghai Jiao Tong University

<sup>4</sup> Graduate School of Information Science and Technology, The University of Tokyo <sup>5</sup> National Institute of Informatics

{ybh1998, wangfeishi, csherryliang, shiboxin}@pku.edu.cn

jiejiren@sjtu.edu.cn, jinhan@nii.ac.jp



#### pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction

David Charatan<sup>1</sup> Sizhe Lester Li<sup>1</sup> Andrea Tagliasacchi<sup>2</sup> Vincent Sitzmann<sup>1</sup>  
<sup>1</sup>Massachusetts Institute of Technology <sup>2</sup>Simon Fraser University, University of Toronto  
{charatan, sizheli, sitsmann}@mit.edu andrea.tagliasacchi@sfu.ca

#### Abstract

We introduce pixelSplat, a feed-forward model that learns to reconstruct 3D radiance fields parameterized by 3D Gaussian primitives from pairs of images. Our model features real-time and memory-efficient rendering for scalable training as well as fast 3D reconstruction at inference time. To overcome local minima inherent to sparse and locally supported representations, we predict a dense probability distribution over 3D and sample Gaussian means from that probability distribution. We make this sampling operation differentiable via a reparameterization trick, allowing us to back-propagate gradients through the Gaussian splatting representation. We benchmark our method on wide-baseline novel view synthesis on the real-world RealEstate10k and ACID datasets, where we outperform state-of-the-art light



Figure 1. **Overview.** Given a pair of input images, pixelSplat reconstructs a 3D radiance field parameterized via 3D Gaussian primitives. This yields an explicit 3D representation that is renderable in real time, remains editable, and is cheap to train.

## Best Paper Awards (6/6)

### Best Paper

#### Generative Image Dynamics

Zhengqi Li   Richard Tucker   Noah Snavely   Aleksander Holynski  
Google Research

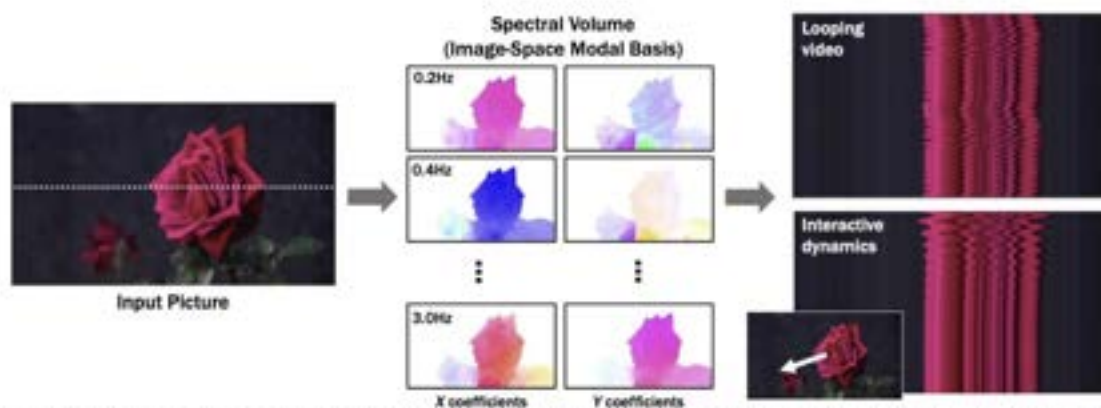


Figure 1. We model a generative image-space prior on scene motion: from a single RGB image, our method generates a spectral volume [23], a motion representation that models dense, long-term pixel trajectories in the Fourier domain. Our learned motion priors can be used to turn a single picture into a seamlessly looping video, or into an interactive simulation of dynamics that responds to user inputs like dragging and

#### Rich Human Feedback for Text-to-Image Generation

Youwei Liang<sup>\*11</sup>, Junfeng He<sup>\*12</sup>, Gang Li<sup>\*12</sup>, Peizhao Li<sup>15</sup>, Arseniy Klimovskiy<sup>2</sup>, Nicholas Carolan<sup>2</sup>, Jiao Sun<sup>13</sup>, Jordi Pont-Tuset<sup>2</sup>, Sarah Young<sup>2</sup>, Feng Yang<sup>2</sup>, Junjie Ke<sup>2</sup>, Krishnamurthy Dj Dvijotham<sup>2</sup>, Katherine M. Collins<sup>14</sup>, Yiwen Luo<sup>2</sup>, Yang Li<sup>2</sup>, Kai J Kohlhoff<sup>2</sup>, Deepak Ramachandran<sup>2</sup>, and Vidhya Navalpakkam<sup>2</sup>

<sup>1</sup>University of California, San Diego

<sup>2</sup>Google Research

<sup>3</sup>University of Southern California

<sup>4</sup>University of Cambridge

<sup>5</sup>Brandeis University

#### Abstract

Recent Text-to-Image (T2I) generation models such as Stable Diffusion and Imagen have made significant progress in generating high-resolution images based on text descriptions. However, many generated images still suffer from issues such as artifacts/implausibility, misalignment with text descriptions, and low aesthetic quality. Inspired by the

in our GitHub repository: <https://github.com/google-research/google-research/tree/master/richhf.18k>.

#### 1. Introduction

Text-to-image (T2I) generation models [12, 17, 41, 42, 56, 58, 59] are rapidly becoming a key to content creation in

## CVPR PAMI TC Awards (1/4)

- Longuet-Higgins Prize (Test-of-Time Award)
  - 目安10年前のCVPRから、学术界に対して最も影響を与えた論文を選定
  - CVPR 2014からは「R-CNN 論文」が選定; 深層学習による物体検出に対して最も多く引用され、その後の研究にも多大なる影響を与えている

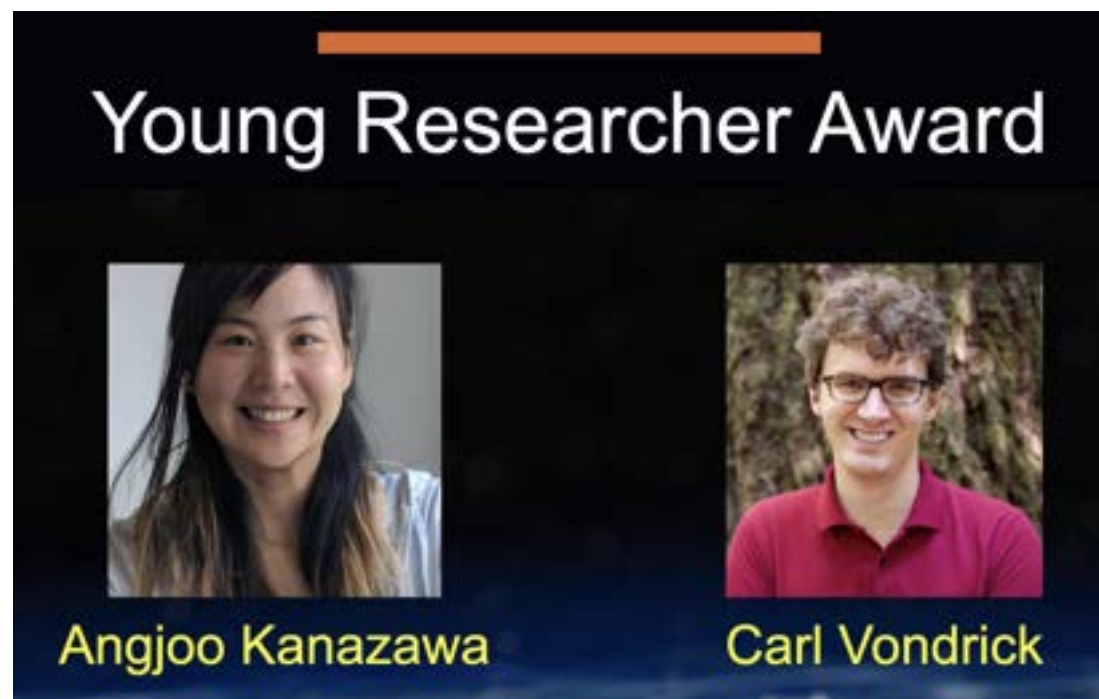
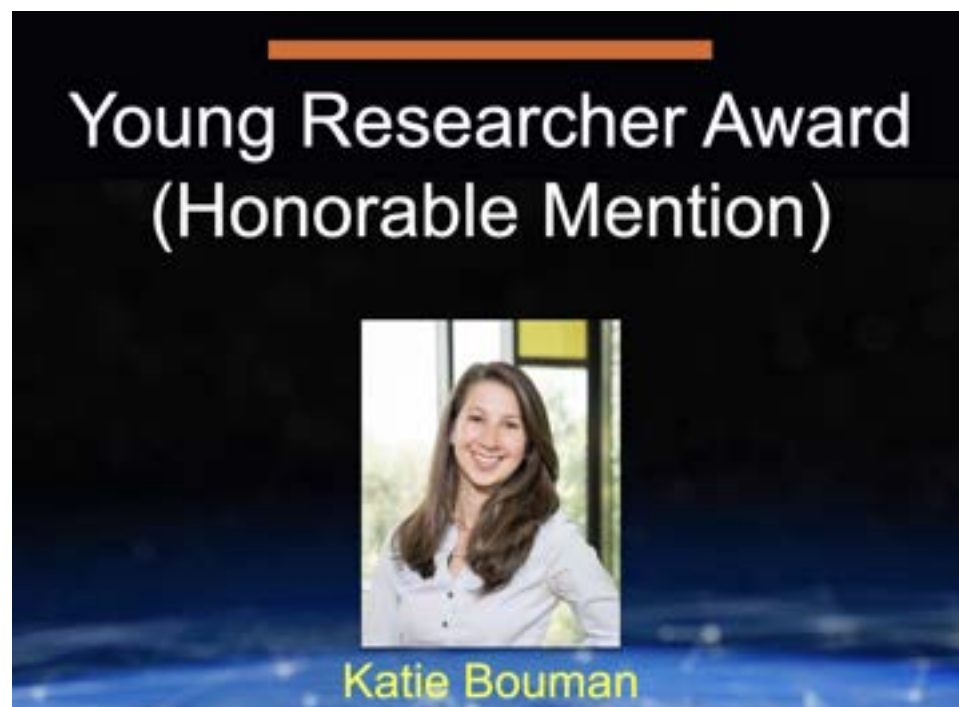
## Longuet-Higgins Prize

*Rich Feature Hierarchies for Accurate Object  
Detection and Semantic Segmentation*

Ross Girshick, Jeff Donahue, Trevor Darrell and  
Jitendra Malik

## CVPR PAMI TC Awards (2/4)

- ❑ Young Researcher Award
  - ❑ 博士号取得後, 7年以内の若手研究者に対して表彰
  - ❑ Katie Bouman: ブラックホール画像撮影のための手法開発にて成果
  - ❑ Angjoo Kanazawa: 人物・環境の動的な3次元再構成にて成果
  - ❑ Carl Vondrick: V&L, フォトグラフィ, 動画理解など多彩な研究にて成果



## CVPR PAMI TC Awards (3/4)

- Thomas Huang Memorial Prize
  - 研究者の育成・教育において貢献が大きい研究者を表彰
  - Andrea Vedaldi: Oxford VGGにおいてAward論文始め多くの重要論文を研究コミュニティに対して提案, 同時に多数のPh.D.を研究室から輩出



## CVPR PAMI TC Awards (4/4)

- PAMI TC Awardは「長い目を見た」研究成果に対する受賞
- 論文 → 10年間で観測
- 研究者 → 若手研究者はPh.D.取得後7年間, 教員は無差別級

「瞬間最大風速」で獲得するAwardも良いけど,  
長い目で見る CVPR PAMI TC Awardは本物という感じ

# CVPR 2024 の動向・気付き(22/116)

## Exhibits (1/2)

- 今年も企業によるdemoが賑わっていた

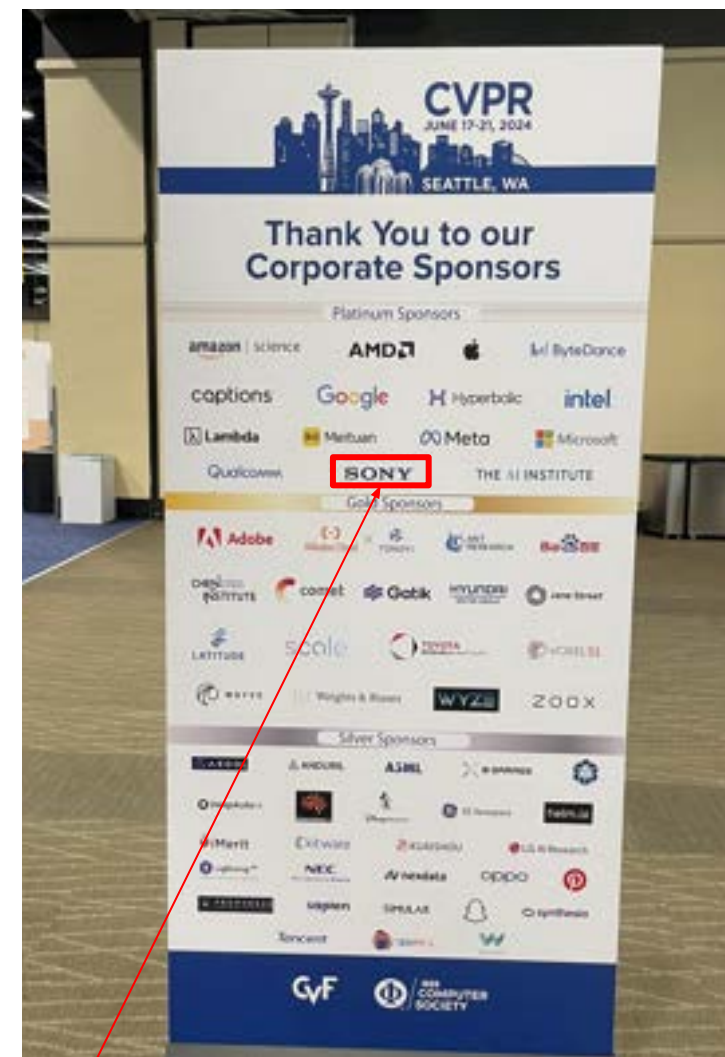




# CVPR 2024 の動向・気付き(23/116)

## Exhibits (2/2)

- SONYのブースが目立つ位置にあり勢いを感じた

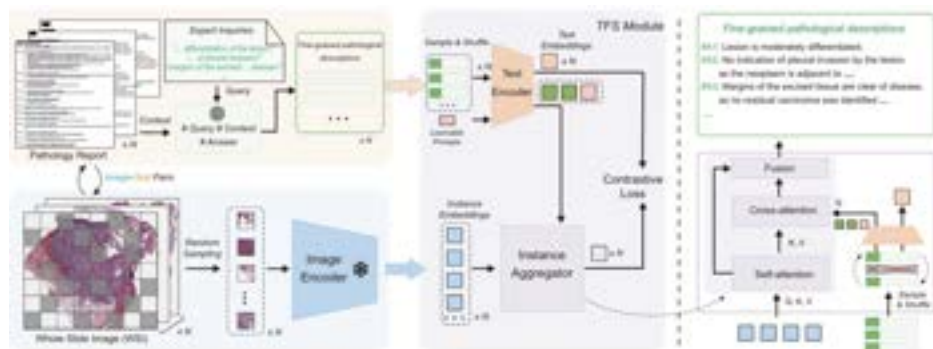


日本企業唯一のPlatinum Sponsor

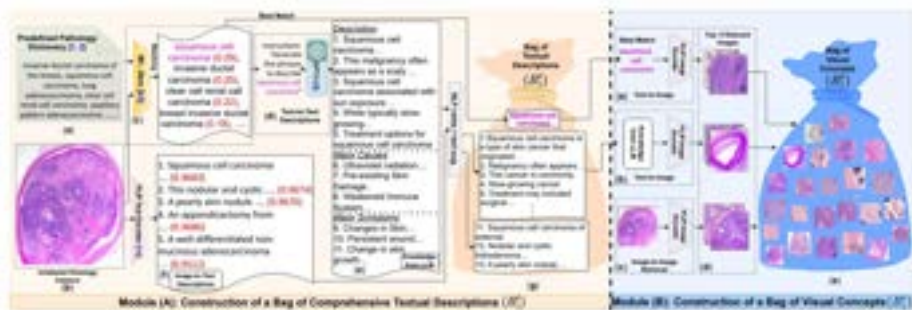
# CVPR 2024 の動向・気付き(24/116)

## 病理画像 x 言語による fine-grained な分類を目指す研究の進展

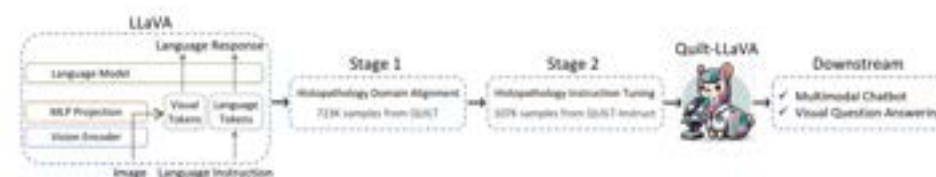
- 一般画像の流れに追従し、vision and language が急速に発展
- 従来設定: がん非がんの2クラス分類がメイン
- 新規設定: 言語情報を用いて multi-class や few-shot classification へと発展



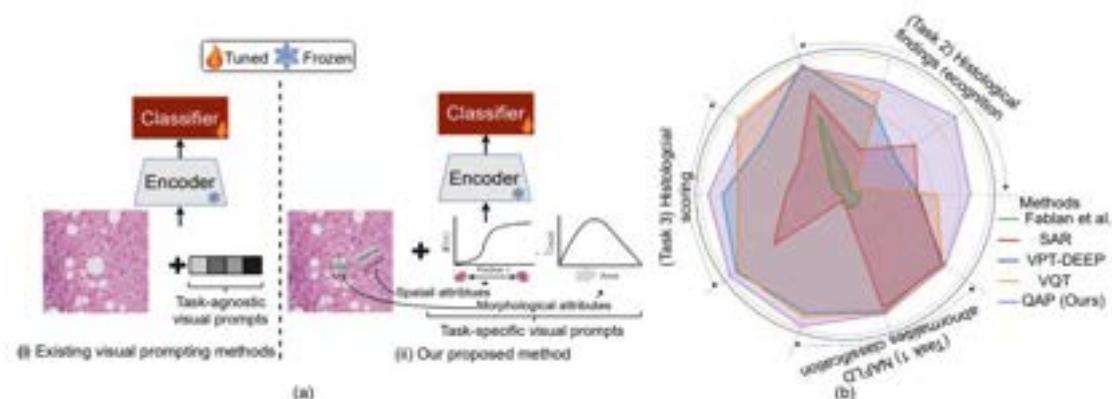
FiVE



CPLIP



Qulit-LLaVA



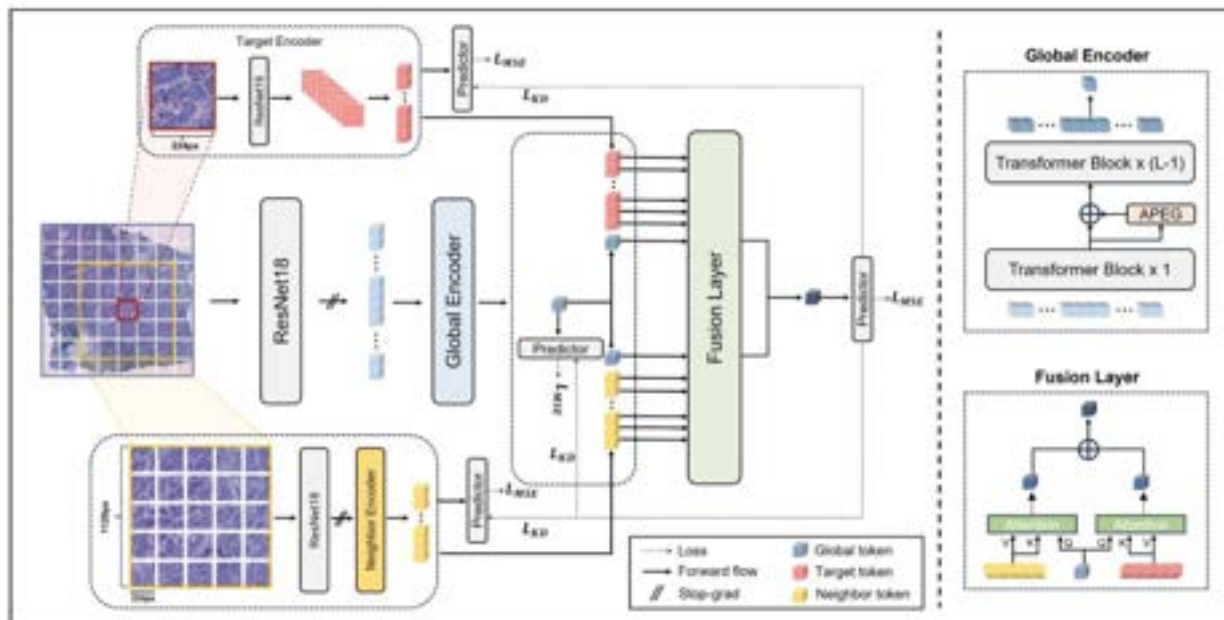
NAFLD

# CVPR 2024 の動向・気付き(25/116)

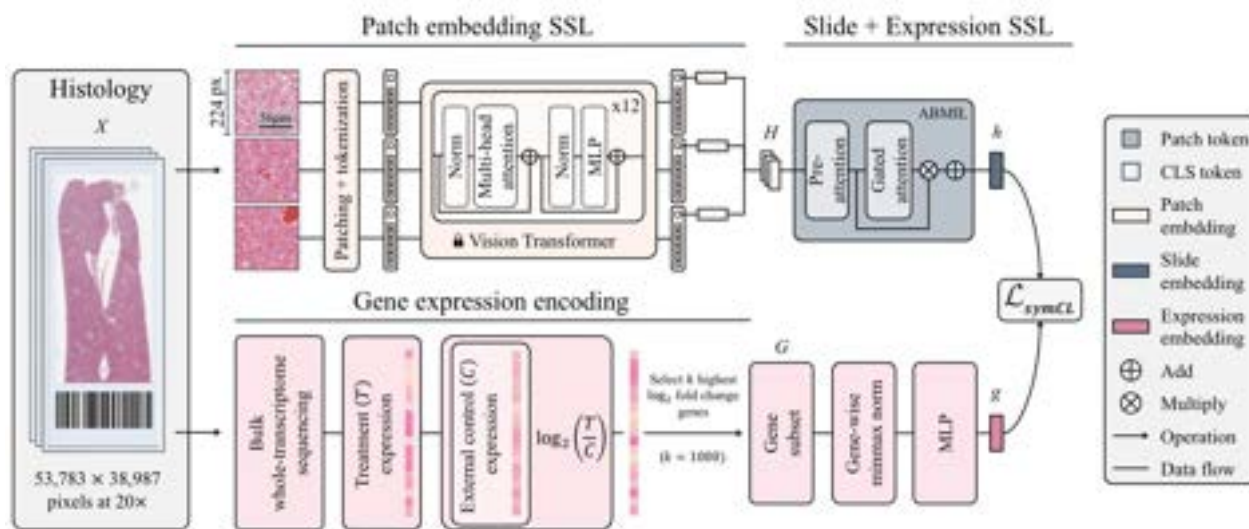
## 病理画像 x Transcriptomics (遺伝子発現データ) 研究の発展

- 画像だけに着目した研究から画像+新規モーダルの研究が登場
  - 病理画像独自のモデルであるTranscriptomics(遺伝子発現)が注目

画像→遺伝子発現の予測を Multi-channel の regression 問題として扱う

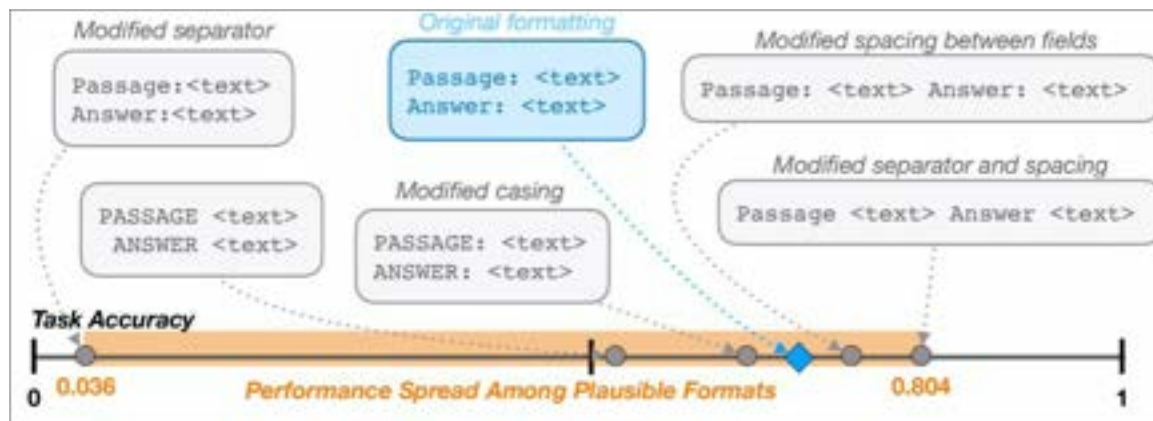


遺伝子発現データを用いて画像の representation learning を強化



## WS Prompting in Vision: LLMs as Agent ([Alane Suhr](#))

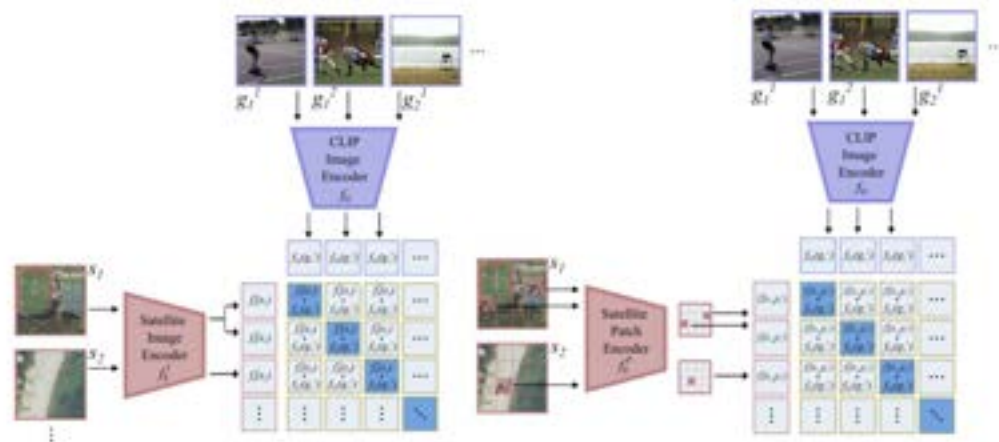
- ❑ NLVR, NLVRv2の著者
- ❑ LLMs agentの評価をメインで紹介
  - ❑ Behavior cloningなどを活用して、評価の結果でAgentの性能を向上可能
  - ❑ LLMs agent評価の際に気をつけるべきポイント: FormatSpread(下図)
    - ❑ FormatSpread: 入力Promptにちょっとずつ微小のフォーマット差を入れると性能が変わる
    - ❑ FormatSpreadの重要性: モデルのLower boundとUpper boundを知る
    - ❑ 論文: Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting



- ❑ 紹介したモデル:
  - ❑ QWen-VL: オープンソースVLM
  - ❑ CogAgent: GUI理解のVLM

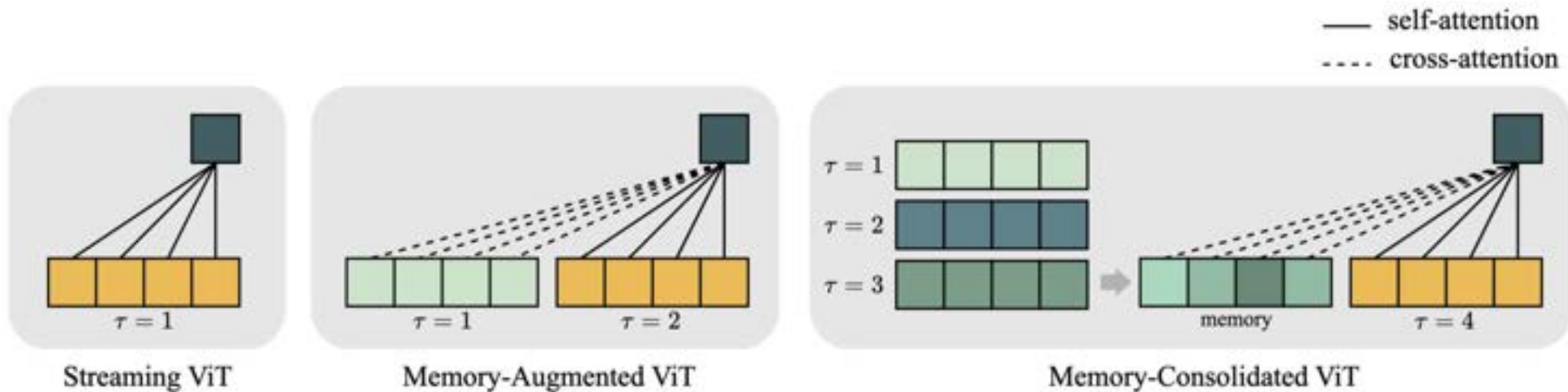
## WS Prompting in Vision: Easter eggs in pre-trained vision models ([Bharath Hariharan](#))

- ❑ Pre-trainedモデルから、知識を取得する方法をメインで紹介
- ❑ Visual-Prompt Tuning
  - ❑ 事前学習済みのTransformerを低コストでVisionの下流タスクへ適応
- ❑ 画像認識の知識をsatellite画像認識へ
  - ❑ Internet画像とsatellite画像のGeotagの関係性でリンクさせる(下図)
  - ❑ 論文: Remote Sensing Vision-Language Foundation Models without Annotations via Ground Remote Alignment
- ❑ Diffusionモデルから
  - ❑ Implicit correspondencesを認識可能
    - ❑ Diffusion featuresのnearest neighborは、教師なしでgeometric と semantic correspondencesを認識可能
    - ❑ 論文: Zero-shot image-to-image translation
  - ❑ 2D diffusion から3D の知識を取得
    - ❑ Correspondence-based seed selection
    - ❑ 論文: RealFill: Reference-Driven Generation for Authentic Image Completion



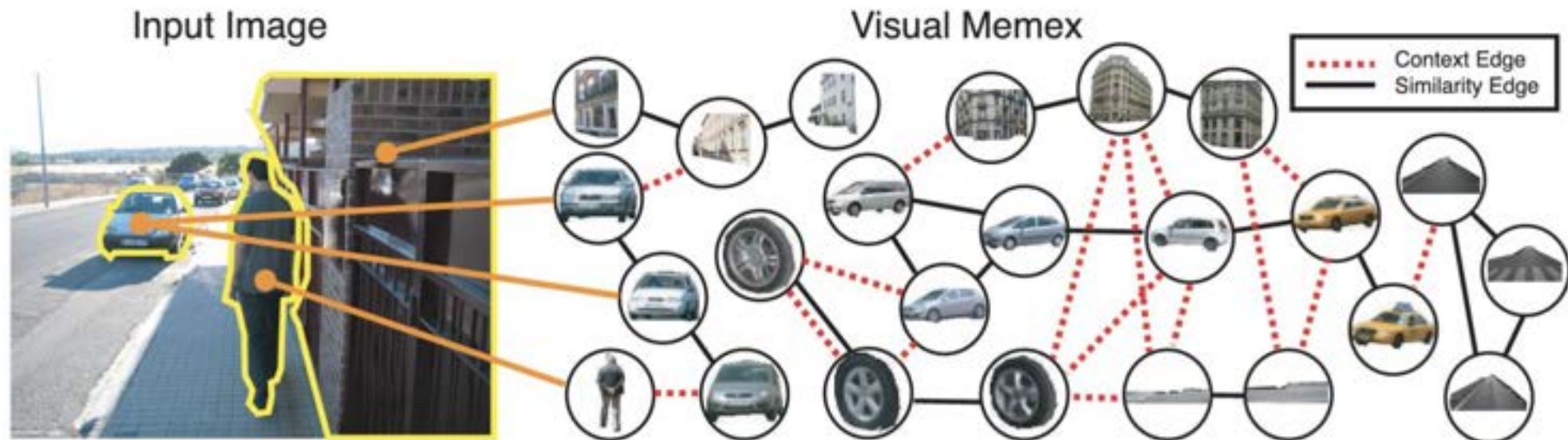
## WS Prompting in Vision: Towards Effortless Adaptation of Image and Video Models ([Ivana Balazevic](#))

- ❑ 画像とビデオ認識のadaptation手法を一本ずつ紹介
- ❑ 画像認識モデルのAdaptation
  - ❑ 論文: Towards in-context scene understanding
  - ❑ アイデア: in-context learningでシーンを理解し、汎化性・データ効率性とadaptationのスピードを向上。デコーダーでは nearest neighbor retrieval を使用する。
- ❑ long-context動画認識のAdaptation
  - ❑ 論文: Memory Consolidation Enables Long-Context Video Understanding
  - ❑ アイデア: メモリー consolidationを既存のvideotransformersに適応し(下図)、Long-contextの理解力を向上。



## WS Prompting in Vision: Visual Prompting Then and Now ([Alexei A. Efros](#))

- ❑ VisionのPromptの歴史
  - ❑ Raven's Progressive Matrices, Markov, Shannon (n-grams)
- ❑ 現在のEfros先生グループの関連研究の紹介 (1/3)
  - ❑ Beyond Categories: The Visual Memex Model for Reasoning About Object Relationships
    - ❑ 物体インスタンス間の関係性(数の実線—アピアランス的关系性、点線—コンテキスト的关系性)をグラフ(Visual Memex)で表現する提案。

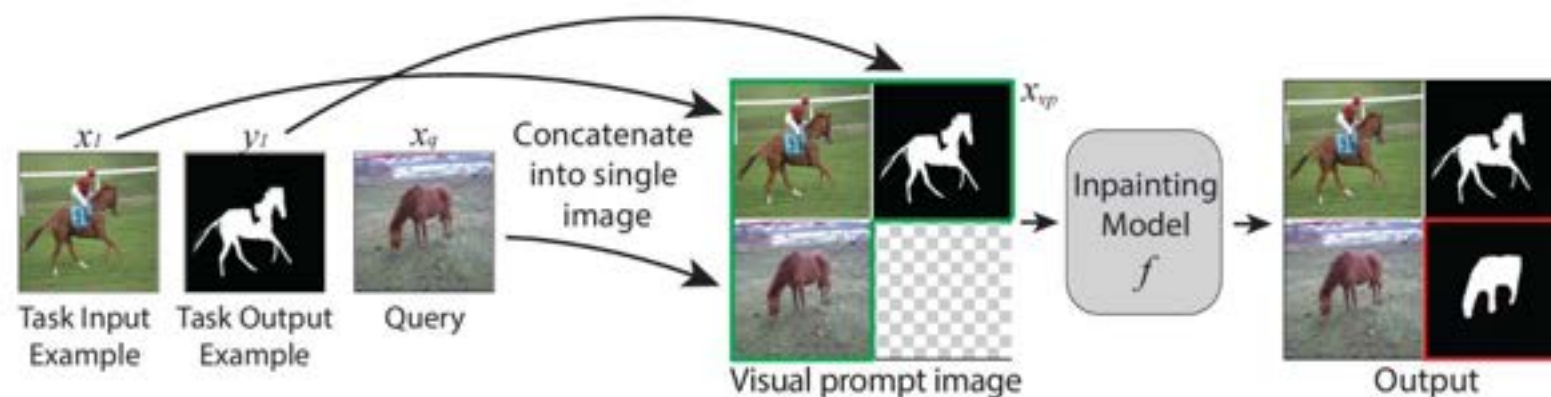


## WS Prompting in Vision: Visual Prompting Then and Now ([Alexei A. Efros](#))

- 現在のEfros先生グループの関連研究の紹介 (2/3)

- Visual Prompting via Image Inpainting

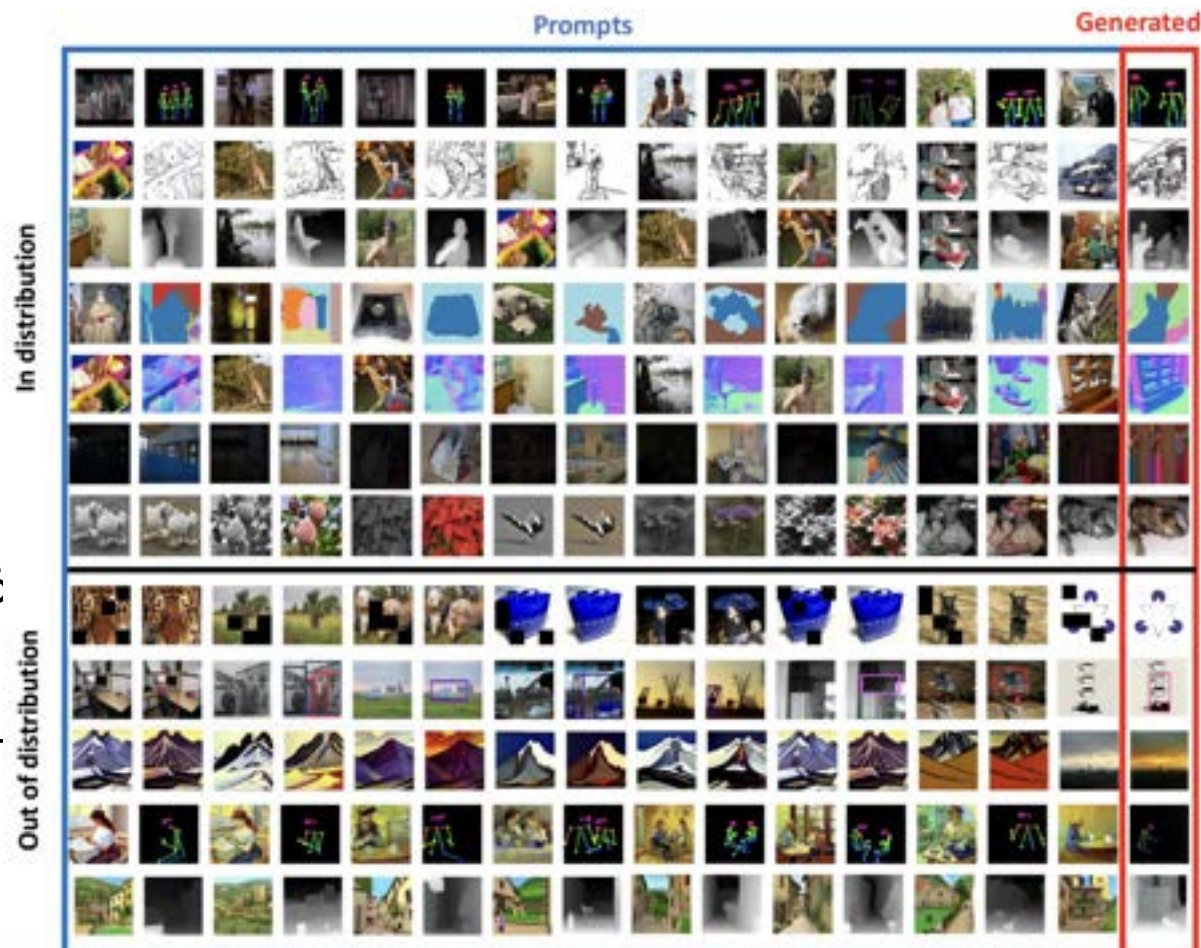
- 画像Inpaintingタスクの形式で複数の画像タスクを扱えるようにしたVisual Prompt手法の提案。入力がexample画像と対応するタスクspecific出力、およびターゲット画像。MAE-VQGANでinpaintingを行う。





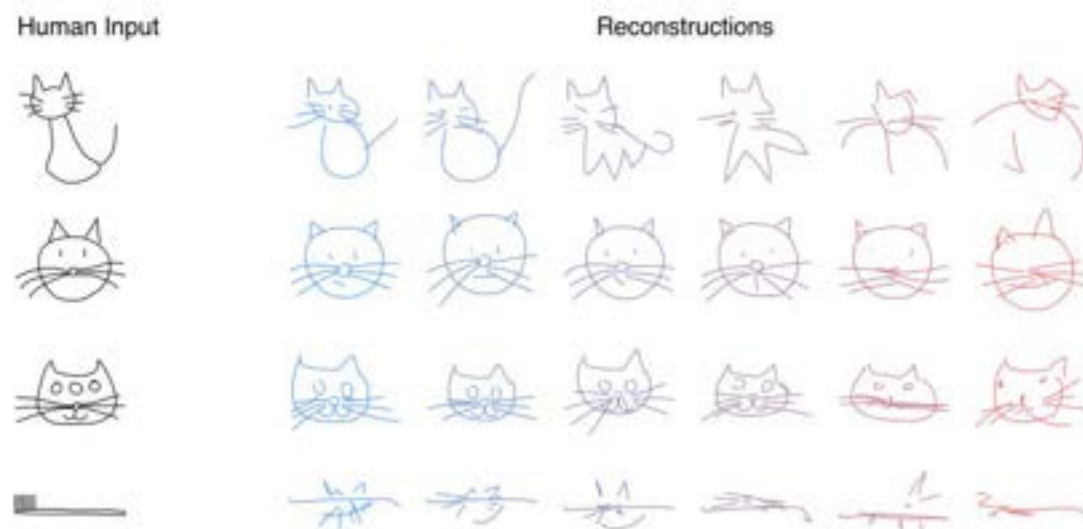
## WS Prompting in Vision: Visual Prompting Then and Now ([Alexei A. Efros](#))

- 現在のEfros先生グループの関連研究の紹介 (3/3)
  - Sequential modeling enables scalable learning for large vision models
    - 画像を言語のようなSequential 関係性を学習可能にするLarge Vision Model 仕組みの提案。
    - 入力のVisual Prompts: 画像Sequence / video frames (visual sentence)。出力は入力に応じてピクセルレベルの画像出力(1枚/複数枚)。様々なVisionタスクを対応可能。また、同一構造で、教師ありと教師なし両方対応可能。
    - 提案の学習済みモデルが認識タスクへの適応や言語と組み合わせたモデルなどが検討できそう。また、複数画像間の関係性を考慮したCorrespondenceの認識はどこまでできるのかが気になる。



## WS Prompting in Vision

- ❑ Visual Prompting in Multimodal Model by [Yong Jae Lee](#) (1/5)
  - ❑ Visual Promptingの変遷
    - ❑ Pre-deep learning
      - ❑ Content-Based Image Retrieval
      - ❑ ShadowDraw (A Neural Representation of Sketch Drawings)
    - ❑ 大規模基盤モデル
      - ❑ LLM, ViT, image-text models
      - ❑ 1つのモデルで様々なタスクに対応



## WS Prompting in Vision

- ❑ Visual Prompting in Multimodal Model by [Yong Jae Lee](#) (2/5)
  - ❑ LLaVAの特徴
    - ❑ 視覚データの理解に優れる、ユーザーの指示に従う、自然言語でコミュニケーション
  - ❑ LLaVAの能力
    - ❑ Extreme Ironing • strong visual understanding • Emergent Multilingual Capability

Visual input example, Extreme Ironing:



Source: <https://www.burrowson.com/wp-content/uploads/2016/12/03-Overstating-Pictures.jpg>

User  
LLaVA

What is unusual about this image?

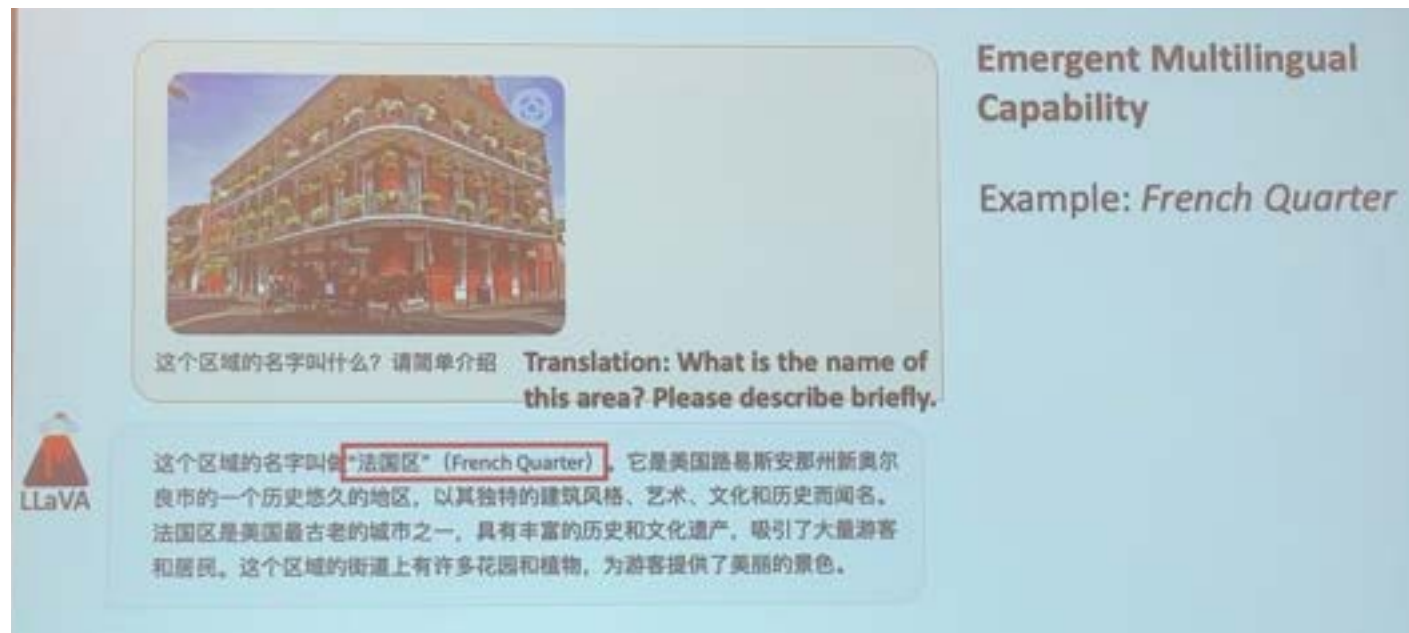
The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

User  
LLaVA

[Start a new conversation, and clear the history]

What's happening in the scene?

The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.



Emergent Multilingual Capability

Example: *French Quarter*

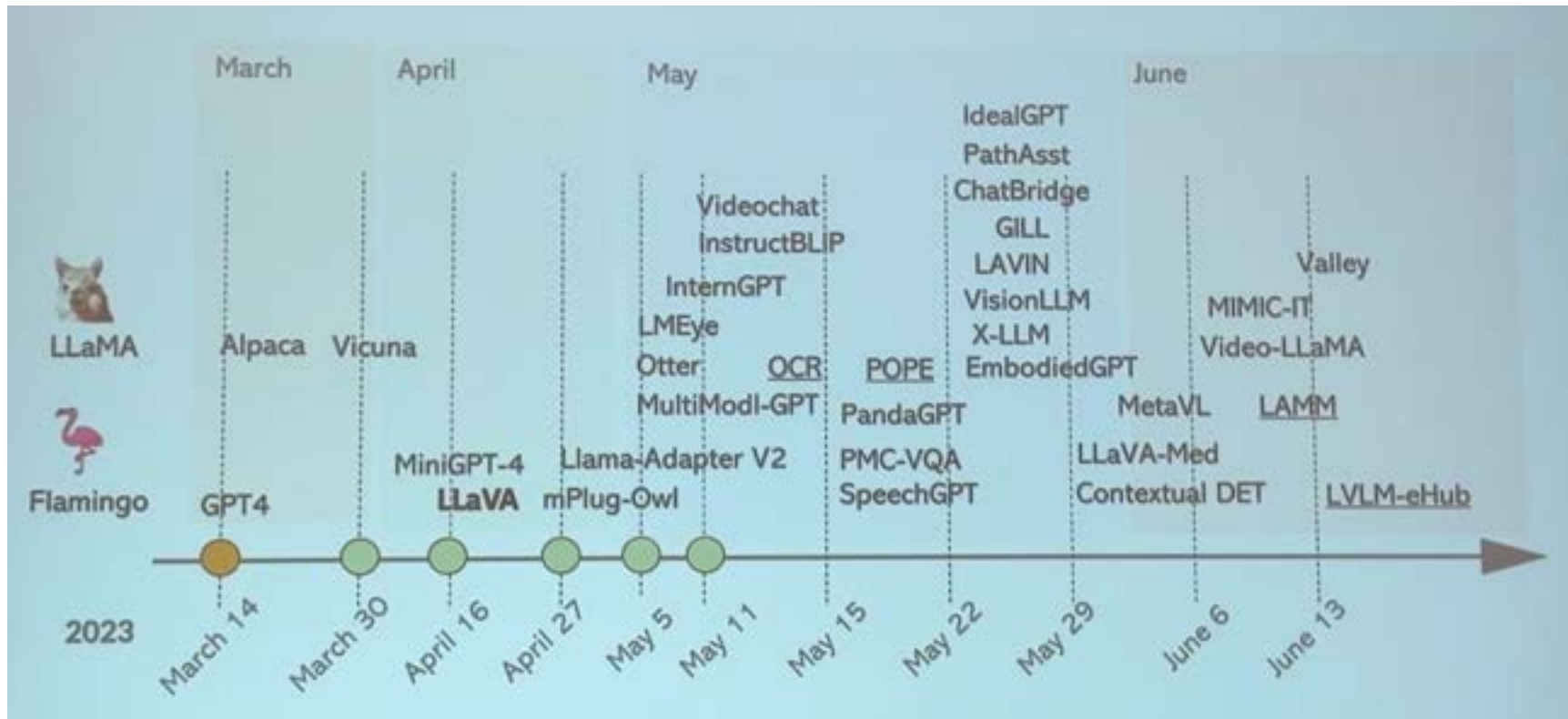
这个区域的名字叫什么? 请简单介绍 Translation: What is the name of this area? Please describe briefly.

这个区域的名字叫做“法国区” (French Quarter)。它是美国路易斯安那州新奥尔良市的一个历史悠久的地区, 以其独特的建筑风格、艺术、文化和历史而闻名。法国区是美国最古老的城市之一, 具有丰富的历史和文化遗产, 吸引了大量游客和居民。这个区域的街道上有许多花园和植物, 为游客提供了美丽的景色。

# CVPR 2024 の動向・気付き (34/116)

## WS Prompting in Vision

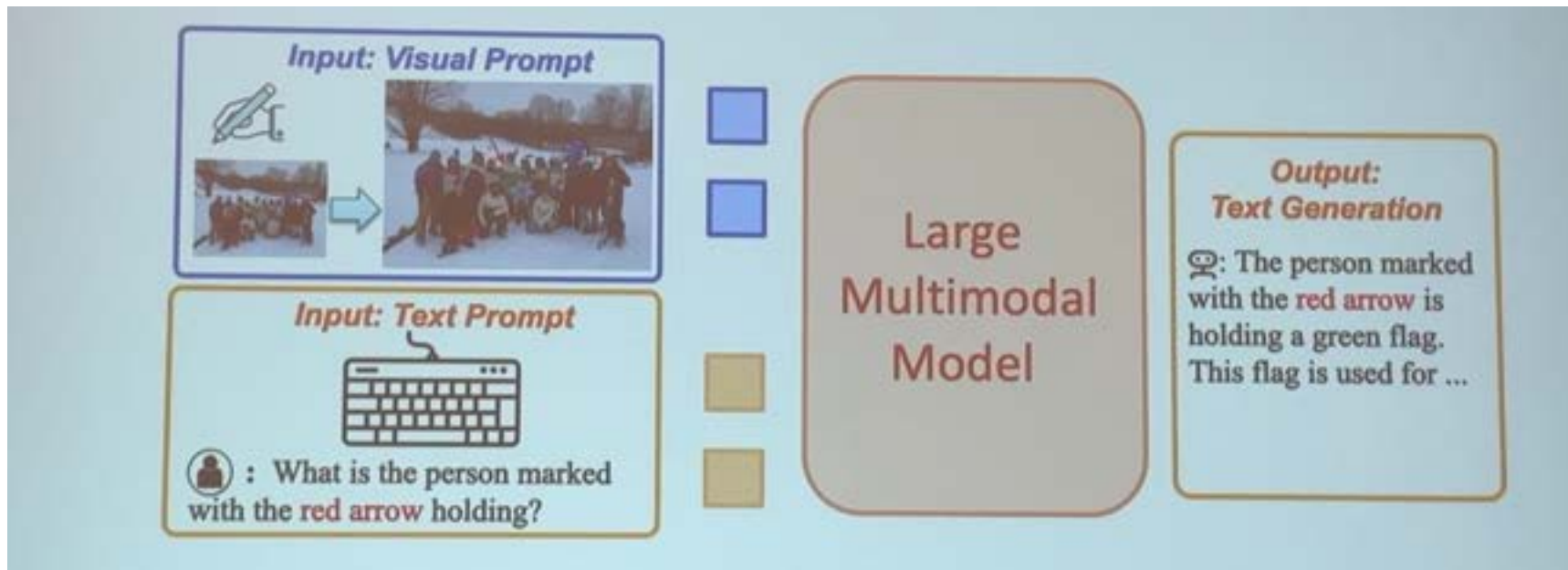
- Visual Prompting in Multimodal Model by [Yong Jae Lee](#) (3/5)
  - LLMの開発・改善に取り組むコミュニティは多く、近年の発展は著しい
    - GPT-4の登場がこの流れに火をつけた



## WS Prompting in Vision

- ❑ Visual Prompting in Multimodal Model by [Yong Jae Lee](#) (4/5)
  - ❑ 既存のLarge Multimodal LLM (LMM) について
    - ❑ 既存のLMMは画像理解において大きな役割を果たしている
    - ❑ LLMに領域やvisual prompt (矢印や手書きの線) を理解させることは可能か？
      - ❑ テキストで複雑な視覚的コンセプトを表現するのは難しい

→ ViP-LLaVA の開発



## WS Prompting in Vision

- ❑ Visual Prompting in Multimodal Model by [Yong Jae Lee](#) (5/5)
  - ❑ Visual Promptingに関する論文紹介
    - ❑ Visual Instruction Inversion: Image Editing via Visual Prompting
      - ❑ InstructPix2Pixからインスパイア
    - ❑ Segment Everything Everywhere All at Once
      - ❑ テキストと視覚的な入力をプロンプトとする汎用的なセグメンテーション
    - ❑ Looking Forward: Multimodal AI Agents
      - ❑ LLMベースのエージェントに関するサーベイ論文から紹介
      - ❑ 自身の行動の振り返り、ツールの使用、計画、他エージェントとの協力が可能
    - ❑ PIVOT: Iterative Visual Prompting Elicits Actionable Knowledge for VLMs
      - ❑ Visual Promptingをロボットの学習に適用
  - ❑ まとめ
    - ❑ 視覚的なコンセプトをテキストで表すのは難しい
    - ❑ Visual PromptはLMMにおける画像理解/生成のコントロール性を高める

## WS Prompting in Vision

### ❑ Old Lessons on Legible Prompts by [Phillip Isola](#) (1/7)

#### ❑ Promptingの新規性

❑ システムへの入力を介してシステムをコントロール することが可能

❑ システムが明示的にコントロール可能なものとしてデザインされていなくても

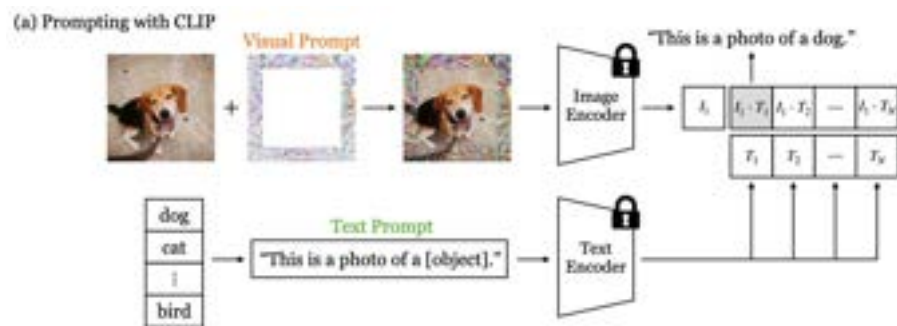
❑ 忘れられている昔の考えは？

❑ システムを明示的にコントロール可能なものとしてデザイン すること

#### ❑ 最近のプロンプティングシステムの例

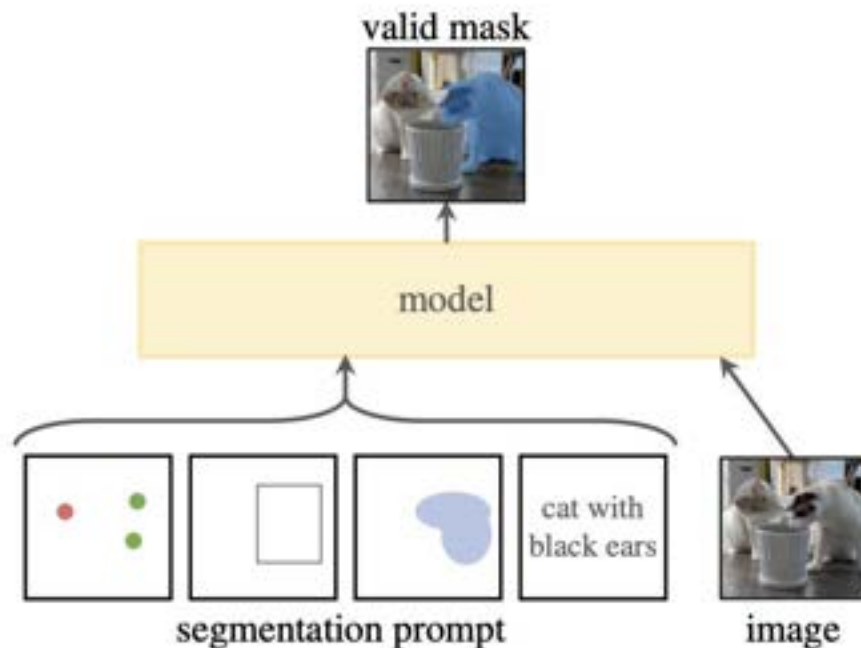
❑ Exploring Visual Prompts for Adapting Large-Scale Models (2022)

❑ Visual Prompting via Image Inpainting (2022)



## WS Prompting in Vision

- ❑ Old Lessons on Legible Prompts by [Phillip Isola](#) (2/7)
    - ❑ 現代のpromptシステムは物体検出などの複数のブランチの統合によってできている
      - ❑ 人間には理解しづらい
        - ❑ Segment Anything Model (SAM) のように人間が理解しやすいpromptを利用することで、システムの学習やデザインが簡単に行える
- 昔のシステムから現代のpromptingにも通じる考え方を学ぼう！



**SAMのprompt**  
点・ボックス・マスク・テキスト



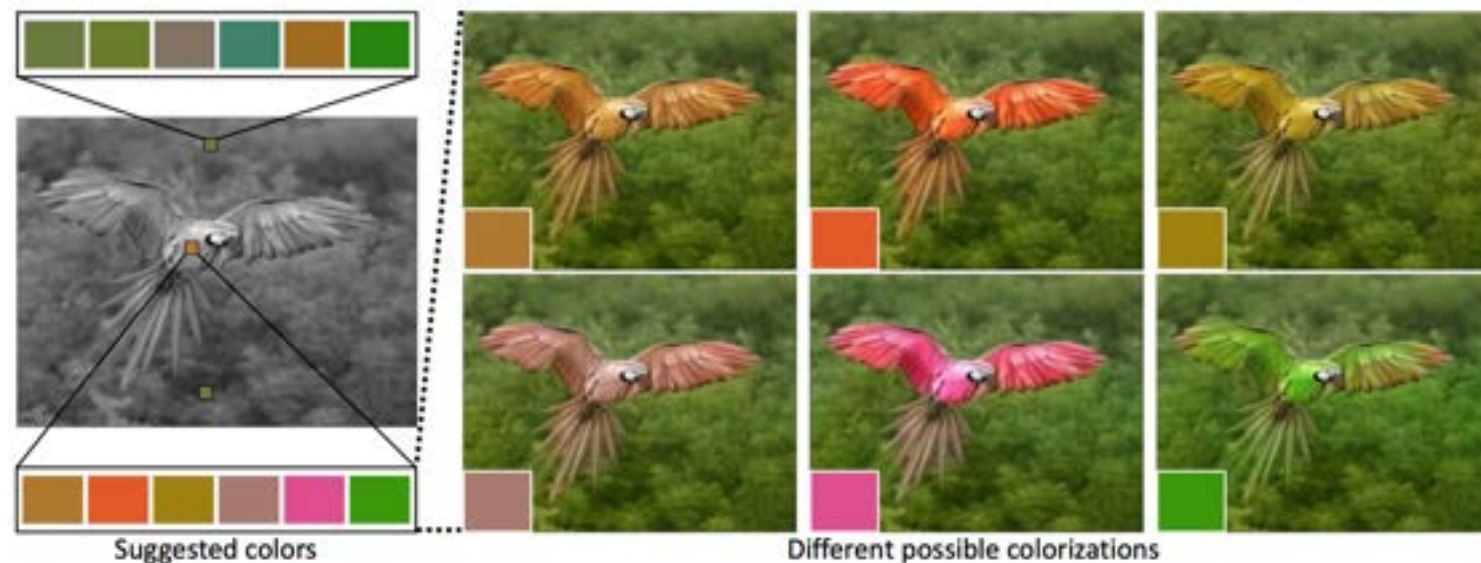
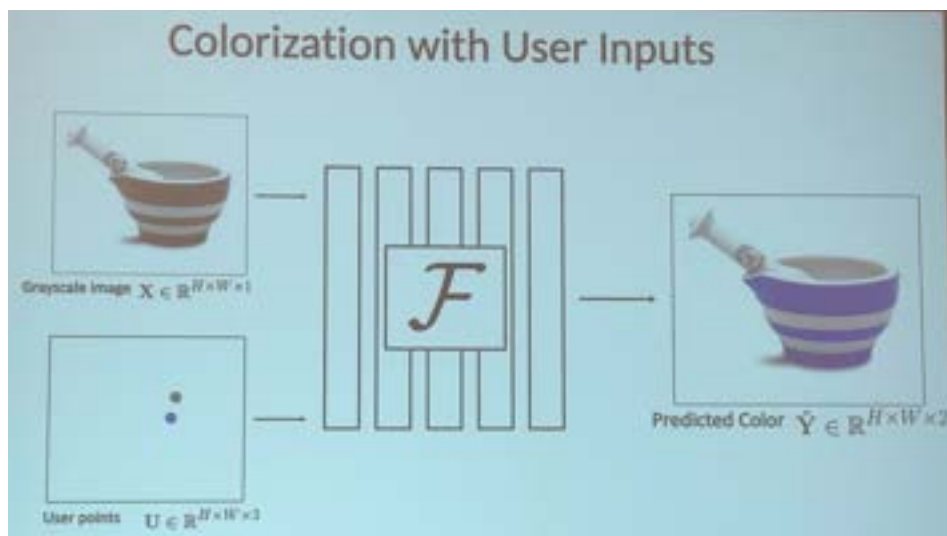
## WS Prompting in Vision

### ❑ Old Lessons on Legible Prompts by [Phillip Isola](#) (3/7)

#### ❑ user-guided colorization

❑ ユーザーが画像と色を入力し、システムはユーザーの入力と近傍ピクセルの色で補完

❑ 関連論文: Real-Time User-Guided Image Colorization with Learned Deep Priors



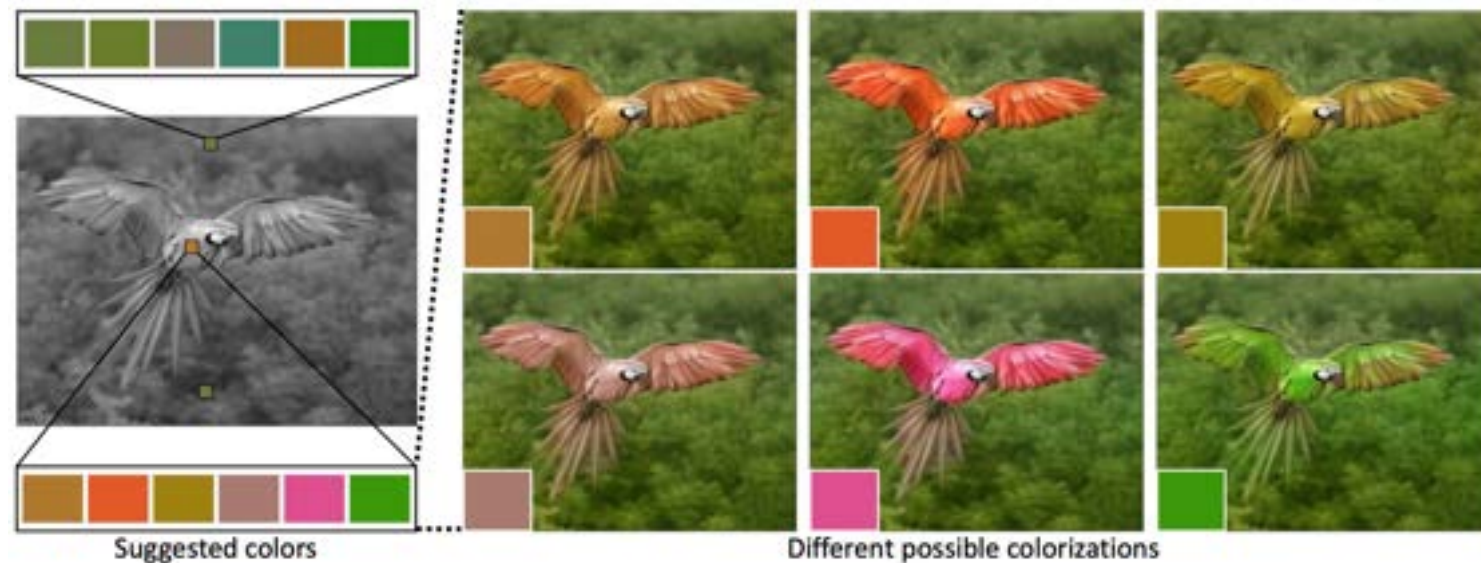
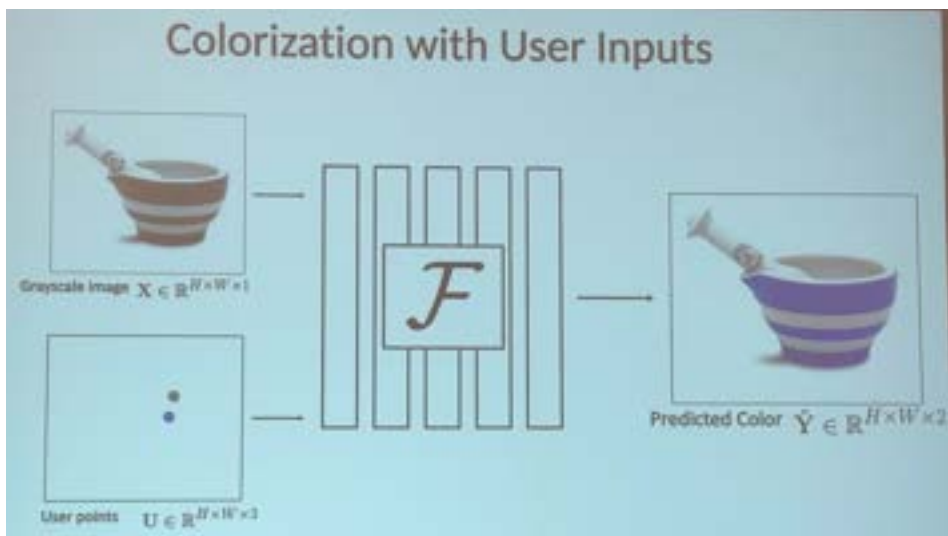
❑ 教訓1: Promptの影響は直感的で予測可能なものであるべき

## WS Prompting in Vision

### ❑ Old Lessons on Legible Prompts by [Phillip Isola](#) (4/7)

#### ❑ user-guided colorization

- ❑ ユーザーがクリックする場所 (User Points) は獲得しづらい
  - ❑ → ランダムにシミュレーション



- ❑ 教訓2: ユーザーがシステムとインタラクションするのと近い方法でデータを取得(シミュレート)する

## WS Prompting in Vision

### ❑ Old Lessons on Legible Prompts by [Phillip Isola](#) (5/7)

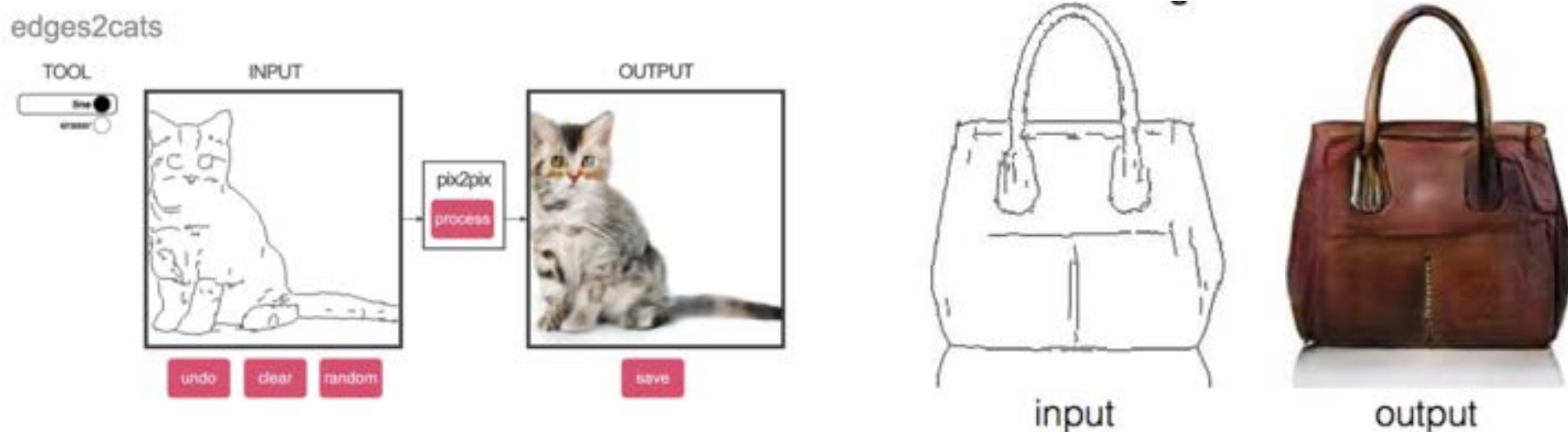
#### ❑ data-driven recommendations

❑ 類似データの統計から色を提案

❑ **edges2cats** : ユーザーのスケッチから色を予測

❑ シンプルな処理ゆえ、目を何個も書いた場合でも目であると認識されて塗りつぶされてしまう

❑ pix2pix: 画像からエッジを取得することでスケッチベースの色塗り用のデータを作成



❑ **教訓3: ユーザーが良いpromptを設定できるようなガイドを提供すべき**

## WS Prompting in Vision

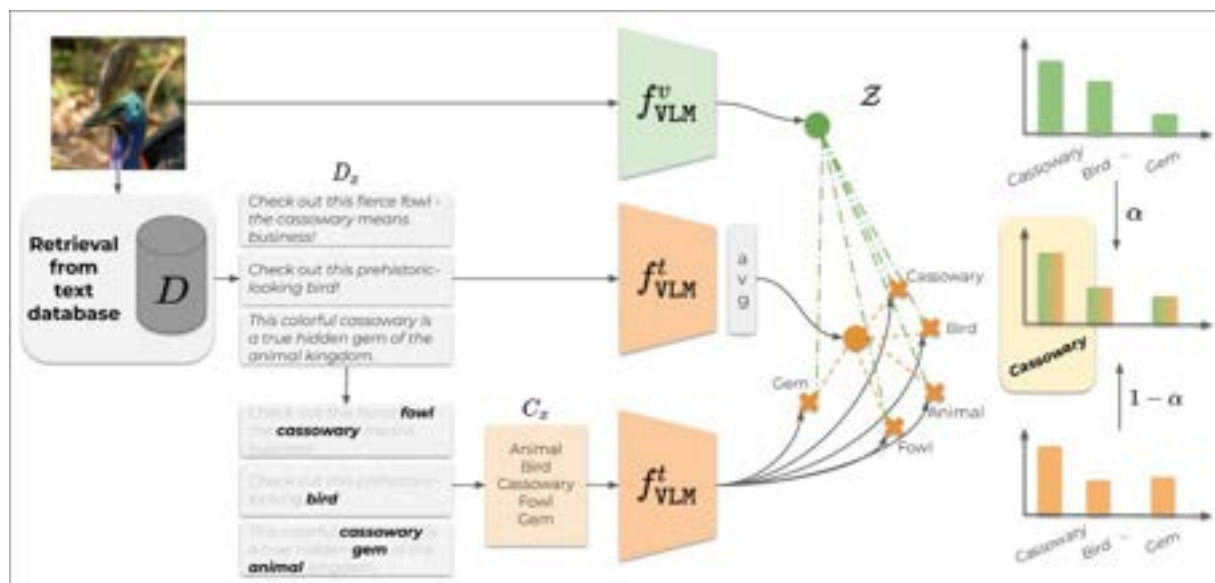
- ❑ Old Lessons on Legible Prompts by [Phillip Isola](#) (6/7)
    - ❑ **Word2Vec**
      - ❑ 言語埋め込みは単語レベル
    - ❑ **GANのコントロール性**
      - ❑ “zoom”, “shift”, “brighten”, “darken”のような単語が上手く反映されない場合がある
    - ❑ **最近の画像生成ツール**
      - ❑ promptを与える形式の生成ツールでは意図した結果になるとは限らない
        - ❑ “impossible”のような曖昧な単語を使うと宇宙を含むような画像になりがち
        - ❑ “make it more”といったpromptでは上手くいきづらい
- テキストは離散的
- ❑ **教訓4: promptはknobのように動作すべき(連続的・細かい操作が可能)**

## WS Prompting in Vision

- ❑ Old Lessons on Legible Prompts by [Phillip Isola](#) (7/7)
  - ❑ 教訓まとめ:
    - ❑ Promptの影響は直感的で予測可能なものであるべき
    - ❑ ユーザーがシステムとインタラクションするのと近い方法でデータを取得(シミュレート)する
    - ❑ ユーザーが良いpromptを設定できるようなガイドを提供すべき
    - ❑ promptはknobのように動作すべき(連続的・細かい操作が可能)
  - ❑ 個人的に考える今後のpromptingの方向性
    - ❑ ユーザーのvisual promptに対応可能なモデルの開発
      - ❑ 物体の指定や編集後の輪郭の入力など
    - ❑ 編集指示を数値やスライダーで連続的に反映

## WS 7th Multimodal Learning and Applications ([Massimiliano Mancini](#))

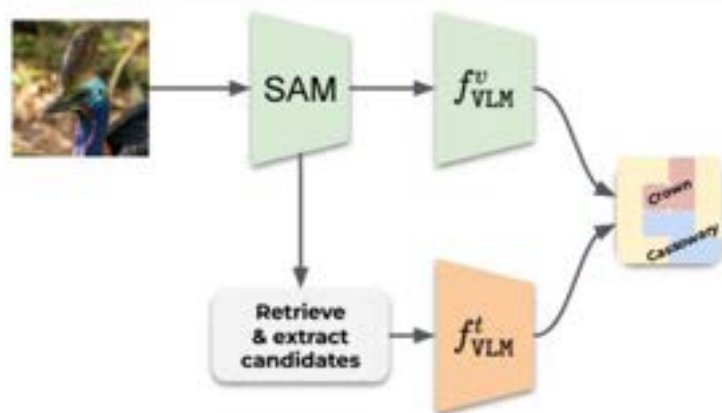
- ❑ Title: Opening multimodal doors with language: from recognition to bias detection
- ❑ Door 1: LLMのsemantic spaceをVisionタスクへの応用 (1/4)
- ❑ Fine-grained認識タスクへの応用
  - ❑ 論文: vocabulary-free image classification
  - ❑ メインアイデア: Unconstrained language-induced semantic spaceから、画像に物体クラスのラベルを推定するタスクと手法CaSEDの提案。CaSED(下図)は画像からテキストをretrievalして、その上テキストからクラスラベルを抽出する。最後に同じVLMで画像と最もマッチングされたクラスラベルを入力画像に与えて物体識別を行う。



# CVPR 2024 の動向・気付き (45/116)

## WS 7th Multimodal Learning and Applications ([Massimiliano Mancini](#))

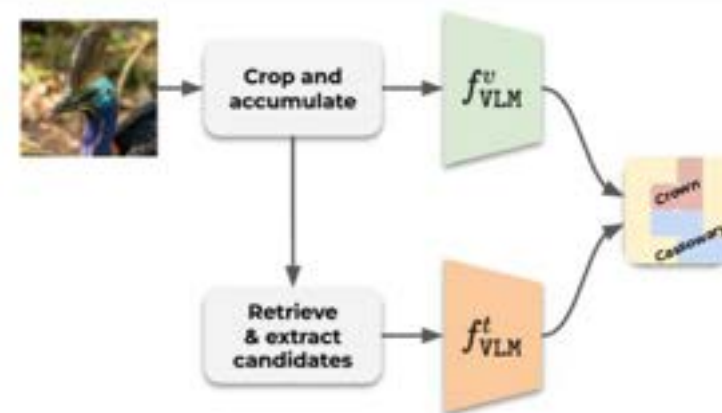
- ❑ Title: Opening multimodal doors with language: from recognition to bias detection
- ❑ Door 1: LLMのsemantic spaceをVisionタスクへの応用 (2/4)
- ❑ Segmentationタスクへの応用
  - ❑ 論文: vocabulary-free image classification and semantic segmentation
  - ❑ メインアイデア: CaSEDとclass-agnostic segmentationモデルと結合して、Vocabulary-freeなsemantic segmentationを行う。class-agnostic segmentationモデルにより画像をsegmentし、CaSEDでsegmentした領域ごとのラベルを付与する。特にSAMモデルと相性が良さそう(下図(a))。



(a) SAM-based method



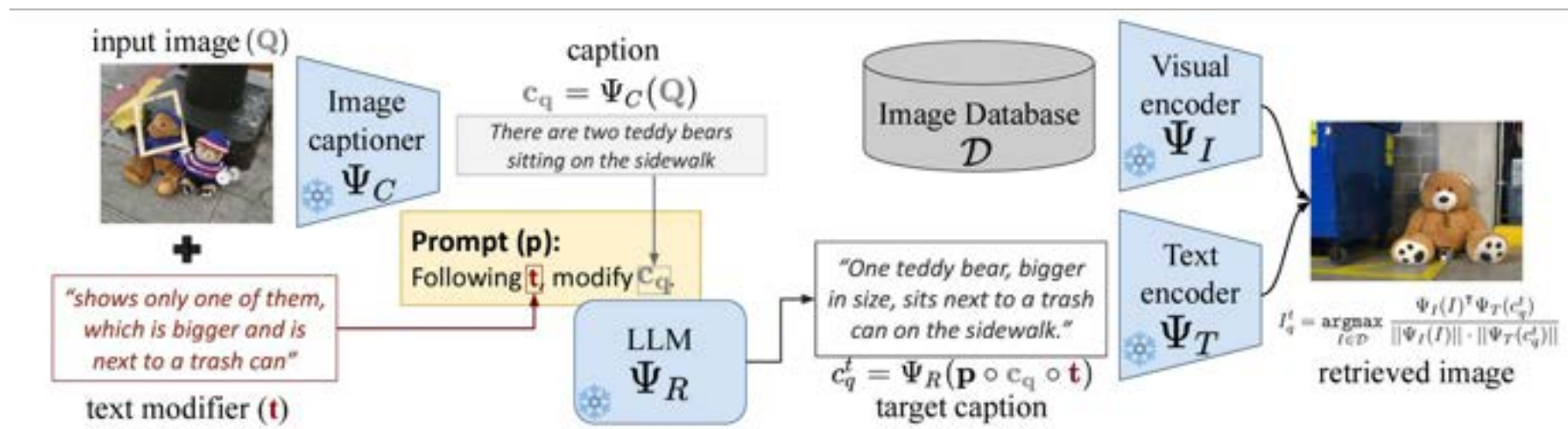
(b) SAN-based method



(c) DenseCaSED

## WS 7th Multimodal Learning and Applications ([Massimiliano Mancini](#))

- ❑ Title: Opening multimodal doors with language: from recognition to bias detection
- ❑ Door 1: LLMのsemantic spaceをVisionタスクへの応用 (3/4)
- ❑ Retrievalへの応用
  - ❑ 論文: Vision-by-language for training-free compositional image retrieval
  - ❑ メインアイデア: 画像とtext modifierから、text modifierを反映した類似した画像をretrievalするタスクと手法 ZS-CIR(下図)の提案。ZS-CIRは画像からVLMによりCaptionを得て、それをベースに text modifierをさらに修正し、最後に修正された後のテキストより画像をretrievalする。ZS-CIRがゼロショットで学習のモデルよりも高い SoTA性能を得た。

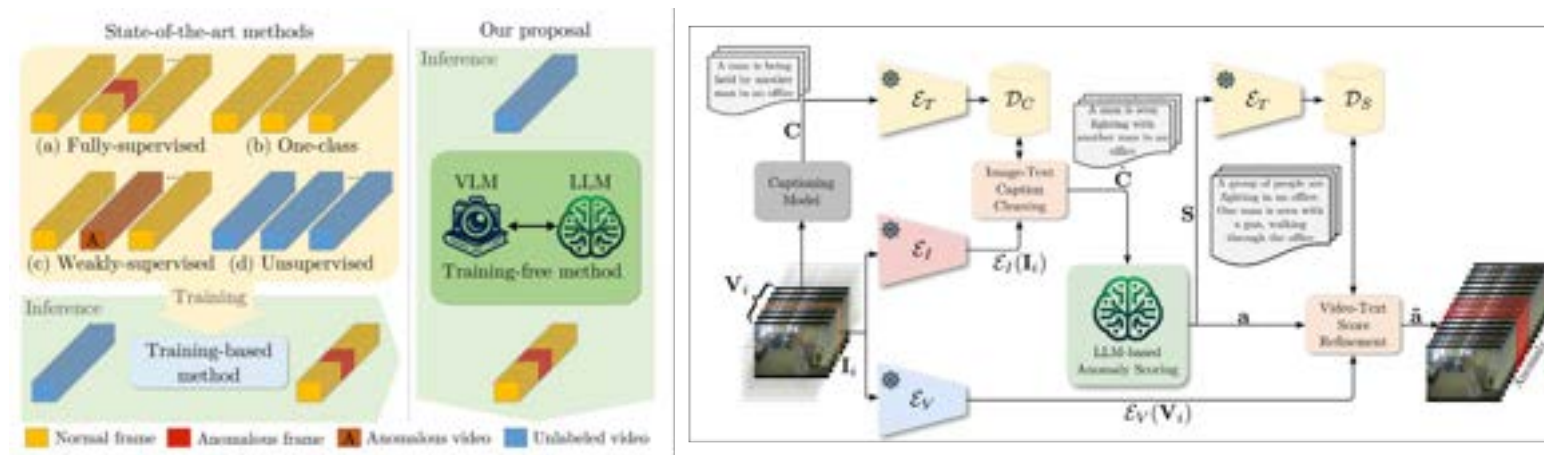




# CVPR 2024 の動向・気付き(47/116)

## WS 7th Multimodal Learning and Applications ([Massimiliano Mancini](#))

- ❑ Title: Opening multimodal doors with language: from recognition to bias detection
- ❑ Door 1: LLMのsemantic spaceをVisionタスクへの応用 (4/4)
- ❑ Video anomaly detectionへの応用
  - ❑ 論文: Harnessing large language models for training-free video anomaly detection
  - ❑ メインアイデア: ビデオからTraining-freeでanomalyを検出するタスクおよび手法を提案。
  - ❑ 手法: VLMとLLMを結合し3段階で行なっている。ステップ1では、画像キャプションできるVLMでビデオから複数枚の画像をキャプションする。その上画像とテキストの類似度でキャプションをクリーニングする。ステップ2ではLLMを利用し、クリーニングされたキャプションから粗末なAnomaly検出を行う。ステップ3では、ビデオとテキストの類似度を評価可能なVLMでAnomalyをリファインする。



## WS 7th Multimodal Learning and Applications ([Massimiliano Mancini](#))

- ❑ Title: Opening multimodal doors with language: from recognition to bias detection
- ❑ Door 2: LLMをanalyze modelsへの応用
- ❑ LLMをbias detectionへの応用
  - ❑ 論文: OpenBias: Open-set Bias Detection in Text-to-Image Generative Models
  - ❑ メインアイデア: Text-to-Imageモデルのバイアス(下図)をOpen-setで検出するタスク・手法OpenBiasの提案。
  - ❑ 手法: LLMからテキストに含まれる可能なバイアスの項目をOpen-setでまず認識し、その後生成モデルでText-to-Imageを実行する。最後に、VQAモデルで生成された画像群に関してあらゆるバイアスの項目でバイアスを検出・定量的評価を行う。



## WS 7th Multimodal Learning and Applications ([Massimiliano Mancini](#))

- ❑ Title: Opening multimodal doors with language: from recognition to bias detection
- ❑ Take-home message
  - ❑ Multimodal LLMs / LLMsは豊かな情報(私たちが知らない情報まで)を持つ
    - ❑ without constraintsでモデルをデザインすべき
  - ❑ Visionと言語をリンクさせることで、semantic hypothesis が得られる
    - ❑ Transparent predictionsを行うモデルデザインが望ましい
  - ❑ LLM reasoningを使って、Biases / failuresを分析可能
    - ❑ モデル設計時、Evaluationを行う際に、FormulatedされたQuestionsを使う(例えば、visual programmingなど)と有利

# CVPR 2024 の動向・気付き (50/116)

## First Joint Egocentric Vision (EgoVis) Workshop

- 一人称視点映像分析コミュニティの議論の活性化のための共同ワークショップ
- データセット別で行われてきたWSを1つに統合し、共同でチャレンジを実施
  - EvalAIを用いた31のチャレンジ
  - 34本のポスター発表(本会議の招待含む)
- keynote: 社会的インタラクション (Jim Rehg)、3次元理解 (Diane Laurus)、幼児の学習 (Frank & Long)、表情計測 (Fernando De La Torre)



八木がWS委員として運営に参加

## First Joint Egocentric Vision (EgoVis) Workshop

- オーガナイザ(実行機関)の裏方にEgoVis Board(アドバイザ機関)を設置
- 各チャレンジの運営役も別途選定することで多数のチャレンジを実施可能に
- 多数のポスターをinviteしチャレンジにて最高性能を達成するための知見を収集→アイデアの深化と精度向上を両立

EgoVis Organizer



Antonino Furnari  
University of Catania



Angela Yao  
National University of Singapore



Xin Wang  
Microsoft Research



Tushar Nagarajan  
FAIR, Meta



Huiyu Wang  
FAIR, Meta



Jing Dong  
Meta



Jakob Engel  
FAIR, Meta



Siddhant Bansal  
University of Bristol



Takuma Yagi  
National Institute of Advanced  
Industrial Science and Technology



Dima Damen  
University of Bristol



Jitendra Malik  
UC Berkeley



Yoichi Sato  
University of Tokyo

EgoVis Board



Giovanni Maria Farinella  
University of Catania



Richard Newcombe  
Reality Labs Research



Kristen Grauman  
UT Austin



Marc Pollefeys  
ETH Zurich



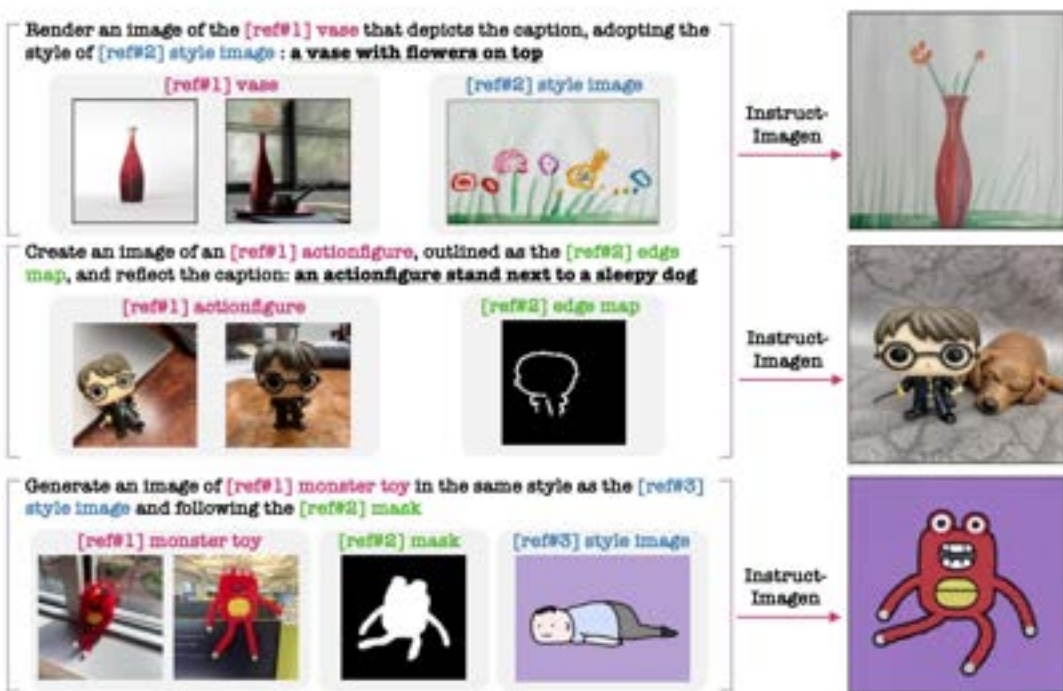
David Crandall  
Indiana University

## EgoVis 2022/2023 Distinguished Paper Awards

- 一人称ビジョンコミュニティの発展に貢献した10本の論文をWSで表彰
- 手順理解/音声理解などのデータセットのほか、EgoVLPといった汎用基盤モデルに関する論文が選定
- Sener et al. Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities. CVPR 2022
- Li et al. Ego-Body Pose Estimation via Ego-Head Pose Estimation. CVPR 2023
- Grauman et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video. CVPR 2022
- Lin et al. Egocentric Video-Language Pretraining. NeurIPS 2022
- Khirodkar et al. EgoHumans: An Egocentric 3D Multi-Human Benchmark. ICCV 2023
- Pramanick et al. EgoVLPv2: Egocentric Video-Language Pre-training with Fusion in the Backbone. ICCV 2023
- Huh et al. EPIC-SOUNDS: A large-scale dataset of actions that sound. ICASSP 2023
- Wang et al. HoloAssist: An Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. ICCV 2023
- Zhao et al. Learning Video Representations from Large Language Models. CVPR 2023
- Chen et al. SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning. NeurIPS 2022

## WS Tool-Augmented Vision (1/21)

- ❑ マルチモーダルな指示を反映した画像生成 by [Xuhui Jia](#)
  - ❑ Instruct-Imagen: Image Generation with Multi-modal Instruction
    - ❑ 画像・マスク・エッジなどを参照し、それらの情報を反映した画像を生成
  - ❑ Subject-driven Text-to-Image Generation via Apprenticeship Learning
    - ❑ 3~5枚の入力画像から、最適化なしで被写体を含む新たな画像を生成



## WS Tool-Augmented Vision (2/21)

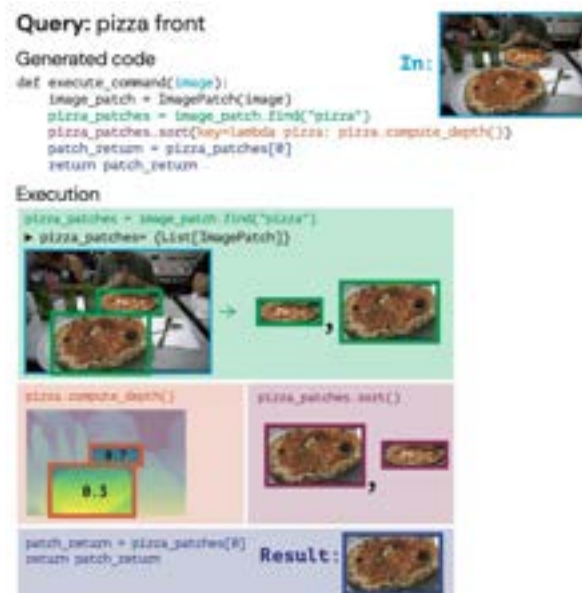
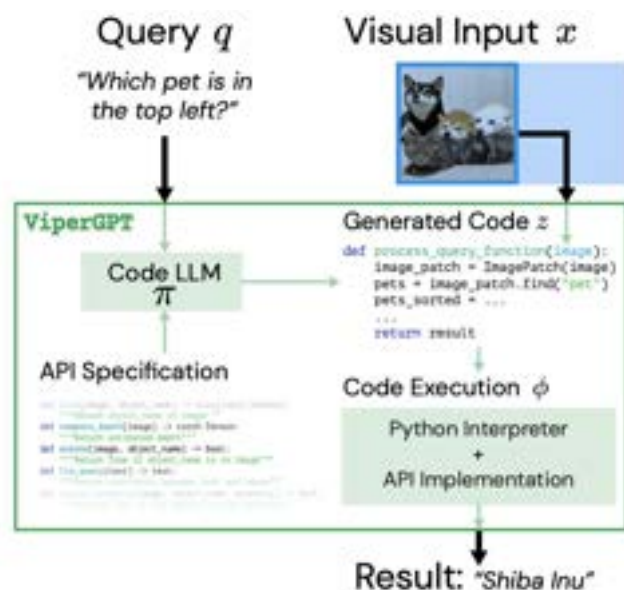
### □ 言語モデルと視覚モデル by [Sachit Menon](#) (1/3)

#### □ End-to-Endモデルの課題

- 計算や構造を掴むことが苦手、解釈性が低い、ハルシネーション
- 例えばBLIP-2では、物体の個数を正確にカウントすることができない
- → 複雑なタスクはサブタスクに分割して解く

#### □ ViperGPT: Visual Inference via Python Execution for Reasoning

- コード生成LLMによって、クエリごとにカスタマイズされたプログラムを生成
- 複雑な視覚タスクを、既存のモジュールの利用によって段階的に解く





## WS Tool-Augmented Vision (3/21)

- ❑ 言語モデルと視覚モデル by [Sachit Menon](#) (2/3)
  - ❑ Multimodal LLMは言語・論理・プログラムベースの推論において強力
    - ❑ → 視覚的な要素が強いタスクにもMLLMは適用可能か？
  - ❑ visual thinking タスクへの Whiteboard-of-Thought の適用
    - ❑ コードによって質問内容を可視化し、可視化した結果を参照することで質問に回答

which lowercase letter is a circle with a vertical line touching it to the right?

Whiteboard-of-Thought


```
Visualization Code
import turtle

# Create a turtle object
pen = turtle.Turtle()
# Draw the circle
pen.penup()
pen.goto(0, -50) # Move to the starting position
pen.pendown()
pen.circle(50) # Draw a circle with radius 50

# Draw the vertical line
pen.penup()
pen.goto(50, 0) # Move to the right edge of the circle
pen.pendown()
pen.goto(50, -100) # Draw the vertical line downwards
```


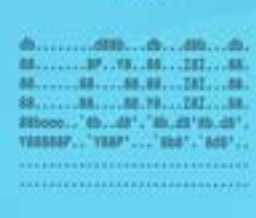
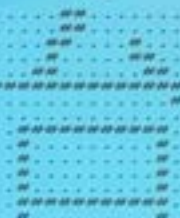


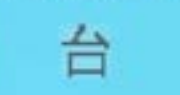
Python Interpreter

Visualization Image



Answer: "q"

Evaluation: ASCII understanding

	MNIST	Words	Kanji
Query			
WoT			

	MNIST	Word	Kanji
Direct	19.6	24.8	1.1
CoT	21.6	27.2	1.1
WoT (ours)	66.0	66.4	73.8

# CVPR 2024 の動向・気付き (56/116)

## WS Tool-Augmented Vision (4/21)

- 言語モデルと視覚モデル by [Sachit Menon](#) (3/3)
  - Multimodal LLMは言語・論理・プログラムベースの推論において強力
    - → 視覚的な要素が強いタスクにもMLLMは適用可能か？
  - Whiteboard-of-Thought は Spatial navigation などにも適用可能
    - 言葉の位置関係を視覚化することで回答を得る

### Spatial navigation

**Query:**  
You have been given a pointy-topped regular hexagonal tile map consisting of 2 rows, where the first row has one tile and the second row has two tiles.

Starting from a vertex, you will move along the edges of these tiles.

Initially, you are positioned at the top corner of the map, where you find a soup bowl. You move down-right by one step, where you find a Chihuahua. You move down by one step, where you find a dome. You move down-left by one step, where you find a peacock. You move down-right by one step, where you find a Norwich Terrier. You move down-right by one step, where you find a barrel. You move up-right by one step, where you find a sulphur-crested cockatoo. You move up by one step, where you find an American robin. You move up-left by one step.

What do you find at the final position?

**Whiteboard-of-Thought**

**Answer: "dome" ✓**

	Non-Grid Structures			2D Grids		Overall
	Circle	Hexagon	Triangle	Square	Rhombus	
Direct	14	3	16	66	63	33
CoT	25	8	26	98	51	42
WuT (ours)	41	61	55	50	52	52

**Query:**

```

G
RE
ETIN
GS
E
A R
T H
OF
WHICH
I AM THE E
LOQU
ENT TONGUE
WHICH POKES
FROM YOUR MOUTH
O PARIS
AND WILL POKE
FOR EVER
A T THE
GER MANS

```

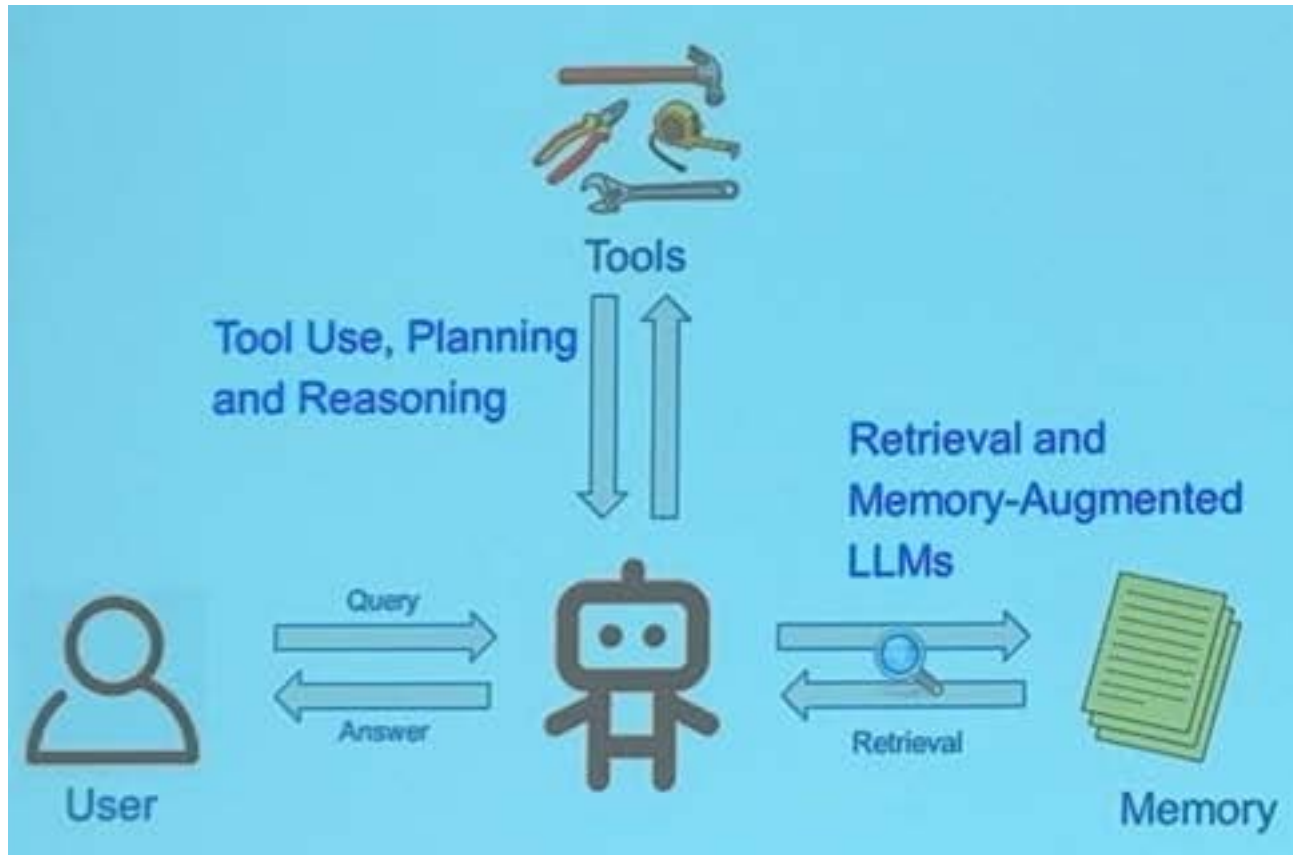
**Chain-of-Thought**  
The poem is arranged in a way that creates a visual image. The lines are centered and vary in length, creating a shape that resembles a tongue or a speech bubble.

**Whiteboard-of-Thought**  
The visual layout of the poem is not just a random arrangement of words but a deliberate design to evoke the image of the Eiffel Tower, symbolizing Paris. This adds a layer of meaning to the poem, connecting the text to the iconic structure and the cultural significance of Paris.

What is the visual meaning of this poem?

## WS Tool-Augmented Vision (5/21)

- Tool-use & Memory (Introduction) by [Cordelia Schmid](#)
  - ユーザーとAIのインタラクションはツールやメモリによって補完されている
  - タスクに応じて複数のシステムを組み合わせることで解釈性が高まる



Tool-use:

- LLMによるChain of Reasoning
- コード生成とAPIの利用
- LLMにも必須？

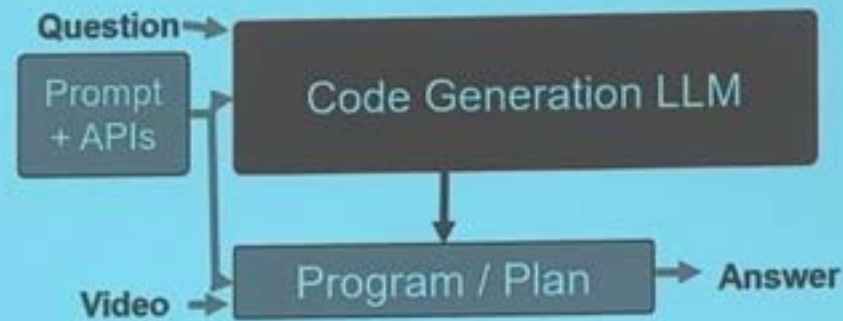
Memory:

- 詳細な記述の獲得
- レアなクラスに関する知識の獲得
- 効率的

## WS Tool-Augmented Vision (6/21)

- ❑ Video Question Answering with Reasoning by [Cordelia Schmid](#) (1/3)
  - ❑ 従来手法について
    - ❑ 動画から取得したフレームのサブセットを基盤モデルに入力してQAを実施
    - ❑ 質問(+API/具体例)をプログラムにマッピング: 全てを1ステージで行う
      - ❑ 複雑性が高く、質問とプログラムの対応を取るのが難しい
      - ❑ 言語の曖昧性に対して視覚的なデータで解決することができない
      - ❑ → モデルの出力や挙動の脆弱性につながる

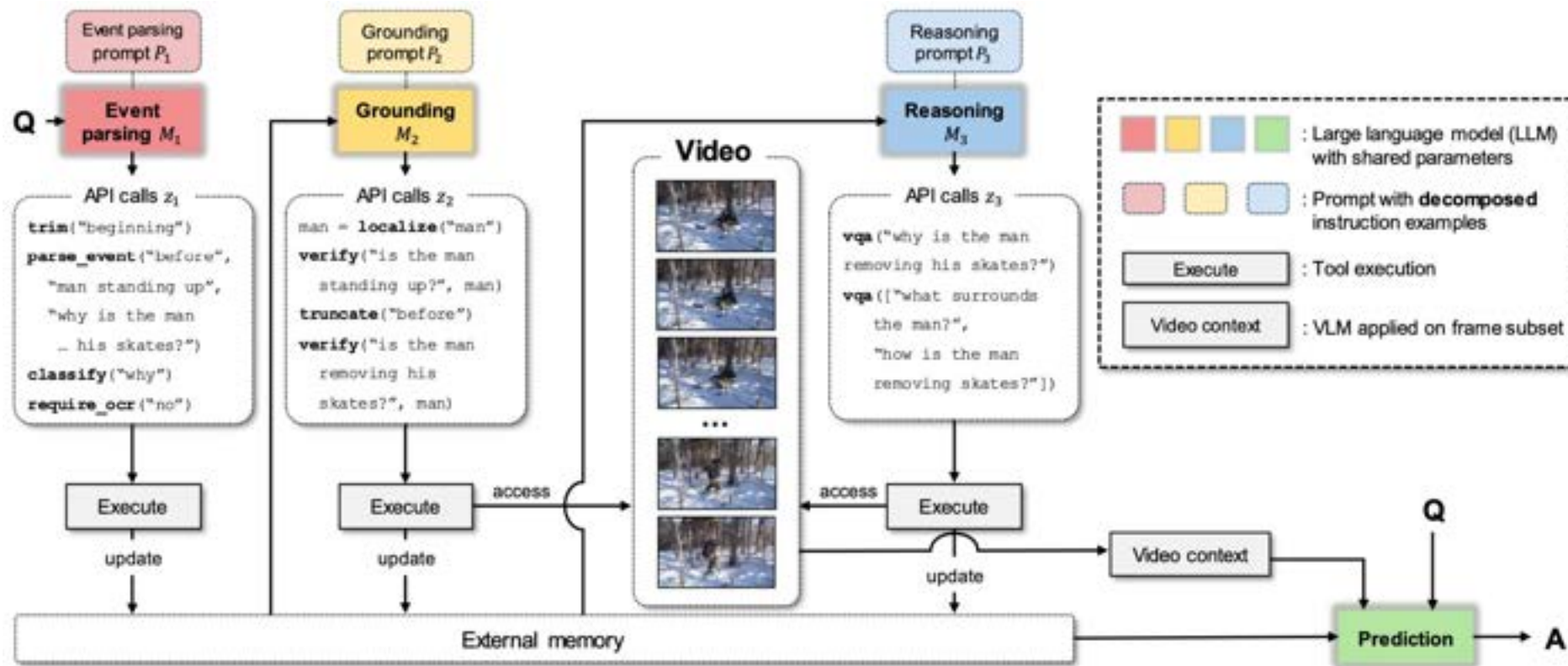
### VideoQA: Modular Models (Single-Stage)



Model has to map question (+ API/examples) to a full program in a single-stage

## WS Tool-Augmented Vision (7/21)

- Video Question Answering with Reasoning by [Cordelia Schmid](#) (2/3)
  - MoReVQA: Exploring Modular Reasoning Models for Video Question Answering
    - 学習不要なmulti-stageシステムを構築し、fine-tuningを要する手法を上回る精度を実現
    - 分解によってタスクを簡略化・ステージごとに解釈可能な中間出力を生成



## WS Tool-Augmented Vision (8/21)

- ❑ Video Question Answering with Reasoning by [Cordelia Schmid](#) (3/3)
  - ❑ Open questions
    - ❑ 長時間データセットの必要性
      - ❑ 既存のvideoデータセットはまだまだ短い
      - ❑ 既存のVideo QAは長時間のコンテキストの推論を必要としない
    - ❑ 動画内の要素をどのように関連付けるか、グラウンディングするか
      - ❑ ストーリー全体の解釈に必要
      - ❑ 時空間を同時にlocalizeする方法がない
      - ❑ 更に複雑な質問に回答するには、Tool-use + reasoning が必要
    - ❑ この分野における prompt engineering については改善の余地がある
      - ❑ 全てのデータに対して同一のpromptを使用している

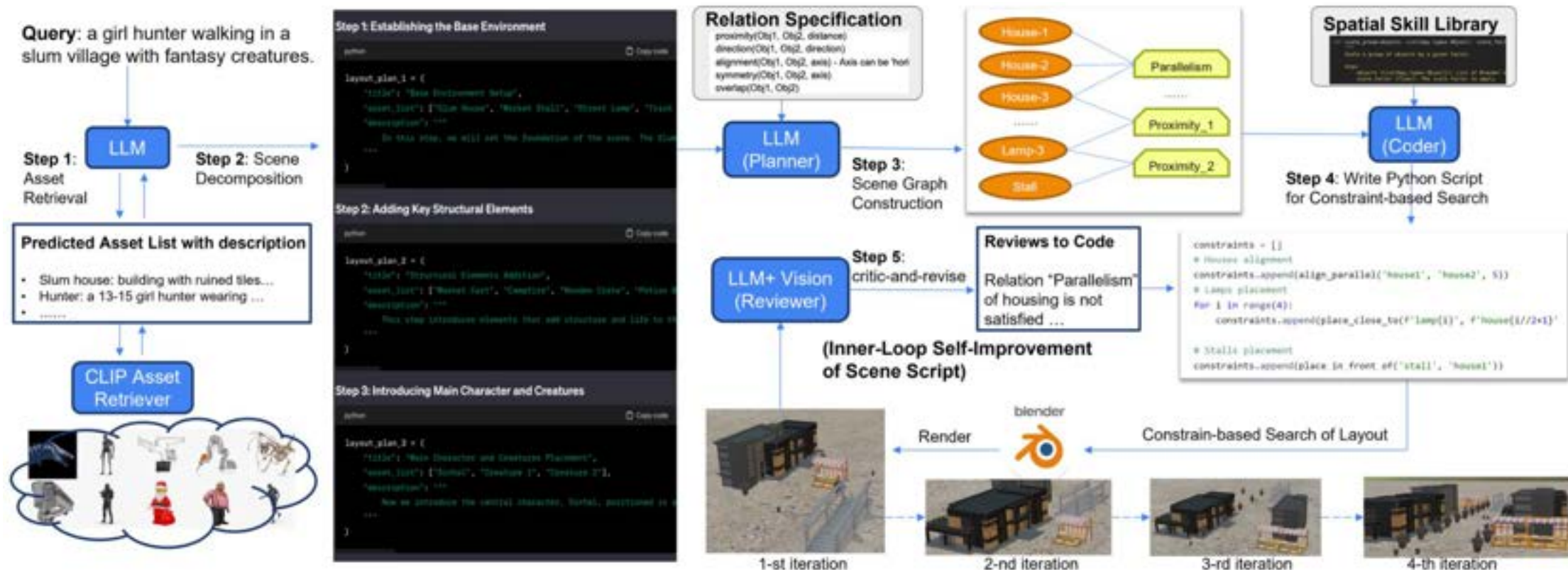
## WS Tool-Augmented Vision (9/21)

□ 3D scene synthesise via LLM generated Blender code by [Cordelia Schmid](#)

□ SceneCraft: 2つのループ構造を導入したシステムにより複雑な3Dシーンを生成

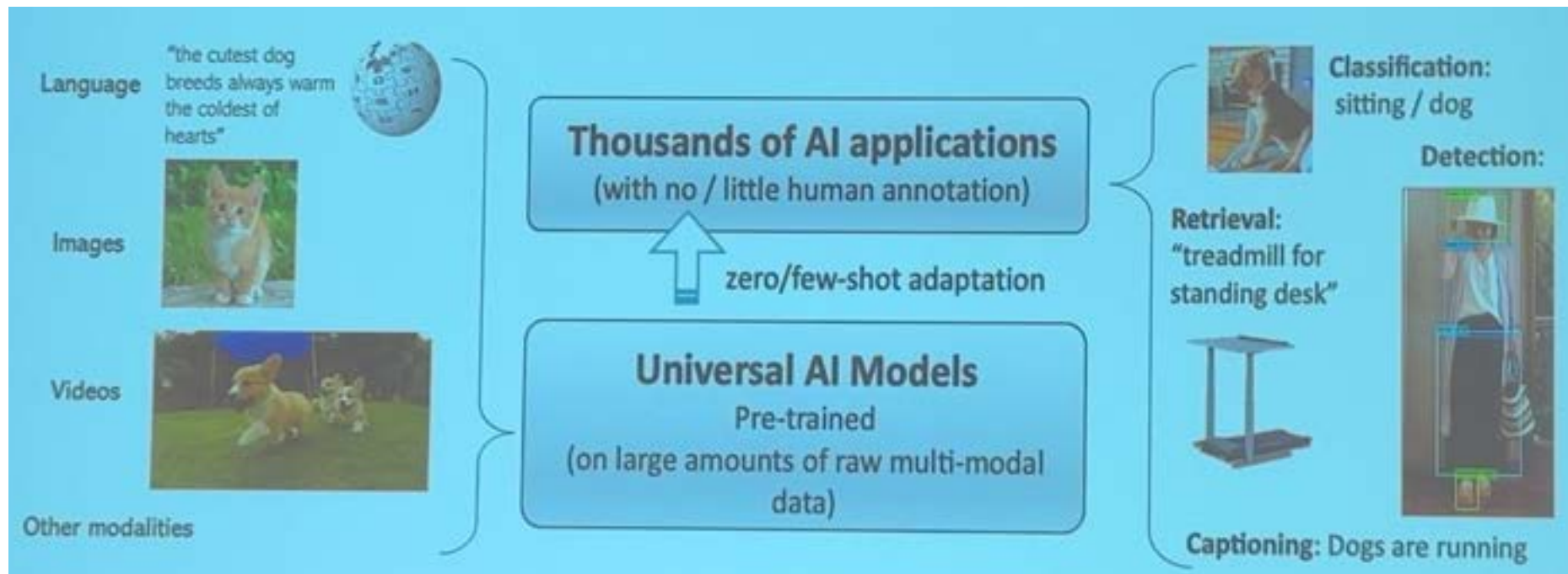
□ inner-loop: 画像生成 → 生成画像をGPT-Vで評価し、スクリプトを修正 → 画像生成

□ outer-loop: inner-loopの結果を基にスキルセットを拡張



## WS Tool-Augmented Vision (10/21)

- ❑ LLaVA An Open Large Multimodal Model by [Chunyuan Li](#) (1/8)
  - ❑ Computer Vision in the Wild
    - ❑ open-set/domainな画像認識とタスクレベルの効率的な転移学習に焦点を当てている
  - ❑ キーポイント
    - ❑ 転移可能なタスクの幅広さ・転移学習におけるコストの低さ

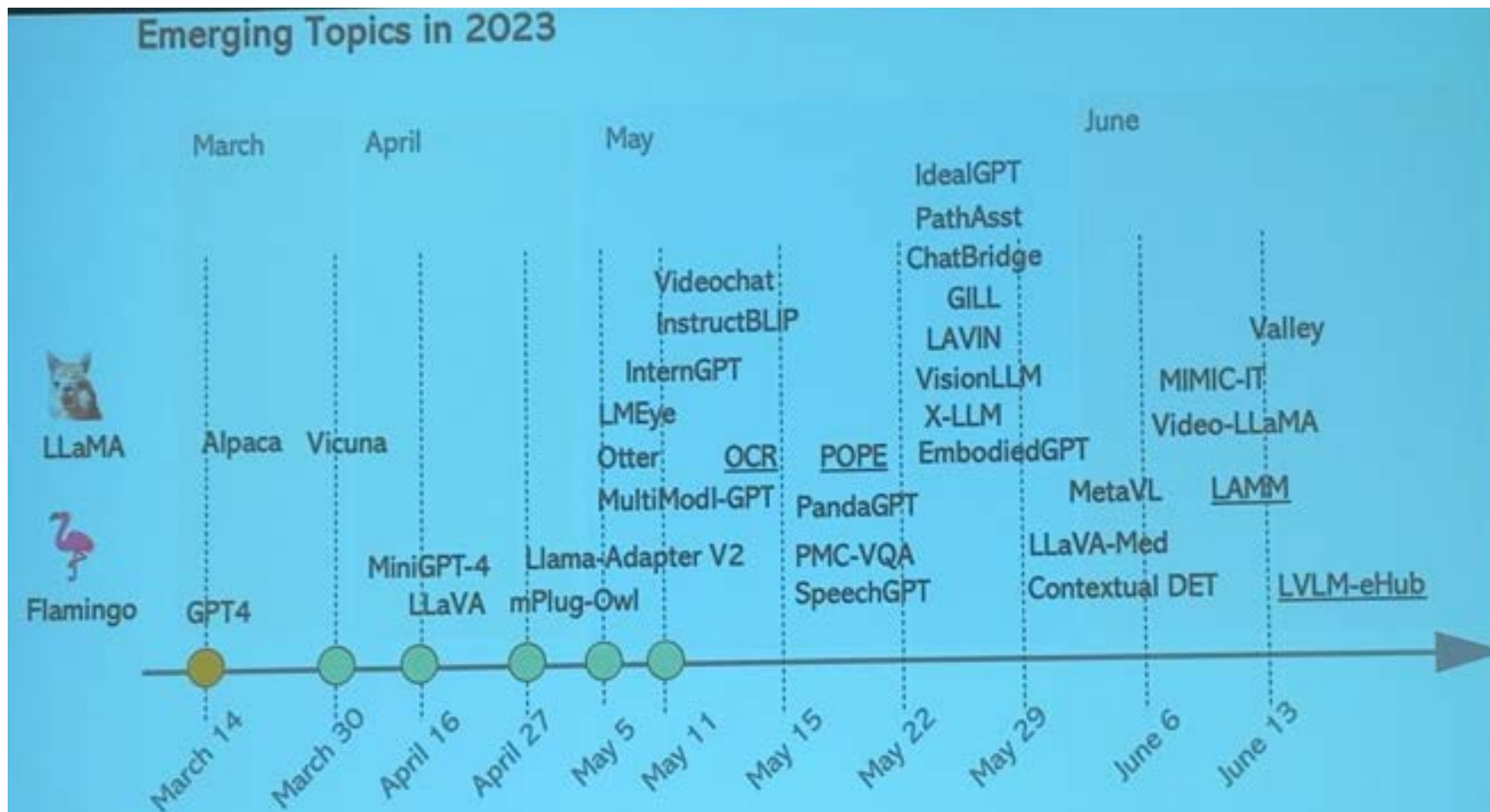




# CVPR 2024 の動向・気付き(63/116)

## WS Tool-Augmented Vision (11/21)

- LLaVA An Open Large Multimodal Model by [Chunyuan Li](#) (2/8)
  - GPT-4の登場以降、分野が急速に発展



## WS Tool-Augmented Vision (12/21)

### □ LLaVA An Open Large Multimodal Model by [Chunyuan Li](#) (3/8)

- Visual Instruction Tuning: **LLaVA** の提案 <https://llava-vl.github.io/>
  - マルチモーダルモデルに対するinstruction tuningの導入
  - GPTを用いたlanguage-image instruction-following dataの生成
- 2ステージの学習
  - 画像特徴と言語特徴のalignmentを取る事前学習: 射影行列のみ学習
  - 射影行列とLLMに対するEnd-to-Endのfine-tuning

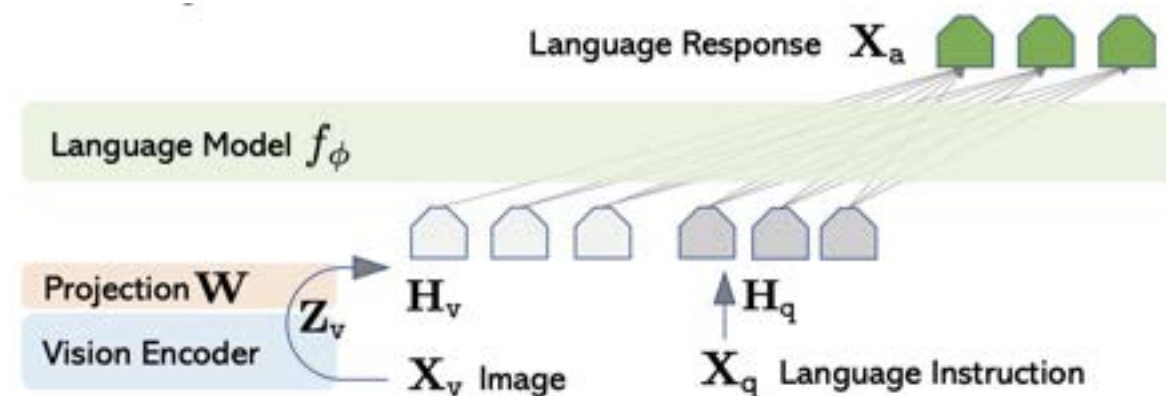


ICHIRAN Ramen [source]

Annotation A close-up photo of a meal at **ICHI-RAN**. The chashu ramen bowl with a spoon is placed in the center. The ramen is seasoned with **chili sauce**, **chopped scallions**, and served with **two pieces of chashu**. Chopsticks are placed to the right of the bowl, still in their paper wrap, not yet opened. The ramen is also served with **nori** on the left. On top, from left to right, the following sides are served: a bowl of **orange spice** (possibly garlic sauce), a plate of **smoke-flavored stewed pork with chopped scallions**, and a cup of **matcha green tea**.

Question 1 What's the name of the restaurant?

Question 2 Describe this photo in detail.



# CVPR 2024 の動向・気付き (65/116)

## WS Tool-Augmented Vision (13/21)

- LLaVA An Open Large Multimodal Model by [Chunyuan Li](#) (4/8)
  - LLaVA系列のモデルも次々と開発されている
  - 発展の方向性は、**Tool-Augmented Systems** と**End-to-End Models** に大別される
    - **Tool-Augmented Systems**
      - ○: 学習なし、低コスト、短い開発サイクル
      - ×: システムの性能向上における柔軟性に欠ける、新たな能力の発現が少ない
    - **End-to-End Models**
      - ○: 軽い学習で新たな能力の発現や新たなシナリオへの適応を実現できる
      - ×: 計算コスト、性能がLLMの性能によって制限される



## WS Tool-Augmented Vision (14/21)

### □ LLaVA An Open Large Multimodal Model by [Chunyuan Li](#) (5/8)

#### □ Tool-Augmented Systems: **LLaVA-Interactive**

- マルチモーダルなhuman-machine interactionを実現
- 特定のタスクに特化した3つの既存モデルを統合することで、汎用的なインタラクションを可能に



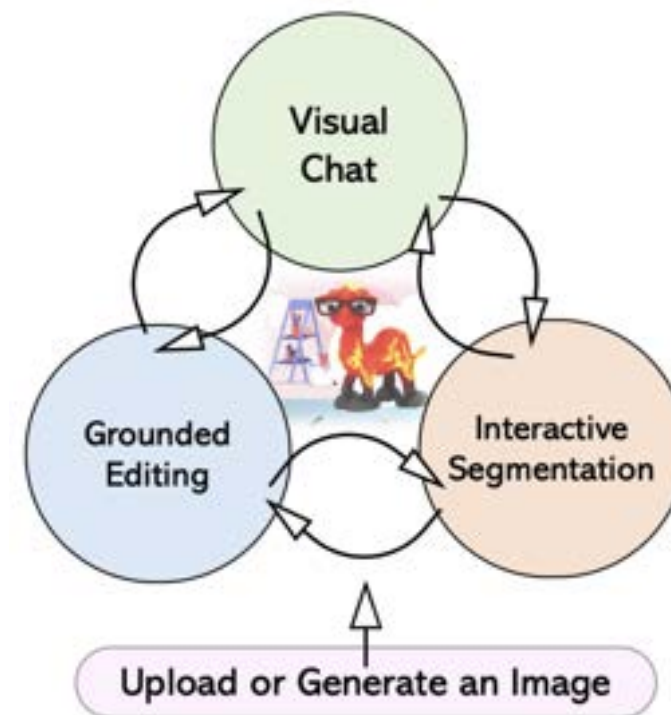
(b) Remove or Change Objects



(c) Inpaint New Objects



(d) Generate New Image



# CVPR 2024 の動向・気付き(67/116)

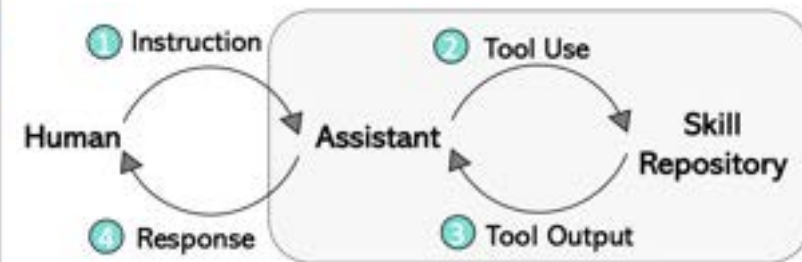
## WS Tool-Augmented Vision (15/21)

□ LLaVA An Open Large Multimodal Model by [Chunyuan Li](#) (6/8)

□ Tool-Augmented Systems: **LLaVA-Plus**

□ LLaVAを拡張した汎用的なマルチモーダルアシスタント

□ ユーザーの入力を受けて関連するツールを起動し、その場でタスクを実行



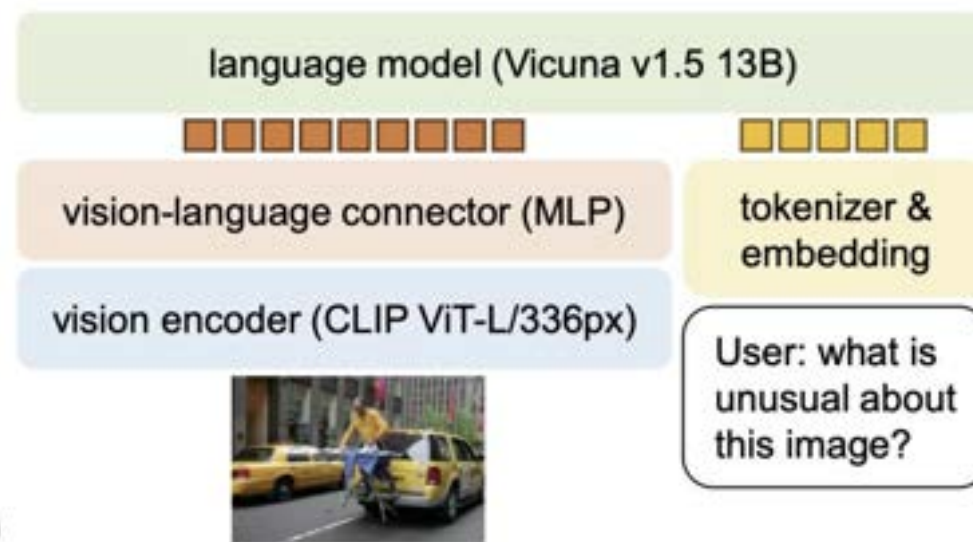
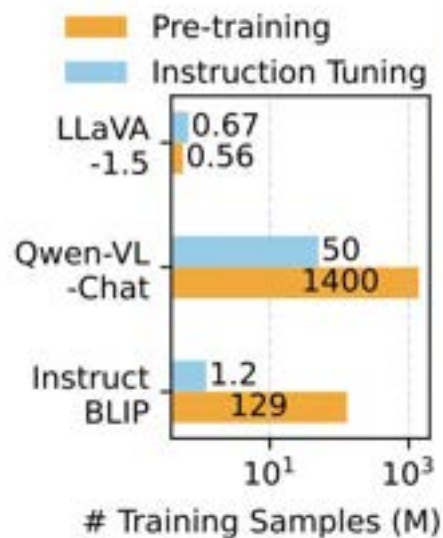
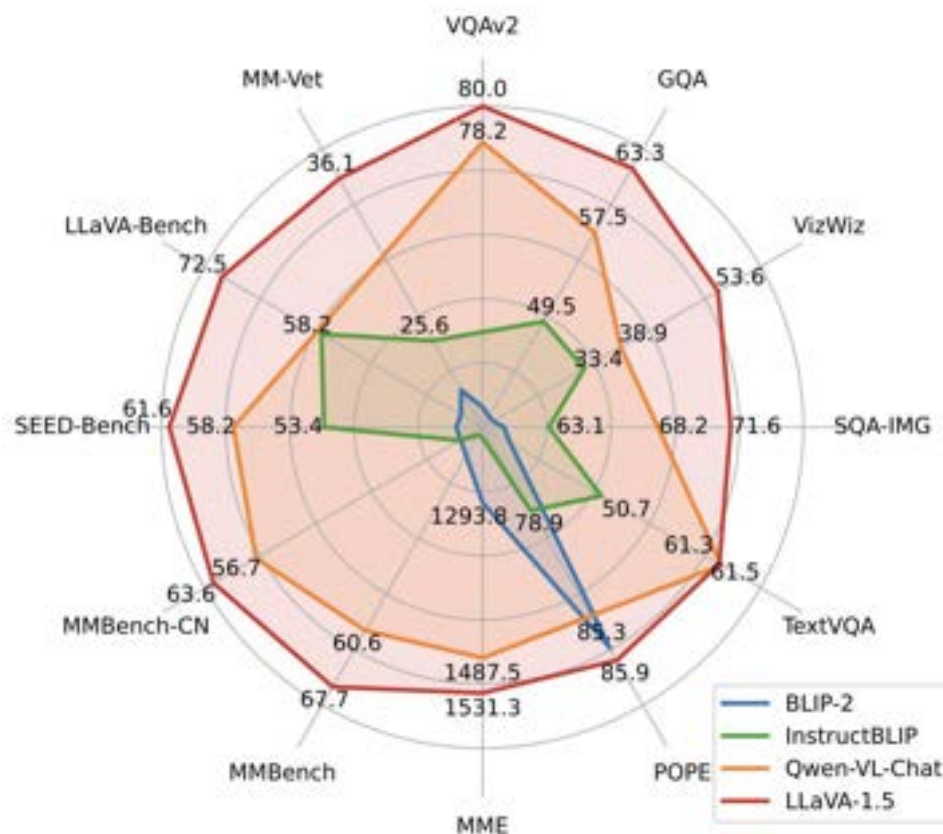
## WS Tool-Augmented Vision (16/21)

### □ LLaVA An Open Large Multimodal Model by [Chunyuan Li](#) (7/8)

#### □ End-to-End Models: **LLaVA-1.5**

□ LLaVAよりもデータ効率のいいベースラインを提案し、11のベンチマークでSoTA。

□ LLaVAに対して、MLP projectionの改善とVQAデータの追加。



## WS Tool-Augmented Vision (17/21)

### ❑ LLaVA An Open Large Multimodal Model by [Chunyuan Li](#) (8/8)

#### ❑ End-to-End Models: **LLaVA-NeXT**

- ❑ 入力画像の高解像度化
- ❑ visual instruction tuning dataの改善
- ❑ LLM backboneのスケールアップ
- ❑ → 推論能力, OCR, 非言語知識の向上

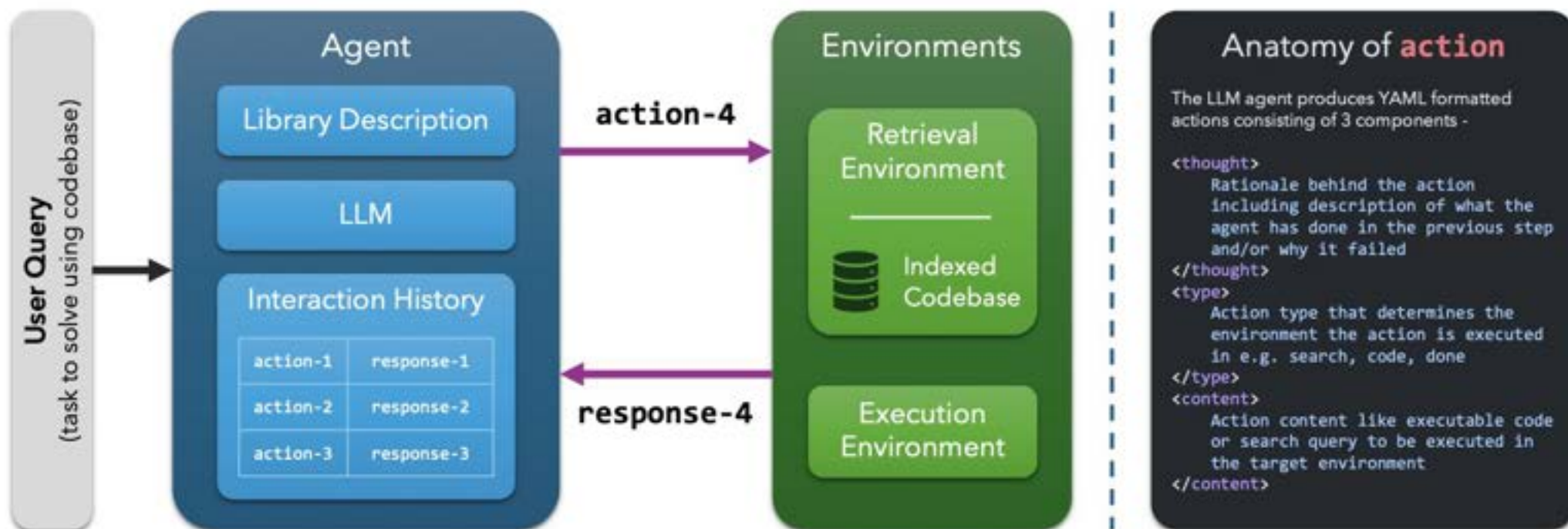
#### ❑ End-to-End Models: **LLaVA-NeXT (Video)**

- ❑ LLMへの入力を複数のパッチから複数のフレームに変更
- ❑ length generalization
- ❑ Direct Preference Optimizationによるモデル応答とAIフィードバックの整合
- ❑ → 優れたzero-shot性能・動画理解能力、長時間動画への対応

# CVPR 2024 の動向・気付き (70/116)

## WS Tool-Augmented Vision (18/21)

- ❑ CodeNavの紹介 by [Aniruddha Kembhavi](#) (1/3)
  - ❑ CODENAV: Beyond tool-use to using real-world codebases with LLM agents
    - ❑ tool-useをツールや使用方法に関して拡張した手法
    - ❑ コードベースのナビゲーション・導入・使用によってユーザークエリを解決
    - ❑ 実行結果のフィードバックによって使用するツールや使用方法を改善





## WS Tool-Augmented Vision (19/21)

### □ CodeNavの紹介 by [Aniruddha Kembhavi](#) (2/3)

#### □ 研究背景

##### □ 近年、tool-useの発展が著しい

- ToolformerやProgPromptなど

##### □ tool-useにおける課題

- 手動記述によるツールの登録が必要
- 全てのツールに対応する単純な関数インターフェースが必要
- ツールが単純化されており、機能や関連情報が限定されている

##### □ 現実世界での応用を考慮すると...

- 新たなコードは日々生み出される → LLMの再学習やfine-tuningは現実的でない
- LLMが未見のコードベースに対応できるようなスケーラブルな方法が必要

#### □ tool-useからの変更点

- 関数のリスト → 実世界のコードベース
- 関数の呼び出し → 自由形状なコードスニペットの生成

# CVPR 2024 の動向・気付き (72/116)

## WS Tool-Augmented Vision (20/21)

### CodeNavの紹介 by [Aniruddha Kembhavi](#) (3/3)

#### 結果

- tool prompt なしで tool-use に匹敵する性能を達成した
- コードベースに含まれる追加のコンテキストを活用できる
- 使用するLLMとしてはGPT-4が最も良い精度となる

method	<i>m&amp;m's</i>				M <sup>3</sup> TOOLEVAL		API-BANK			
	precision	recall	f1	steps	accuracy	steps	precision	recall	f1	steps
tool-use	82.9 ± 4.5	81.7 ± 0.4	79.6 ± 2.3	4.9 ± 0.1	83.7 ± 2.8	6.6 ± 0.5	86.6 ± 0.8	93.6 ± 1.1	88.5 ± 0.7	3.4 ± 0.1
code-use	88.0 ± 6.1	78.2 ± 4.5	80.6 ± 5.1	7.2 ± 0.2	81.7 ± 4.9	7.8 ± 0.4	84.0 ± 0.7	89.3 ± 0.6	85.3 ± 0.3	5.3 ± 0.0

tool description	length	f1	steps
w/o desc	0	74.1 ± 1.9	7.0 ± 0.1
tool names	694	78.1 ± 3.5	6.7 ± 0.2
+ desc	3680	80.8 ± 0.4	6.9 ± 0.1
+ prototypes	4627	80.7 ± 5.0	6.1 ± 0.1
library desc (CODENAV)	2061	80.6 ± 5.1	7.2 ± 0.2

LLM	precision	recall	f1	steps
gpt-4-1106-preview	88.0 ± 6.1	78.2 ± 4.5	80.6 ± 5.1	7.2 ± 0.2
gpt-3.5-turbo-0125	54.36 ± 2.3	15.77 ± 1.4	22.96 ± 0.8	9.08 ± 1.07
Mixtral-8x22B-Instruct-v0.1	82.50 ± 2.1	62.31 ± 1.9	67.91 ± 1.3	9.06 ± 0.3
Qwen1.5-110B-Chat	78.15 ± 3.1	38.49 ± 5.5	48.84 ± 5.1	10.00 ± 0.2

## WS Tool-Augmented Vision (21/21)

### □ まとめ

#### □ tool-useの背景・概要

- End-to-Endの大規模モデルでは複雑なタスクが解けない
  - fine-tuningでは特定のタスクに関する性能しか向上させられない
  - → **タスクを分割、タスクに特化した既存のモデルを使用する**
- 具体例を与えることで、LLMが登録された既存モデル(ツール)を使用して  
ドを生成

タスクを解くコー

#### □ tool-useにおけるトレンド

##### □ フィードバック(ループ構造)の導入

- 中間出力の生成による解釈性の高さを活かしている
- 生成結果の評価と使用ツールの探索に対する評価の区別

##### □ 自動化・対応可能タスクの増加

- 動画などの新たなドメインに対してtool-useを使用
- フィードバックの導入によるコード修正の自動化
- コードベースの導入によるツール登録の省略、選択可能ツールや関連情報の増加

## WS Foundation models for autonomous systems (1/6)

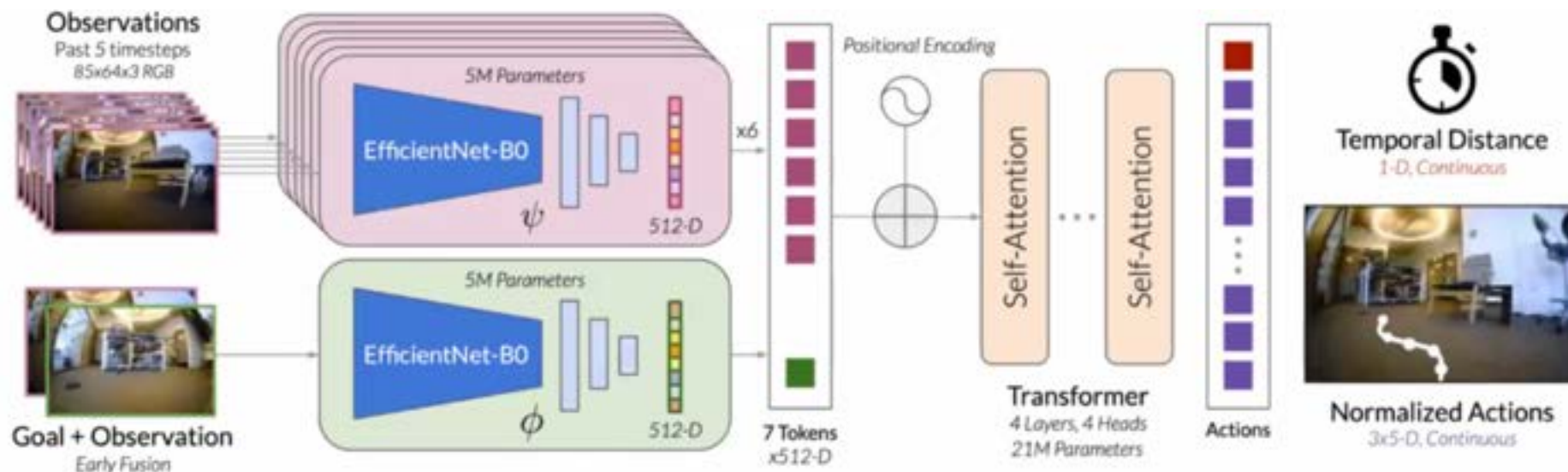
### ❑ Robotic Foundation models

❑ ロボティクスFMの構築において、Navigation、Manipulation、cross-embodied learningの視点からの研究を紹介した。

### ❑ Navigation:

#### ❑ ViNT: Vision NaVision Navigation Transformer

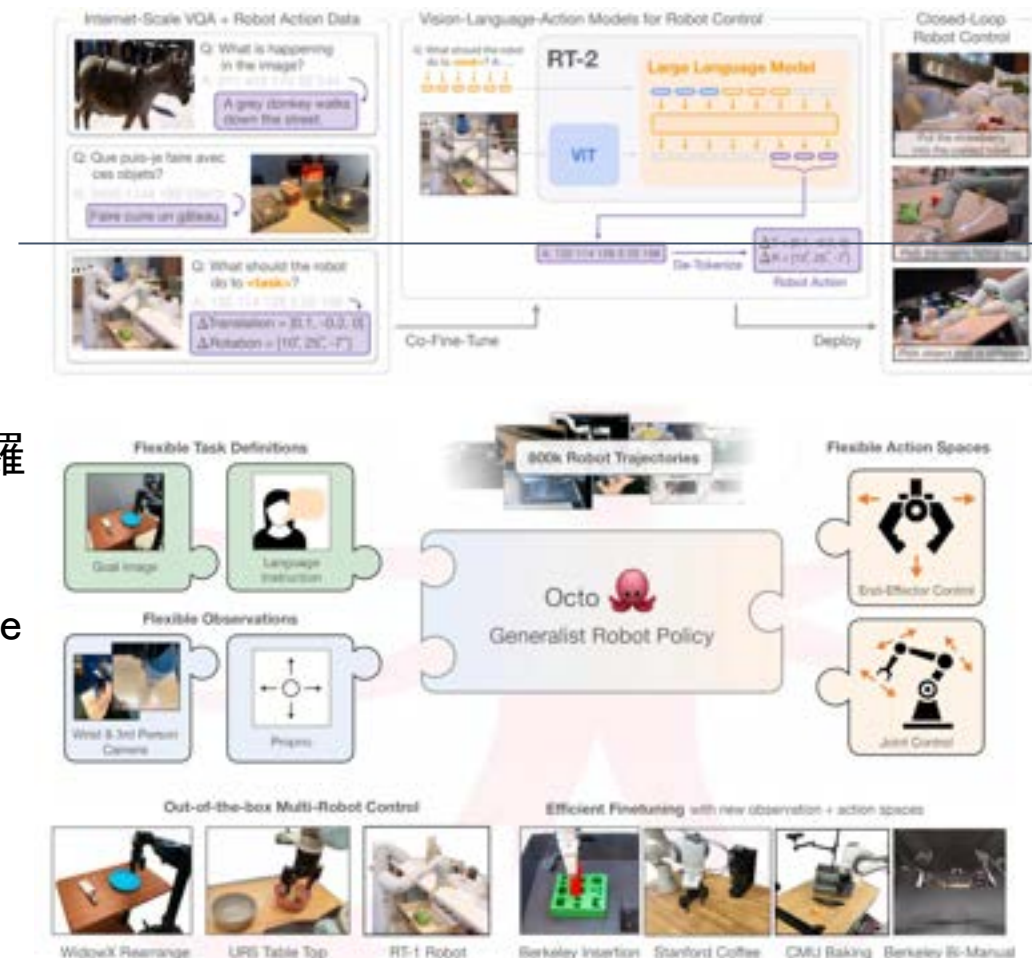
❑ EfficientNetで過去と現在の観察とゴールをエンコードし、トポロジカルグラフベースのプランナーで未知な場所にも探索できるモデルです。このモデルは時間的距離と動作を予測し、長距離ナビゲーションなどに適応します。



## WS Foundation models for autonomous systems (2/6)

### ❑ Robotic Foundation models

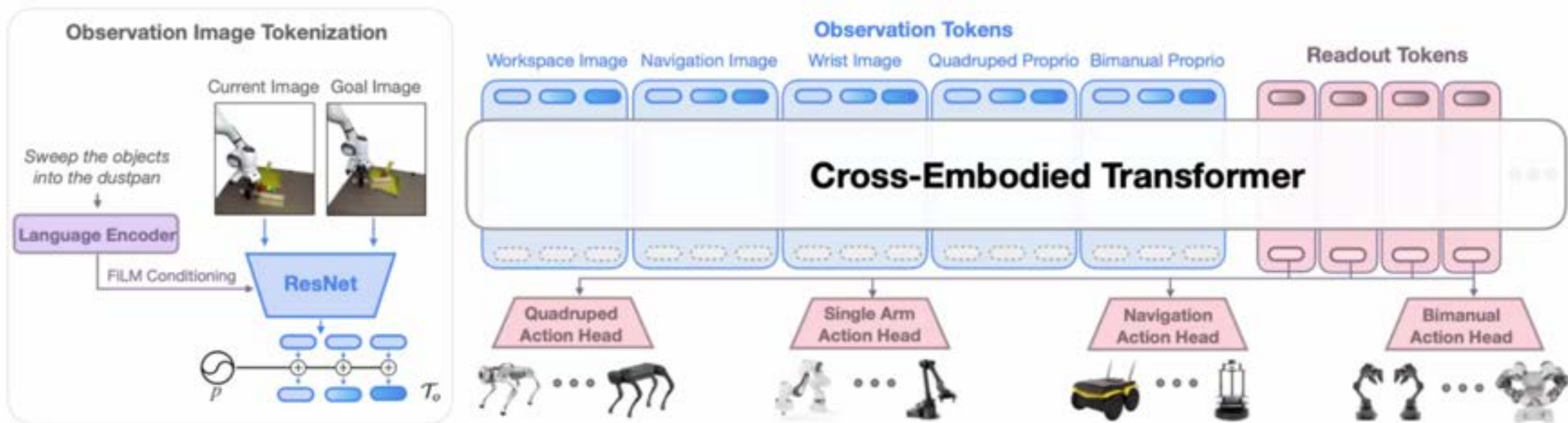
- ❑ ロボティクスFMの構築において、Navigation、Manipulation、cross-embodied learningの視点からの研究を紹介した。
- ❑ Manipulation:
  - ❑ RT-2: ウェブとRoboticsデータを学習したVision, Language, Action (VLA) の大規模モデル
  - ❑ Open X-Embodiment: タスク、ロボット種類、環境が網羅している膨大ロボティクスデータセット
  - ❑ Octo: OpenEmbodiedXデータで学習したFlexibleなタスク、観察、動作をできるロボティクス関連のOpen Source
  - ❑ OpenVLA: 大規模なVLAモデル



## WS Foundation models for autonomous systems (3/6)

### ❑ Robotic Foundation models

- ❑ ロボティクスFMの構築において、Navigation、Manipulation、cross-embodied learningの視点からの研究を紹介した。
- ❑ Cross Embodied Learning:
  - ❑ CrossFormer: transformerベースで構築された多様なロボットと多様なタスに対応したモデル



## WS Foundation models for autonomous systems (4/6)

### ❑ Foundation Models as Real-World Simulators

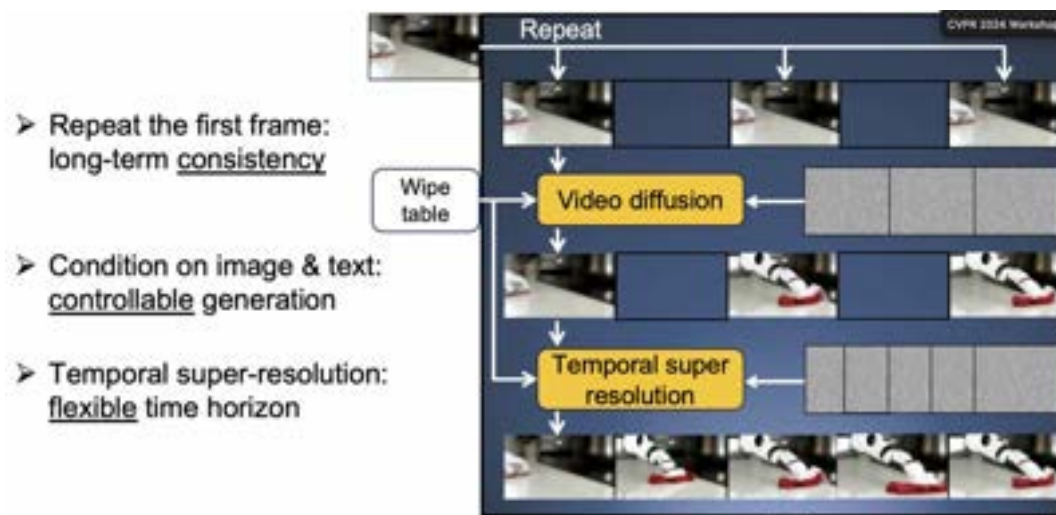
❑ ロボティクスFMの学習には、良いデータ/シミュレーター+良いアルゴリズムで達成できる

### ❑ UniSim: Learning Interactive Real-World Simulators

❑ Diffusionモデルを用いて、Interactiveなシミュレーター

❑ 実験には、シミュレーションからのデータを学習して、Inverse Dynamic経路で、実際のロボットにも実行できる。

❑ UniSimのデータでは、Internetスケール(大量)、Temporal flexibility, 各難易度のタスクのメリットがある



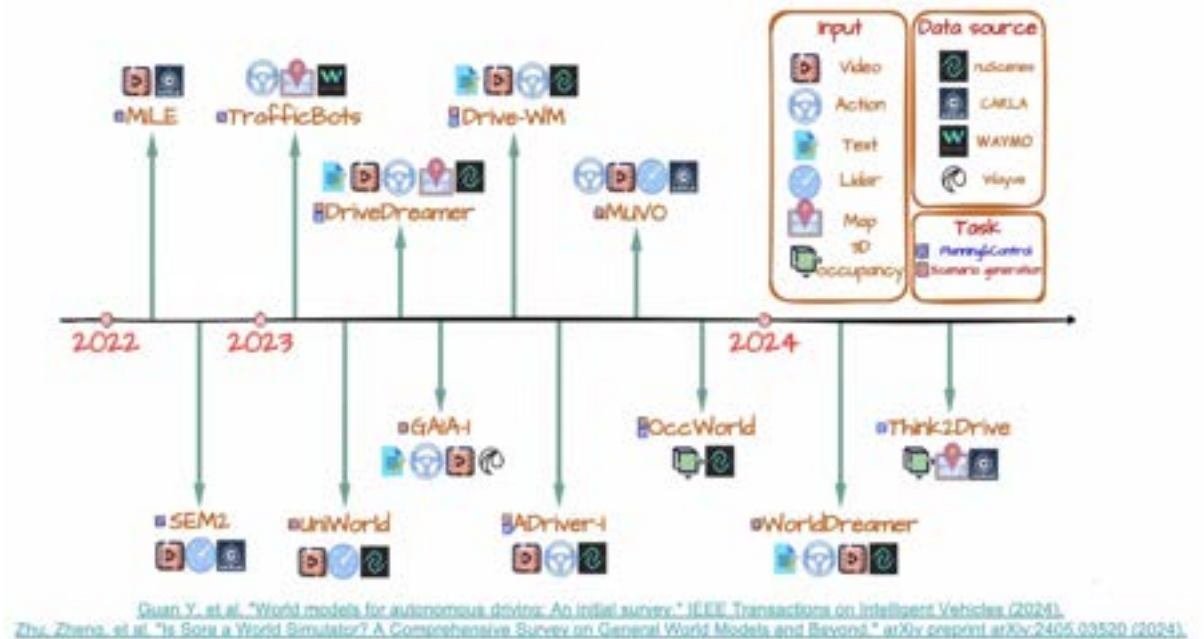
# CVPR 2024 の動向・気付き (78/116)

## WS Foundation models for autonomous systems (5/6)

- ❑ The road to Embodied AI: Building Embodied AI to be Safe and Scalable
- ❑ End-to-endの手法で自動運転とかを実現するためには、simulation, multimodal, dataの観点から研究を紹介した。
- ❑ Ghost GYM、PRISM-1、wayscenes101
  - ❑ 時間のDynamicを含め、実シーンとほぼ類似ようなシミュレータ

- ✓ Closes the domain gap in perception and behaviour
- ✓ Simulates dynamic and deformable scenes
- ✓ Simulates full environment and embodied platform
- ✓ Data driven and scalable
- ✓ Controllable over long-tail

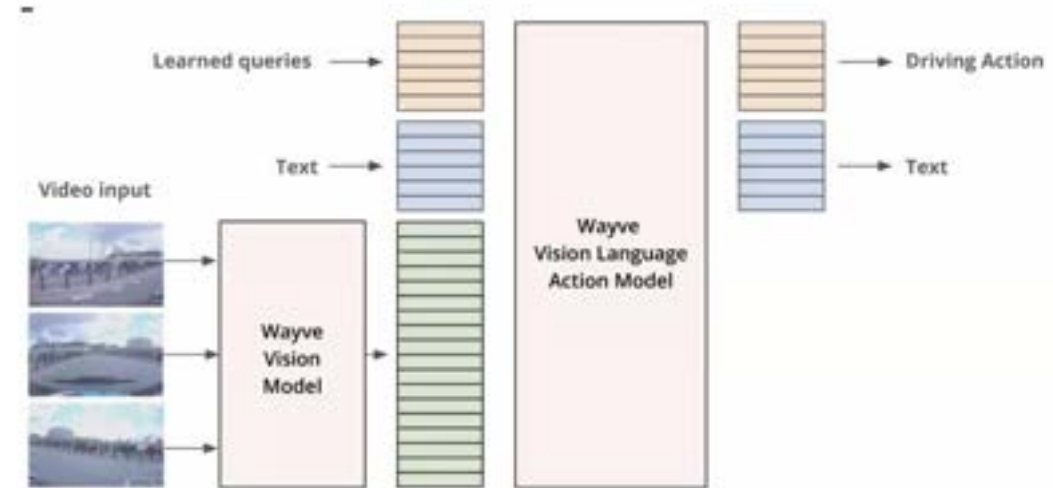
優れたADASシミュレータを持つべき





## WS Foundation models for autonomous systems (6/6)

- ❑ The road to Embodied AI: Building Embodied AI to be Safe and Scalable
- ❑ End-to-endの手法で自動運転とかを実現するためには、simulation, multimodal, dataの観点から研究を紹介した。
  - ❑ Multimodal
    - ❑ LLM4drive
    - ❑ Lingo-1
    - ❑ Lingo-2
- ❑ これから自動運転するための工夫
  - ❑ Data:
    - ❑ 質: ウェブだけではbiasがある
    - ❑ タスク、応用、モダリティを含めて多様性を考慮
  - ❑ training compute
    - ❑ 高次元のビデオデータ
  - ❑ physical embodiment
    - ❑ 安全性、ノイズロボット、コントロールなしの環境



## CVPR 2024 LIMIT Workshop



cvpaper.challenge

## CVPR 2024 Workshop on Representation Learning with Very Limited Images

**-Zero-shot, Unsupervised, and Synthetic Learning in the Era of Big Models-**

- ❑ 「日本発」のCVPR Workshop採択！
- ❑ データ・ラベル・計算機など限られた(LIMIT)リソースから深層学習を実施するというコンセプトのワークショップ



## CVPR 2024 LIMIT Workshop



cvpaper.challenge

## CVPR 2024 Workshop on Representation Learning with Very Limited Images

**-Zero-shot, Unsupervised, and Synthetic Learning in the Era of Big Models-**

- **大企業のみでなく大学のリソースでも基盤モデル構築できるように！**
  - 基盤モデル時代になり、リソースを持つ限られた研究機関のみが基盤構築するのではなく、限られたリソースからでも基盤モデル構築できるように、という願いを込めた
  - 「限られた」データ・ラベルからの学習については、権利・安全上の全数把握がしやすいということでもある



# CVPR 2024 の動向・気付き(82/116)

## CVPR 2024 LIMIT Workshop

- 限られたリソースからの深層学習を扱うワークショップ
- {データ, 計算機, ラベル}など, 何かが不足していながら機械学習を実施



- オーガナイザも下記の研究を実施するメンバーが集結
  - 合成データによる学習
  - ゼロ実画像学習
  - 自己教師あり学習
  - 教師なし学習
  - ロバスト性AI
  - など

<https://hirokatsukataoka16.github.io/CVPR-2024-LIMIT/> より参照

# CVPR 2024 の動向・気付き(83/116)

## CVPR 2024 LIMIT Workshop

- 拡大し続けるデータ・モデルパラメータの中でリソース側が限界
- 巨大リソースを持っていなくても学習できる, が重要

**Models**

**Vision Transformer (ViT)**

ViT Models	Params
22B	22B
Giant	2B
Huge	632M
Large	307M
Base	86M
Small	22M
Tiny	5M

**Datasets**

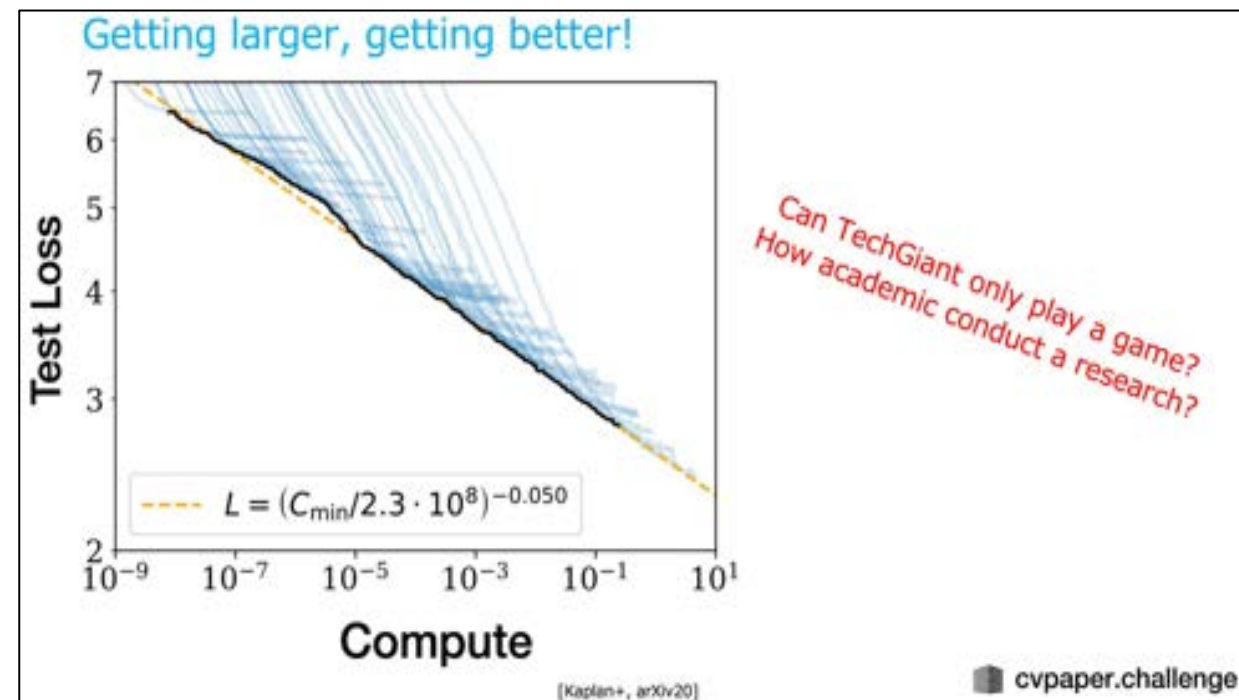
**JFT-300M/3B/4B**  
[Sun+, ICCV17][Zhai+, CVPR22][Dehghani+, ICML23]

**IMAGENET**  
[Deng+, CVPR09]

**LAION**  
<https://laion.ai/>

Large size is 'JUSTICE'

cvpaper.challenge

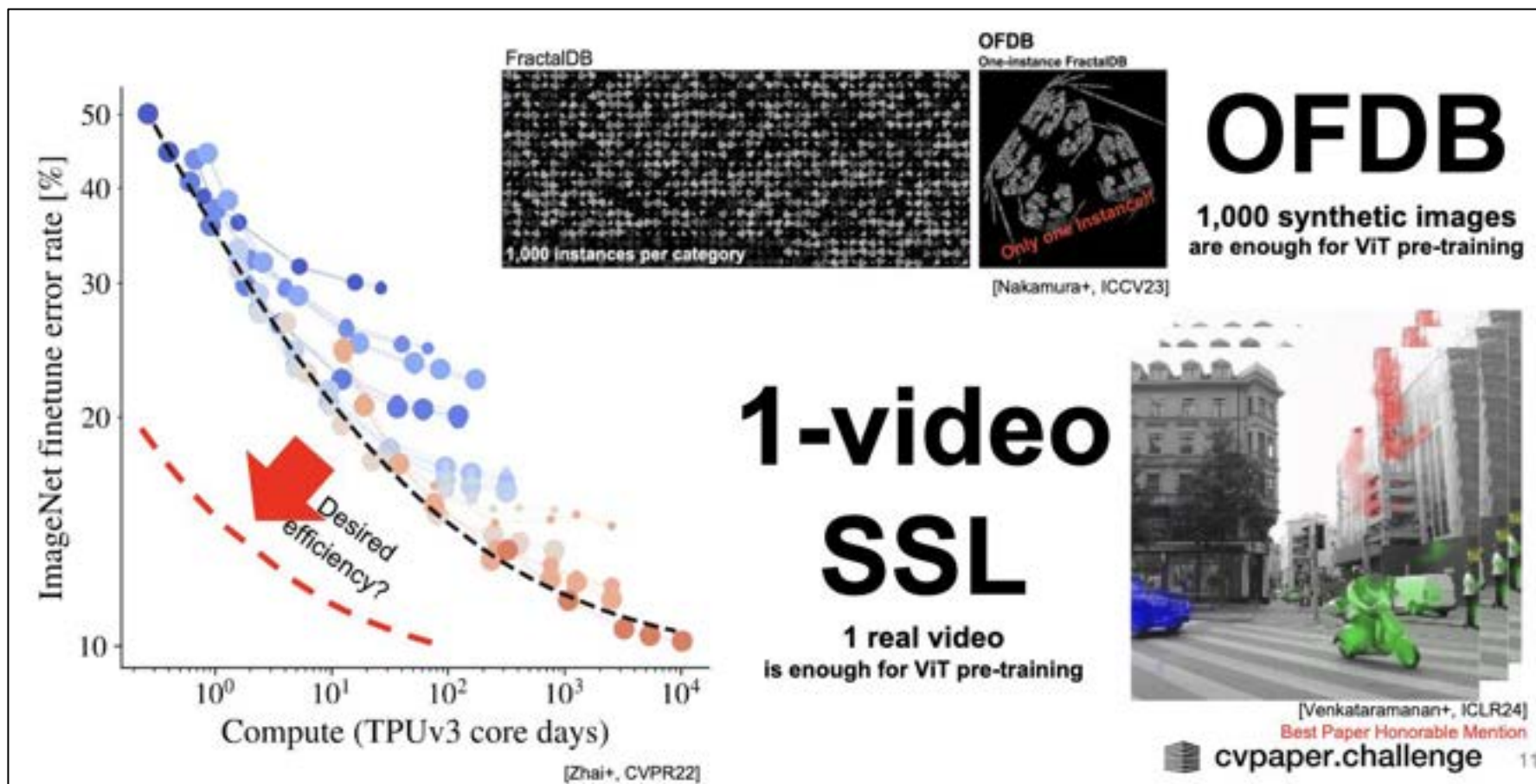


<https://hirokatsukataoka16.github.io/CVPR-2024-LIMIT/> より参照

# CVPR 2024 の動向・気付き(84/116)

## CVPR 2024 LIMIT Workshop

- CVのスケーリング則 (scaling laws)
  - 1,000枚の合成画像 / 1実動画でViT学習成功でき, CVのスケーリング則は崩せる?
  - 低リソースで基盤モデルが構築できる



# CVPR 2024 の動向・気付き(85/116)

## CVPR 2024 LIMIT Workshop

- 多くの技を駆使すれば限られたリソースからの深層学習が可能！

**Sim-to-Real**  
[Shrivastava+, CVPR17]

**Domain Randomization**  
[Sundermeyer+, ECCV18]

**Learning to See by Looking at Noise**  
[Baradad+, NeurIPS21]

**Representation Learning from synthetic images**  
[Tian+, CVPR24]

**Formula-driven Learning**  
[Kataoka+, IJCV22]  
[Anderson+, WACV22]

**1-image SSL**  
[Asano+, ICLR20, ICLR23]

**Visual Prompt Tuning**  
[Shtedritski+, ICCV23]

The cow that is the smallest

Keyword: Synthetic Data, Adversarial Learning, Data Augmentation, Domain Randomization, Formula-driven Supervised Learning, Visual Prompt Tuning

cvpaper.challenge 12

## CVPR 2024 LIMIT Workshop

- Invited talk 1: Phillip Isola (MIT)
- N=0: Learning Vision with Zero Visual Data
  - 実画像がいない視覚モデルの学習について説明
  - (1) ノイズからの視覚機能学習, (2) プログラミングコードからのGenerative Art生成と視覚機能学習, (3) LLMによる言語や画像の修正, について説明
  - 詳細は <http://web.mit.edu/phillipi/www/bio.html> をご確認ください

→ (1, 2, 3)からの視覚モデル構築は可能であることを実証！

※Invited talk 1では[Kataoka+, ACCV2020]への引用もあり, 元ネタであることを示してくれて嬉しい！



## CVPR 2024 LIMIT Workshop

- ❑ Invited talk 2: Zeynep Akata (TUM / Helmholtz Munich)
- ❑ Learning with Small Number of Images in Multimodal Large Language Models
  - ❑ Zero-shot LearningやExplainable AIの先駆者
  - ❑ 基盤モデルが流行するずっと前から Zero-shot Learning について研究してきたパイオニア
  - ❑ 「言語」による説明性について説明 → 中にはDecision TreeやLLMによる説明を加えていることもあり, なるべく人間にわかりやすい形で推論経過を説明

→ 真に人間に分かりやすい形はやはり言語！？ (特徴マップやヒートマップは分かりづらい?)

## CVPR 2024 LIMIT Workshop

### □ 2回目のワークショップ運営を終えて

- 招待講演・オーラル・ポスターなど質の高いご発表のおかげで大盛況で終えた; 常時席が埋まっていて、招待講演中は地面座り見・立ち見・会場外から聴講など凄い熱気を感じた
- ワークショップ中は特にトラブルもなく, Chairも運営メンバー全員のリレーで乗り切る
- ポスターも招待枠含め大盛況で, 発表してくれた皆さんに感謝!



※夜はオーガナイザ・招待枠のみんなで懇親会をして次回連携への足がかりを作った!  
国内外の繋がりを強化することが国際連携成功へのキー?

# CVPR 2024 の動向・気付き (89/116)

## WS Urban Scene Modeling(1/3)

- 都市管理などで使用される建物の3次元モデリングに関するWS
- 論文発表の他、他視点写真からの建物モデリングのコンペティションも開催
- オーガナイザはCV・CG・測量分野で構成、**分野横断の兆し**

### Organizers



**Ruisheng Wang**  
Professor  
University of Calgary



**Jack Langerman**  
Sr. Applied Scientist  
HoloVR Inc.



**Ilke Demir**  
Senior Research Scientist  
Intel Corporation



**Qixing Huang**  
Associate Professor  
University of Texas at Austin



**Shangfeng Huang**  
Researcher  
University of Calgary



**Daoyi Gao**  
Researcher  
Technical University of Munich



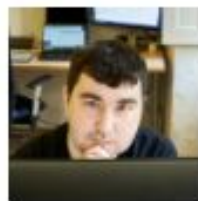
**Xiang Ma**  
Head of Research  
Amazon Web Services



**Hanzhi Chen**  
Researcher  
Technical University of Munich



**Florent Lafarge**  
Researcher  
INRIA



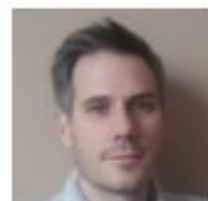
**Dmytro Mishkin**  
Researcher  
Czech Technical University  
in Prague



**Tolga Birdal**  
Assistant Professor  
Imperial College London,  
UK



**Hui Huang**  
Professor  
Shenzhen University



**Clement Mallet**  
Research Scientist  
LASTIG



**Caner Korkmaz**  
Researcher  
Imperial College London



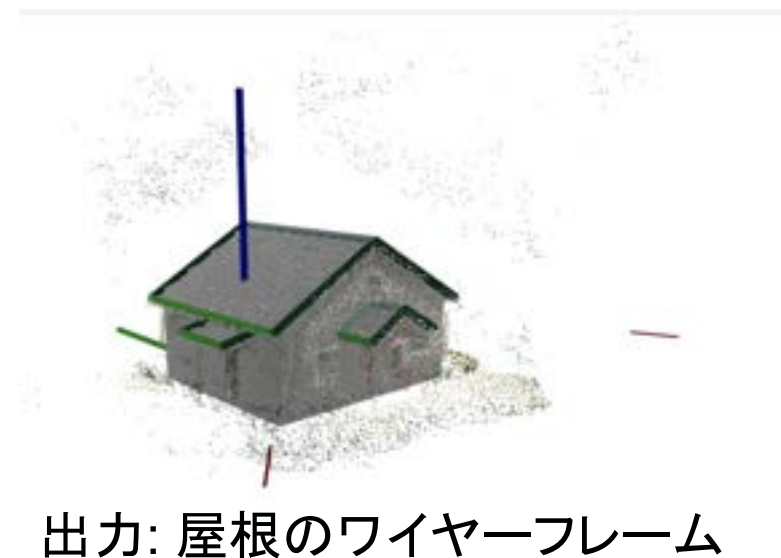
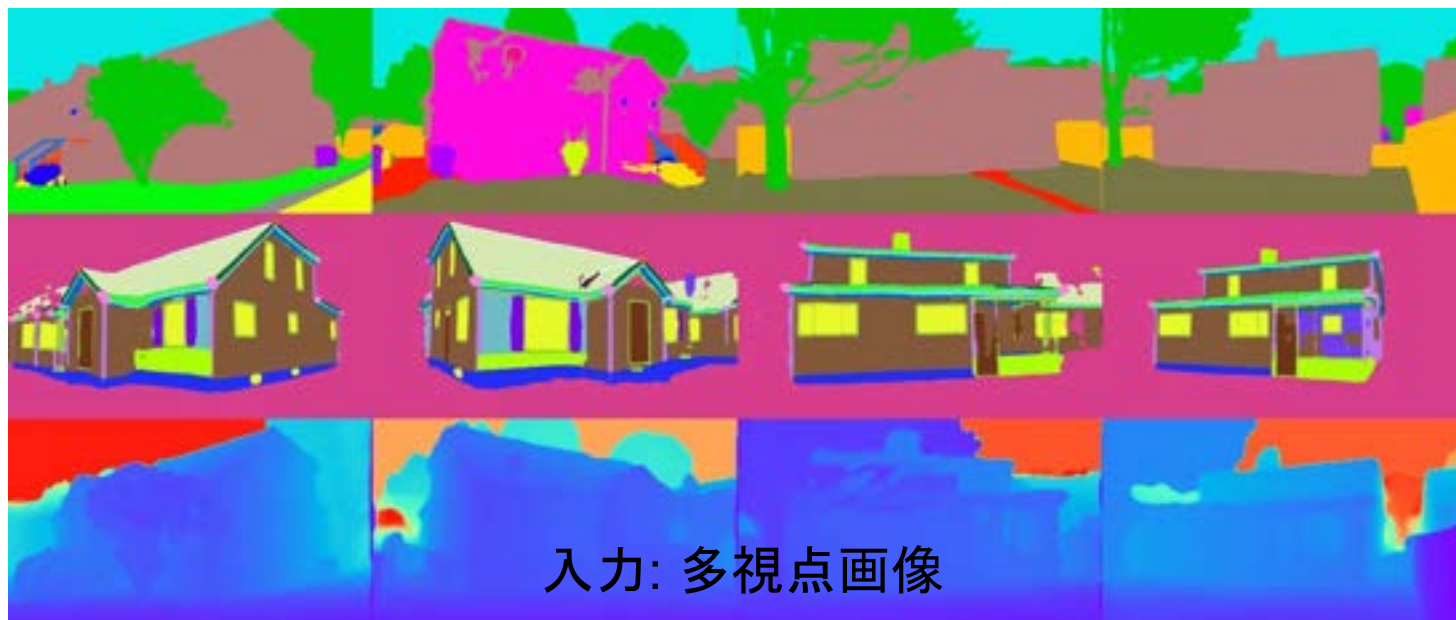
**Yang Wang**  
Associate Professor  
Concordia University



**Marc Pollefeys**  
Professor  
ETH Zurich

## WS Urban Scene Modeling(2/3)

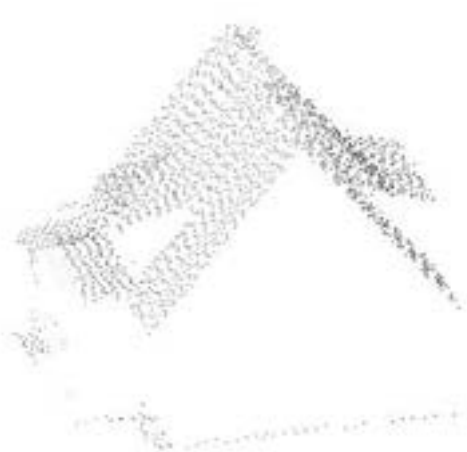
- コンペ①屋根構造のワイヤースケルトンを推定するタスク
  - 多視点画像から屋根のワイヤースケルトンを抽出する
  - コンペティションの結果では、以下の流れの手法が上位
    - ①画像ごとの頂点の抽出
    - ②各画像から推定した頂点の対応付け
    - ③頂点からワイヤースケルトンを作成する



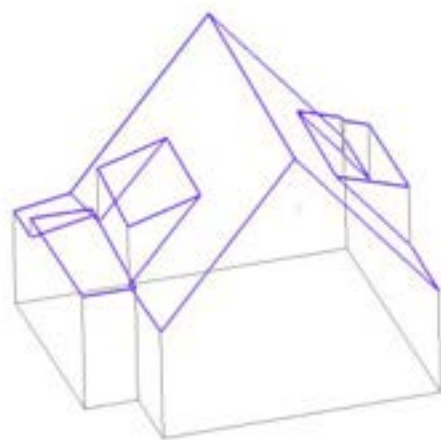
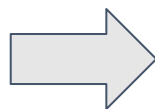
## WS Urban Scene Modeling(3/3)

### □ コンペ②建物モデルを作成するタスク

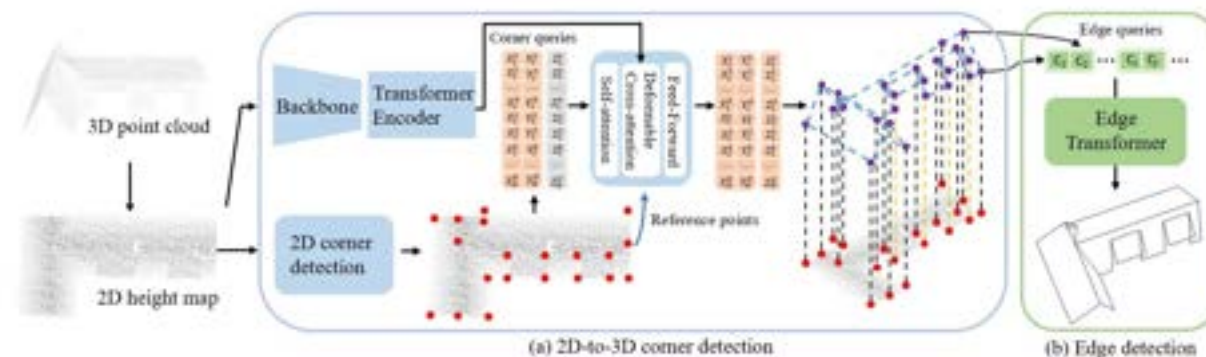
- 点群から建物モデルをワイヤースケルトンに基づいて作成する
- コンペティションの結果では、以下の流れの手法が上位
  - ①点群を2次元高さ画像に変換してコーナー推定
  - ②点群の特徴量に高さ画像から推定したコーナー情報を追加してもう一度コーナー推定
  - ③コーナー推定結果をTransformerに入力してエッジを推定



入力: 点群



出力: 建物のワイヤースケルトンモデル



上位の手法

## WS Computer Vision for Science(1/3)

- ❑ 科学分野でもコンピュータビジョン技術を活用している研究者は多い
- ❑ AI4scienceに関する基調講演が2023年にCVPRとICCVでなされるなど、コミュニティとしてはコンピュータビジョンと科学分野の融合に関心があるがCVPRで実際に発表される件数は少ない
- ❑ ワークショップを開くことで、実務者間での議論を活発化する狙いがあり、今後、AI4scienceがCV分野でも広く認知される可能性が高い



Mike Wainwright  
Dunlap Fellow  
University of Toronto



Tali Treibitz  
Head of Marine Imaging Lab  
University of Haifa



Ellen Zhong  
Assistant Professor of CS  
Princeton University



Mark Cheung  
Science & Deputy Director  
Space & Astronomy, CSIRO



Sara Beery  
Assistant Professor  
AI + Decision Making & CSAIL MIT



Aviad Levi  
Assistant Professor (2024)  
University of Toronto



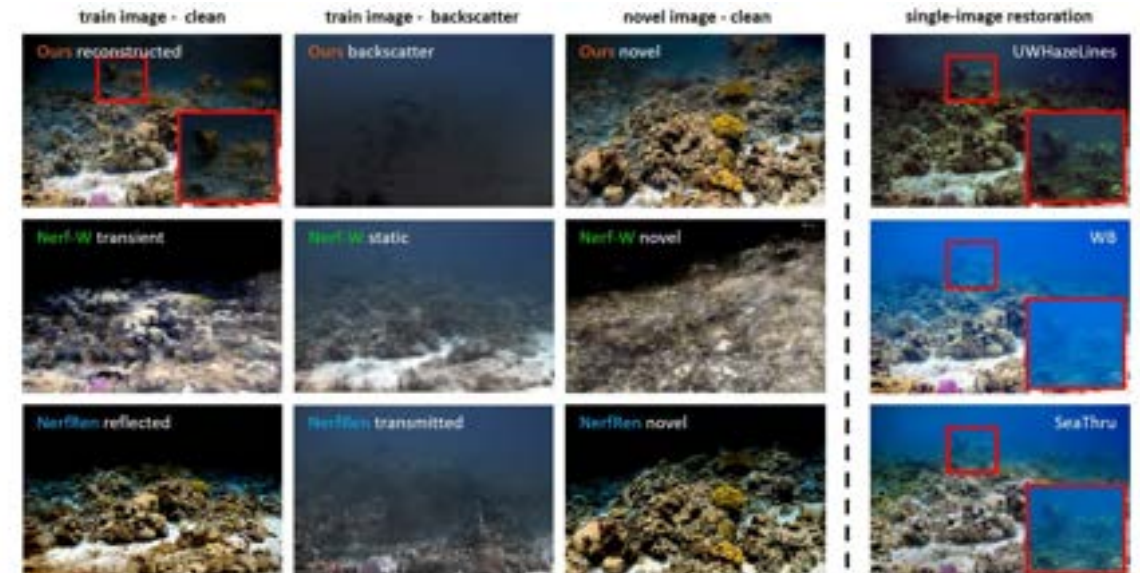
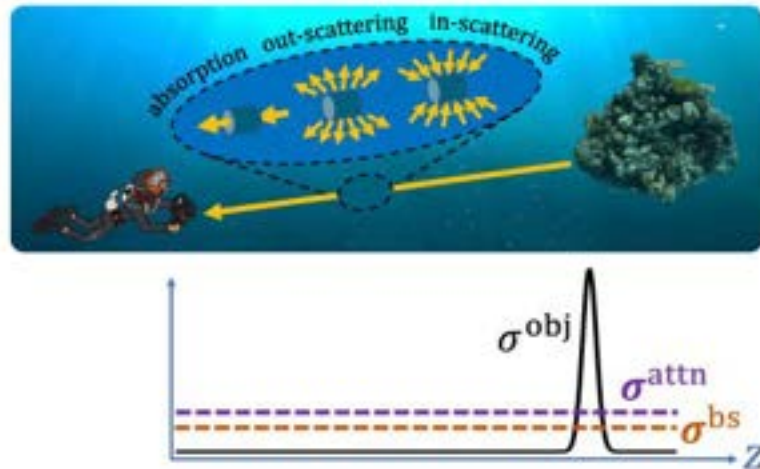
Yoav Schechner  
Professor  
Technion



Jacob Berv  
Schmidt AI in Science Postdoc  
University of Michigan

## WS Computer Vision for Science(2/3)

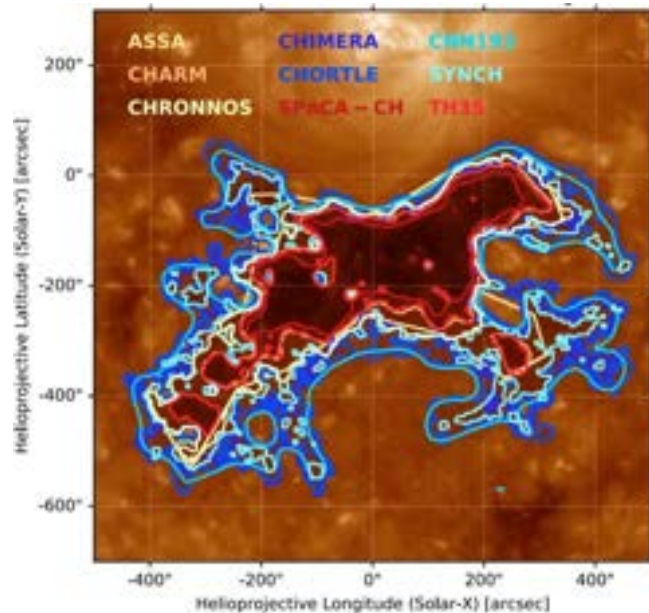
- 招待講演をピックアップ: Computer Vision for Coral Science
- 水中で撮影した画像は、CV分野で扱われる一般的な自然画像と比較すると色や濁りなどの影響を強く受ける
- 例えば、水中写真でNeRFを実行する場合には、光線空間のモデリングを工夫すると再構成の質が向上する  
=>単にCV技術をscienceに応用するだけでなく、ドメイン特有の工夫が重要



[https://openaccess.thecvf.com/content/CVPR2023/html/Levy\\_SeaThru-NeRF\\_Neural\\_Radiance\\_Fields\\_in\\_Scattering\\_Media\\_CVPR\\_2023\\_paper.html](https://openaccess.thecvf.com/content/CVPR2023/html/Levy_SeaThru-NeRF_Neural_Radiance_Fields_in_Scattering_Media_CVPR_2023_paper.html) より参照

## WS Computer Vision for Science(3/3)

- ❑ 招待講演をピックアップ:  
Computer Vision for Heliophysics & Astrophysics
- ❑ 太陽物理学の分野では機械学習技術が日常的に使われてきており、深層学習がブームになってからはよりそれが顕著になった
- ❑ コロナホールの抽出にはセグメンテーション、磁場データからの太陽活動領域抽出には物体検出が用いられる



一定時間ごとに毎日観測される宇宙のデータとCV技術による自動化は相性が良い

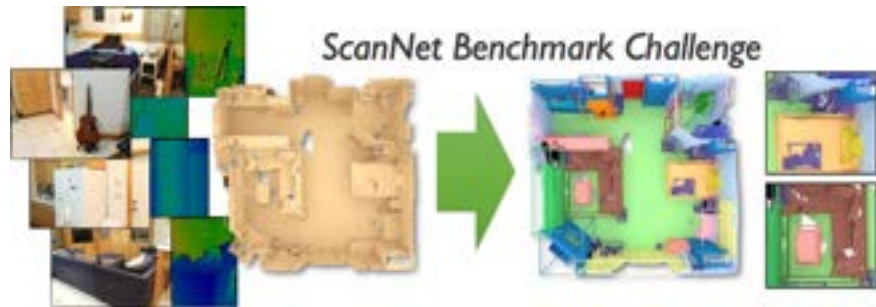


# CVPR 2024 の動向・気付き(95/116)

## 3次元環境の理解に関するワークショップが3つ開催

- ❑ 1st Workshop on Multimodalities for 3D Scenes
- ❑ Open-Vocabulary 3D Scene Understanding
- ❑ ScanNet++ Novel View Synthesis and 3D Semantic Understanding Challenge

2018年は20クラス程度の  
点群セグメンテーション



現実空間  
をより意識

2024年のCVPR

マルチモーダル化

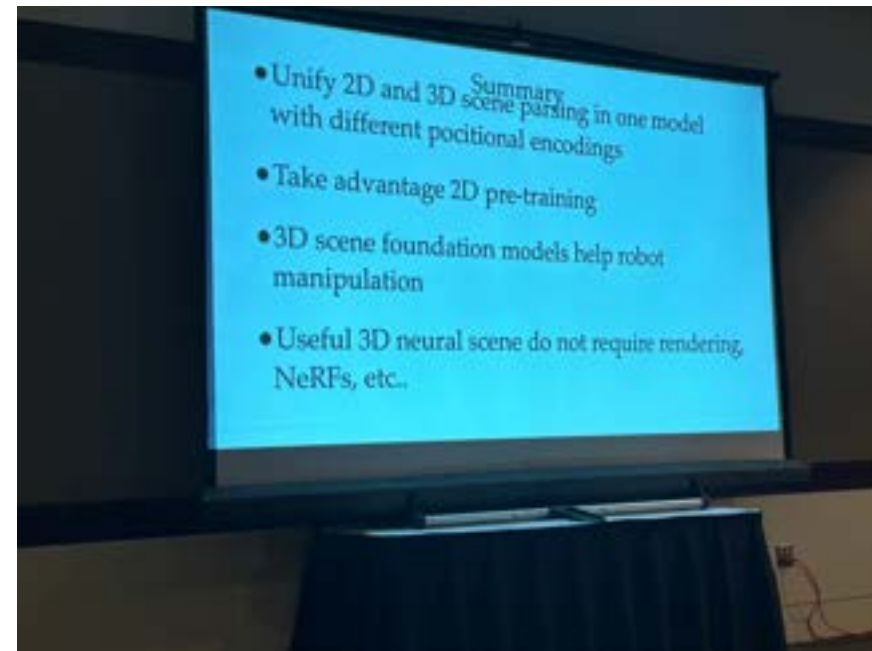
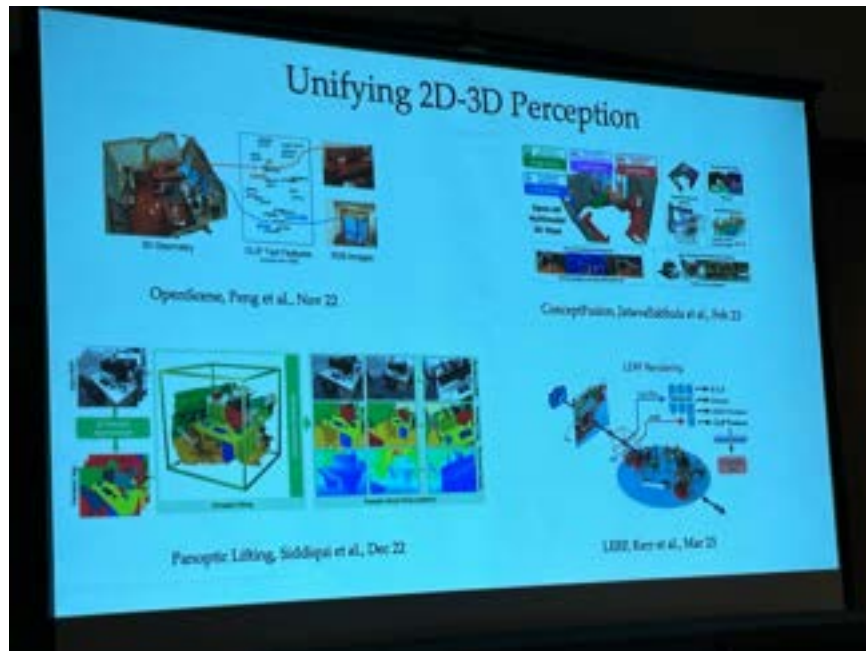
オープンボキャブラリー化

1000クラス化

## 1st Workshop on Multimodalities for 3D Scenes

- Title: Unified 2D/3D Foundational Models across Images, Language and Actions
- Speaker: Katerina Fragkiadaki

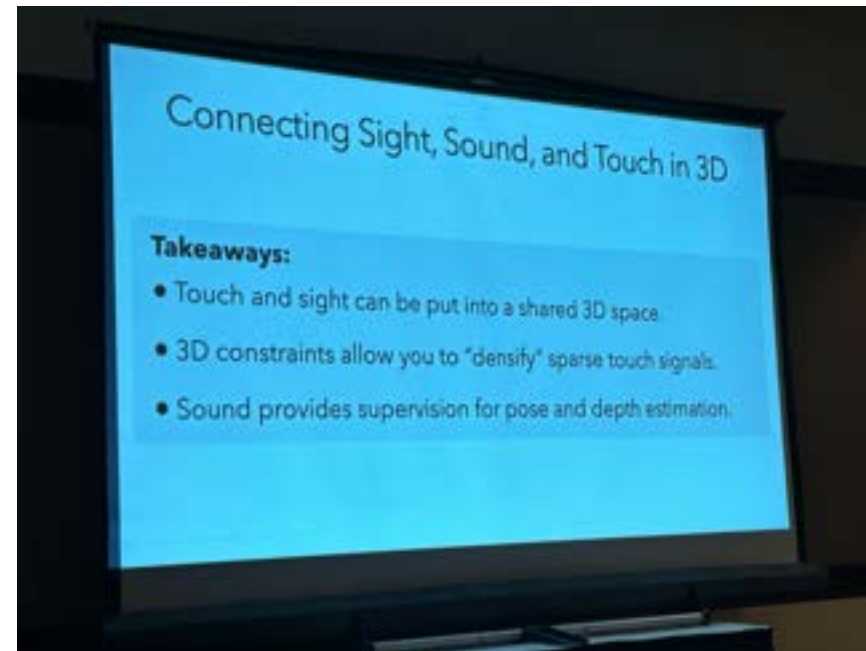
これまで別々に扱われることが多かった 2Dと3Dを統合する試み



## 1st Workshop on Multimodalities for 3D Scenes

- ❑ Title: Tactile-augmented Radiance Fields
- ❑ Speaker: Andrew Owens

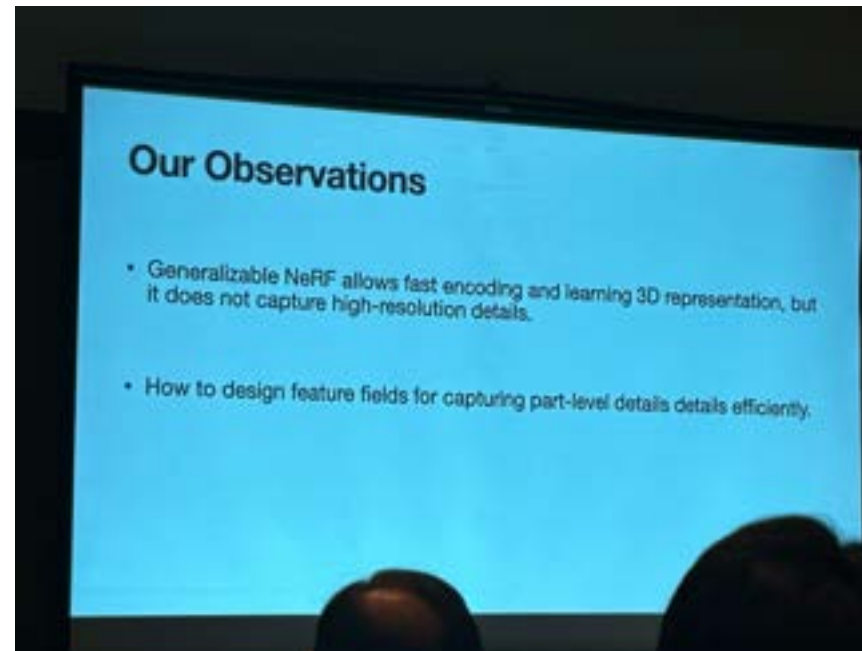
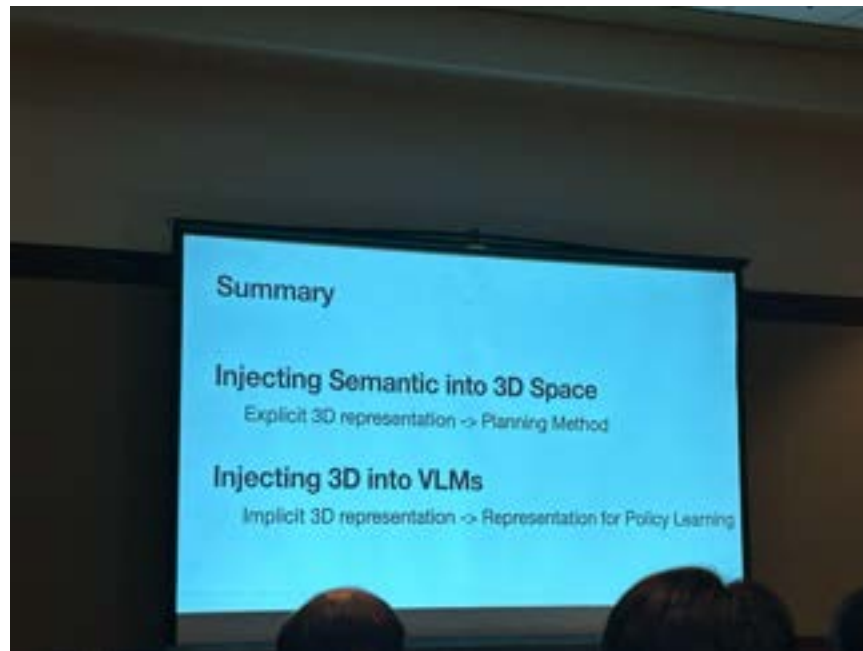
What does **\*\*tactile\*\*** tell you about 3D structures??



## 1st Workshop on Multimodalities for 3D Scenes

- Title: Tactile-augmented Radiance Fields
- Speaker: Generalizable 3D Spatial Perception and Control

### 3Dシーンを言語表現を介してどのように理解するか



## 2nd Workshop on Open-Vocabulary 3D Scene Understanding

- コンペティションでは、Open-vocabulary 3D object instance searchと Open-vocabulary 3D functionality groundingのタスクが開催

### 我々の生活をサポートする ARデバイスやロボットを意識

Open-vocabulary 3D object instance search  
自然言語で指示した物体を抽出



[https://opensun3d.github.io/cvpr24-challenge/track\\_1/](https://opensun3d.github.io/cvpr24-challenge/track_1/)  
より引用

Open-vocabulary 3D functionality grounding  
ある動作を実行するために必要な物体を抽出



[https://opensun3d.github.io/cvpr24-challenge/track\\_2/](https://opensun3d.github.io/cvpr24-challenge/track_2/) より引用

# CVPR 2024 の動向・気付き(100/116)

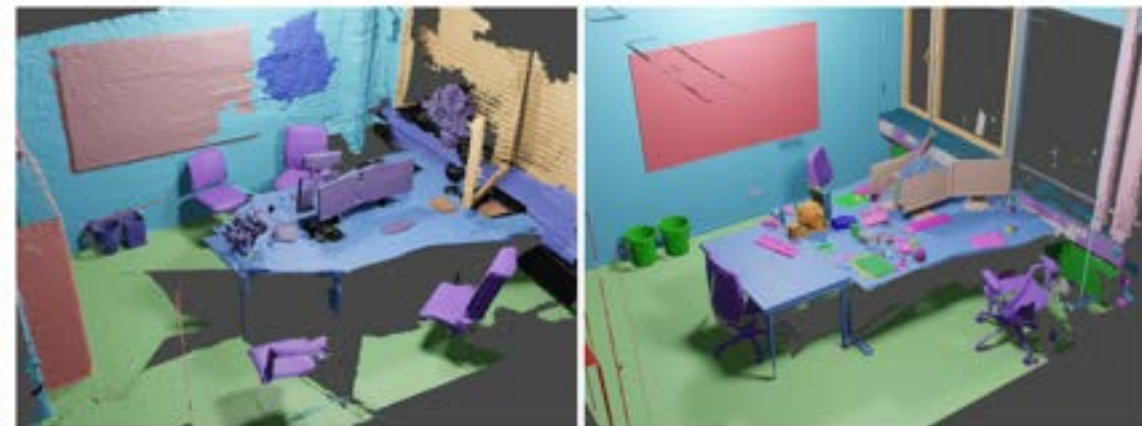
## ScanNet++ Novel View Synthesis and 3D Semantic Understanding Challenge

- コンペティションでは、新視点合成とセマンティックセグメンテーションインスタンスセグメンテーションタスクが開催
- 3Dセグメンテーションでは、従来のデータセットよりも桁の多い1000クラスを扱い、新視点合成ではより難しい訓練に使用したカメラ軌道以外を用いて合成結果の質を評価

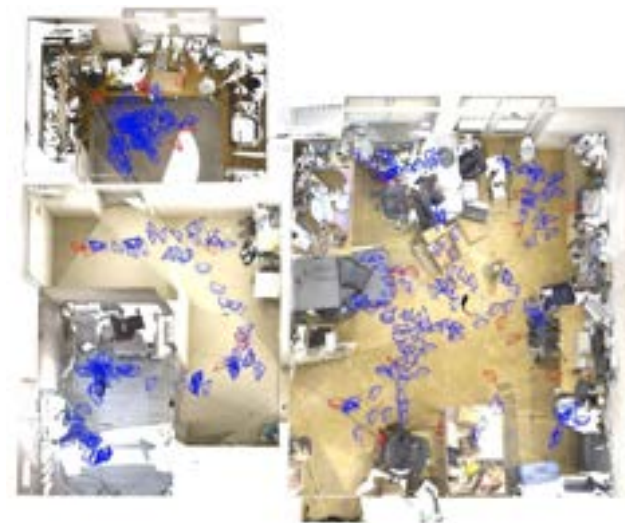
**実応用を目指す際に必要な評価ができるようにデータセットを整備**

ScanNet

ScanNet++  
クラス数増加・点群の質が向上



新視点合成では訓練に使用していない軌道外でテストを実行



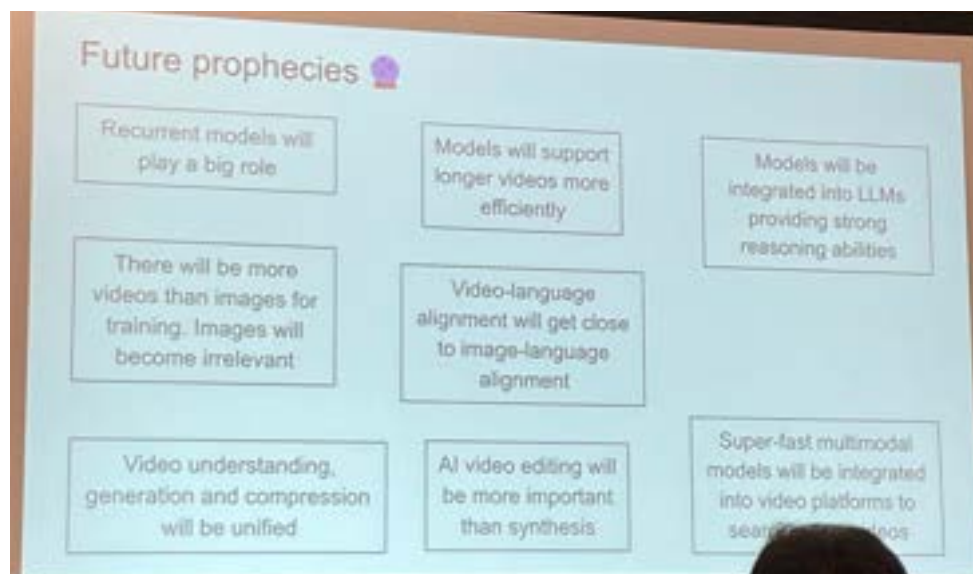
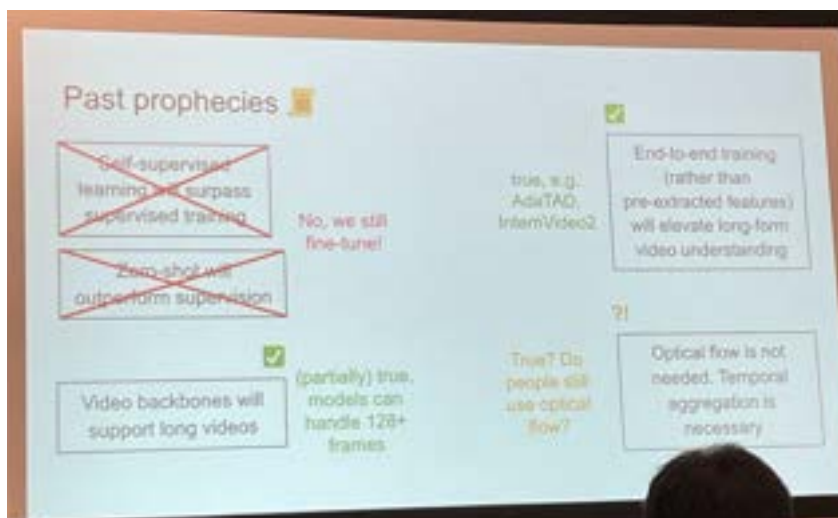
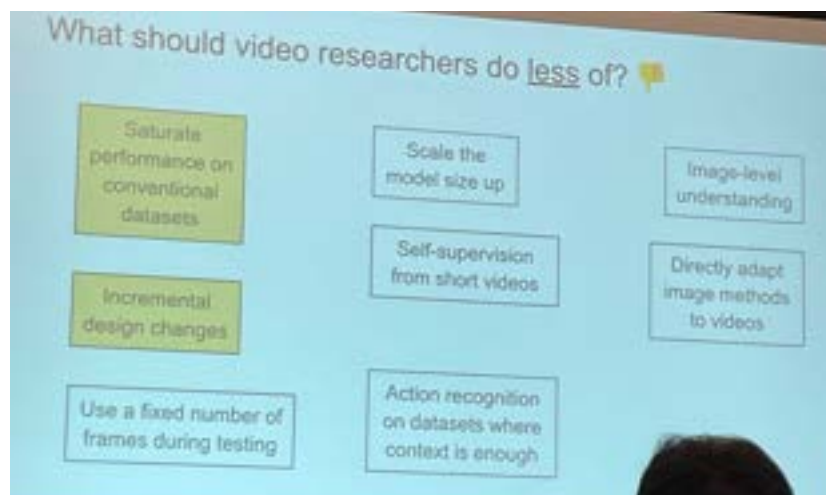
赤がテスト  
青が訓練

## 動画認識系のワークショップ

- ❑ [Learning from Procedural Videos and Language: What is Next?](#)
- ❑ [What is Next in Video Understanding?](#)
- ❑ 気づき
  - ❑ LLMに関連するものは多い
    - ❑ LLMでナレーション情報を変換してうんぬん
    - ❑ MLLMの弱点の分析と改善
    - ❑ LLMの要領でNext Token PredictionでVideo Taskを解く
  - ❑ Vision/LLMをスケールさせるうえでのデータバイアス
  - ❑ あんまり細かいことを考えてないTop Researcherもいる
    - ❑ トップ研究者でも、データで殴ればOKみたいなことしか考えてないような人がいると思うと、Fine-grainedな部分に取り組むことでの勝ち筋もあるように感じる。
    - ❑ → fine-grainedにいくか、データで殴るかの2極化？
  - ❑ 現状のモデルを使ってみて、取り組むべき/取り組める課題を見つけ出す重要性を再確認 (MLLMが何でもできそうだからこそ、諦めずに何ができないかを探す.)

## WS What is Next in Video Understanding

### □ 動画認識のこれからに関するワークショップ





## WS What is Next in Video Understanding?

- ❑ Angela Yao のトーク: Video Question-Answering in the Era of LLMs
- ❑ やはりMLLMがそれっぽくタスクに取り組めるので、そのなかでもできていないのはどういった部分になるのか、の仮説・実験から初めて、それに対処する手法を提案している。(ただし提案は多くの場合がデータの準備)
  - ❑ Temporal Understanding
    - ❑ Distinguish events that happen before vs. after concurrently?
  - ❑ Visual Grounding
    - ❑ Are predictions anchored on relevant video content?
  - ❑ Multimodal Reasoning
    - ❑ Reasoning based on language and visual inputs?
  - ❑ Robustness
    - ❑ Perturbations on video and question

## CVPRにおけるRobotics(1/2)

- ❑ WS: Embodied AI(未知の実体を持つAI)の最新の成果を共有し議論
- ❑ Roboticsでは「Embodiment gapを克服する事がロボットAIの拡張の肝
- ❑ 例) Embodied Mobile Manipulation:
  - CV研究者の視点からMobile Manipulator(4足やHumanoidを想定)を運用について講演 ※「操り」を考える上では「未知」の取り扱いに議論がある。
  - 例) S. Song氏「In-the-wild robot teaching without in-the-wild robots」
  - ← ICRA (Roboticsのトップカンファ)にもこの思想が流入しつつある



<https://embodied-ai.org/>

Human Demonstration



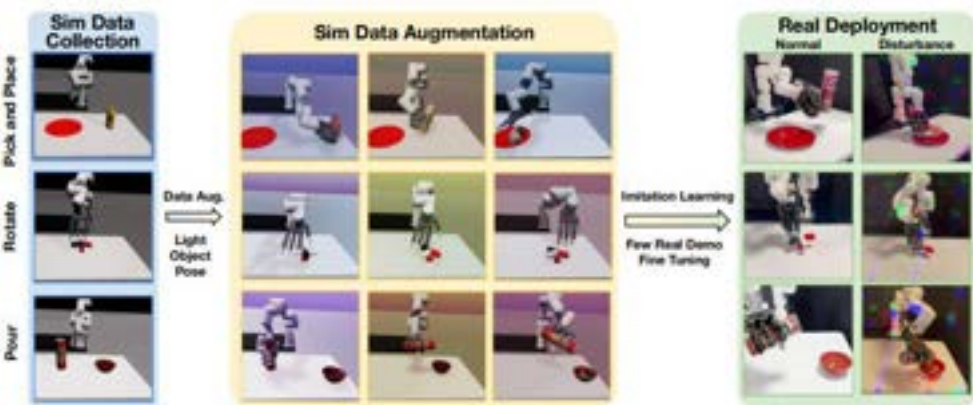
身体性の克服には、ロボットの種類を増やして学習データを拡張することが一般的。  
→ 一方で一つのアイデアとして、ロボットを用いない提案(ここではカメラの固定と相対的な位置の学習)によって、そのギャップは埋まったものと想定できる。

<https://arxiv.org/abs/2402.10329>

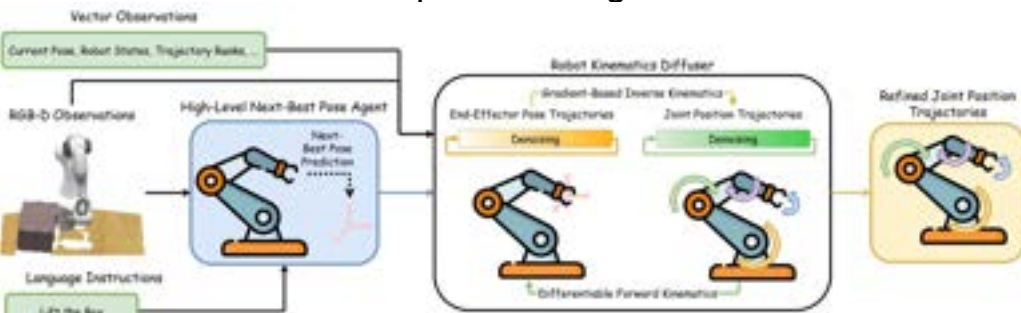
# CVPR 2024 の動向・気付き(105/116)

## CVPRにおけるRobotics(2/2)

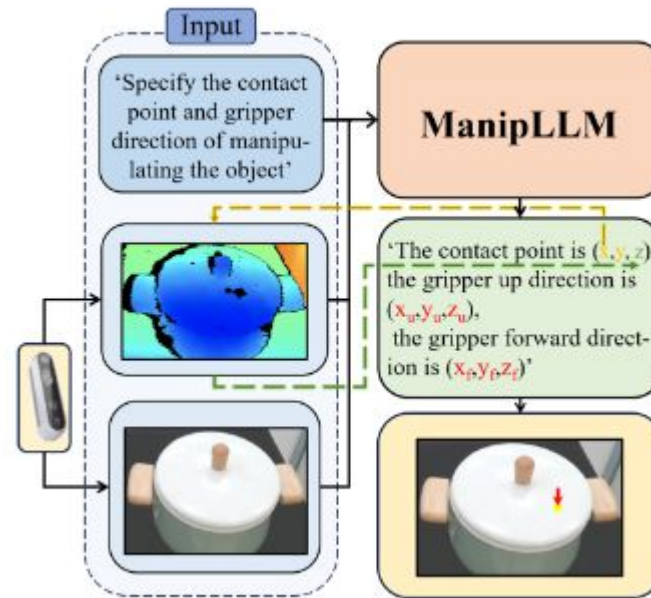
- ロボット応用についてはRobotics/CVのそれぞれの専門家が目指す所が同じになった印象—CVPRはロボット実験は不要だがレベル的には十分可能な水準
  - ポスターや講演においてロボットが目立つ/ Diffusion modelの軌道生成・画像処理が多い
  - 時期的にICRA(5月)→CVPR(6月)であるため、どれに出すかは悩ましくなるだろう。



<https://arxiv.org/abs/2402.14795>



<https://yusufma03.github.io/projects/hdp/>



<https://sites.google.com/view/manip-llm>



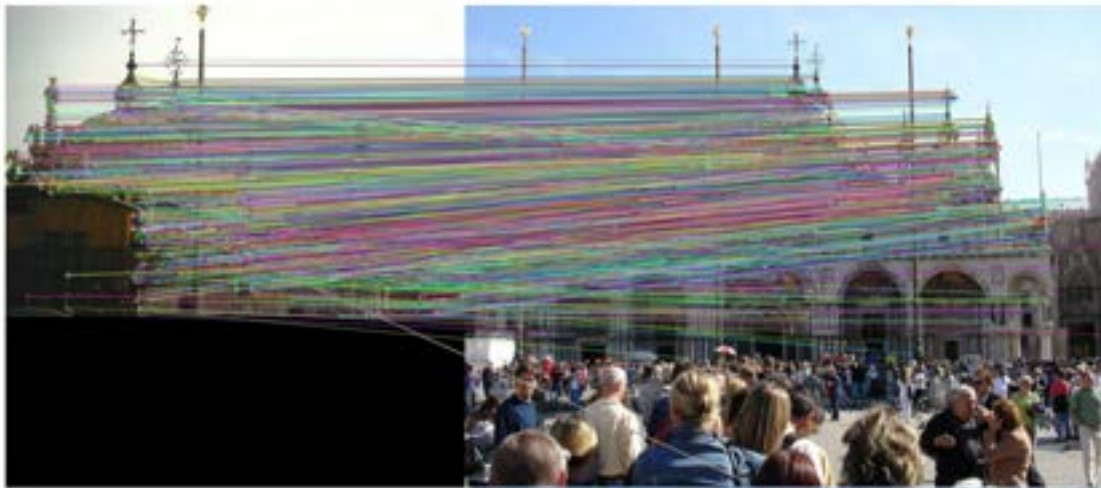
<https://sites.google.com/view/diffusion-edfs>

## WS EarthVision: Large Scale Computer Vision for Remote Sensing Imagery

- 国連SDGsの目標達成にCV×リモセンがどのように貢献できるのかを議論するWS
  - Keynote 2件、Panel Discussion 1件、ポスターのライトニングトークで構成されオーラルは技術としての新しさという観点ではなく、アプリケーションとしての観点のWS
    - ポスターに関しては技術的な内容も多いためあえて分離していると推察
  - Keynote 1⇒EUDRというRegulationできたけどEOとしてどう貢献すべきか
  - Keynote 2⇒オーストラリアの森林火災モニタリングについてCVを組み合わせでどう行っているか
  - 一般的なCV文脈であまり語られない技術としては、衛星観測をCVで分析を実施し、その結果をさらにシミュレーションに投入、未来予測を行うCV×データ同化が行われている点(Paper award: Deep Generative Data Assimilation in Multimodal Setting(Y. Qu et al.) など)
  - 一方でここでもMMFM(Multi-Modal Foundation Model)が猛威を振るっている(Paper award :Sat2Cap: Mapping Fine-Grained Textual Descriptions from Satellite Images(A. Dhakal et al.)など)
  - Panel DiscussionもMMFMの文脈であったが、結局SDGsに関する環境監視はできるとして誰がそのモデル作り、アップデートしていくのかという課題提起で終わった

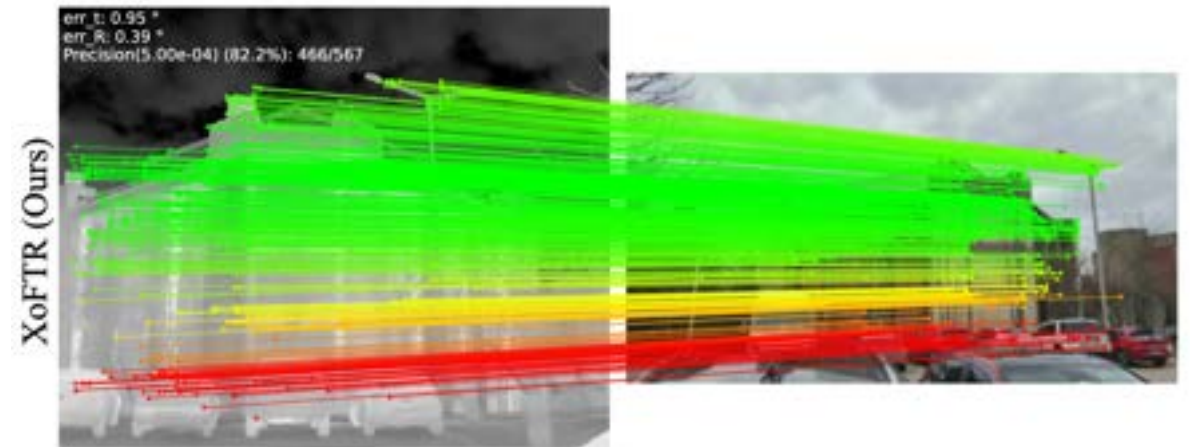
## WS Image Matching: Local Feature and Beyond

- 6回目となる局所特徴量をテーマとしたワークショップ.
  - LLMの時代になっても局所特徴量は進化を続けている.
  - RGB同士だけでなく,異なるモダリティの画像(IR, Thermal)とのマッチングがトレンド
  - 実はKaggleのIMCコンペもこのワークショップの一部.



角度も天候も時間帯も異なる画像同士のマッチング

J. Edstedtet al. "DeDoDe v2: Analyzing and Improving the DeDoDe Keypoint Detector", in CVPR 2024.

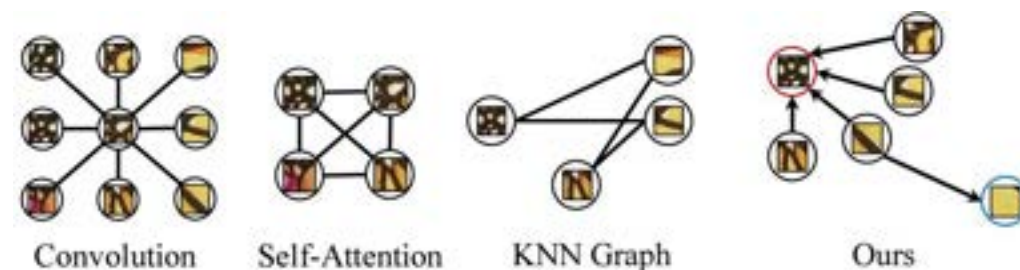
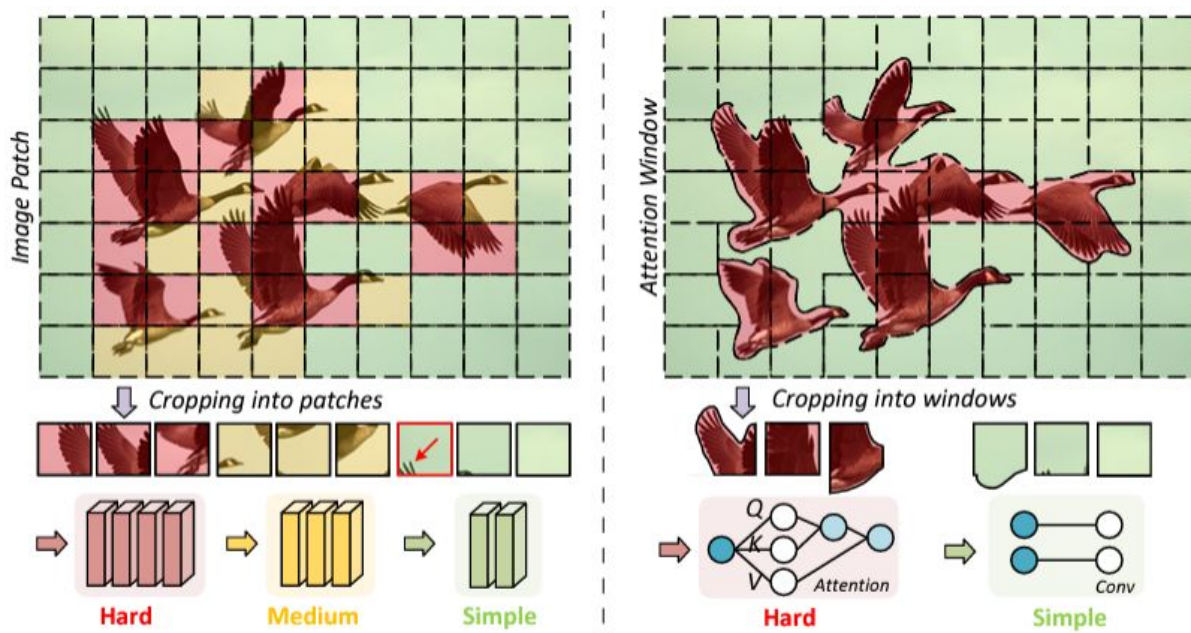


RGBとIR画像の異なるモダリティ同士でのマッチング

O.Tuzscuoglu et al. "XoFormer: Cross-modal Feature Matching Transformer", in CVPR 2024.

## 超解像のメタ話

- 一枚画像中で異なる復元の難易度に着目する話が多くなってきているのかも？
- 超解像では、復元が難しい領域と簡単な領域が一枚画像中に混在している。
  - パッチごとにHard/Medium/Simpleで難易度分けするClassSRを代表として、一枚画像中で難易度分け → 割り当てるパラメータ量を変化させることにより超解像処理の計算量を削減する試みが増えてきている



[Image Processing GNN: Breaking Rigidity in Super-Resolution](#)

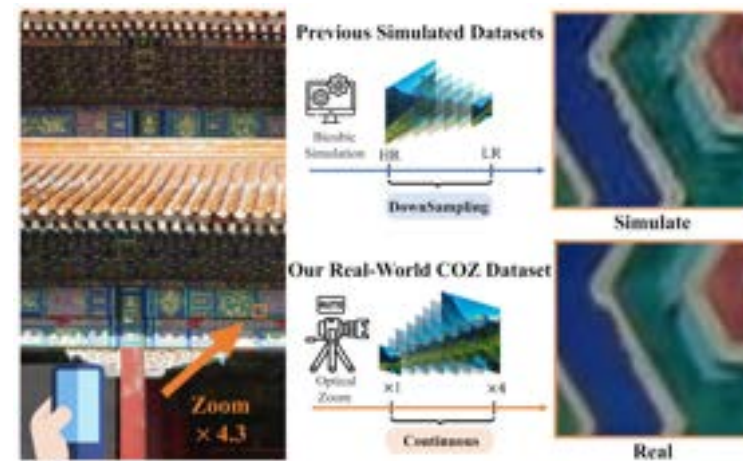


Figure 4. Visualization of adaptive bit-mapping of AdaBM. Results are from EDSR-AdaBM (w8 a6MP).

[AdaBM: On-the-Fly Adaptive Bit Mapping for Image Super-Resolution](#)

## 超解像のメタ話

- ❑ 従来の2倍/4倍といった固定のスケールではなく、**任意スケール**に対応した超解像が注目されていると感じる
- ❑ CVPR2024でもベンチマークの提案があった: COZデータセット

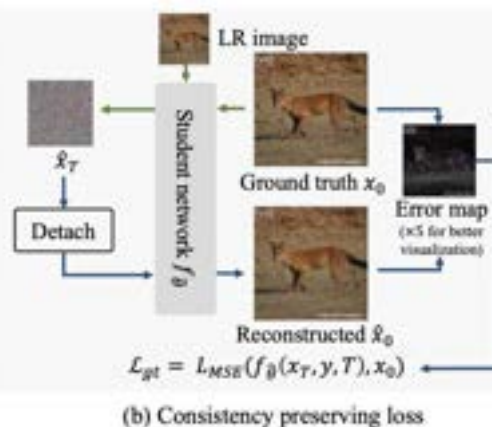
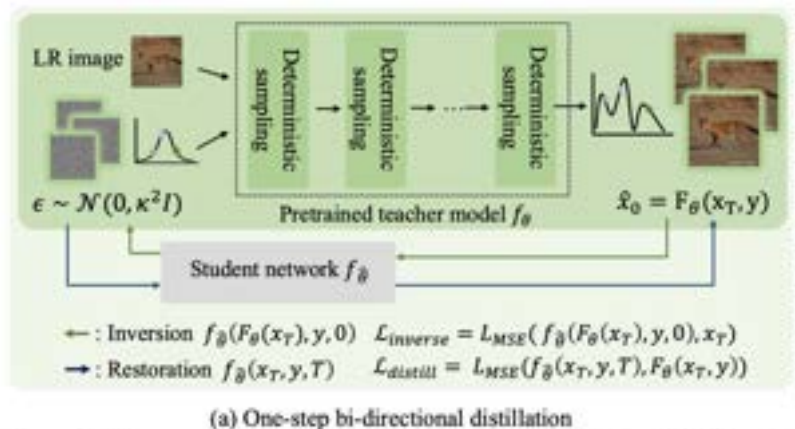


COZ データセット論文より引用

<https://arxiv.org/pdf/2403.10255v1>より引用

## 超解像のメタ話

- Diffusion ベースの手法の増加
  - ステップ数の削減
  - LRとの一貫性の保持
  - 任意スケールの超解像
- LR画像に映っていないものを生成させない工夫が重要



<https://github.com/wyf0912/SinSR> より引用

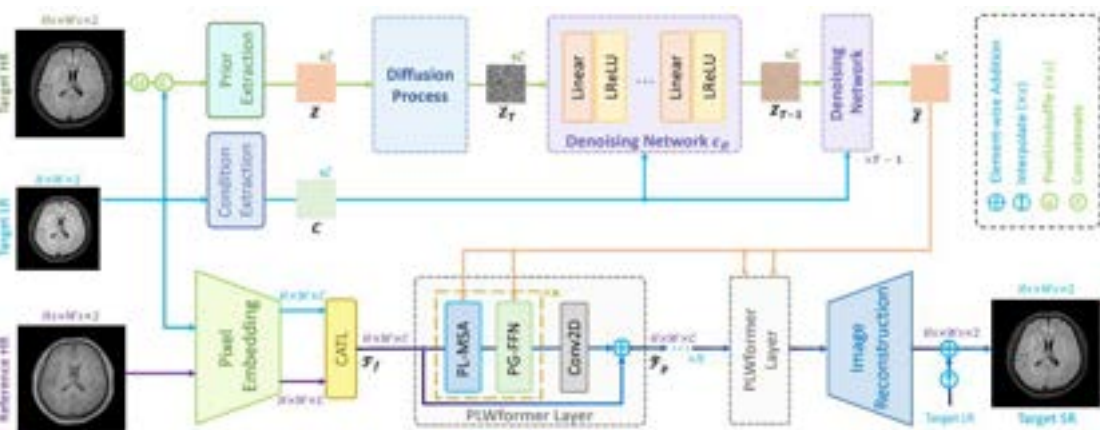
[https://openaccess.thecvf.com/content/CVPR2024/papers/Wu\\_SeeSR\\_Towards\\_Semantics-Aware\\_Real-World\\_Image\\_Super-Resolution\\_CVPR\\_2024\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2024/papers/Wu_SeeSR_Towards_Semantics-Aware_Real-World_Image_Super-Resolution_CVPR_2024_paper.pdf) より引用



# CVPR 2024 の動向・気付き(111/116)

## 超解像のメタ話

- Diffusionモデルによる生成画像を参考にする手法が増加傾向
- 長所: 意味的に正しく、かつ写真のようにリアルなディテールを復元できる



- テキストでガイドを入れる手法も最近人気



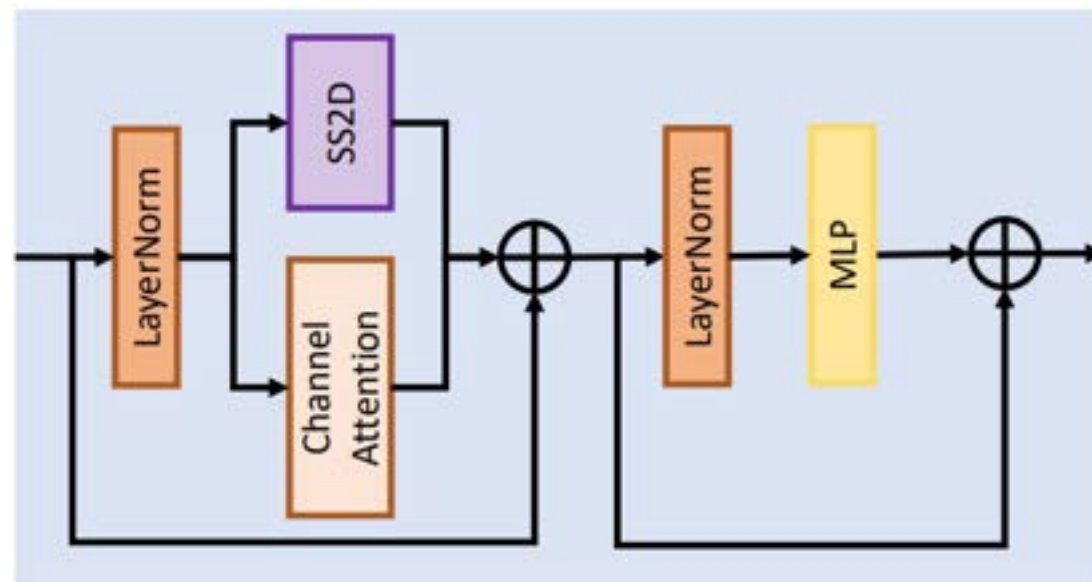
LR input      DPS [12] using diffusion model trained on Imagenet      Ours  
"A statue of Walt Disney holding Mickey Mouse hands is showing in front of Cinderellas castle."

# CVPR 2024 の動向・気付き(112/116)

## 超解像のメタ話 CVPR Workshop : [NTIRE 2024](#)

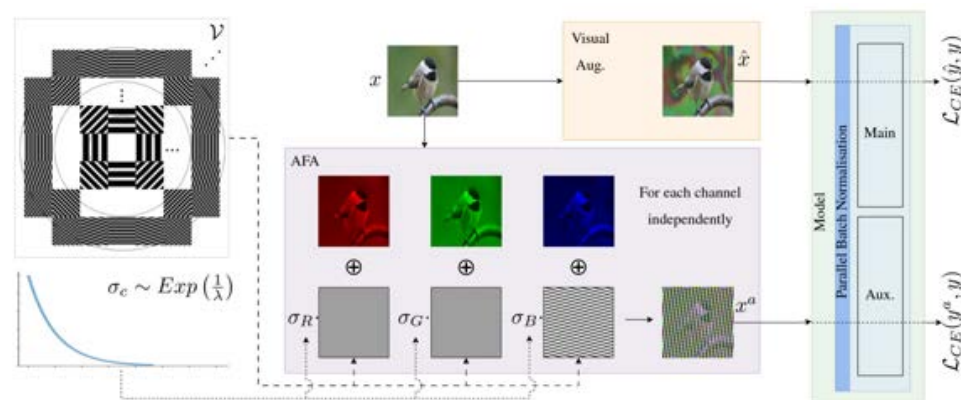
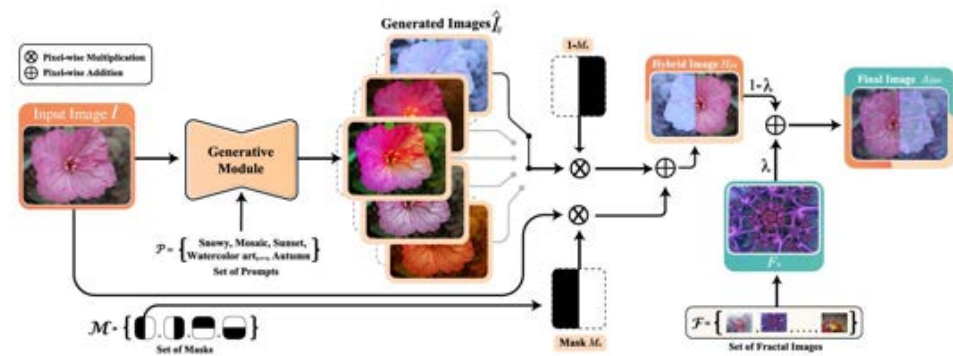
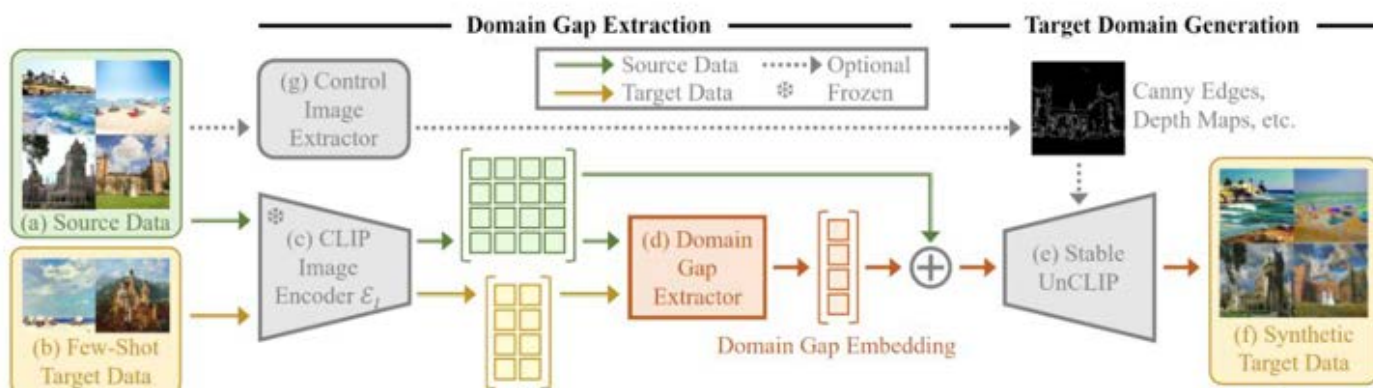
- 新しいトレンドは**Mamba**かも？
- MambaSRがNTIRE 2024 Challenge on Image Super-Resolution (×4)にて最高性能を達成！
- Mamba: 状態空間モデルの利用
  - MambaSRではState Space Transformer (SSFormer)を導入
  - HATをベースとして、Hybrid Attention BlocksをSSFormerによって置換

Team	Rank	PSNR (primary)	SSIM
XiaomiMM	1	31.94	0.8778
SUPSR	2	31.41	0.8711
UCAS-SCST	3	31.28	0.8666
SYSU-SR	4	31.19	0.8660
Jasmine	5	31.18	0.8665
ACVLAB	6	31.18	0.8655
mandalinadagi	7	31.13	0.8648
SKDADDYS	8	31.11	0.8643
KLETech-CEVI	9	31.03	0.8633
SVNIT-NTNU	10	31.03	0.8633
ResoRevolution	11	31.01	0.8647
BetterSR	12	30.97	0.8621
Contrast	13	30.69	0.8563
BFU-SR	14	30.55	0.8560
SCU-VIP-LAB	15	29.78	0.8506
Nudter	16	30.17	0.8446
JNU-620	17	30.43	0.8426
LVGroup-HFUT	18	29.98	0.8380
Uniud	19	29.97	0.8440
SVNIT-NTNU-1	20	29.34	0.8199



## データ拡張のメタ話

- ❑ 学習画像に対して、別の画像を重ね合わせる方式のデータ拡張手法がトレンドかも。
- ❑ データ拡張の目的: 各タスクの純粋な性能向上、ロバスト性向上
- ❑ 提案手法によって実現できること
  - ❑ データ量のかさまし
  - ❑ データのドメイン分布の幅を広げる (学習画像の特徴に幅を持たせる)
- ❑ 言語を絡めた生成モデルを利用してデータ拡張用の画像を生成する論文が増えてきているように感じる。



## WS AI4Space 2024

- 宇宙分野の課題解決にAIを活用
  - 自立宇宙システムのための画像認識などの問題(訓練データの不足, 未知の運用環境など)

極端な環境でのミッションにAIが役立つ



引用: <https://aiforspace.github.io/2024/>

## WS AI4Space 2024

- 特に人工衛星や宇宙船の姿勢推定の論文が多い(1/2)
- Exploring AI-Based Satellite Pose Estimation: from Novel Synthetic Dataset to In-Depth Performance Evaluation by Fabien Gallet
  - 宇宙空間では大量の実データの収集にコスト莫大  
→人工の宇宙空間画像120,000枚のDBを作成

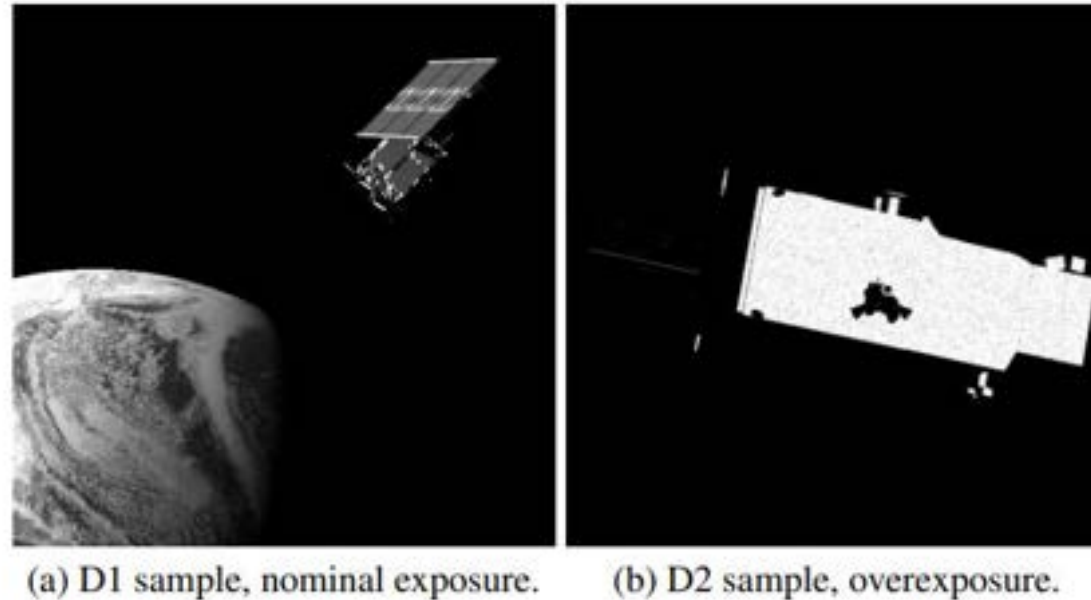


Figure 1. Examples of RAPTOR dataset images.

## WS AI4Space 2024

- 特に人工衛星や宇宙船の姿勢推定の論文が多い(2/2)
- Monocular 6-DoF Pose Estimation of Spacecrafts Utilizing Self-iterative Optimization and Motion Consistency by Yunfeng Zhang
  - 3Dモデルを構築し, MKPNetの自己反復最適化により姿勢推定精度の向上
  - 宇宙船の軌道推定のコンペ SPARK 2024 Challenge で3位
- Revisiting the Domain Gap Issue in Non-cooperative Spacecraft Pose Tracking by Kun Liu
  - オンライン学習, オフライン学習, フライトの3フェーズ
  - SPARK 2024 Challenge で2位

Team	Translation_error	Orientation_error	Pose_error
csu_nuaa_pang	0.0252	<b>0.0187</b>	<b>0.0252</b>
lucca	<b>0.0243</b>	0.0448	0.0508
juanqilai (ours)	0.0335	0.0843	0.0934
igodrr	0.0823	0.7214	0.7417
yanlj	0.0335	1.0362	1.0454
dwoiwode	0.0739	1.4460	1.4651
nalixvignola	0.1303	2.1407	2.1741

SPARK 2024 Challenge Spacecraft Trajectory Estimation

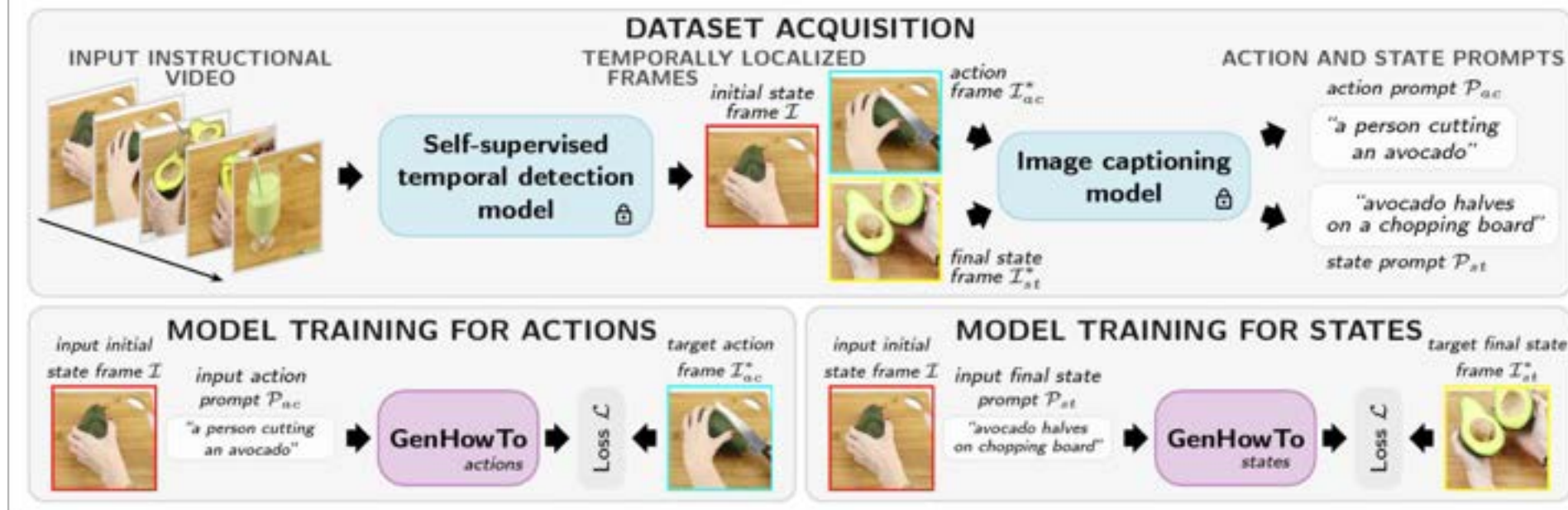
# CVPR 2024 論文サマリ

---

- CVPR 2024 1ページ 1論文サマリ

## GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos

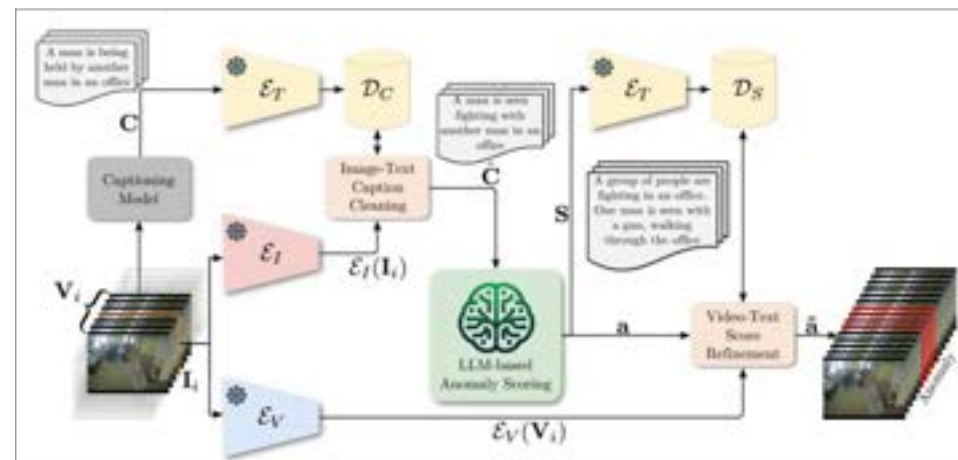
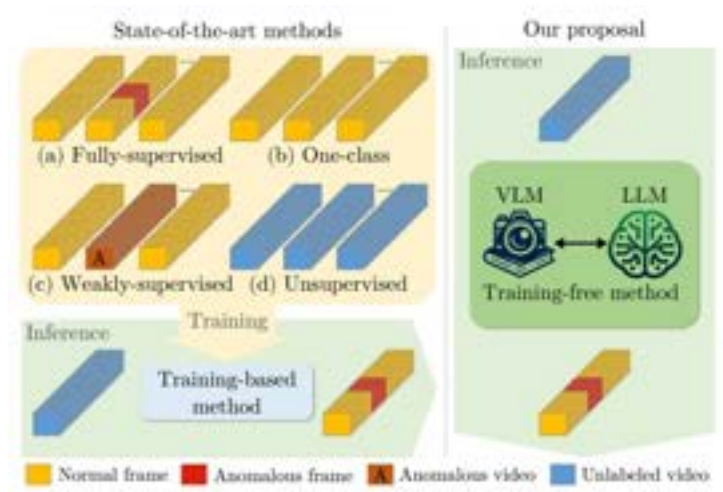
- ❑ **概要:** 画像、テキストから、画像の情報を保ちながら、Actionsや物体状態変更の画像を生成するタスクと手法GenHowToの提案。SeenとUnseenにおいてかなり精度が高いSoTAを実現。
- ❑ **手法:** HowToビデオデータセットから、初期の物体状態、状態変更Action、最後の物体状態の画像とテキストのTripletsを取得し、大規模学習データセットを得た。画像とテキストのペアから、ControlNetをベースとしたDiffusionモデルでターゲット画像を生成する。
- ❑ **感想:** Modular的なモデルを使うと、Hallucinationに対してもっと強くなりそう？ステップ数が増えた場合、中間の状態はどれくらいConsistentなのか気になる。





## Harnessing Large Language Models for Training-free Video Anomaly Detection

- **概要:** ビデオから Training-free で anomaly を検出するタスクおよび手法を提案。Anomaly 検出の学習データの準備が人手かかる。また、Anomaly がスパースで学習もコストが高い。ここで、LLM に含まれる異常行動を識別する能力を Anomaly 検出に活用し、学習用いずに複数のベンチマークで学習済みのモデルよりも高い SoTA を達成。
- **新規性:** 問題定義が新規で面白い。既存の手法が主に学習ベースであった。LLM を Anomaly 検出タスクへの活用も新しいポイント。
- **手法:** 手法は VLM と LLM を結合し 3 段階で行なっている。ステップ 1 では、画像キャプションできる VLM でビデオから複数枚の画像をキャプションする。その上画像とテキストの類似度でキャプションをクリーニングする。ステップ 2 では LLM を利用し、クリーニングされたキャプションから粗末な Anomaly 検出を行う。ステップ 3 では、ビデオとテキストの類似度を評価可能な VLM で Anomaly をリファインする。
- **感想:** 画像ベース VLM とビデオベース VLM の照合で粗末 → リファインされた検出を行なっている。VLM と LLM の賢い活用。



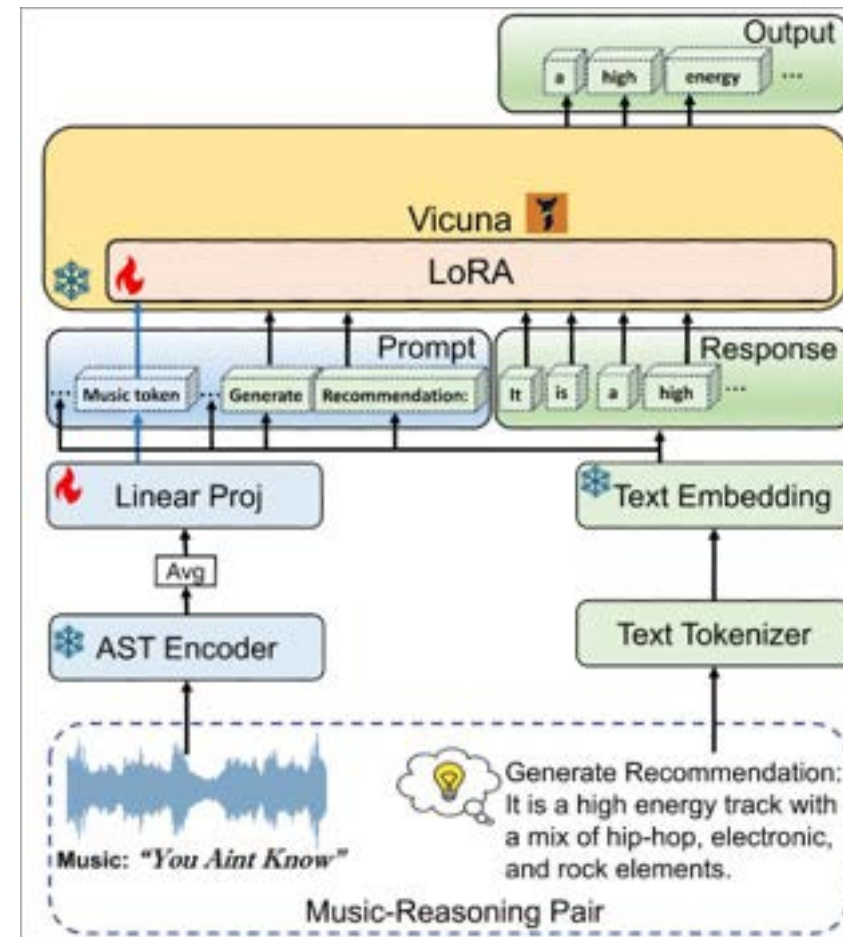
## OpenBias: Open-set Bias Detection in Text-to-Image Generative Models

- **概要:** Text-to-ImageモデルのバイアスをOpen-setで検出するタスク・手法OpenBiasの提案。OpenBiasで検出したバイアスが人間の判断と近い。実験により、Stable Diffusion 1.5, 2, XLに含まれるバイアスを発見した。
- **新規性:** 既存のText-to-Imageモデルのバイアス検出はClose-set設定が多く、検出できるバイアスが限定的であった。
- **手法:** LLMからテキストに含まれる可能なバイアスの項目をOpen-setでまず認識し、その後生成モデルでText-to-Imageを実行する。最後に、VQAモデルで生成された画像群に関してあらゆるバイアスの項目でバイアスを検出・定量的評価を行う。
- **感想:** Open-setでバイアスを検出するアイデアが面白い。手法はシンプルでStraightforward。どうやって有利なバイアスとよくないバイアスを区別するのが気になる。



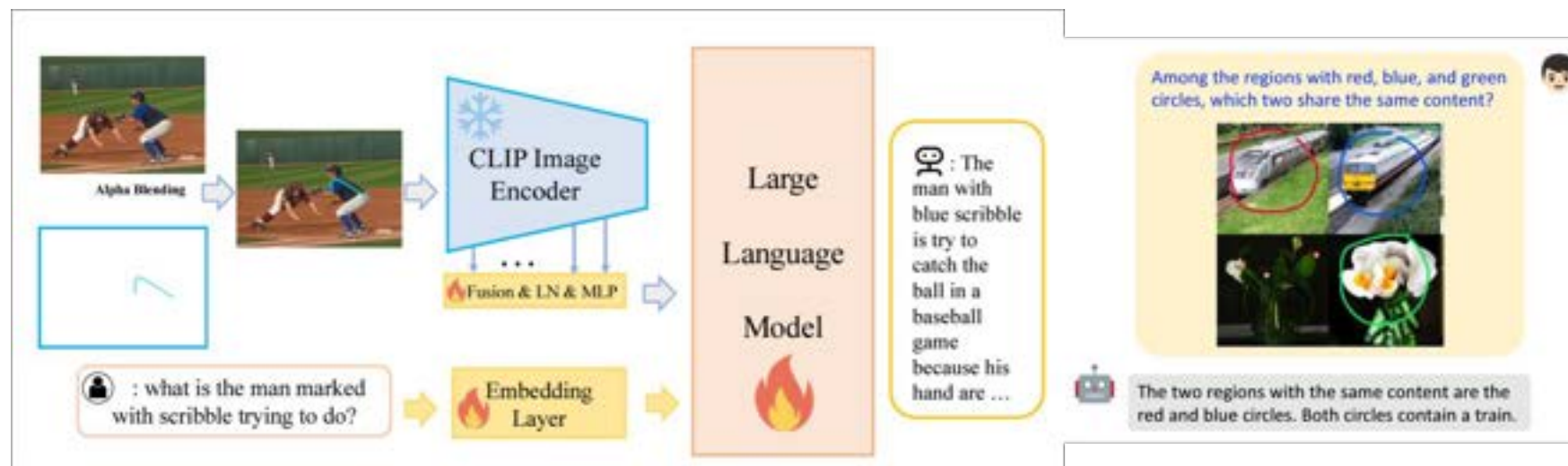
## MuseChat: A Conversational Music Recommendation System for Videos

- ❑ **概要:** Human userと対話しながら、入力ビデオに音楽をRecommendする手法MuseChatの提案。既存の大規模ビデオデータセットとLLMを利用して独自の学習データセットを提案。
- ❑ **新規性:** 既存手法が人間のpreferenceを考慮しないものが多い。MuseChatでは対話の形式でuser preferenceに従ったRecommendationができる。MLLMの力で、音楽をRecommendする際の理由などを可視化・解釈可能となる。
- ❑ **感想:** Applicationとしては面白くて実用性が高い。音楽に関して対話で色々理解できるMusic-LLMができると面白そう。データセット構築の部分が面白そう。



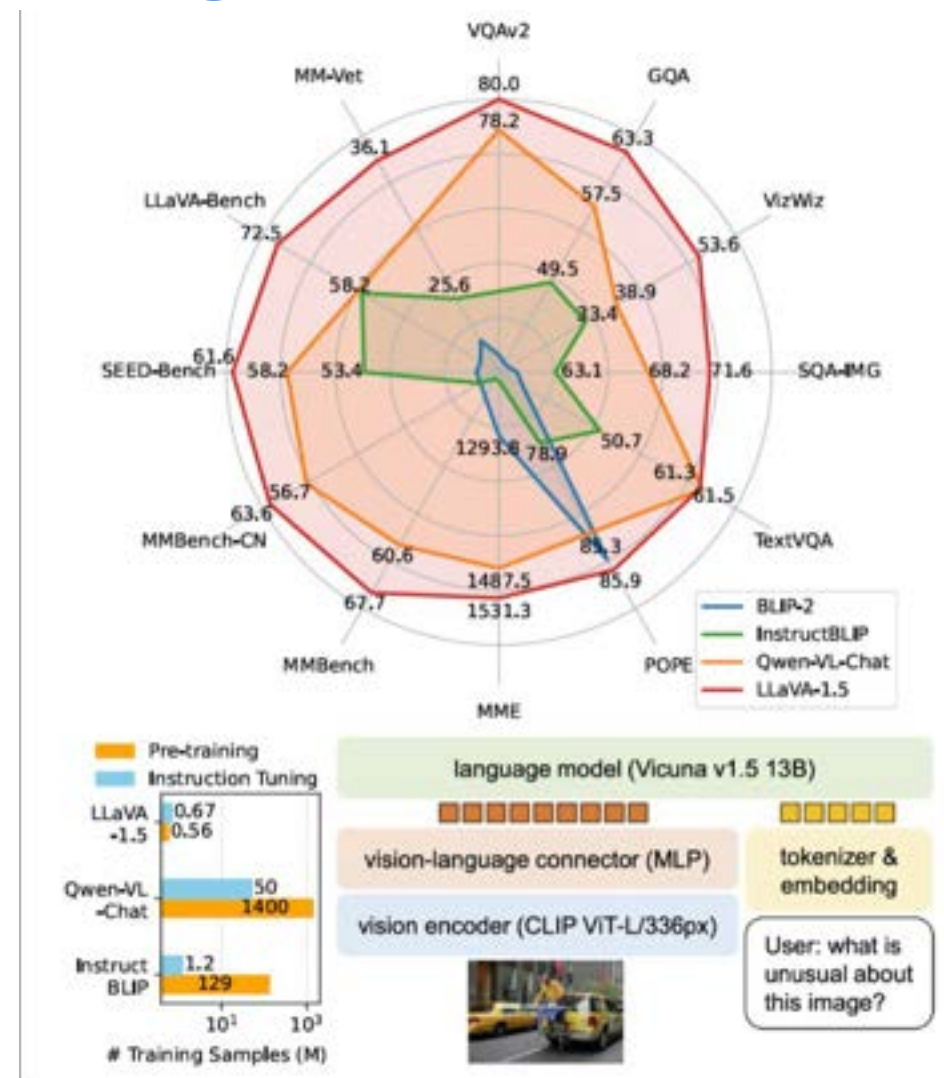
## ViP-LLaVA: Making Large Multimodal Models Understand Arbitrary Visual Prompts

- ❑ **概要:** LMMsの画像Regionベース認識能力を高める仕組みの提案。あらゆるRegionを指定するPromptsを扱える。Visual Promptsを元画像の上に重ねて(alpha blending) CLIPエンコーダーに入力。その後得られたMulti-layerの特徴量をMLPによりvisual tokensを得る。Instruction tuningの段階でtext embedding layers, LLMなどをfinetuneする。提案のViPがVisual7Wなど複数のRegionベース認識ベンチマークでSoTAを実現。
- ❑ **新規性:** 既存のLLaVAやMiniGPTなどのLMMモデルがRegionベースの認識能力が不足している。また、既存のRegionをテキストで指定する手法のRegion表示が座標やbounding boxなどでユーザが自由に扱えない場合がある。ViP-LLaVAではRegionを指定することで、SOTAなRegion認識能力を得た。また、線、矢印、サークルなど様々な形式の入力Visual Promptを扱う。
- ❑ **感想:** 複数の画像の領域や同一画像中の複数領域についてどこまで細かくCorrespondenceの認識ができるのかが気になる。



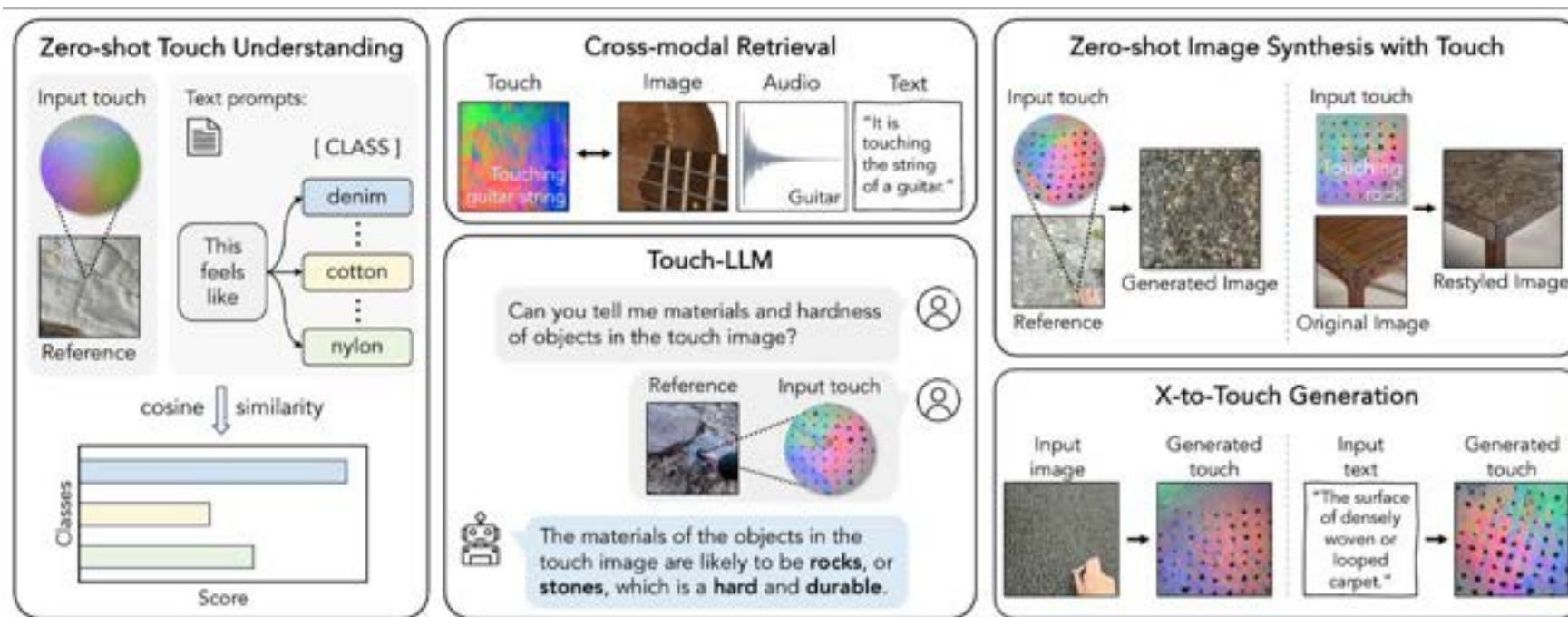
## Improved Baselines with Visual Instruction Tuning

- **概要:** 新しいLMM LLaVA-1.5を提案。11ベンチマークデータセットでSoTAな精度を達成。また、既存のIMMの学習と比べて必要な学習データが規模小さい(600K画像テキストペア)かつ学習時間が短い(A100一枚で1日程度)。
- **新規性:** Controlled設定で、LMMsの色々なデザインを実験した。下記のいくつか面白い発見があった。画像をGridsに分割することで、LLaVAがもっと高解像度の画像を適応可能。LLaVAの学習データの75%の解像度を下げてもモデルの性能があまり落ちない。モデルサイズとデータの質が学習の結果と比例する。
- **手法:** LLaVA-1.5構造上LLaVAと比べて主に二つの違いがある: MLP cross-modal connectorとacademic task related dataの使用。
- **感想:** LLaVA-1.5の複数画像認識能力が不足している。



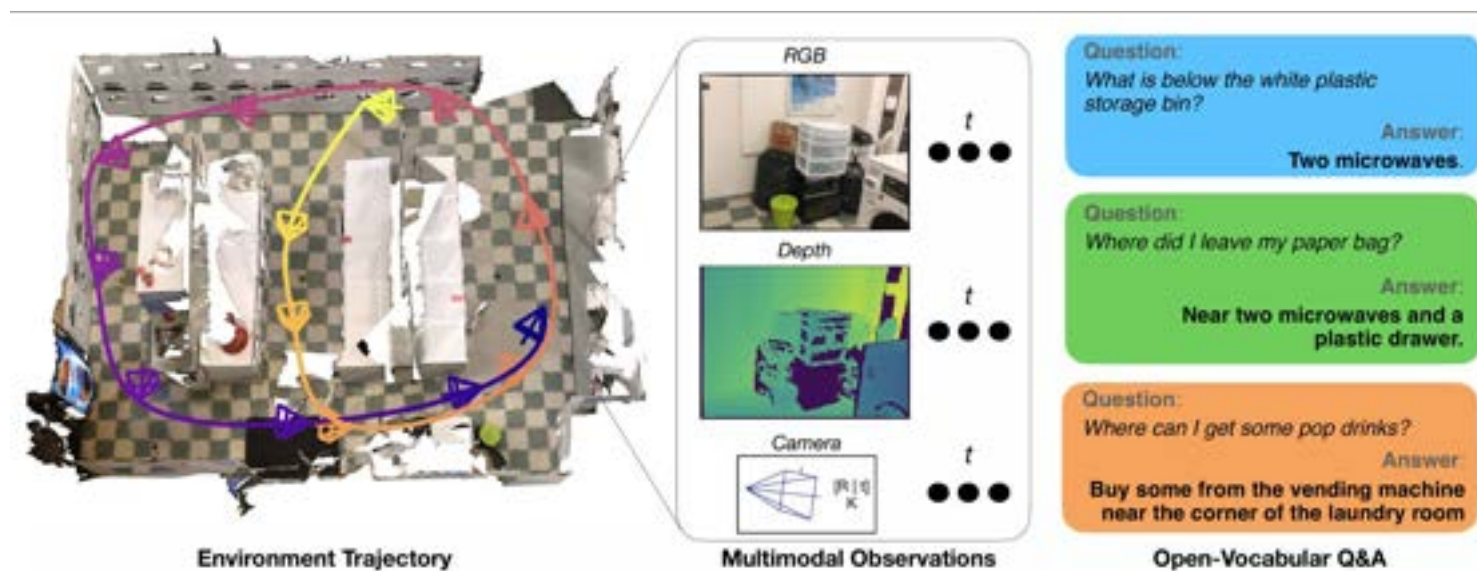
## Binding Touch to Everything: Learning Unified Multimodal Tactile Representations

- ❑ **概要:** 触覚を視覚、言語、音声とリンクさせるモデルUniTouchを提案。また、異なる触覚センサーを学習できるsensor-specific tokensを導入。UniTouchで様々な触覚理解のタスクを行える: Zero-shot触覚認識; 触覚表現から画像・音声・テキストのRetrieval、画像生成、X-to-Touch生成。特に、Touch-LLMで触覚から複雑な推論なども可能。
- ❑ **新規性:** 触覚表現のCLIP。様々な触覚タスクを統一した。また異なる触覚センサーを対応可能。
- ❑ **手法:** 触覚表現と視覚表現のContrastive Learningで触覚表現を学習。その後、視覚表現を通して他のモダリティとリンクさせる。
- ❑ **感想:** Touch-LLMが面白い。Robot manipulationに活用できる。触覚が重要だが、こういったアプリケーションで使えるか想像しづらい。



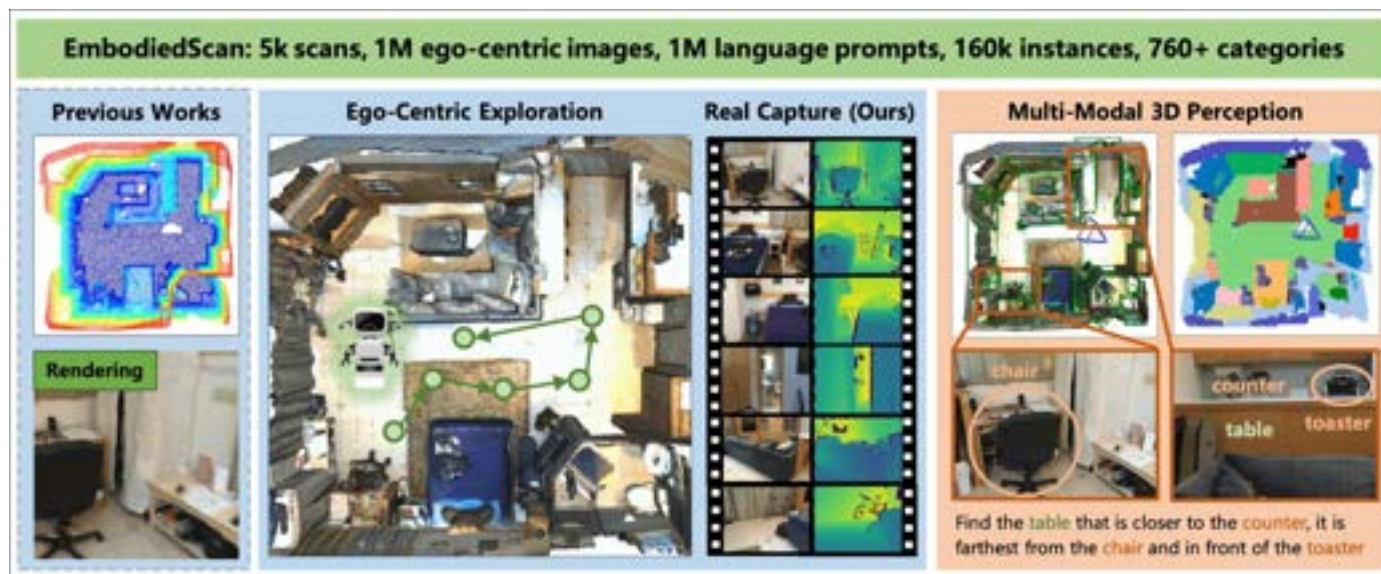
## OpenEQA: Embodied Question Answering in the Era of Foundation Models

- ❑ **概要:** Embodied Agentのための大規模データセットの提案。二つのモードを可能にした。モード1: Episodicな動画から質問を回答; モード2: Activeで環境探索しながら質問を回答。
- ❑ **新規性:** データセットの新規性がメイン。まず既存データセットは主にシミュレーション環境のみで、OpenEQAデータセットは実環境Episodic動画入力も可能にした。また、Open-Vocabularyで質問回答するところも新しい。
- ❑ **実験:** GPT-4Vを代表とした強いMLLMを使っても、人間レベルと差が大きい。特に、Spatial Relationshipsなどはチャンスレベル。
- ❑ **感想:** 同じくCVPR2024のEmbodiedScanと類似する。EmbodiedとLLMはトレンド。GPT-4Vなどを使っても、Embodied環境の問題がチャンスレベルなところが面白い。



## EmbodiedScan: A Holistic Multi-Modal 3D Perception Suite Towards Embodied AI

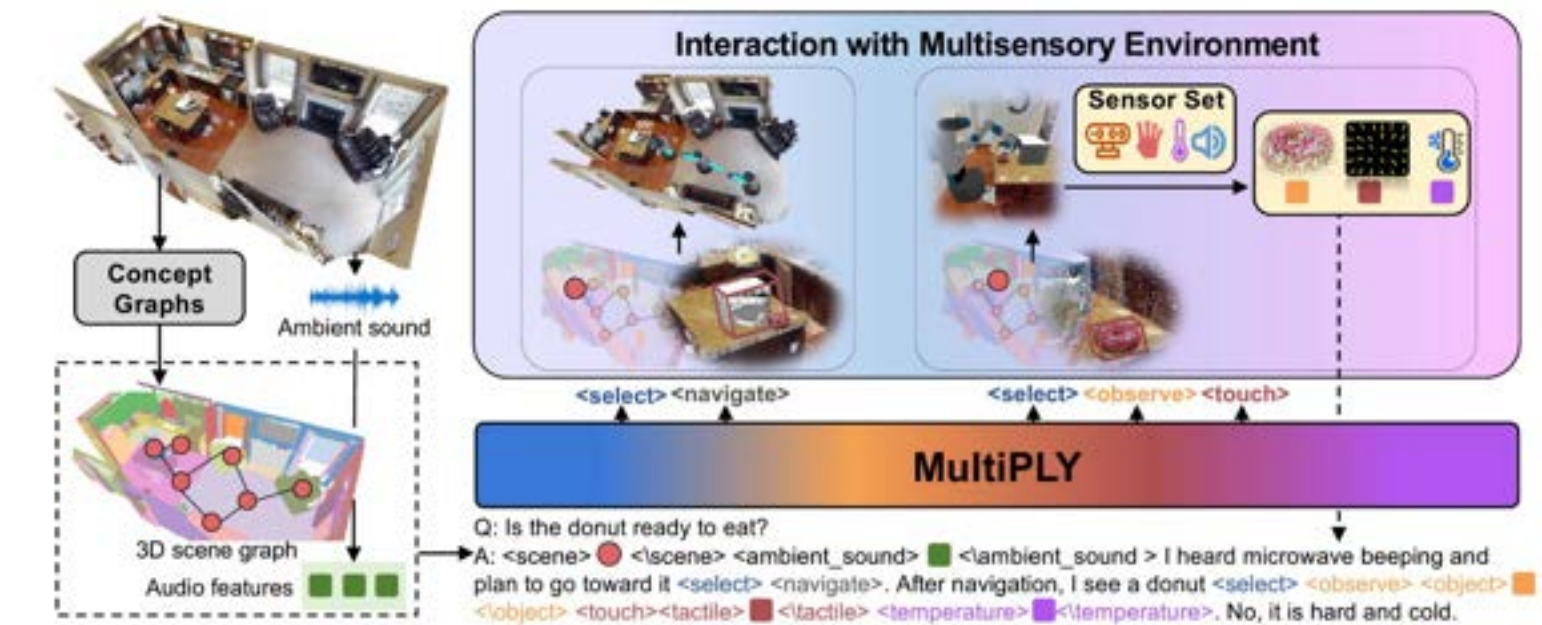
- ❑ **概要:** 大規模Indoor Embodied Agentの学習のためのデータセットEmbodiedScanを提案。EmbodiedScanでは、First Person ViewのRGBD動画入力から、3次元シーンのObject Detection、Segmentation、そして、Languageから3D Groundingなどのタスクを行う。
- ❑ **新規性:** タスクおよびデータセットの新規性がメイン。既存タスクでは、入力がHolistic的な3次元シーン、それと比較して、EmbodiedScanではFPVのRGBD動画。
- ❑ **手法:** 基本的なベースラインを構築した。構造的は、Multimodalを扱うEncoderとタスク別のDecoder。
- ❑ **感想:** 人間みたいなFirst Person View入力のEmbodiedとLLMの結合がトレンド。





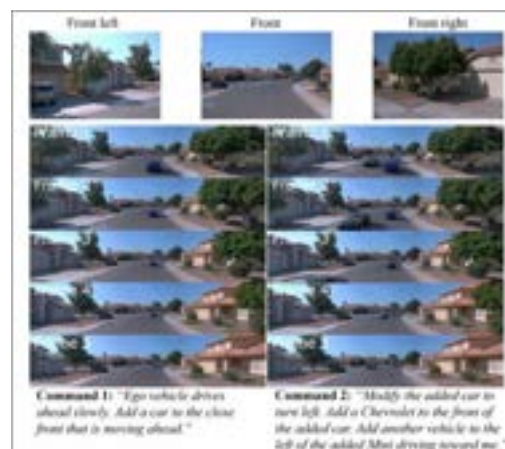
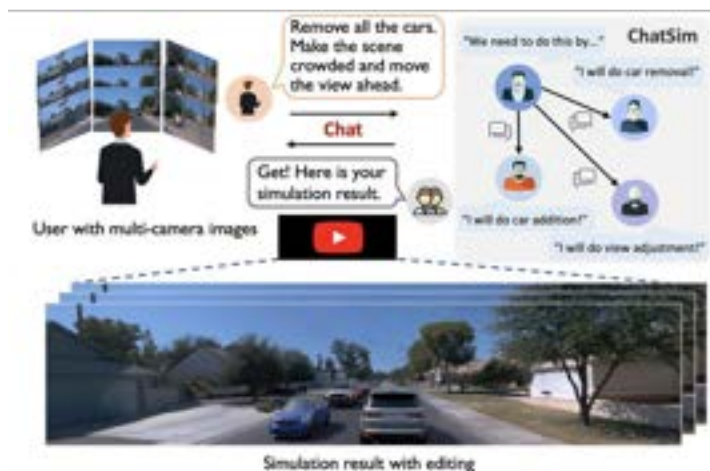
## MultiPLY: A Multisensory Object-Centric Embodied Large Language Model in 3D World

- ❑ **概要:** Embodied環境中、Multisensoryの情報を理解・Interactiveで自主的に必要なSensoryの情報取得できるMLLMモデルの提案。また、ChatGPTを利用し大規模Multisensory Interactionデータセットを提案。
- ❑ **新規性:** 既存のMLLMはMultisensoryのデータを理解できるモデルが多いが、ActiveでMultisensoryの情報を使って対話する、必要に応じて自主的にMultisensoryのデータをEmbodied Sceneから取得できるモデルがほとんどない。提案の仕組みは上記を実現。
- ❑ **手法:** ChatGPT、既存の3次元シーン環境HM3Dデータセット、3次元Multisensory物体データセットObjectFolderとObjaverseなどを活用してデータセットを構築。3D Scene Graphを介して、Pretrained LLMのInstruction Finetuningを有効的に行う。



## ChatSim Editable Scene Simulation for Autonomous Driving via LLM-Agent Collaboration

- ❑ **概要:** シミュレーションシーンを対話で編集可能な手法の提案。提案手法はLLMとMulti-camera NeRFの結合で、高精度かつflexibleでシーンを編集可能にした。
- ❑ **新規性・結果:** 自動運転の学習データ(視覚)の作成の新しい手法の提案;これまでの既存手法(例:3次元再構築、NeRF系手法、3次元シミュレーター)と比べて、生成したシーンのリアリティ・質が高い。また、言語入力により概要的な編集、細かい詳しい編集、編集したシーンを再編集などを可能にした。生成系の定量的評価も既存手法を大幅に上回った。
- ❑ **手法:**
  - ❑ 複数のAgent(LLMベース)がCollaborateし、入力されたコマンドを解析しながらシーンの編集を行う。また、それぞれのAgentは特定なタスクに集中しより高精度を達成。
  - ❑ Multi-cameraに対応可能な新しいNeRF手法を提案。
  - ❑ 外部リソースをシーンに追加可能にするために、Lightingを編集可能な新しいNeRF系手法も提案。



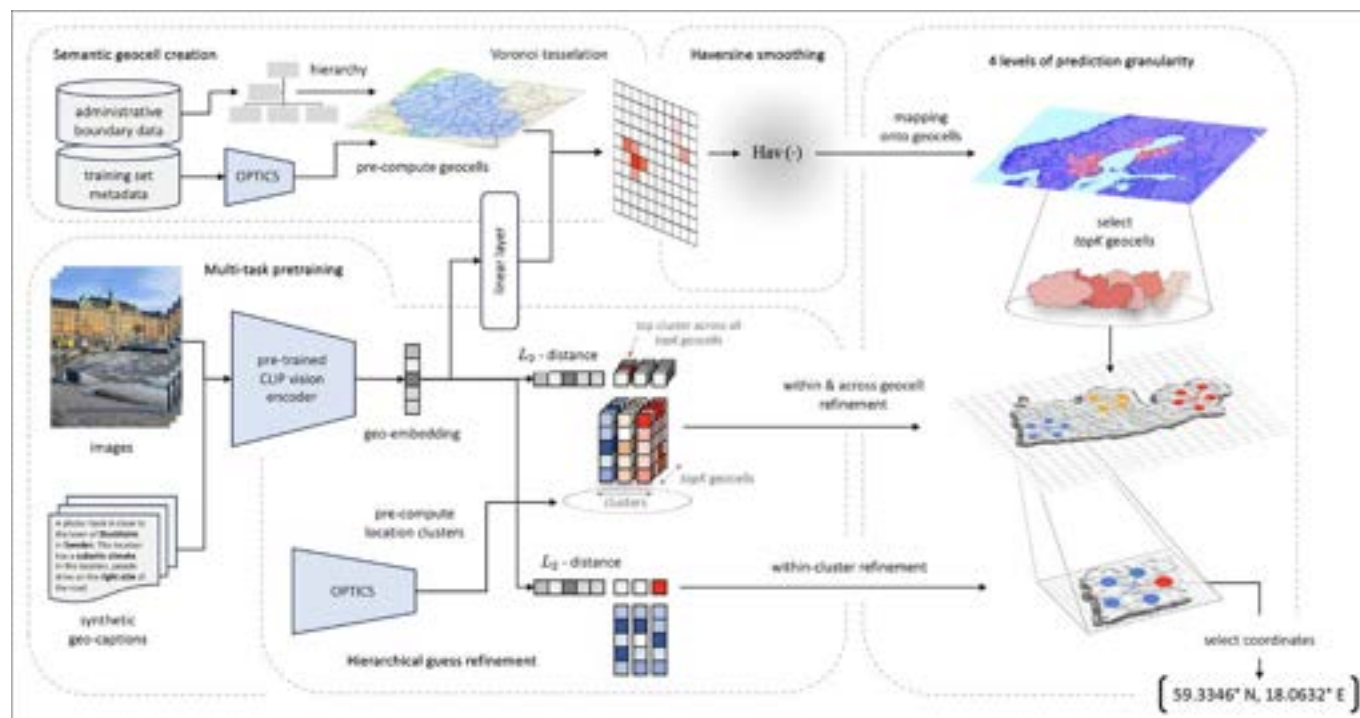
## SPOC: Imitating Shortest Paths in Simulation Enables Effective Navigation and Manipulation in the Real World

- ❑ **概要:** シミュレーション環境でShortest Path PlannersをImitation Learningをする手法を提案。提案のImitation Learningを大規模なシミュレーションデータセットで学習した結果、Unseen sceneと実環境で高い精度を達成(実環境50%以上の精度)。
- ❑ **新規性:** 大規模のリアリティ性が高いデータセットでImitation Learningで高いSim2Realの精度を達成することを示した。高いSim2Real性能を出すための知見を示した: ①シミュレーション環境内のAssetsの多様性; ②性能が良いVisual Encoderの使用; ③TransformerのLong context Windowsを採用; ④学習データの大規模化。また、データセットCHORESを公開。
- ❑ **手法:** 提案手法がRGBのみ使う。また、最近の手法でよく使われているLLMやMapping Moduleなどを使わない。Human demonstrationも使用せずに、Heuristic shortest path PlannersからImitation Learningを行う。
- ❑ **感想:** 大規模データセットと有効的な学習Policyが良かった。特に規模の効果が大きそう。シミュレーションのみで学習で、リアル環境の部屋で50%以上の精度を実現したことが不思議に思う。



## PIGEON: Predicting Image Geolocations

- ❑ **概要:** 場所の複数枚の画像/1枚画像から、その場所の座標を推定する手法の提案。地理情報、CLIPの利用で提案手法が人間の専門家より高い精度を達成。
- ❑ **新規性:** 既存手法は主にClusteringや対比学習などを用いて画像ベースでGeolocalizationを行う。そのため、unseen場所の推定精度が低い。提案手法は地理情報の利用およびCLIPの利用で、精密な地理情報や記述で強いgeolocationの特徴量をマルチタスクで習得した。
- ❑ **感想:** Wikipediaなどの情報で更に精密的にGeolocalizationできそう。



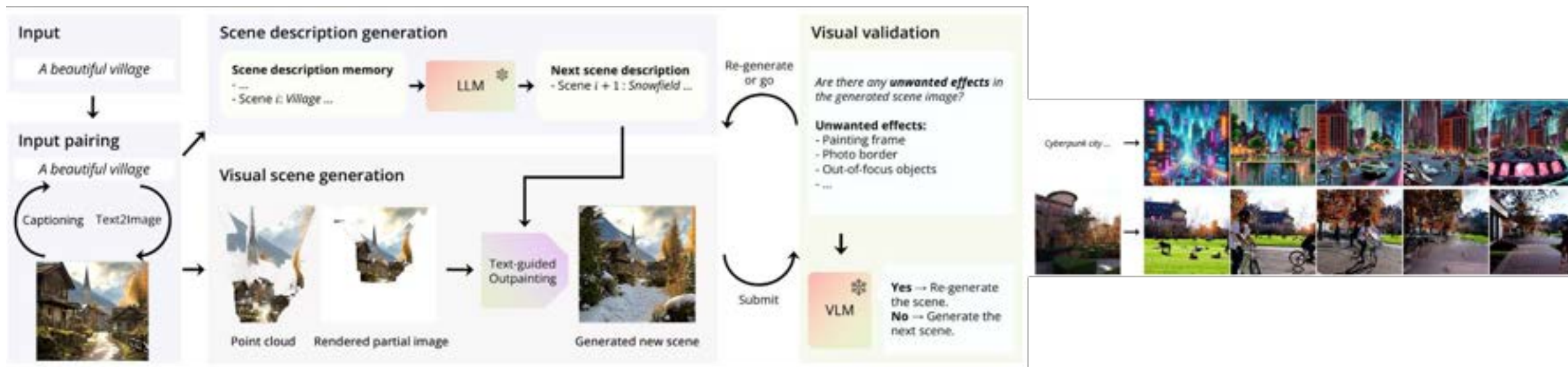
## CogAgent: A Visual Language Model for GUI Agents

- ❑ **概要:** VLM-based GUI AgentとAgentの学習のための大規模データセットの提案。Userの入力とGUI画面から、UserのDemandに合わせてGUIの操作の言語提示、一連操作の計画などが可能。
- ❑ **新規性:** GUIの画面とUserのDemandから、高精度でGUIの使用をサポート可能なVLMを提案。既存の手法では言語テキストやHTMLのみを利用するものが多い。密にアノテーションしたデータセットの提案も評価される。
- ❑ **手法:** Encoder-decoderベースのVLMを使用。また、GUIの画面理解のための高解像度画像をエンコードするモジュールを使用。
- ❑ **感想:** AIが進んだ先のGUIがどのようになっていくのかが気になる。



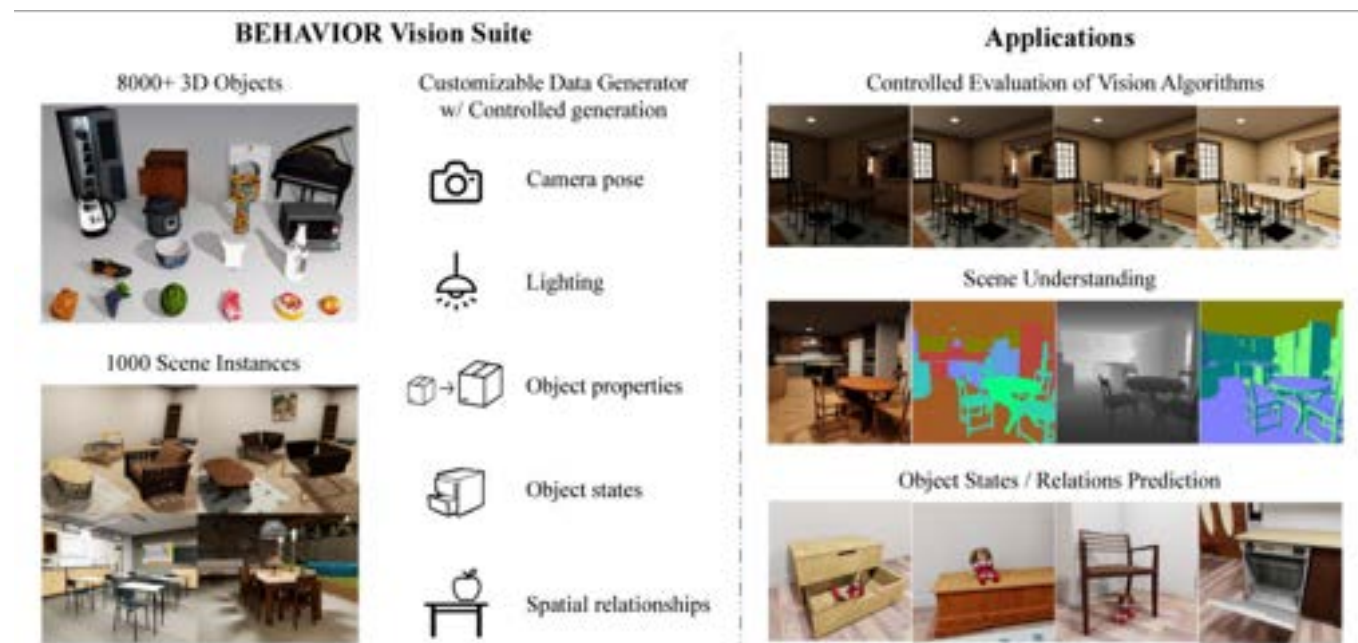
## WonderJourney: Going from Anywhere to Everywhere

- ❑ **概要:** 場所の画像/記述テキストから、3次元的に一致していてストーリー性がある一連の画像を生成するタスクperpetual 3D scene generationと手法の提案。
- ❑ **新規性:** タスクの新規性がメイン。既存の手法が一枚の画像生成や3次元シーンの生成が多い。ここでは3次元一致性を保った複数の画像生成を行っている。
- ❑ **手法:** 複数のモジュールの組み合わせたModular手法。また、LLMでシーンの記述を行う。視覚モジュールでシーンの3次元一致性を確保する。VLMモデルで全体的な生成結果の確認をする。
- ❑ **感想:** 既存のモデル(画像生成モデル、LLMでストーリーモデル、Outpaintingモデル、3次元生成モデル)をうまく組み合わせて綺麗な結果を生成できた。モデルの各パーツは斬新的ではない。モデルが理解しやすい。将来類似した方法で長い漫画・動画が作れそう? 3次元を使って一致性を保つところは評価される。



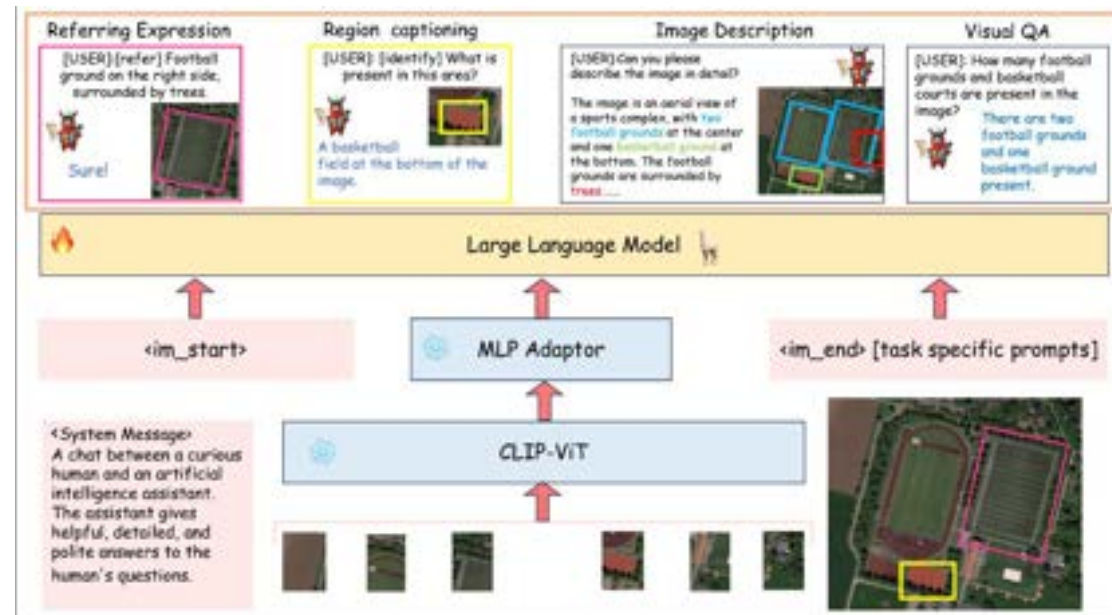
## BEHAVIOR Vision Suite: Customizable Dataset Generation via Simulation

- ❑ **概要:** 様々なCVタスクの学習と評価に活用できるBEHAVIOR Vision Suite (BVS)を提案。BVSではリアリティ性が高い多様なシーンと物体から構成する。環境の照明や、物体の属性・状態・位置関係を操作可能。
- ❑ **新規性:** 規模が大きいかつ、様々なCVタスクを評価可能な面が既存のデータセットより優れる。手法がDomain shiftに対してのロバスト性が評価可能。また、Sim2Real transferも評価可能。
- ❑ **感想:** 物体の属性・状態をシミュレーションできるところが良い。Human modelを入れられるともっと使用場面が広がりそう。Sim2realできる部分が評価される。



## GeoChat: Grounded Large Vision-Language Model for Remote Sensing

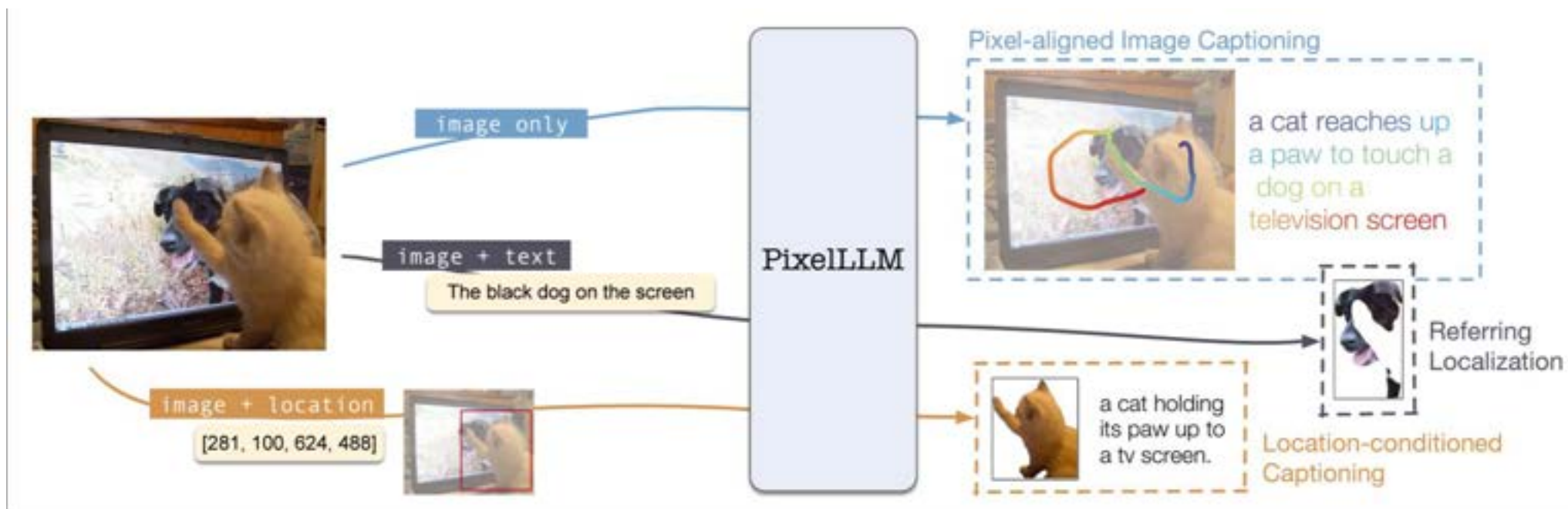
- ❑ **概要:** Remote Sensing画像に適応したMLLMとMultimodal Instruction-Tuningデータセットの提案。
- ❑ **新規性:** GPT-4Vを含めて既存のMLLMがRemote Sensing画像認識に性能が劣る。また、既存のRemote Sensingの研究がFixed-sizedデータセットで行うものが多く、Open-vocabularyで同時にRecognitionとGroundingができるモデルが提案されていない。
- ❑ **手法:** 手法が通常の他のタスクに適応する際のMLLMと類似。まず大きなInstruction-Tuningデータセットを用意。その際にVicuna-v1.5を活用し低いコストで318kのデータセットを構築。LLaVA-1.5とLoRA fine-tuningで自らのMLLMを学習。
- ❑ **感想:** LLMをRemote Sensingに活用した例。既存の複数のRemote Sensingの問題を一つのモデルにより対応可能にした。提案データセットの貢献が大きい。画像を1枚ではなく、複数枚にすると分析できるものが増えそう。





## Pixel Aligned Language Models

- ❑ **概要:** LLMをVisual Groundingに適応した提案。画像や画像と座標/テキストの組み合わせなどの入力から、Captioning、Localization、Word-Groundingなどを行うモデルPixelLLMを提案。複数のベンチマークでSOTAなVisual Grounding精度を達成。
- ❑ **新規性:** 既存VLM手法が画像をテキストに落とすことが可能。提案のPixelLLMがLLMがどれくらいVisual Grounding(テキストやワードから画像の対応領域を指定)に効果があるのかを検討した。
- ❑ **手法:** PixelとWordのAlignmentがされているデータセットで学習。構造がシンプルで、Word特徴の上にMLPを入れて座標を予測する。また、LLMのLoRA Finetuningを使用した。
- ❑ **感想:** モデルがシンプルでStraight-forward。事前学習が十分な場合で、LLMが高いGrounding能力があることを示した。



## Visual Concept Connectome (VCC): Open World Concept Discovery and their Interlayer Connections in Deep Models

- **概要:** 学習済みモデルのVisual Conceptに関する理解を可視化するUnsupervisedな手法の提案。
- **新規性:** 既存の手法がSingle Layerのノード分析がメイン。提案のVCCがMulti-Layerの分析を可能にした。
- **感想:** 学習済みモデルのエラー分析に使えるところが強い。LLMのVisual Conceptsに関する理解でVCCと結合して自動分析が可能？



## MMSum: A Dataset for Multimodal Summarization and Thumbnail Generation of Videos

- ❑ **概要:** Multimodal Summarization with Multimodal Output (MSMO)のための新しいデータセットMMSumを提案。MMSumを使って、複数のタスク (Videoやテキストの単独のSummarizationとMSMO)を評価可能。データセットでは17カテゴリ、170サブカテゴリが定義され、人間がアノテーションした5100ビデオから構成される。
- ❑ **新規性:** 既存のMSMOデータセットはデータセットの規模が小さい、ラベルが整理されていない、そして公開されていないなどの問題がある。MMSumはMSMOタスクの最初の大規模・丁寧に人間によりアノテーション・全公開されているデータセットとなる。また、Summarizedされたビデオとテキストから、Thumbnailをルールベース生成の手法も提案。Thumbnail自動生成がかなり実用性高い。
- ❑ **感想:** 提案されたタスクThumbnail Generationがショートビデオが流行っている今では実用性高そう。Thumbnailどうやって評価するのが気になる。ThumbnailのパーツでLLMの活用やCustomized的な設定を入れると面白そう。



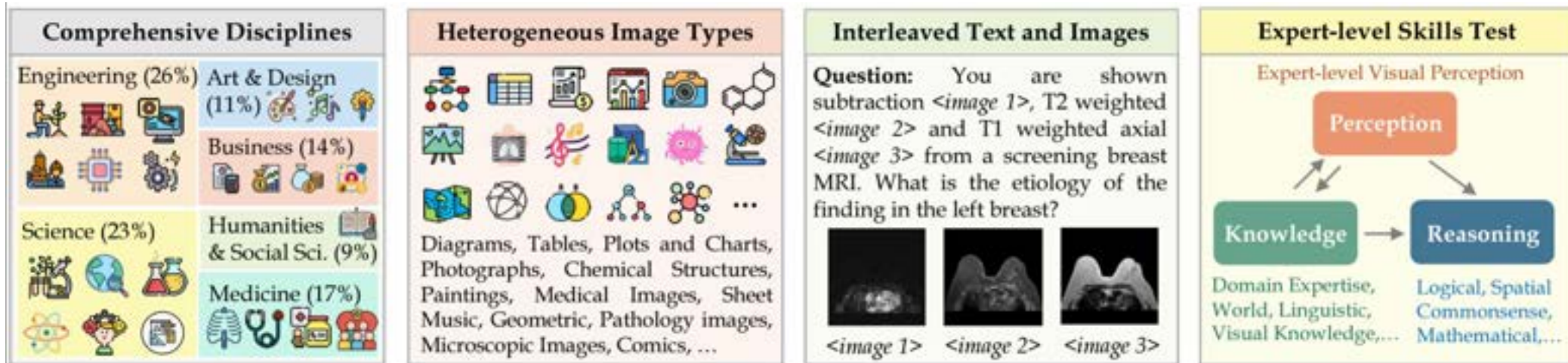
## SOK-Bench: A Situated Video Reasoning Benchmark with Aligned Open-World Knowledge

- ❑ **概要:** Situationに関する背景知識の理解、Open-World Commonsense Reasoningなどを評価できる新しいVideo Question Answeringデータセット SOK-Benchの提案。LLMを活用して、知識グラフの抽出やデータセットQAの自動生成を行った。
- ❑ **新規性:** 従来のVideo QAデータセットでは様々なReasoning能力を評価できるが、実環境の様々なSituationに関してProblem-solvingのためのデータセットがなかった。
- ❑ **データセット生成手法:** 下の図で従来データセットの作成方法と提案のLLMを活用したデータセットの比較を示す。提案のSOK-Benchは知識グラフをベースにしている。VLモデルによりビデオから人物などの認識を行い、LLMが知識グラフを抽出したり、さらにグラフをベースにQAの生成を行う。
- ❑ **感想:** LLMをデータセット生成に使う研究が多くなってきた。Reasoningや構造化知識が必要なデータセット設定において特にLLMが使いやすい。



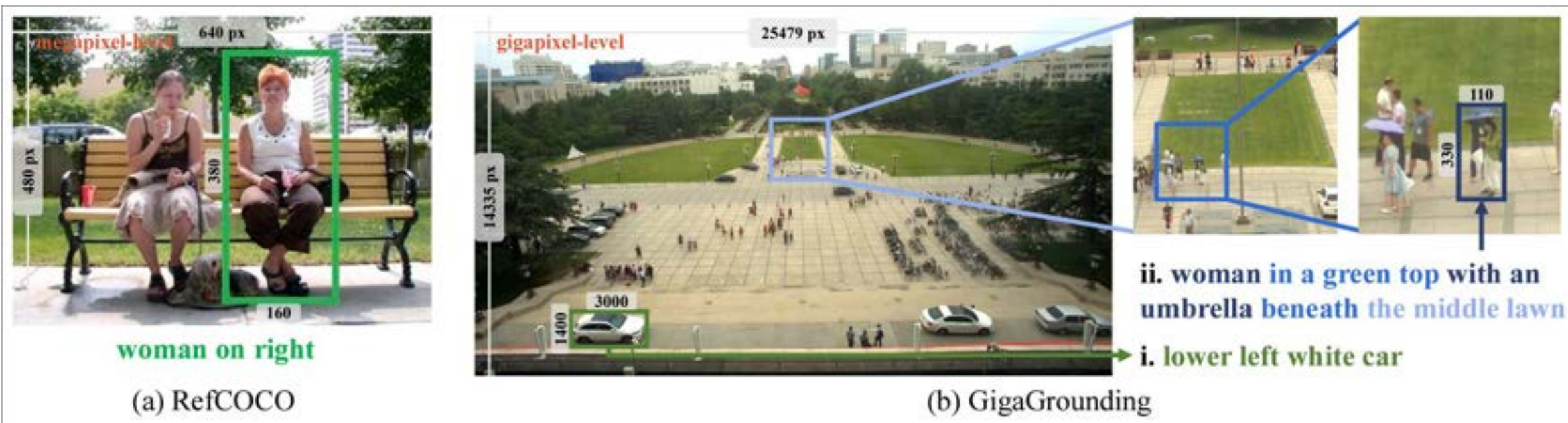
## MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

- ❑ **概要:** Art & Design, Business, Science, Health & Medicine, Social Science, Tech & Engineeringなどの6つの分野のCollege-levelの専門知識の理解を網羅的評価できる大規模マルチモーダルデータセットMMMUの提案。GPT 4VやGemini Ultraなどがそれぞれ56%と59%の精度しか達成できない。今後のMLLMの評価のために活用できる。
- ❑ **新規性:** 異なる分野の専門知識に関するマルチモーダルデータセットの提案があまりなかった。MMMUは今後Expert AGIの評価に活用可能。表、グラフ、写真、絵画など様々な画像内容を扱っている。問題解答するには、テキストと画像を合わせた理解が必要なものが多い。
- ❑ **感想:** MMMUデータセットでGPT 4Vなどの足りていない部分をいくつかデータセットでカバーできた。MMMUが実際の教材やテストから問題を集めたため、GPT 4Vなどのモデルがすでに回答を暗記した可能性があるため、回答の根拠もモデルに説明してもらおうとっと解釈性がある。



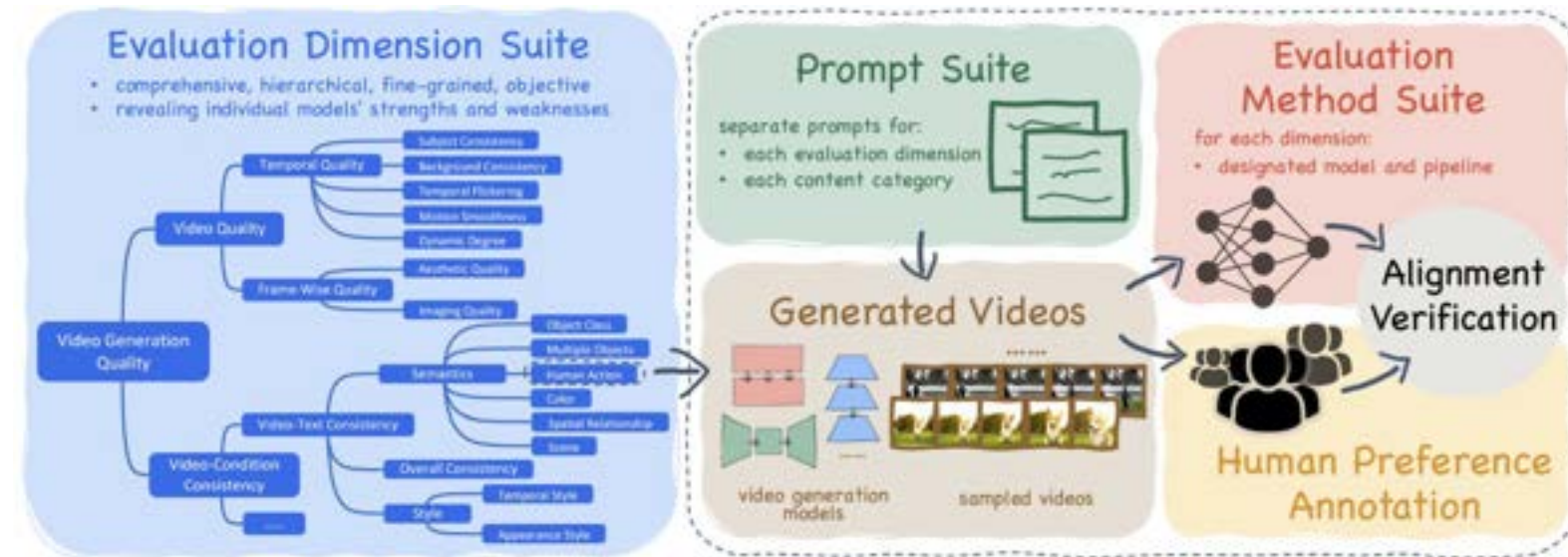
## When Visual Grounding Meets Gigapixel-level Large-scale Scenes: Benchmark and Approach

- ❑ **概要:** 高解像度の大きなシーンの画像から、画像中の人物の領域を記述するテキストから、画像中の人物をGroundingする新たなデータセット GigaGroundingを提案。
- ❑ **新規性:** 既存のVisual Grounding用データセットは解像度普通の画像中に一つ/複数の比較的に大きな人物をGroundingする。提案の GigaGroundingは大きな画像中に複数の小領域を占めた人物をGroundingする。人物を複雑な画像から特定するための手法の評価に特に活用しやすい。
- ❑ **感想:** GigaGroundingデータセットは人間でも特に労力がかかる。Videoバージョンや3Dバージョンも面白そう。



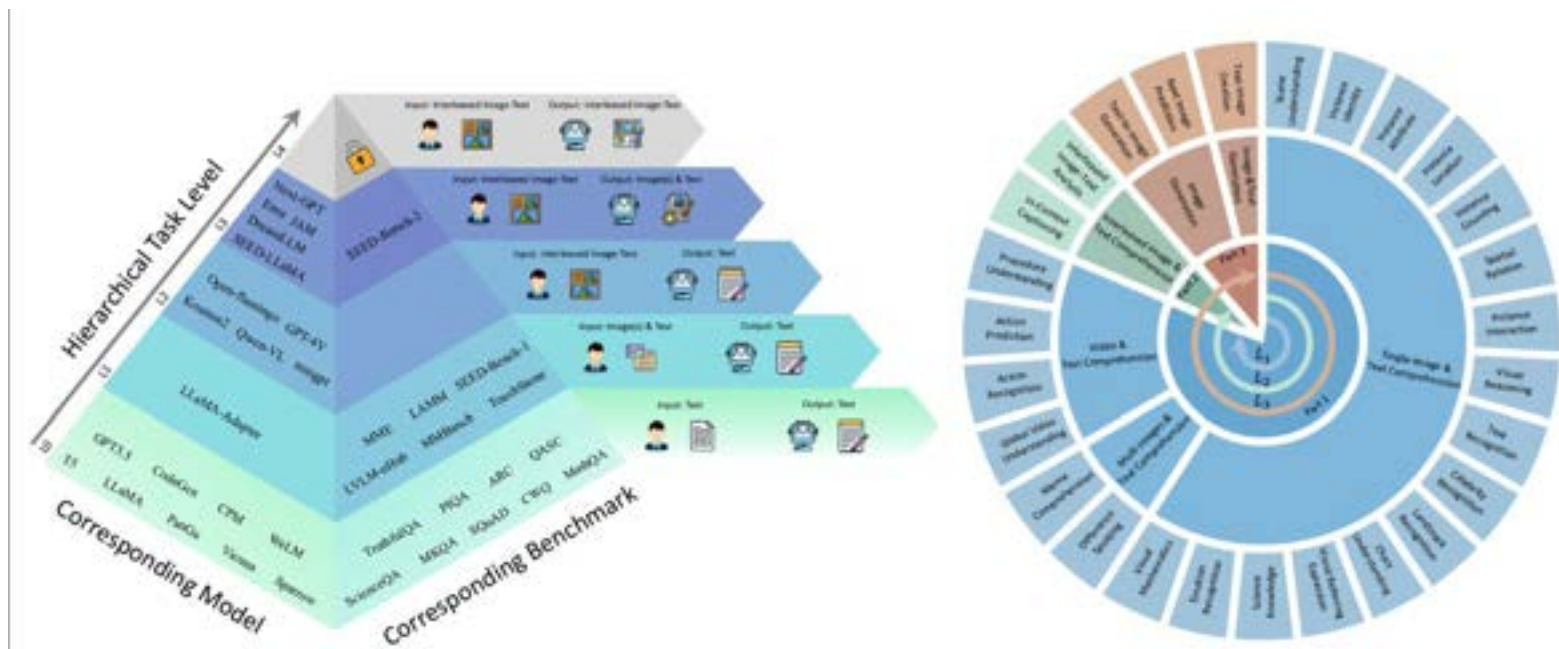
## VBench: Comprehensive Benchmark Suite for Video Generative Models

- ❑ **概要:** ビデオ生成モデルを生成ビデオの質、テキストとビデオの一致性など様々な方面から評価できるBenchmark Suiteを提案した。人間評価とクロスチェックした結果、VBenchが人間の判断に近いとわかった。
- ❑ **新規性:** これまでのVideo Generationの評価指標が人間の評価との一致性が低いと指摘されていて、Vbenchが人間の評価に近い。実用性が高い。また、Vbenchは単独な指標ではなく、16指標を整合しているところも新しい。
- ❑ **手法:** トップダウンで16の評価の軸を指定している。評価軸ごとに100個のテキストPromptを用意し、モデルの生成を評価する。
- ❑ **感想:** エンジニアリング的な要素が強い。複数の指標をよくデザインしたSuiteで評価するアイデアが良かった。今後、他のタスクでも、整合されたBenchmark Suiteで評価されそう。



## SEED-Bench: Benchmarking Multimodal Large Language Models

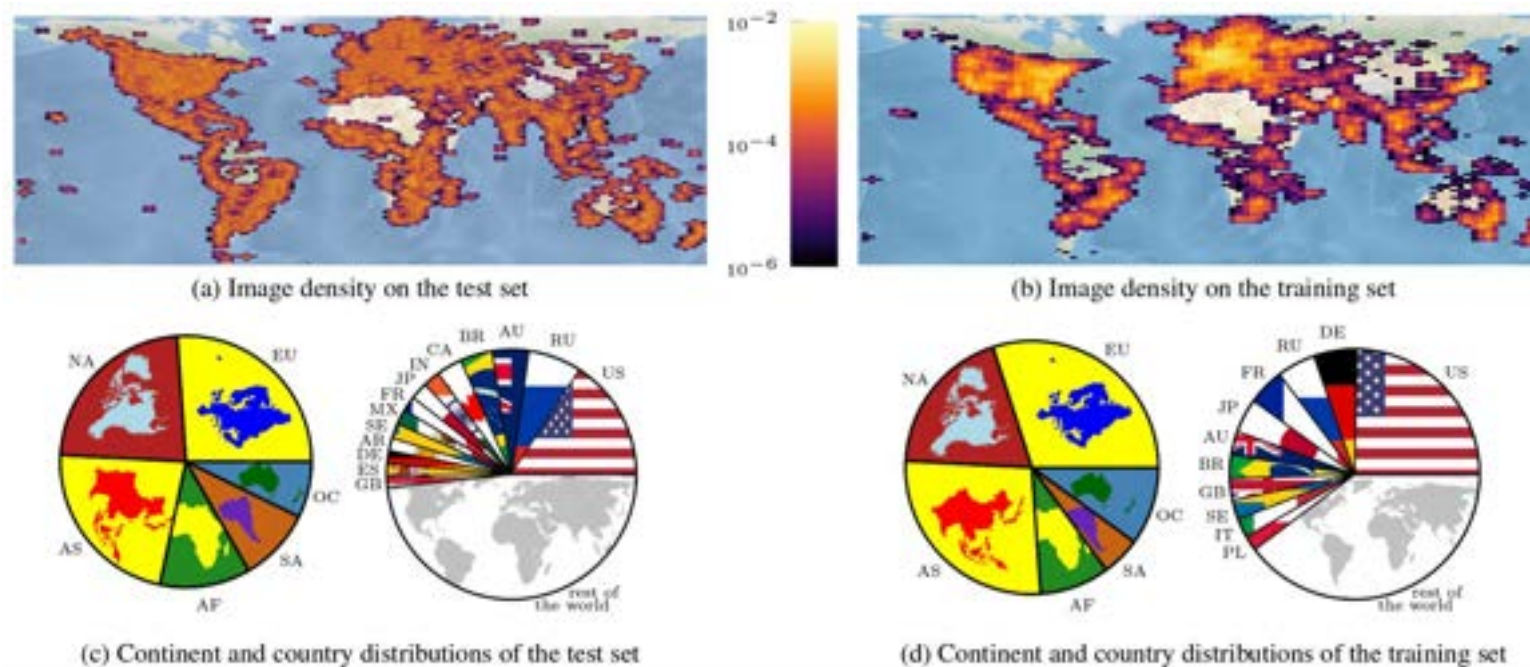
- ❑ **概要:** MLLMを評価するための新たな大規模ベンチマーク用データセットの提案。タスクのレベルを階層的に定義し、インプットがテキスト・画像とテキストの混合、複数画像、出力も画像とテキストの組み合わせの多種類のタスクを評価可能。
- ❑ **新規性:** ひとつの軸ではなく、27種類のタスクをMLLMをより詳細的に評価可能にした。
- ❑ **手法:** 23の最近のMLLMを評価。著者たちが定義したタスクレベルでは最近のMLLMはL1レベルのタスクに対しても不足していて、L1以上のタスクはまだまだ改善のスペースが大きいことを示した。
- ❑ **感想:** LLMやMLLMなどを活用してデータセット作成し、作成コストや労力が減らせた。網羅的なベンチマーク用データセットは最近いくつか提案されてきた。タスクレベルを整理するところが良かった。





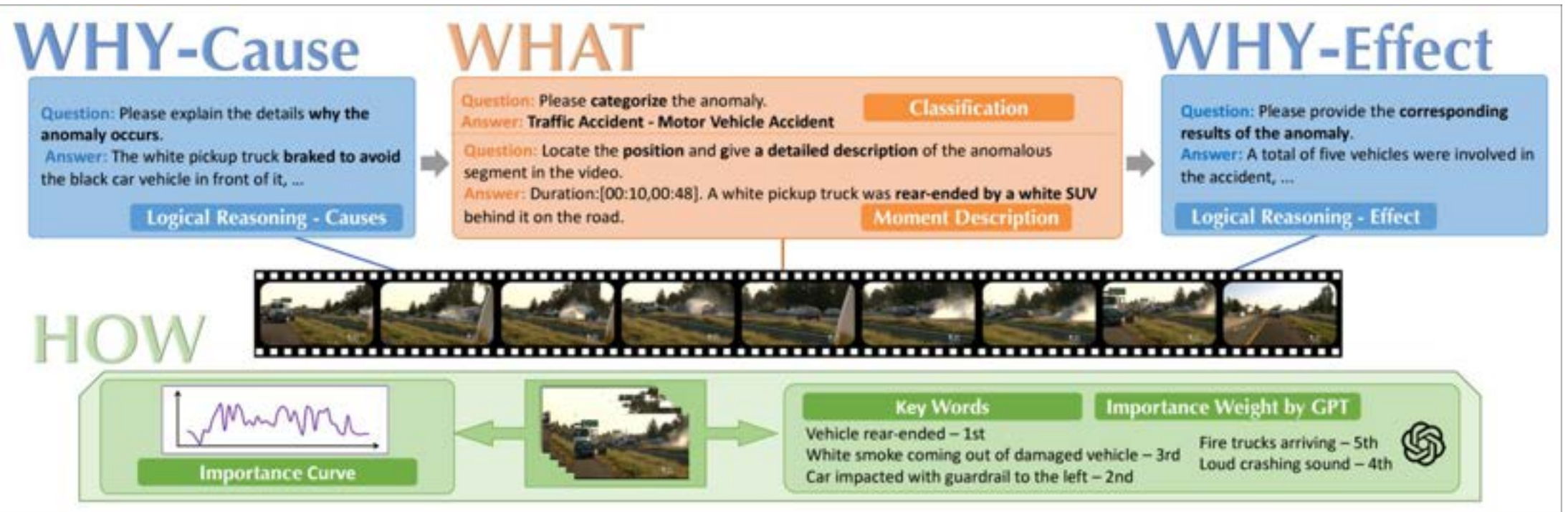
## OpenStreetView-5M: The Many Roads to Global Visual Geolocation

- ❑ **概要:** Open accessの新たなStreetViewデータセットOpenStreetView-5Mを提案。データセットでは228国の5百万以上の場所がAnnotationされた画像から構成する。画像から場所を認識 (geolocation) するタスクに活用できる。
- ❑ **新規性:** StreetViewデータセットは公開されているものが限られている。ここで5Mの画像が公開された。既存データセットと異なり、学習・テストセットでは分布が異なるようにした。
- ❑ **感想:** Geolocationタスクなど、人間でも難しいタスクが多くなってきた。



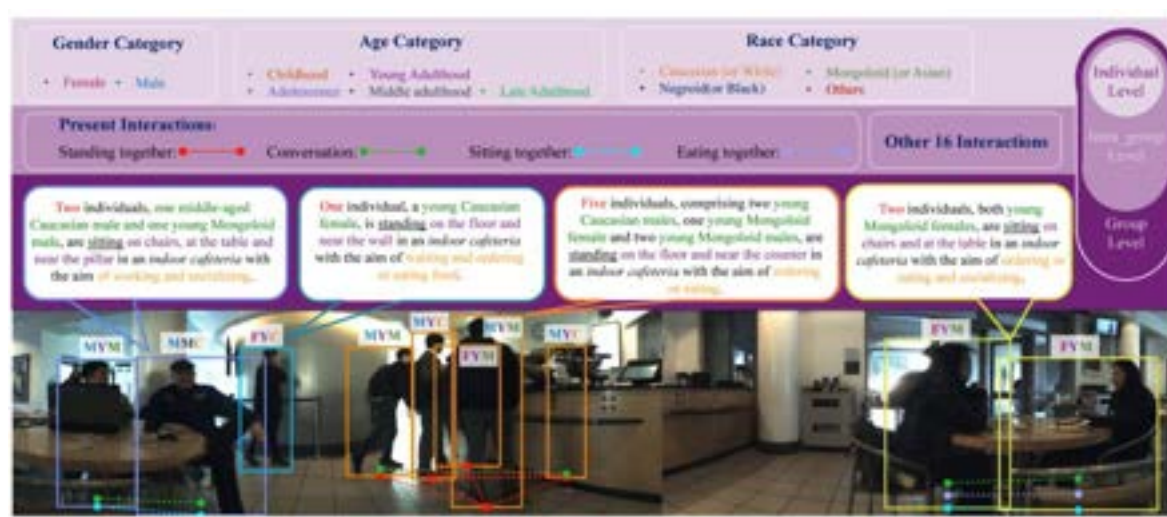
## Uncovering What, Why and How: A Comprehensive Benchmark for Causation Understanding of Video Anomaly

- ❑ **概要:** 交通シーンビデオから交通事故を認識・原因分析するためのデータセットCUVAの提案。CUVAは交通事故(/異常)のタイプ、開始・終了の時間、原因および影響のテキスト記述などをアノテーションしている。また、新たな評価指標MMEvalを提案。
- ❑ **新規性:** 交通シーンのビデオから原因の分析を詳細的に行う既存データセットがなかった。
- ❑ **感想:** 少しVisual Comet論文と似ている。発生する原因となる時間Localizationの部分が面白い。



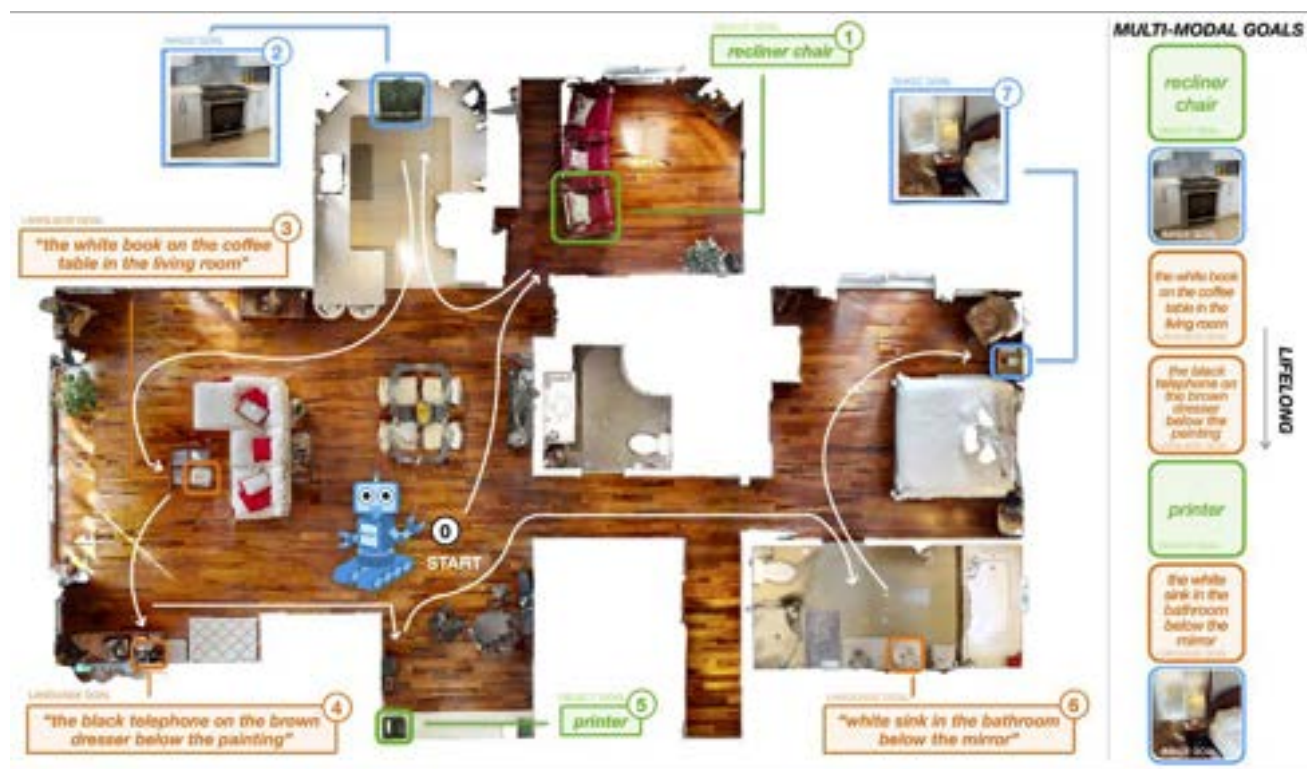
## JRDB-Social: A Multifaceted Robotic Dataset for Understanding of Context and Dynamics of Human Interactions Within Social Groups

- ❑ **概要:** Human social behavior認識のための大規模データセットJRDB-Socialの提案。既存のRobotics用ビデオデータセットJRDBに、3つのレベルのアノテーションを追加。Individual levelでは一人一人の性別、年齢、Raceなどをアノテーション。Intra group levelでは、人間同士間の動的な関係性(一緒に立っている、会話しているなど)を定義。Group levelではグループの人の目的、詳細的な関係、行動などを記述する。既存のVLMがJRDB-Socialで、人間精度との差が大きいことを示した。特にGroup levelの精度が低い。
- ❑ **新規性:** つd。既存のデータセットではIndividual levelやIntra group levelのアノテーションがメインで、もっと高レベルのGroup levelに関する検討が少なかった。
- ❑ **感想:** Human social behaviorをビデオから高精度で認識するのが難しい。さらに対話の内容が入るともっと難しそう。また、JRDB-Socialは人間でも判定が難しいもっとfine-grainedな関係認識が含まれていなさそう。



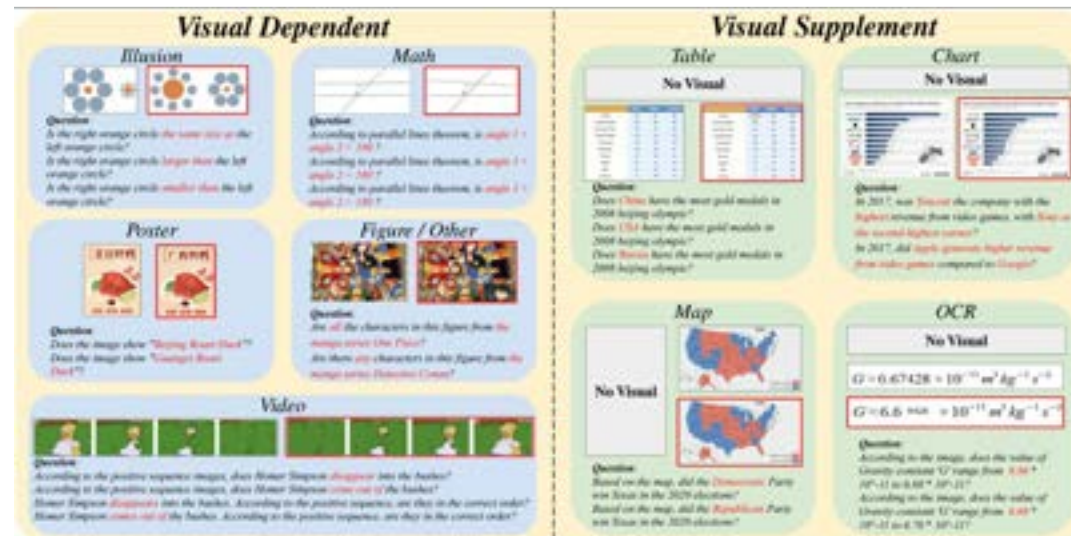
## GOAT-Bench: A Benchmark for Multi-modal Lifelong Navigation

- ❑ **概要:** 物体ラベル、物体記述文、画像の3タイプの入力から、連続の複数のGoalを一つずつナビゲーションしていくタスクとデータセット GOAT-Benchを提案。
- ❑ **新規性:** 既存ベンチマークデータセットでは一つのみの物体を探す。ここで一連の物体を順番に探すLifelong navigationを行う。この設定で、うまく過去のメモリーを利用することで、効率よく物体探しができる。
- ❑ **手法:** 実験でメモリーパーツの重要性を示した。
- ❑ **感想:** 連続で複数のGoalを探すのと、これまでのObjectNavタスクで一つのみの物体を探すのがあまり大差が無いように見える。



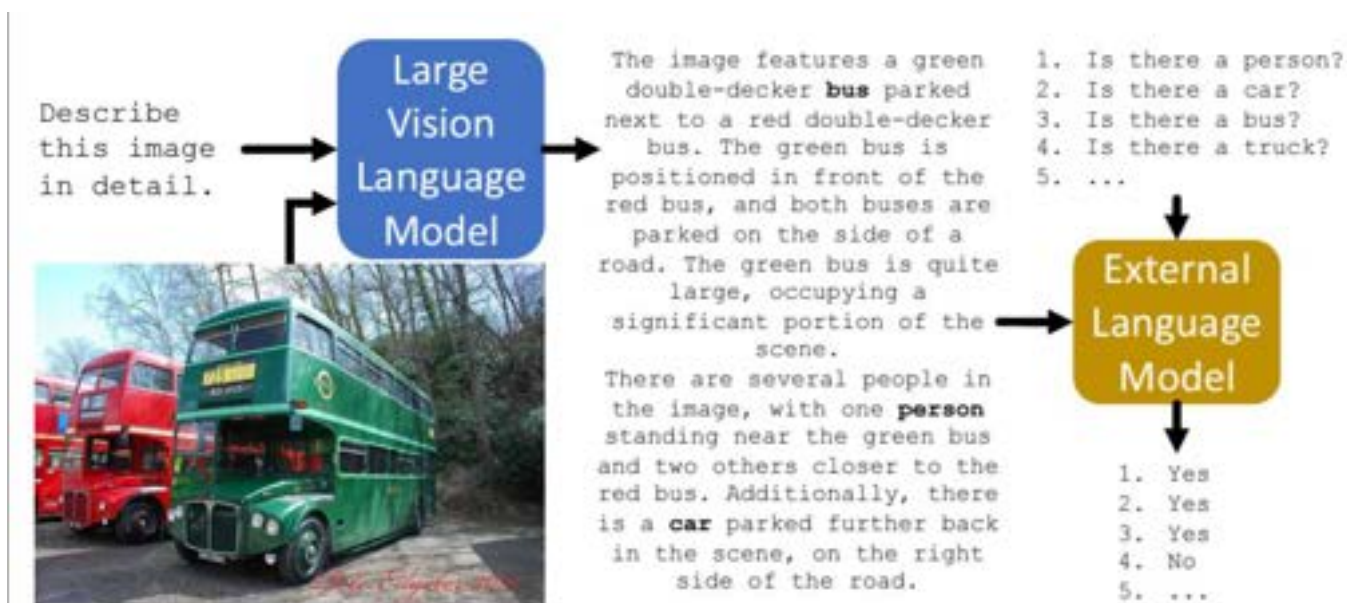
## HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models

- **概要:** LVLMsの言語と画像に対してのHallucinationに関して詳細的に評価するためのデータセットHallusionBenchの提案。微小な差がある画像やテキストペアで、LVLMsがそういった差を理解できるかどうかを検証。また、実験で下記のいくつか面白い知見を得た。このデータセットに関して事前知識がある際に、GPT 4vを含めた既存のLVLMsが言語のHallucinationになる傾向。事前知識がない場合でも、画像側のIllusionになる傾向。GPT 4vを含めた既存のLVLMsが画像や言語に関してのManipulationに弱い。また、複数画像間のtemporal reasoningに弱い。
- **新規性:** 著者たちがHallusionBenchが初めてのHallucinationに関して網羅的に調査するためのベンチマークと主張。
- **感想:** Hallucinationは他にもたくさんある。どうやって全てのHallucinationを検証できるかは気になる。有害なHallucinationを防ぐ仕組みが重要そう。HallusionBenchデータセットがあまり大きくない(346画像、1129質問)。



## THRONE: A Hallucination Benchmark for the Free-form Generations of Large Vision-Language Models

- ❑ **概要:** Open-ended free-form hallucination (type Iと定義)、特に画像Captioningタスクを評価するベンチマークTHRONEを提案。
- ❑ **新規性:** 既存研究は特定の固定したフォーマットの出力(Multichoices、Classificationなど)に関するHallucinationの検討がメイン( type IIと定義)。ここでは、Open-ended free-formのhallucinationを扱う。
- ❑ **手法:** Type I, II hallucinationを有効的に緩和するdata augmentation手法を提案。具体的に、Visual Instruction Tuningでモデルが画像中の物体の存在について解答する。上記により物体の有無に関わるhallucinationを緩和。
- ❑ **感想:** Hallucinationの調査のみではなく、Hallucinationを緩和するdata augmentation手法まで提案したところが良かった。

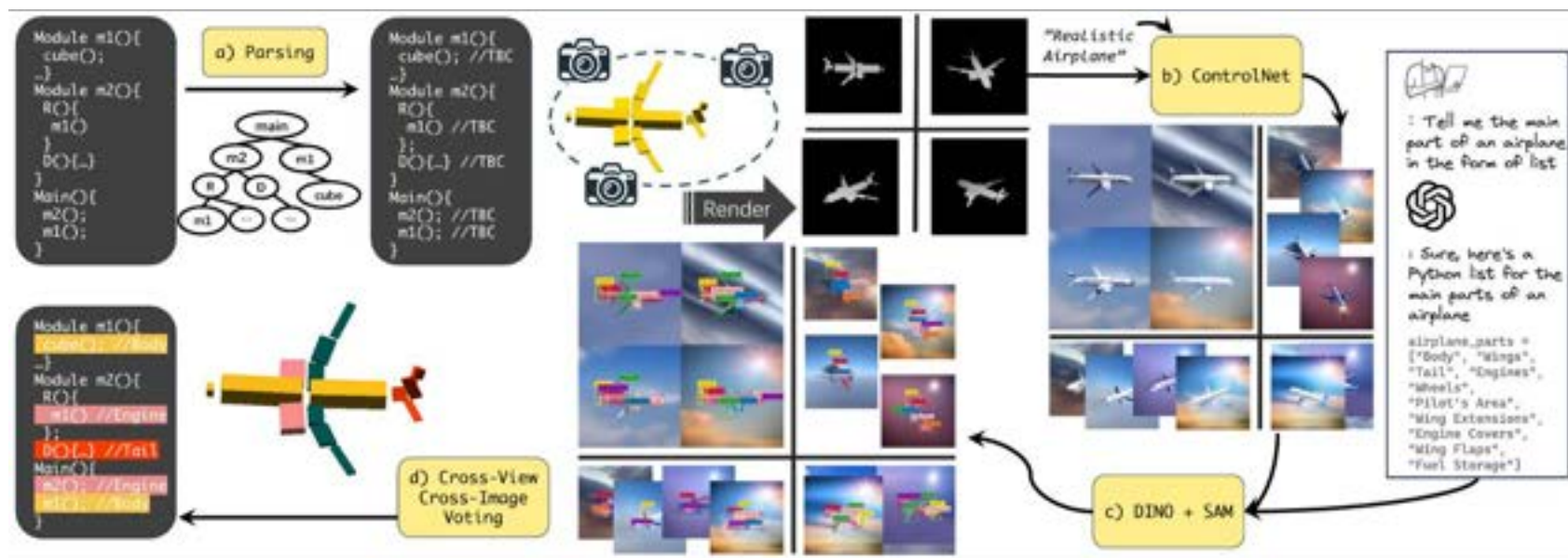


**Type I Hallucination Evaluation**

Instruction	Response
Describe this image in detail.	The image showcases a fruit stand at a grocery store, featuring a variety of fruits on display. There are several bunches of <b>bananas</b> , with some placed in the foreground and others in the background. The <b>bananas</b> are arranged in different sections, creating an appealing presentation for <b>customers</b> . In addition to the <b>bananas</b> , there are also <b>apples</b> and <b>oranges</b> on display. The <b>apples</b> are located towards the left side of the image, while the <b>oranges</b> are placed in the middle and right side of the stand. The fruits are well-organized and presented in an attractive manner, making it an inviting sight for <b>shoppers</b> .
<p> <span style="color: green;">■</span> - GT Class  <span style="color: yellow;">■</span> - Type II Hallucination  <span style="color: red;">■</span> - Hypothetical Content (not a Hallucination)                 </p> <p style="text-align: right;"><i>Type I Hallucinations Present and Found</i></p>	
MSCOCO Object Prediction from Description	
Human	CHAIR
<b>banana</b> apple orange	<b>banana person</b> apple orange
	THRONE (Ours)
	<b>banana</b> apple orange

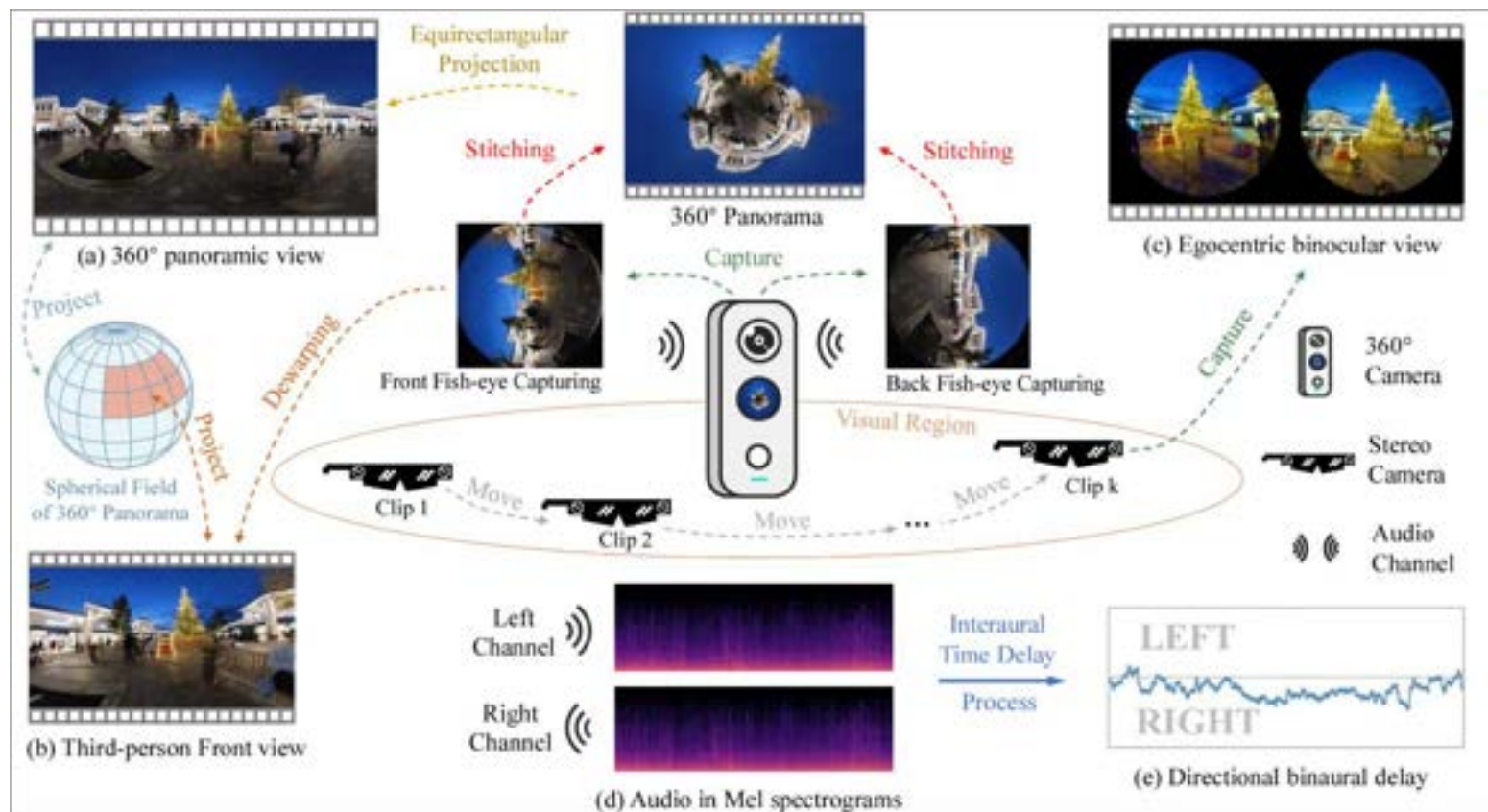
## CADTalk: An Algorithm and Benchmark for Semantic Commenting of CAD Programs

- ❑ **概要:** CADのProgramsに自動的にCommentsを追加するタスク・手法CADTalkとデータセットを提案。タスクが新しい。
- ❑ **手法:** いくつかの最新の手法の組み合わせ。まず入力のプログラムをParseし、コメントを追加する部分を決める。次に、複数の視点からレンダリングを行う。次にImage2Imageで画像を編集し、ChatGPTにより物体のパーツ情報を得て、レンダリングの画像をDINOやSAMでパーツ分割する。最後に投票により、パーツの情報を決めてCommentsを追加する。
- ❑ **感想:** アイデアが面白い。LLMとレンダリングによりデザイン言語の理解を行っている。デザインツールの開発に活用できそう。人間同士のアイデアについての共通理解を高めたり、プログラムのデバッグなど様々なところで使えそう。手法の各々の部分が既存研究だが、手法の全体が面白い。抽象化されるものの理解という面で、いろんな研究へ適応できそう。



## 360+ $x$ : A Panoptic Multi-modal Scene Understanding Dataset

- ❑ **概要:** 複数センサー(360度panoramic view, third-person front view, egocentric monocular & binocularビデオ、multi-channel audio)で様々な場所で撮影した人間行動データセットの提案。また、テキストで場所のアノテーションも人手で追加した。
- ❑ **新規性:** 既存のデータセットより多くのセンサーを持って撮影を行った。より人間のセンシング情報と近いようなデータを取得。
- ❑ **感想:** 多様なセンサーで人間より以上の認識ができそう。密なセンサー+密なアノテーションなどは評価されやすい。





## CLOVA: A Closed-Loop Visual Assistant with Tool Usage and Update

- ❑ **概要:** Visual Programming系の手法(問題をLLMで分割して、小問題をそれぞれVisual Moduleで解く)のContinual Learning仕組みCLOVAを提案。問題を解くプロセスでどのパーツが間違ったかを分析し、有効的に学習データを取得し、間違ったパーツの学習を行う。
- ❑ **新規性:** 既存のVisual Programming手法では、問題を分割するLLMや、小問題を解くVisual Moduleを固定している。CLOVAでは、LLMとVisual Moduleを有効的に更新できる仕組みを提案。
- ❑ **手法:** 3つのプロセスから構成する: Inference段階でLLMを使用して問題を分割し小問題を別々のVisual Moduleで解く。Reflection段階では、回答が間違った例のHuman Feedbackを得る。Learning段階で、学習データを自主的に収集し、間違ったModuleの学習を行う。
- ❑ **感想:** CLOVAはセマンティック的にリッチなhuman feedbackでMLLMの性能を向上する。上記のようなfeedbackやannotationは効率よく(少ない学習exampleで)MLLMなどの性能向上につながる。自己修正・Reflectionが重要。ReflectionモジュールがMLLMでできるとおもしろう。



## SPIN: Simultaneous Perception, Interaction and Navigation

- ❑ **概要:** 環境中に環境をActiveで観察し、自主移動、カメラ角度調整、アームを調整しピッキングができるWhole-Body Mobile Manipulatorを提案。
- ❑ **新規性:** 手法的な新規性がメイン？既存のAgentが複数のモジュールで自主移動とManipulationを扱う。SPINでは一つのモデルで、No MappingかつNo Planningで、Activeで自主移動とManipulationを行う。
- ❑ **手法:** Depth入力を扱う。Ego-visionから同時にperceive, interact, navigateするポリシーを強化学習により学習。
- ❑ **感想:** SPINを含めて、いくつか大規模でSimulationで学習して高い実環境精度を達成したAgentが提案された。ロボティクスでは関連のロボットがどこまでできているのかが気になる。



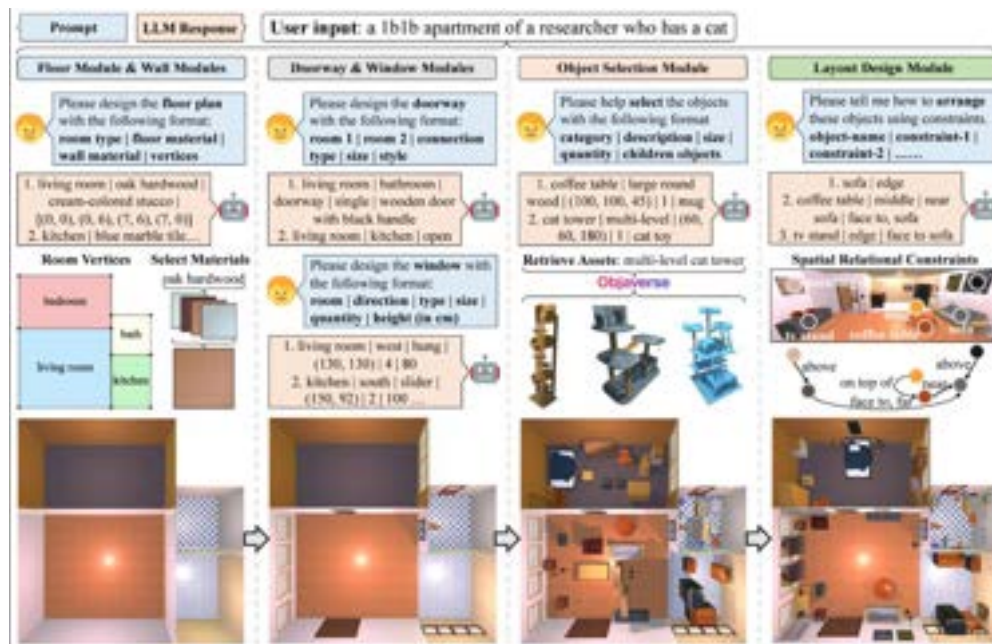
## Habitat Synthetic Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation

- ❑ **概要:** 211 scenesの3次元データセットHSSD-200を提案。既存データセットと比べてHSSD-200のリアリティ性が高く、assetsの多様性も高い。HSSD-200には、211家のSim環境と18,656 unique 3d物体から構成される。Scale vs. realismで実験した結果、既存データセットと比べて小規模で高い性能及びSim2Real汎化性を示した。
- ❑ **新規性:** HSSD-200で実験した結果、sceneのリアリティ性と質が事前学習に大きく影響することがわかった。実験で2桁大きいProcTHORで学習した結果がHSSDより劣る。
- ❑ **感想:** HSSD-200は配置変更や物体の変更が可能なので、うまく設計した場合、リアル3D環境認識の事前学習にも効果でそう。



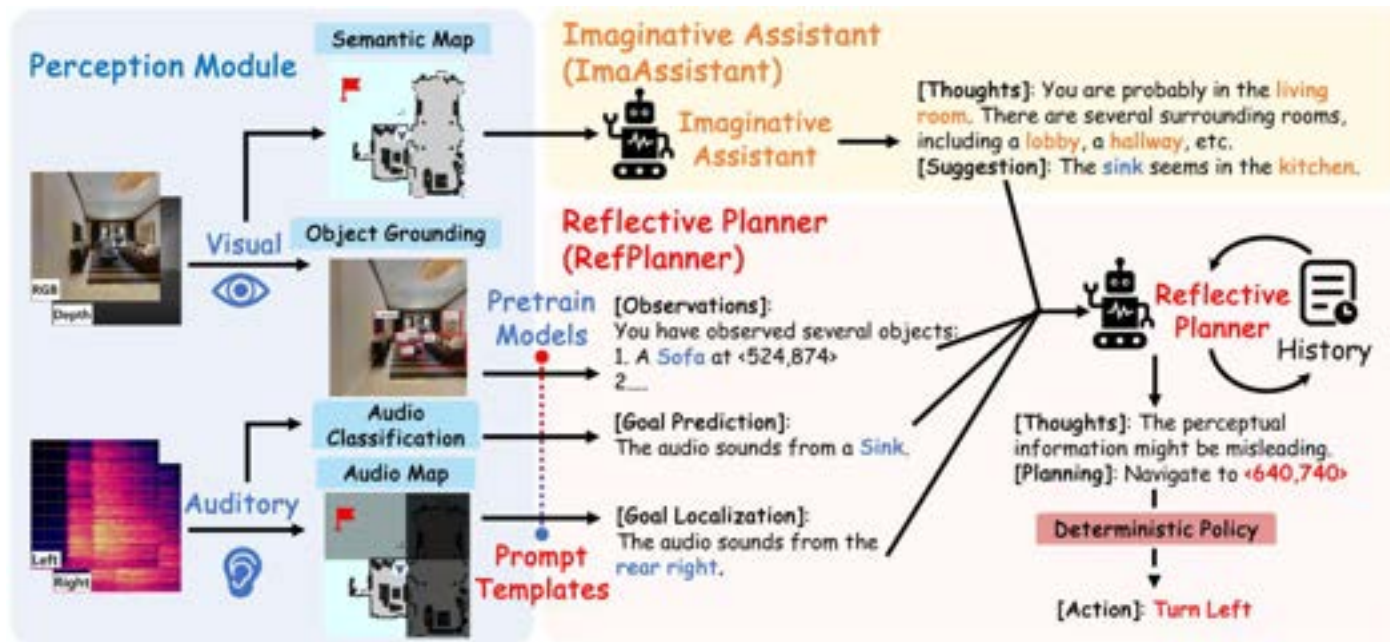
## Holodeck: Language Guided Generation of 3D Embodied AI Environments

- ❑ **概要:** テキストから、3次元シーンを生成する手法の提案。具体的に、まずGPT4で入力の言語から部屋のスペース、物体構成、物体間の位置関係を生成。次に既存のAssetsで部屋を組み立てる。Holodeckで事前学習することで、既存データセットProcTHORより高いZero-shot ObjectNavi精度を達成。
- ❑ **新規性:** GPT4などのLLMを活用することで、既存のデータセットよりもっと人間の部屋に近いかつバリエーションが多い部屋を生成可能。比較的低いコストで3次元シーンを生成可能。
- ❑ **感想:** 手法は基本的にGPT4で部屋のデザインをして、そして既存のデータセットのAssetsで部屋を構成するため、比較的技術上新規性が薄い。事前学習やApplicationの設計の面では面白い。



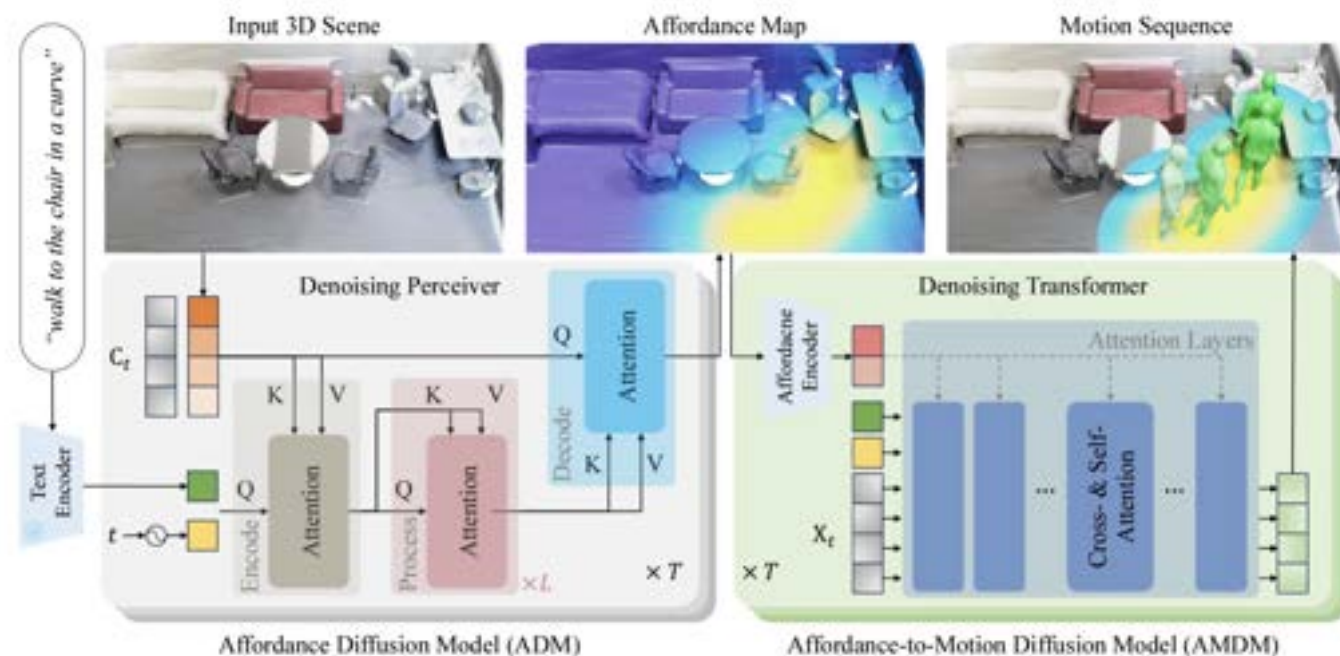
## RILA: Reflective and Imaginative Language Agent for Zero-Shot Semantic Audio-Visual Navigation

- **概要:** LLMを用いてZero-shotでAudio Visual Navigationを行う手法の提案。
- **新規性:** Reflective plannerでは認識結果をそのまま使用するのではなく、誤って認識したかどうかをチェックするところが新しい。
- **手法:** まず、MLLMで異なるセンターの情報をテキストに変換する。Reflective Plannerでは観測した情報によりNavigationのPlanを行う。さらに、観測の結果(テキスト)が正しいかどうかをチェックし、誤った認識を除く。また、LLMベースのImaginative Assistantモデルを提案し、グローバルの認識をベースにNavigationのアドバイスをする。
- **感想:** Reflective Plannerは局所な視覚認識が正しいかどうかをチェックするところが面白い。LLMを使ったEmbodied Agentがたくさんあってそれぞれの良さを精度以外にどうやって評価するのが気になる。



## Move as You Say, Interact as You Can: Language-guided Human Motion Generation with Scene Affordance

- **概要:** 言語の入力から、指定された3次元シーンとinteractするhuman motion生成の手法の提案。3次元シーンとInteractionする精度を高めるために、SceneのAffordanceの予測を挟んだ仕組みを提案。いくつかのベンチマークで高い精度を達成。
- **新規性:** 3次元シーンのAffordanceとConsistentな動作生成が従来難しいとされている。また、3次元シーンにおける動作生成のデータセットの作成コストが高い。提案手法が、比較的容易に計算できるAffordance Mapを使うことで、低コストで3次元を考慮した動作生成を行なっている。
- **感想:** 3次元シーンにおける長いMotionの生成がまだ難しい。短いMotionをうまく組み合わせるところで工夫するともっと大規模で複雑な動作生成できる? program を使うと良さそう。



## Rich Human Feedback for Text-to-Image Generation

- ❑ **概要:** 生成AIの生成画像にミスやartifacts/implausibilityなどが存在する場合が多い。それらを改善するため、18K画像に対しHuman Feedbackを収集しデータセットRichHF-18Kを提案。RichHF-18Kでは、左下図のように、Annotatorsが生成画像中のartifacts/implausibilityとミスを別々で画像中からポインティングしながら、正しく生成できなかった文字部分も記録する。また、生成の質4つの指標から5段階で評価する。また、自動的に画像とテキストから生成が良くない領域のヒートマップ、テキスト、評価値を出力するシンプルなモデルを提案。最後に、提案データセットで既存の手法を再学習してSoTAな生成ができた。
- ❑ **新規性:** 画像生成の質向上をデータ側から改善する新しいアプローチを提案。また、領域・テキスト・評価値など複数の方面からHumanのFeedbackを集めたところが評価される。
- ❑ **感想:** RichなHuman FeedbackやAnnotationで少量データで高い精度を達成する研究が複数あった。規模のみではなく、質の重要性がかなり挙げられている。



(a) Muse [6] before finetuning



(b) Muse [6] after finetuning

## Comparing the Decision-Making Mechanisms by Transformers and CNNs via Explanation Methods

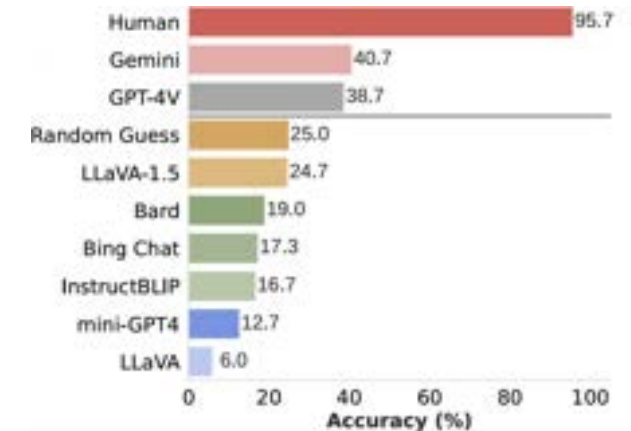
- 概要:** CNNやTransformersなどがクラス識別する際に、画像中の領域にどこを見て判断しているのか、部分的に領域をぼやけた場合にどれくらい認識の結果に影響があるのかについて網羅的に調べた研究。画像を部分的にブラーして認識の結果が少しずつ徐々に影響される現象をCompositional behaviorと定義し、画像中の少ない領域から認識可能で領域をブラーすると認識精度が劇的に低下する現象をDisjunctive behaviorと定義した。あらゆるCNNとTransformerの二つの behaviors を検討した。
- 新規性:** 領域が画像識別への影響を網羅的に調べた。特に二つの新しい behaviors の視点を中心に行った。また、CNNが比較的DisjunctiveでTransformerが比較的Compositionalという傾向を発見した。また、ネットワークを深く、Receptive Fieldを大きくすることで、Compositional性を向上できることを示した。また、Normalizationの手法も影響することがわかった。類似したネットワークが類似する領域で物体を識別する傾向があった。
- 感想:** 類似するネットワークが類似した物体領域を見るところが特に面白い。異なるネットワーク構造が違うタイプのパターンに強い可能性もある。





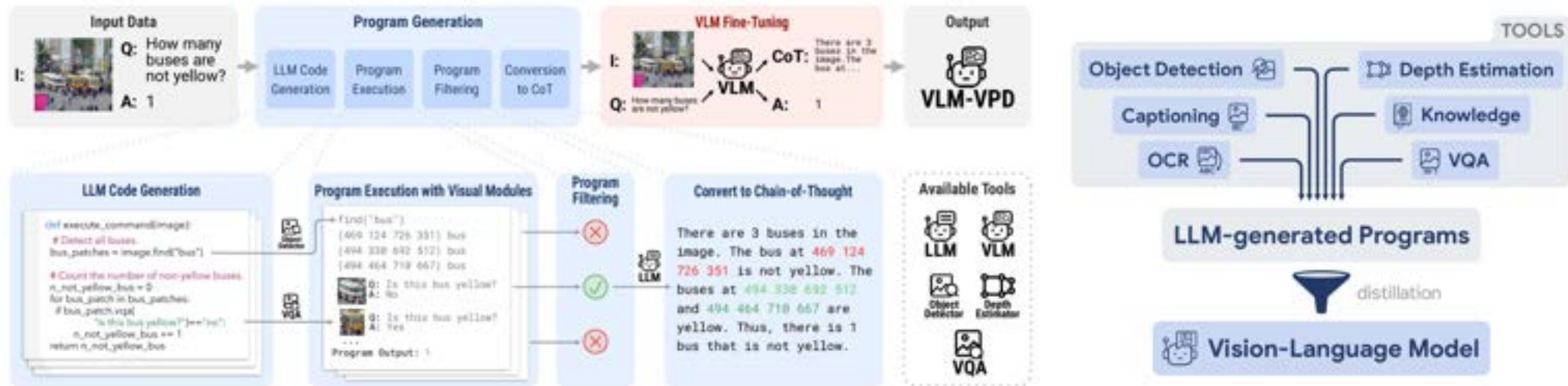
## Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

- ❑ **概要:** GPT4vなどのMLLMが人間がシンプルと思っている問題を間違えてしまう問題点がある。この問題点を調査する方法、データセットを提案し、最近の性能が良いMLLMをベンチマークした。最も性能が良いGemini ProやGPT 4Vなどでも人間と大きなギャップがあった(右図)。
- ❑ **新規性:** GPT4vなどの上記の問題点はよく知られているが、網羅的に調査する手法があまりなかった。
- ❑ **手法:** CLIPの特徴空間で近い画像ペアを収集し、人間に画像ペアの詳細な差をアノテーションした上で、QAのデータセットを構築した。このデータセットで既存のMLLMを評価。CLIPベースのモデルが特に精度が低下することを発見。また、self-supervised learningの特徴量(SSL)やSSLとCLIPのミックスにより精度を多少高められることを発見。
- ❑ **感想:** GPT4vなどは他のsystematic shortcomingsもあるはず。言語で記述できる違いと連続を記述できる(例:特徴量)ものの組み合わせが必要そう。



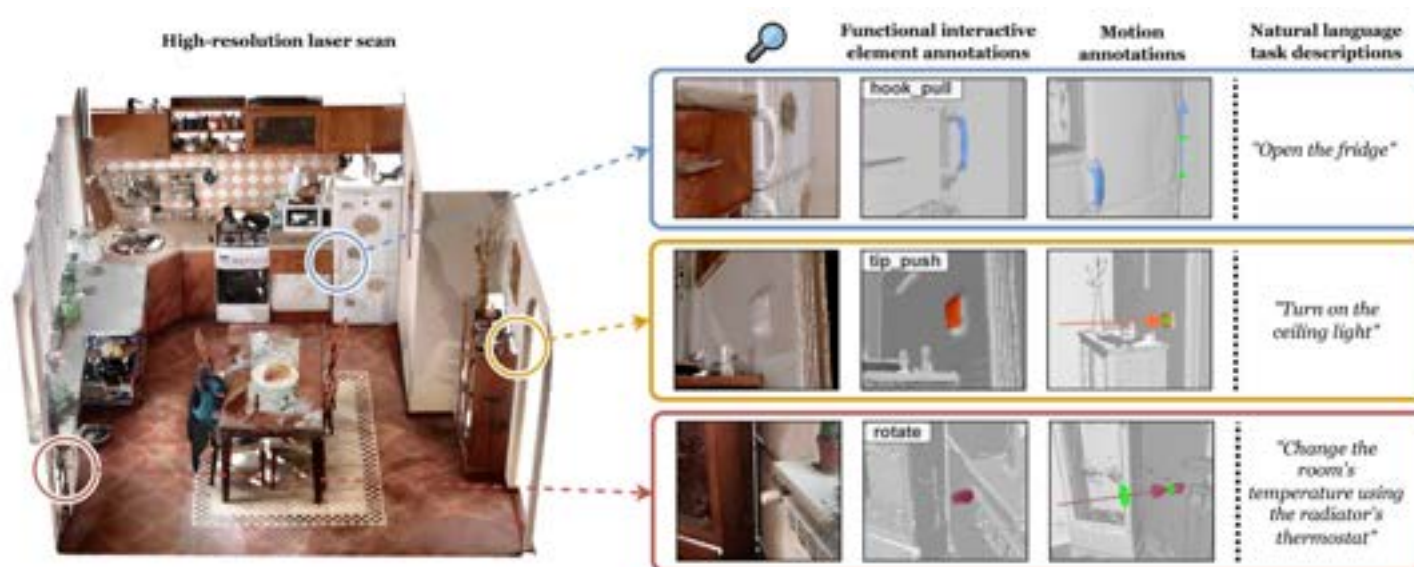
## Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models

- ❑ **概要:**。Visual Programの知識をVLMsに蒸留し、Reasoningが強いかつ学習が容易な手法の提案。
- ❑ **新規性:**。Visual Programmingの手法が色々あったが、直接ではなく蒸留する手法でVisual ProgrammingのReasoning能力を利用するのがあまりなかった。提案の仕組みは既存のVLMsに適応しやすい。
- ❑ **手法:**。メインアイデアは、Visual Programmingの知識をInstruction Tuningの形式でVLMsへ蒸留する。具体的に、まず入力の問題を解くプログラムをLLMにより生成し、あらかじめ持っているModuleで解けないプログラムをFilteringで排除し、最後に良いプログラムをChain-of-Thought形式のInstructionに変換する。
- ❑ **感想:**。手法が面白いが、蒸留よりはVisual Programming手法自体の改善の方がもっと強いReasoningが実現できそう。Visual Programmingとブラックボックス推論両方結合する方が良さそう。



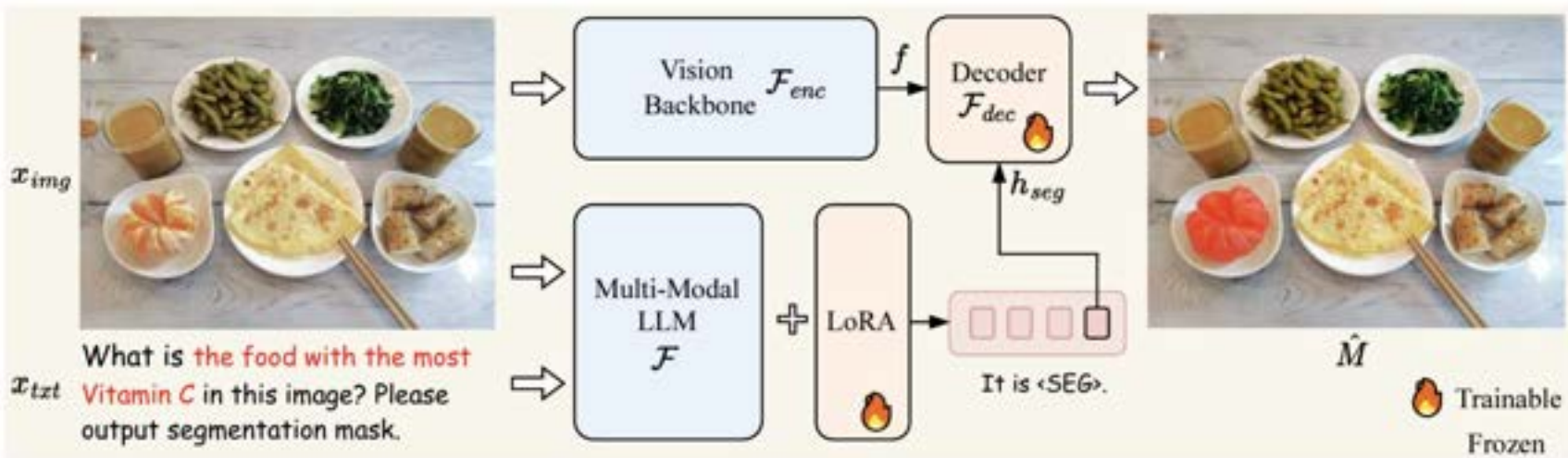
## SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes

- ❑ **概要:** Fine-grained 3D sceneにAffordanceをアノテーションした大規模データセットSceneFun3Dを提案(710シーン、9アフォーダンスカテゴリ、14.8kアノテーション)。SceneごとのInteract可能な物体・モーシオン・そして動作のテキストを手動でアノテーションした。また、3つの新しいタスク: functionality segmentation, task-driven affordance grounding, 3D motion estimationを提案し。SceneFun3Dで既存の手法はまだまだ改善できる余地があることを実験で示した。
- ❑ **新規性:** 3次元のアノテーションが付与したFine-grained affordanceのデータセットがあまりなかった。
- ❑ **感想:**。ロボティクスに活用できそう。この論文と類似したような、実際のロボット応用に近いようなデータが増えてきた。



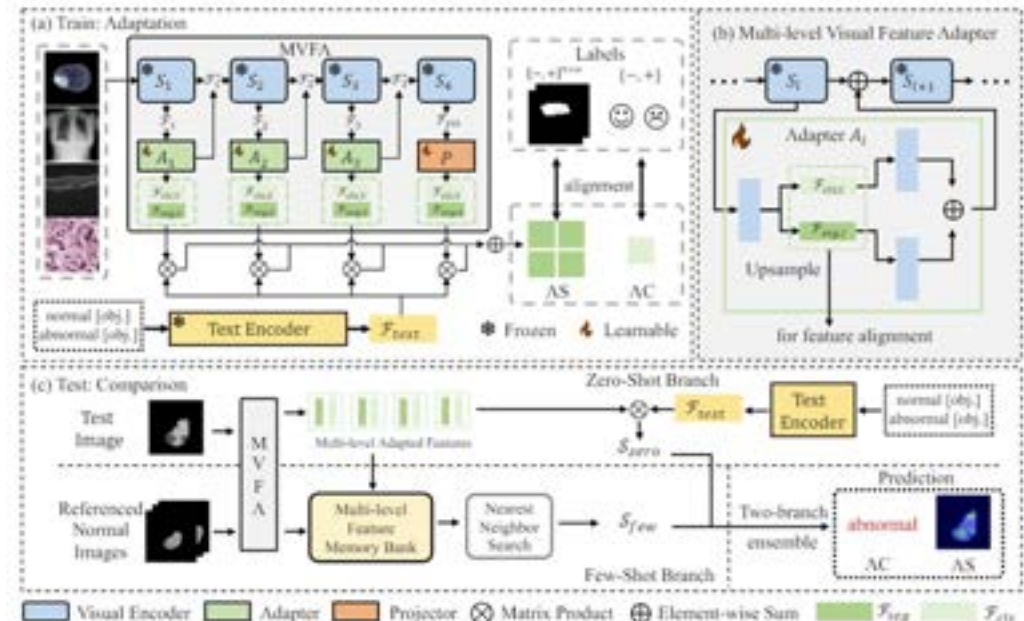
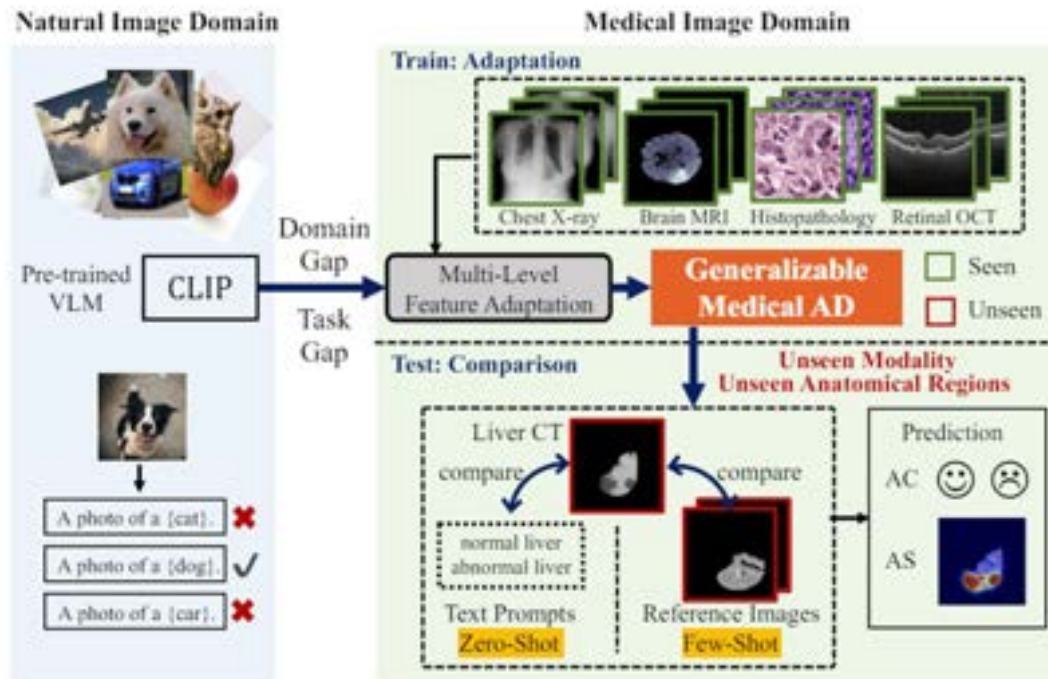
## LISA: Reasoning Segmentation via Large Language Model

- 概要:** 領域を指定するReasoningが必要なテキストと画像から、Reasoningを経由しテキストが対応している領域を画像からセグメントするタスク Reasoning Segmentationと手法(左図)を提案。また、既存のデータセットをLLMでReasoning Segmentationタスク用に变化し学習用データセットを作成。新しい<SEG>トークンを導入し、MLLMの知識をセグメンテーションに活用できるようにした。
- 新規性:**。既存の類似タスクでは、2段階でまずReasoningをしてからセグメントするものが多い。ここで1段階でReasoningとセグメントできる手法を提案。
- 感想:** LISAによりかなりFlexibleで言語により物体分割を行える。類似した形式で、領域が関わるPromptが色々設計できそう。



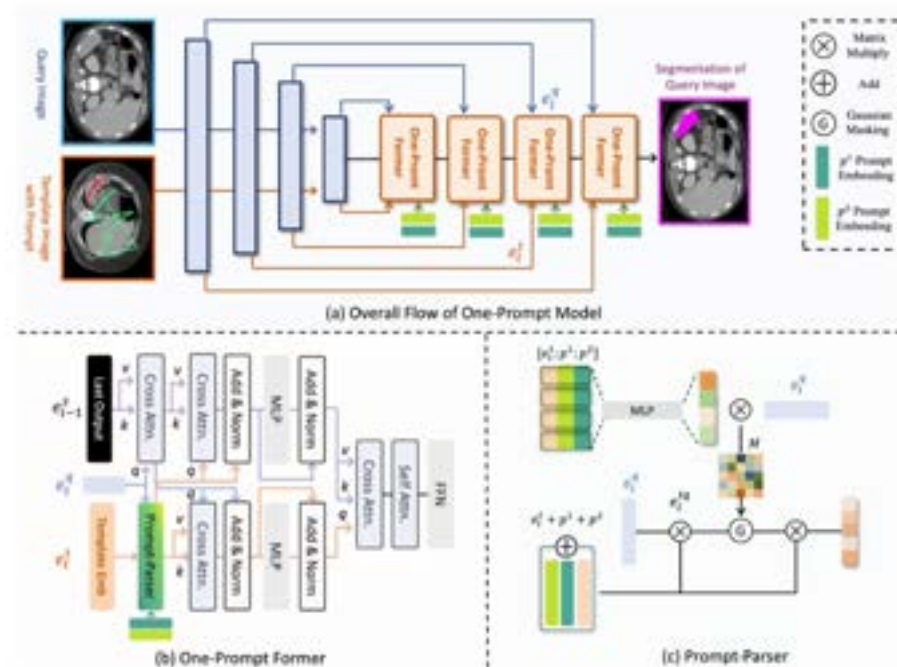
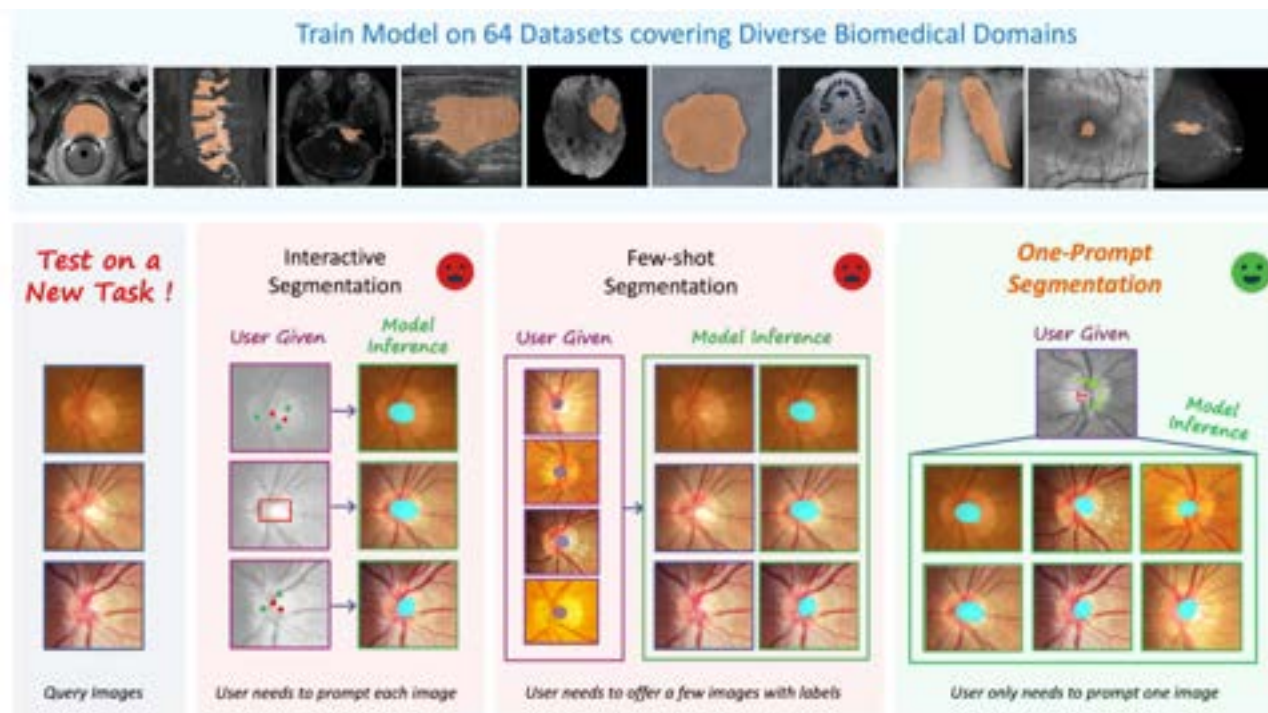
## Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images

- 概要:** 自然画像を対象に学習された CLIP を医療データの異常検知に適応するためのフレームワーク。タスクシフトとドメインシフトを同時に達成している。脳MRI、肝臓CT、網膜OCT、胸部X線 と幅広い領域を対象に検討ゼロショットと Fewショットの異常分類でともに改善を認めた(異常クラス分類平均6.24%と7.33%、異常セグメンテーションで平均 2.03%と2.37%の改善)
- 新規性:** 医療異常検知領域における適応として新規性がある。
- 感想:** 異常クラス分類と異常セグメンテーションにおいて、ゼロショットと Fewショット共に性能が改善されているため、低コストで幅広い領域に応が期待される。プロンプト側の開発も可能であるため、ユーザーによる開発も可能になる。



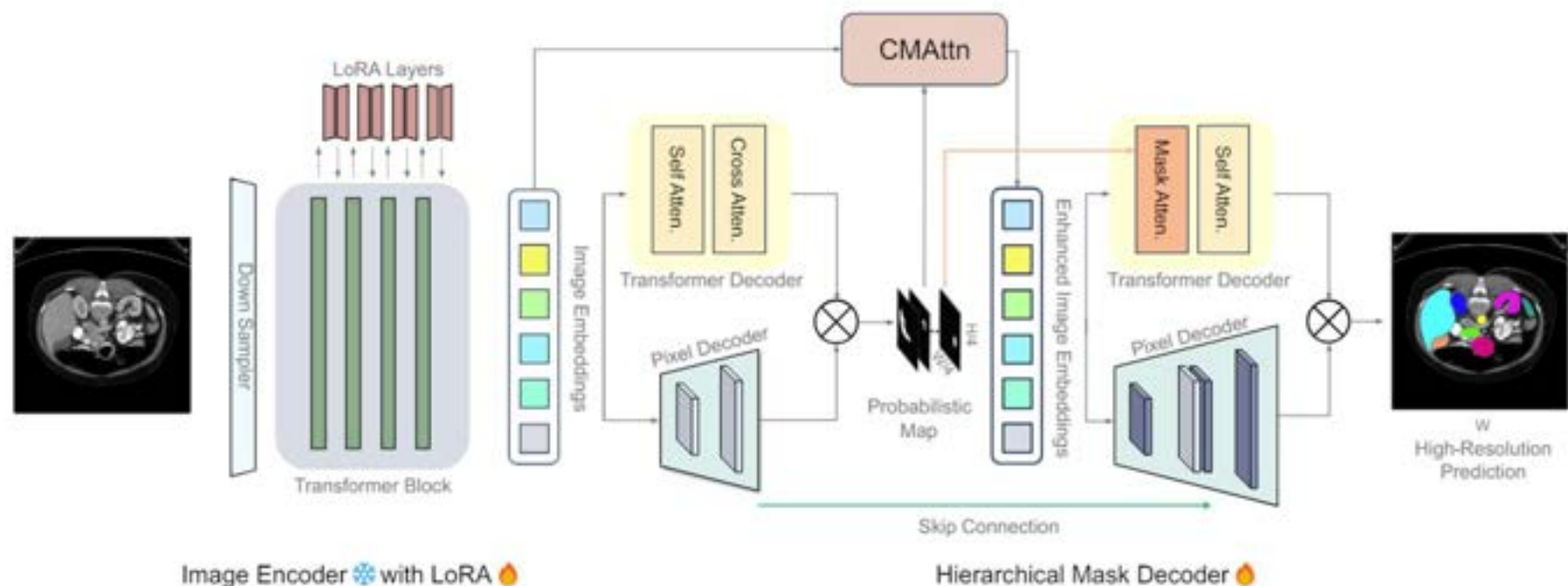
## One-Prompt to Segment All Medical Images

- 概要:** SAM(segment anything)モデルの医療適用。SAMはゼロショットセグメンテーション能力があるが医療領域へのゼロショット能力は低いことが知られている。プロンプトの工夫である程度対応できるが限界があり、多数の医療データが必要とされる。本手法は一つのプロンプトデータのみで、他のデータにも適応できる。さらに Doodleと呼ばれる新しいラインプロンプトを提供する。
- 新規性:** 一つの医療画像プロンプトを他のタスクに応用できる点が新しい。また SAMで利用可能な画像プロンプトは細かいセグメンテーションには不向きだったが、Doodleと呼ばれる新しいラインプロンプトは複雑な形に適応可能である。
- 感想:** SAMの弱点だった複雑なセグメンテーション問題を Doodleという新しいプロンプトで改善しているのが素晴らしい。



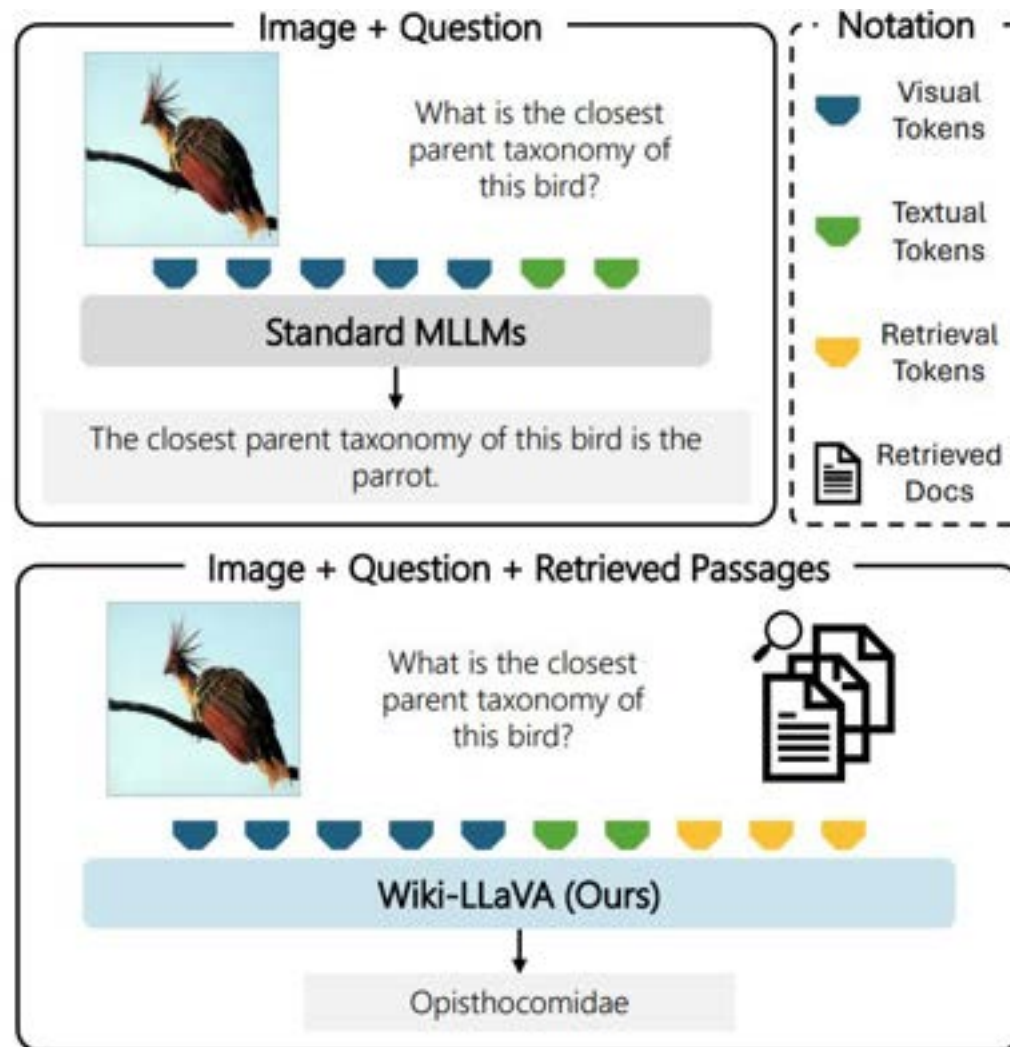
## Unleashing the Potential of SAM for Medical Adaptation via Hierarchical Decoding

- **概要:** SAMは医療ドメインの自然画像に対しては高い性能を発揮するが、学習にデータが用いられていない医療ドメインは苦手。プロンプトを用いた手法は医療知識が必要。本手法はプロンプトなしの医療 SAM改良。
- **新規性:** プロンプトなしの医療 SAM改良である点が新しい。二段階のアプローチ。一段階目で事前マスクを作成し、二段階目で事前マスクに基づき、異なる画像領域間の相互作用を空間的に調節。
- **感想:** 医療SAM改良は追加データを必要とする手法が多いが、本手法はモデル側の改変であるため、根本的な解決になる可能性がある。



## Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs

- ❑ **概要:** マルチモーダルLLMに対して、外部知識を参照することが可能な Retrieval augmentedな枠組みを導入することで、外部知識が必要な質問等にも回答可能なマルチモーダルLLMを実現
- ❑ **手法:** Visual tokenとTextual tokenの入力に加えて、wikipediaから検索した情報を入力するRetrieval tokenを新たに導入することで、外部知識を参照する。外部知識の参照には、CLIPモデルおよびContrieverモデルを利用しており、画像に関連する文章をCLIPにより検索したのち、質問に関連しない文章をContrieverモデルによりスクリーニングしている
- ❑ **感想:** wikipedia等の外部知識を参照することができれば、LLMモデルを利用しなくても質問等に回答することが出来てしまいそう。参照ドキュメントの検索は画像からのみで行うので、CLIPモデルの精度がかなり影響しそう。文章も検索クエリに入れてあげてもよいのでは？



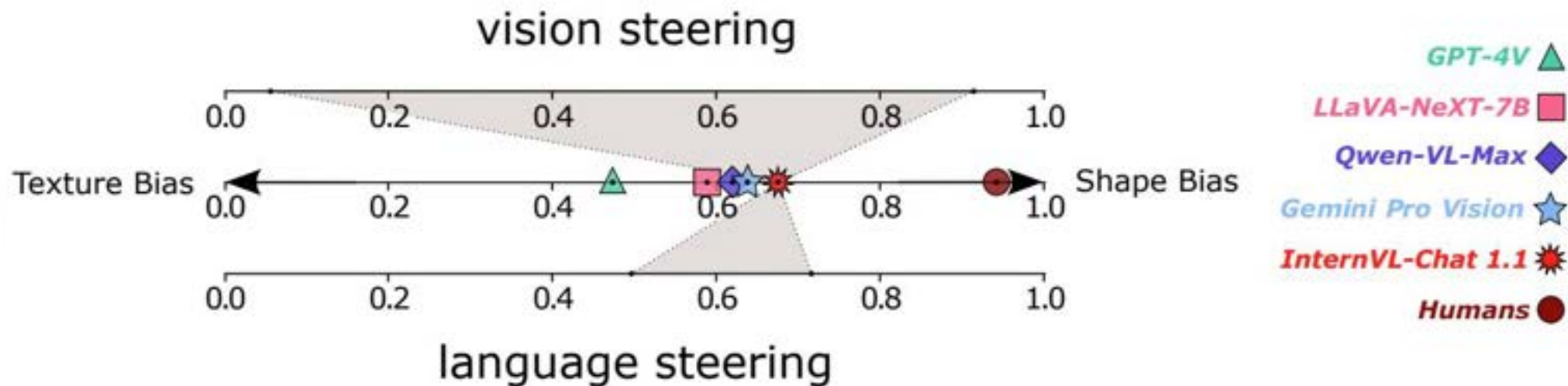


## Are Vision Language Models Texture or Shape Biased and Can We Steer Them?

- **概要:** Vision and languageモデルがtextureとshapeのどちらにbiasを持つのかを検証した論文. VQA, Image captioningを通じた実験により Vision and languageモデルはshapeにbiasを持つことを明らかにしている. また, languageの情報を活用することで, 上記のbiasを緩和することが可能であることも明らかにしている.
- **手法:** ImageNetに対して, style transferを施したデータセットを用いることでbiasを測定している. 具体的には, shapeに基づく正解率とtextureに基づく正解率を比較することで, shape biasとtexture biasを算出している.
- **感想:** 確かに, shape側にbiasが寄ってはいるが0.6ポイント周辺であるため, 顕著な結果であるとは言いにくそうな印象. また, データセット等もImageNetを主体とした検証であることから, 未だ検証の余地はありそう

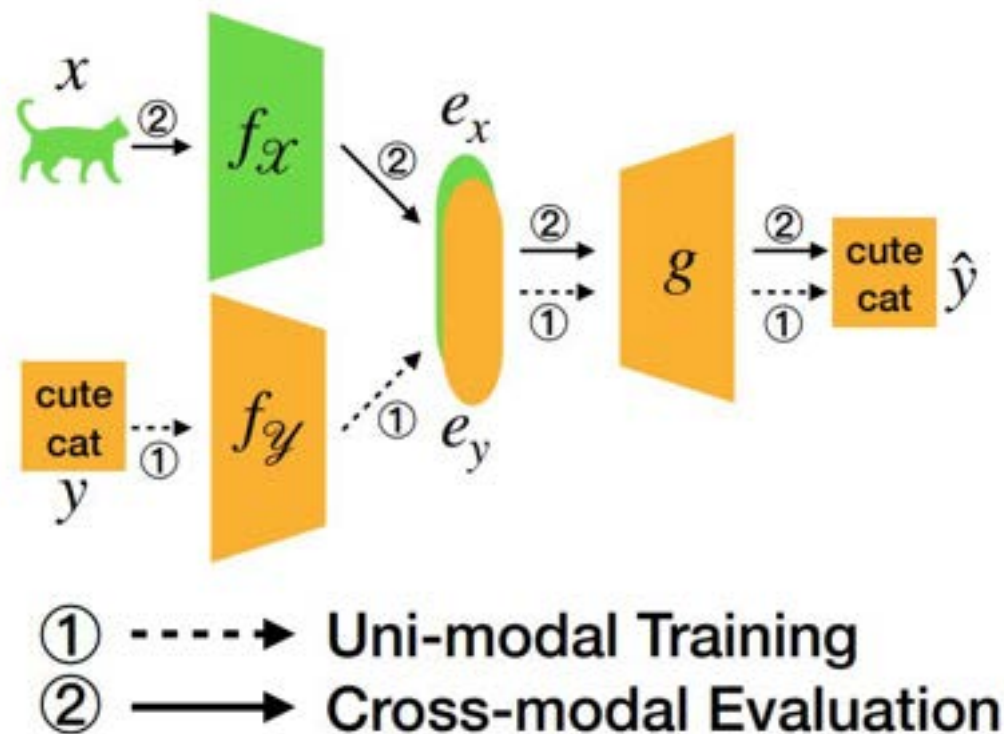


Shape: Elephant  
Texture: Bottle



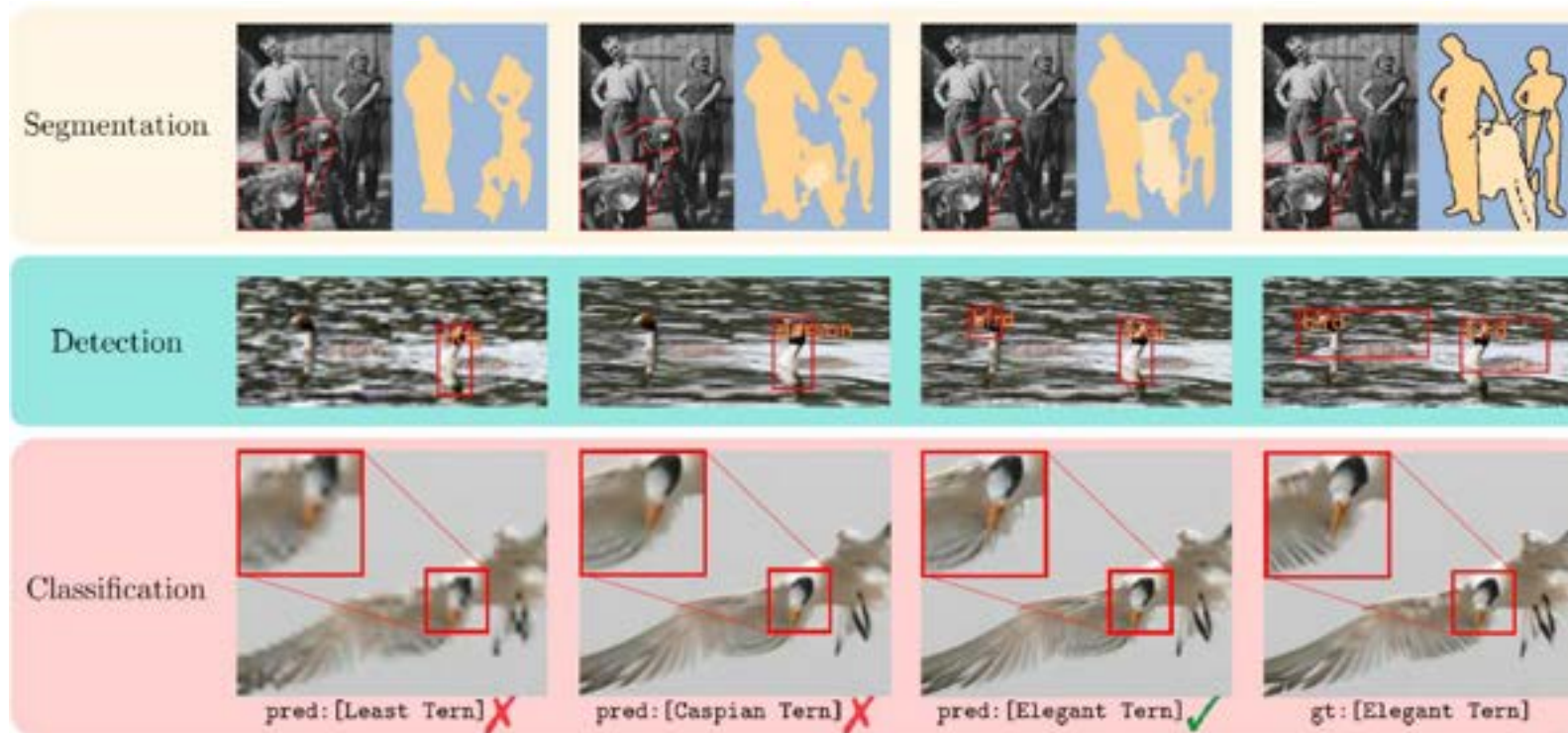
## Connect, Collapse, Corrupt: LEARNING CROSS-MODAL TASKS WITH UNI-MODAL DATA

- 概要:** Uni-modalデータを用いたcross-modal学習の高精度化のために, cross-modal空間のgeometryを解析し, modality gapやalignment noiseを考慮可能な手法を提案した論文. 実験により, uni-modalデータを用いたcross-modal学習を高精度化可能であることを確認している.
- 手法:** Connect step, Collapse step, Corrupt stepの3段階により構成されている. Connect stepにて異なるモダリティのデータをつなげ(従来のCross-modalと同様), Collapse stepにてCross-modal空間におけるmodality分布のミスマッチを解消する. また, Corrupt stepにて微小なノイズを付与することによりcross-modal学習に生じるalignment noiseをuni-modalな学習において模倣している.
- 感想:** 手法としてのcontributionは多くないが, 解析自体を丁寧に行っていることのcontributionは大きそうな印象を抱いた. ただ, 可能であれば, modality gapやalignment noiseが生じる原因についてまで言及して欲しかった.



## Beyond Image Super-Resolution for Image Recognition with Task-Driven Perceptual Loss

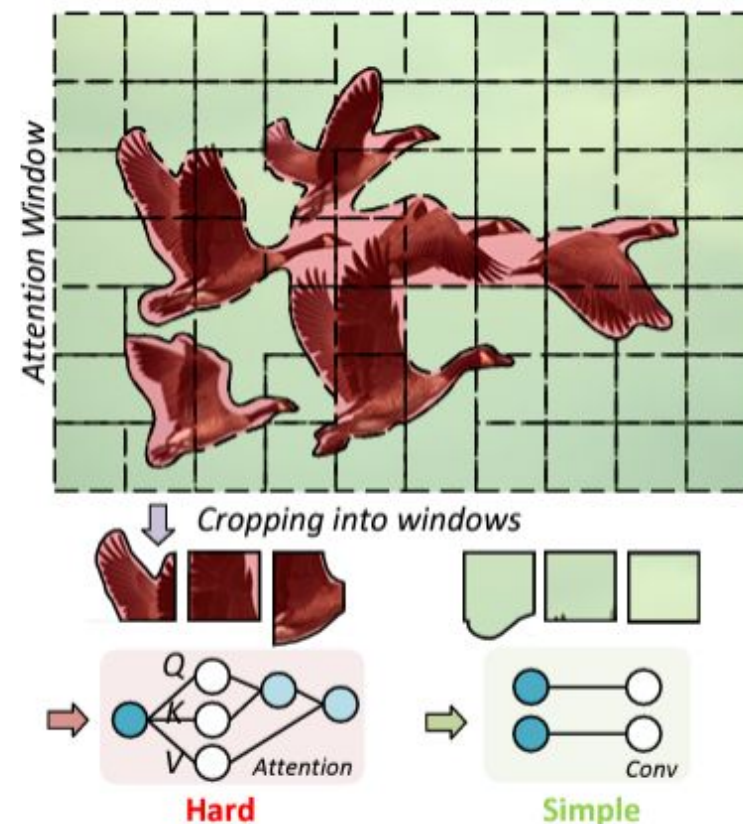
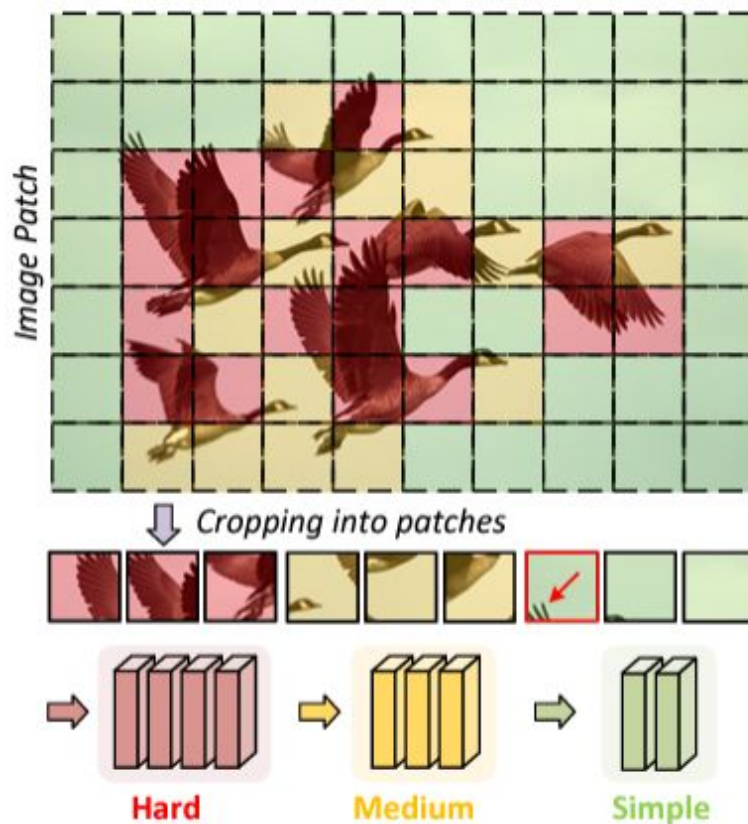
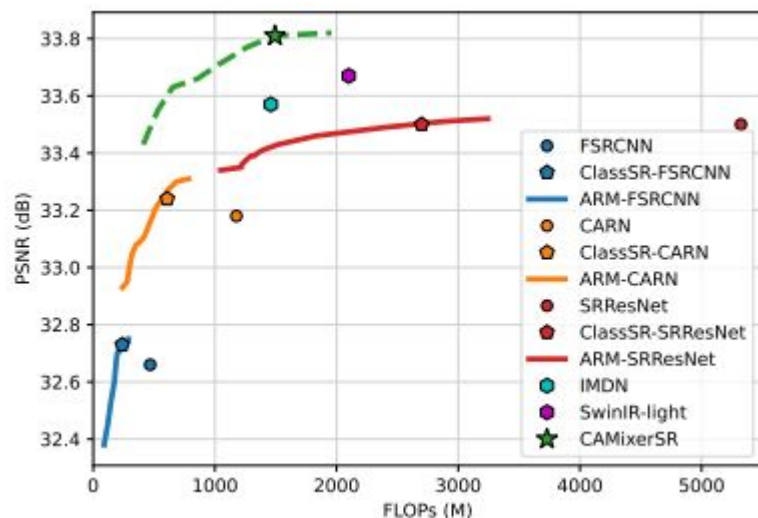
- ❑ 超解像技術の画像理解のタスクへの応用
  - ❑ semantic segmentation、object detection、image classification
- ❑ task-driven perceptual (TDP) lossにより、タスクに特化した知識を獲得可能



## CAMixerSR: Only Details Need More “Attention”

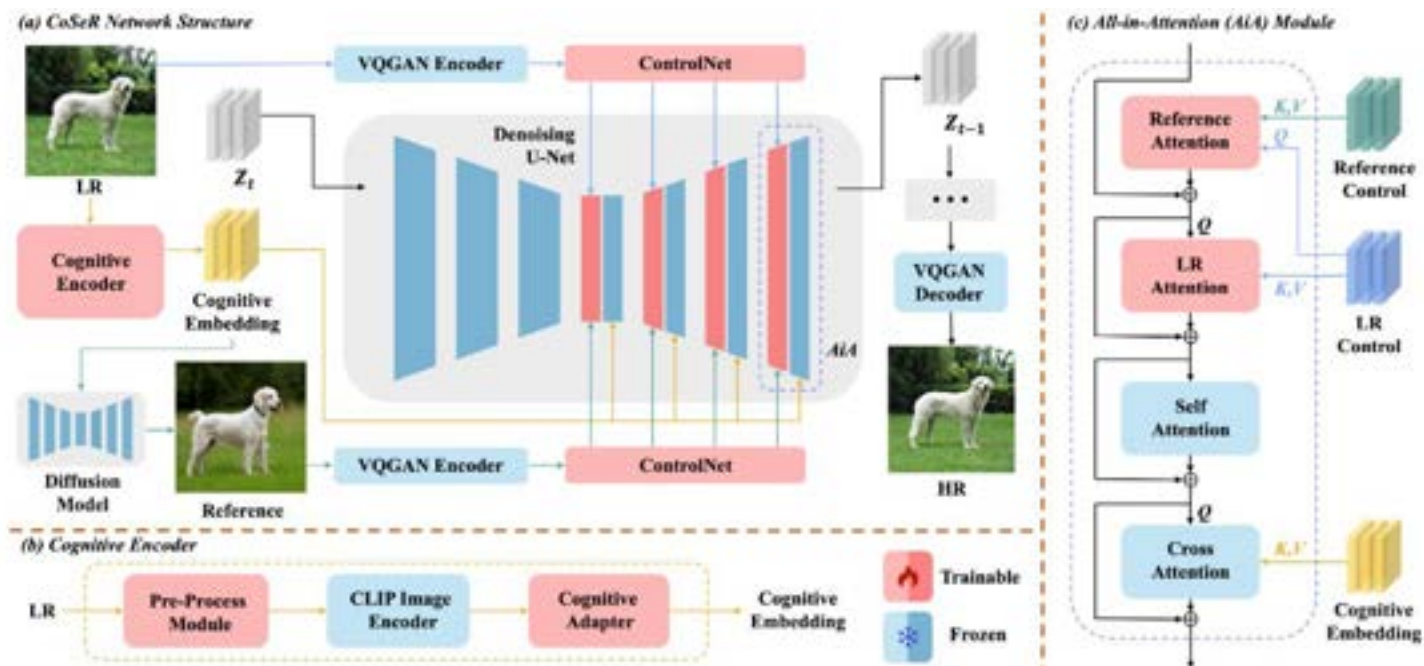
- content-aware mixer (CAMixer)
  - ClassSRよりも細かくsimple contentと複雑な(高周波成分を多く含む)texture領域を分割
  - texture領域に対してはSelf-Attentionを適用、simple contentに対してはConvolution。

- SRの品質と計算コストのトレードオフを改善



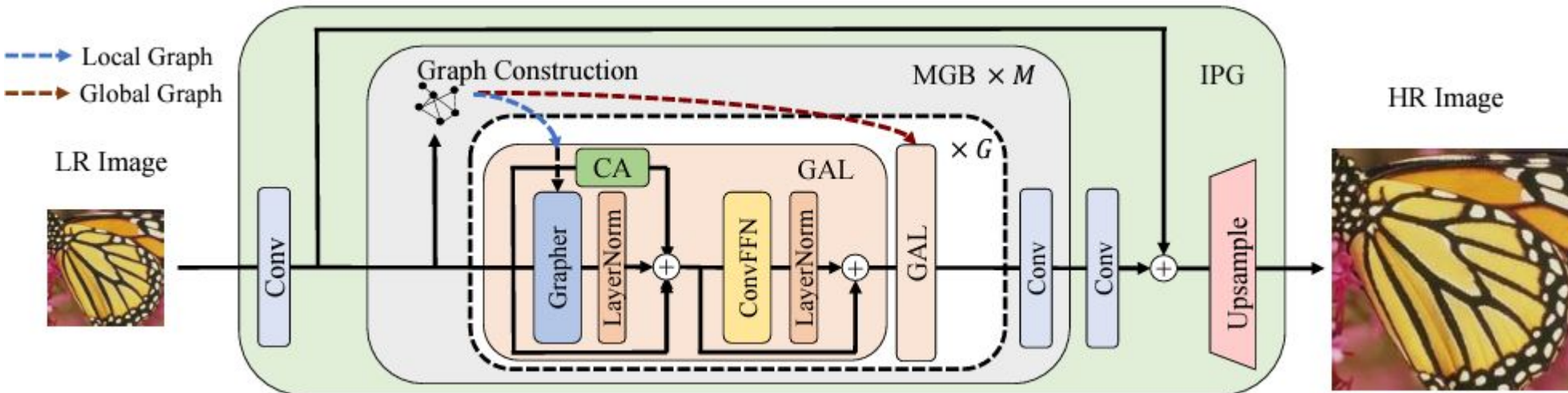
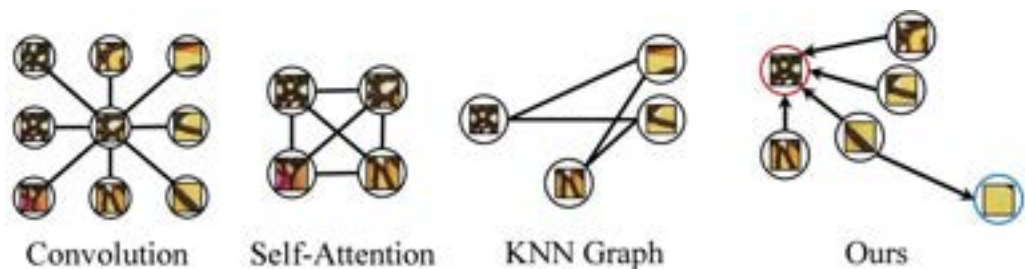
## CoSeR: Bridging Image and Language for Cognitive Super-Resolution

- ❑ 低解像度の入力画像からDiffusionモデルを利用してReference画像を生成
- ❑ 意味的に正しく、かつ写真のようにリアルなディテールを復元可能
- ❑ 感想: テキストガイドに近いことを画像領域で実施した面白い発想



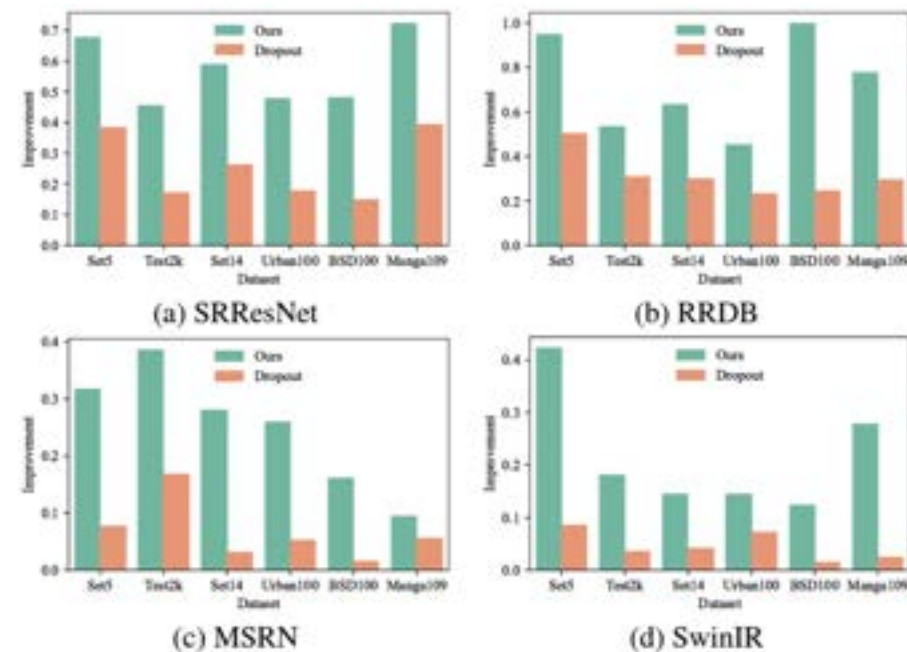
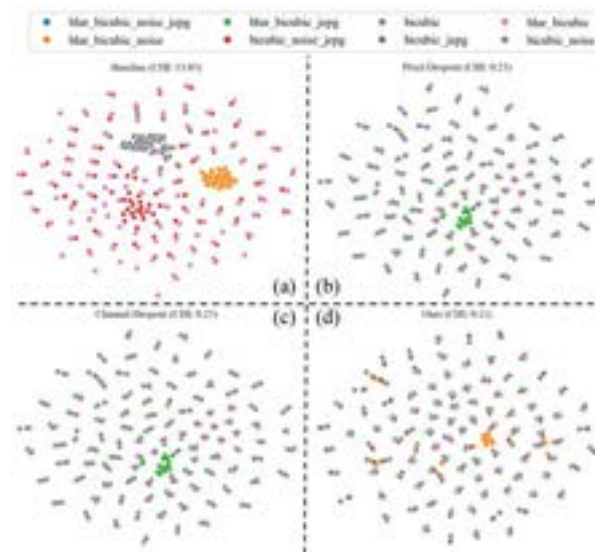
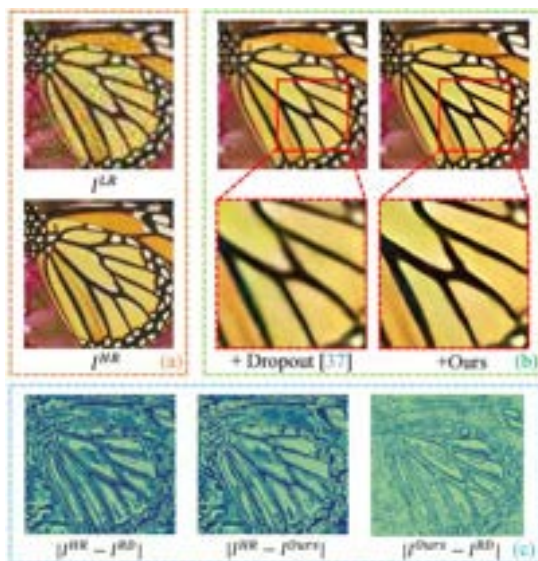
## Image Processing GNN: Breaking Rigidity in Super-Resolution

- CNNやWindow-attentionによる情報処理の代わりにグラフを利用
  - より柔軟な情報処理が可能
  - より細かい処理が必要なところに、より多くのノードを割り当てる。
- 局所画像(近傍画素)のグラフ(local graphs)と画像全体のグラフ(global graphs)を探索
- Image Processing GNNの内部
  - GAL: local グラフを利用したLayer
  - MGB: global グラフと複数層のGALを利用したBlock



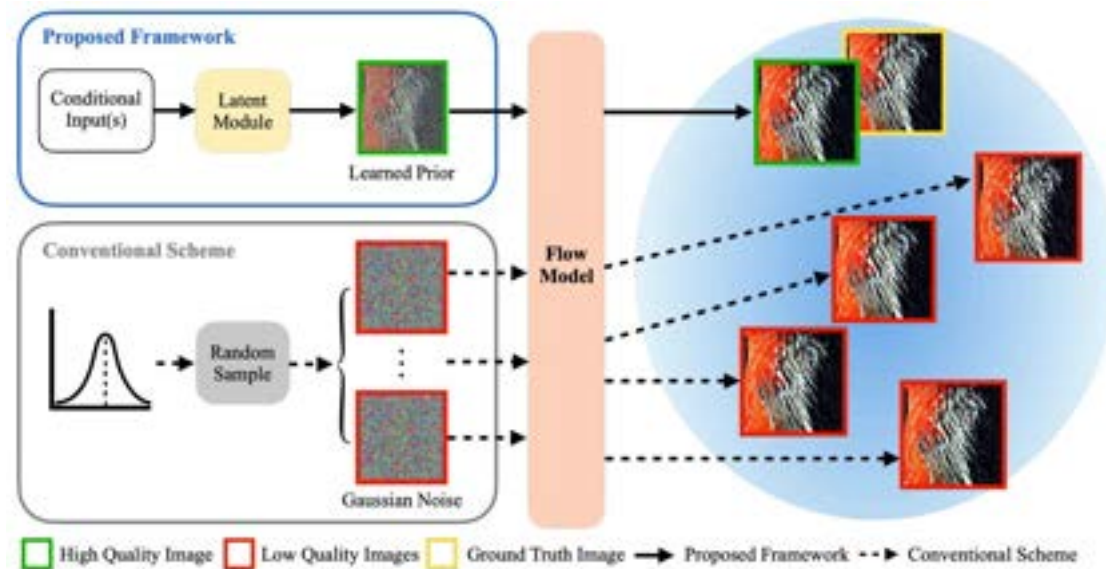
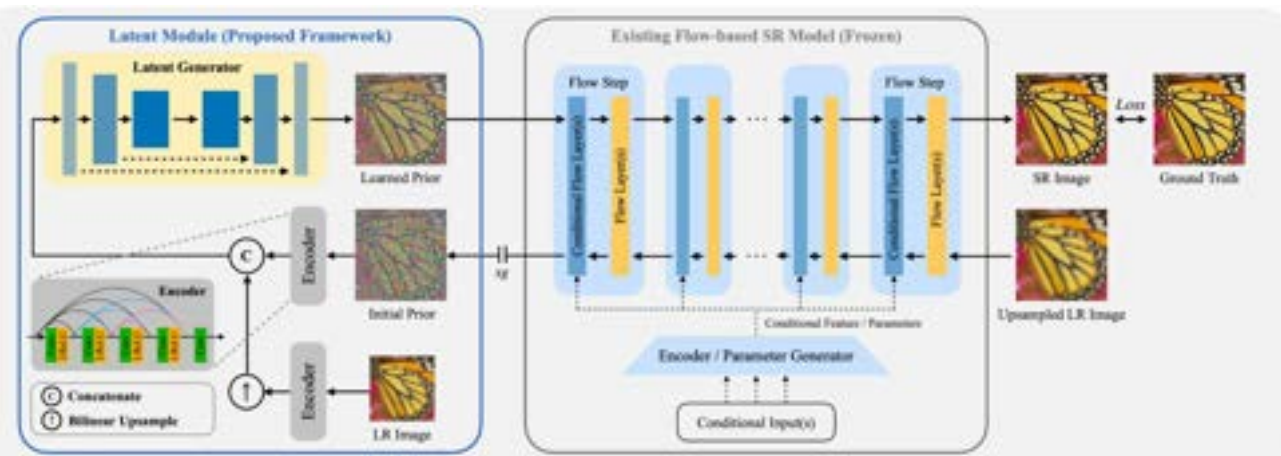
## Navigating Beyond Dropout: An Intriguing Solution Towards Generalizable Image Super Resolution

- メインメッセージ: 正則化の文脈で深掘りする研究を増やすべきだ!
- 新しい正則化の手法を提案。  
Dropoutを使うと、詳細な画像再構成の能力が損なわれることを問題視。
- Dropoutを使う場合と、提案手法を使う場合の復元画像の差分
  - 提案手法によって、より鮮明な復元画像を実現
- どのモデル、どの評価DSでも有意に性能向上



## Boosting Flow-based Generative Super-Resolution Models via Learned Prior

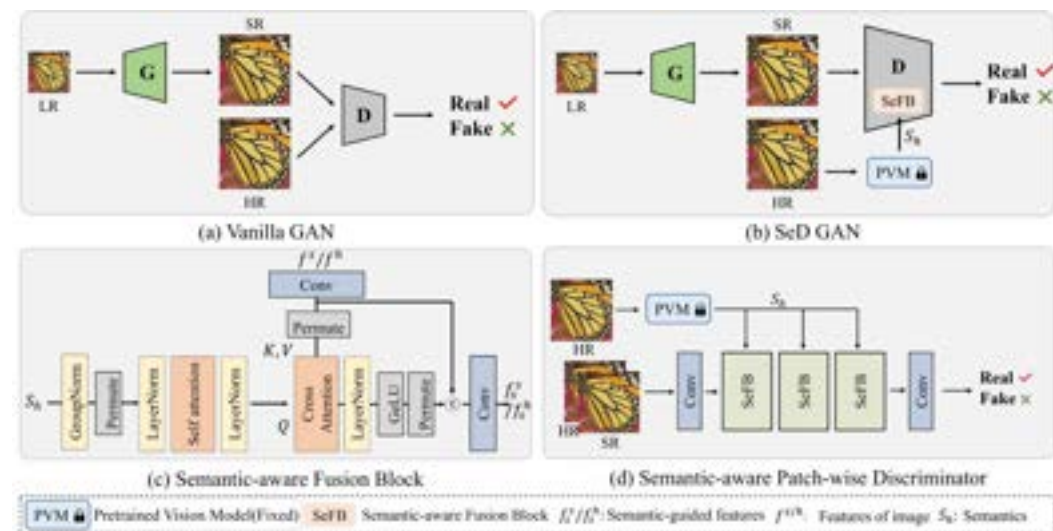
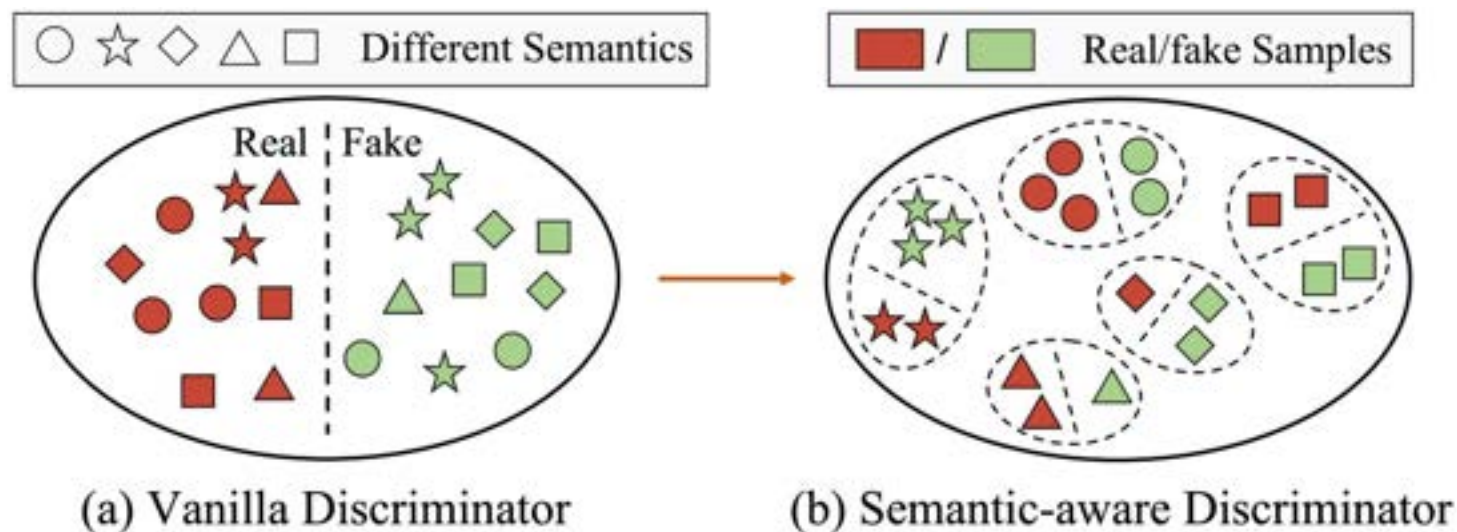
- ❑ **概要:** FlowベースのSRモデルのための条件付き学習手法. 従来のFlowベースのモデルでは、1. グリッドアーティファクト 2. 逆行列の爆発 3. 固定のサンプリング温度による最適でない結果などの課題が生じている. これらの問題に対処するために学習される事前分布の特徴に条件を課している.
- ❑ **手法:** 既存のFlowベースのモデルと潜在モジュールで構成されている. 潜在モジュールでは画像空間と潜在空間の両方から低解像度(LR)画像の特徴を抽出し, 条件付き事前分布を出力する. 目的関数としてVGG19による知覚損失を使用し, 潜在空間損失を正規化項として採用することで分布外の生成を防止する.
- ❑ **感想:** 従来研究のSRFlowでは”事前分布からランダムサンプリングした潜在変数+LR”からSRを生成していたが, 本手法ではLRの特徴を反映した潜在分布を生成している点で新しいと感じた.





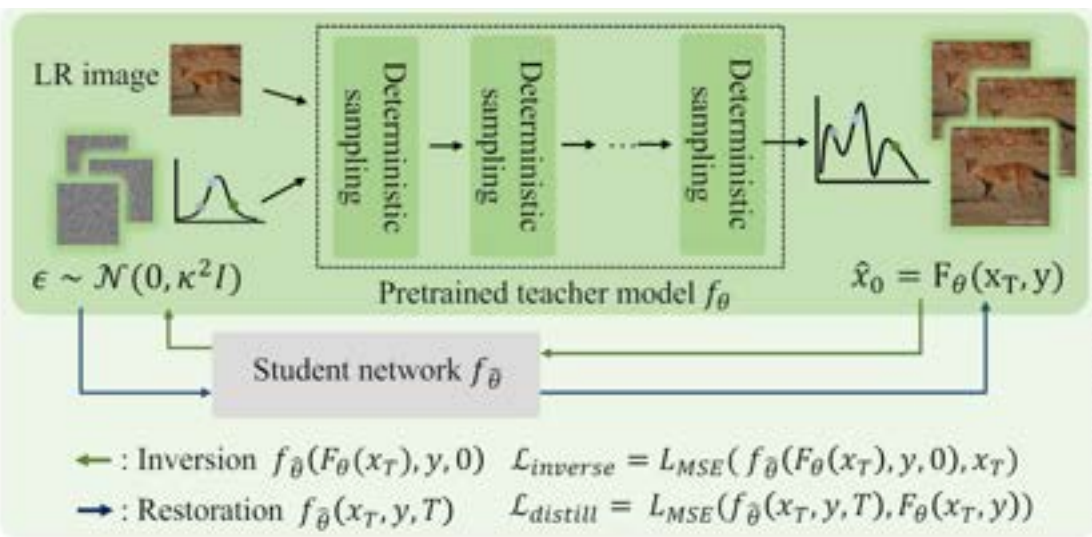
## SeD: Semantic-Aware Discriminator for Image Super-Resolution

- 概要:** セマンティックを考慮したテキストチャの生成を行っている, GANは識別器を利用して現実世界の高解像度画像の分布を学習するが, 分布の学習の粒度が粗いため仮想のテキストチャの影響を受けやすく, 直感に反する生成を行うことがある。画像の意味を条件として導入することにより, ネットワークが細かい分布を学習できるように促している。
- 手法:** 学習されたセマンティック識別器から画像の意味を抽出する。ピクセル単位のセマンティクスを得るために事前学習済みCLIPの中間特徴からセマンティクスを抽出。SeFBブロックを提案しており, 抽出されたセマンティクスをクエリとみなしクロスアテンションを通じてセマンティクスを考慮した画像特徴を識別器に入力する。
- 感想:** 多くのGANベースの超解像手法と組み合わせることが可能である点が強いと感じた。

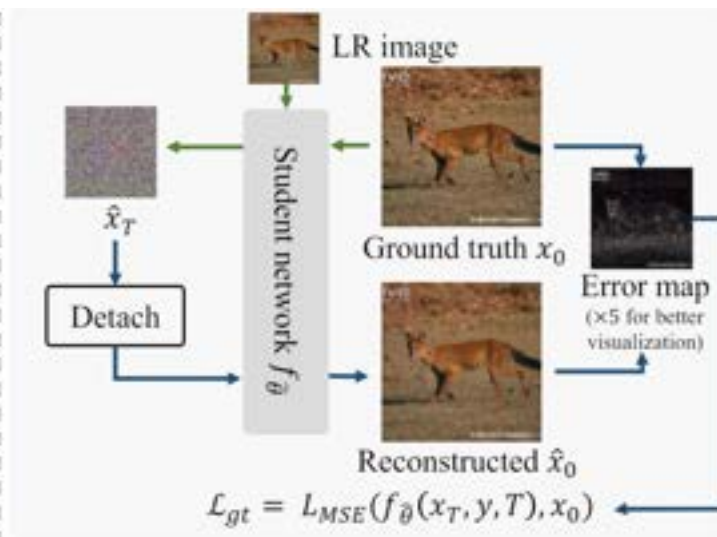


## SinSR: Diffusion-Based Image Super-Resolution in a Single Step

- 概要:** 単一ステップで超解像を実現するDiffusionベースの手法を提案. SoTaモデルと同等またはそれ以上の性能を示すと共に推論時間を大幅に短縮.
- 手法:** 事前学習済みResShiftを利用した決定論的サンプリングを行い,  $x_0 \leftrightarrow x_T$ を単一ステップでマッピングする生徒モデルに蒸留. 蒸留プロセスにGT画像を利用した一貫性保存損失を導入することで, 生徒モデルが教師モデルからの情報を使用するだけでなく, GT画像からも直接学習することを促す.
- 感想:** ステップ数を減らしつつ高い性能を出していることが良い.  $x_t$ と $x_0$ の直接的なマッチングが比較的容易であるという結果は興味深い.



(a) One-step bi-directional distillation



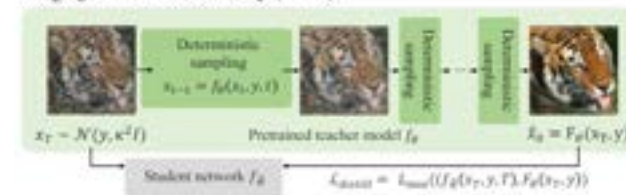
(b) Consistency preserving loss



(a) The inference of SR3 [32] starts from a pure noise, which requires a large number of inference steps ( $T=100$  after using DDIM [36]).



(b) The recent SOTA method ResShift [45] shortens the Markov chain to speed up the inference process by incorporating the information of the LR image  $y$  to the initial state  $x_T$  ( $T=15$ ).

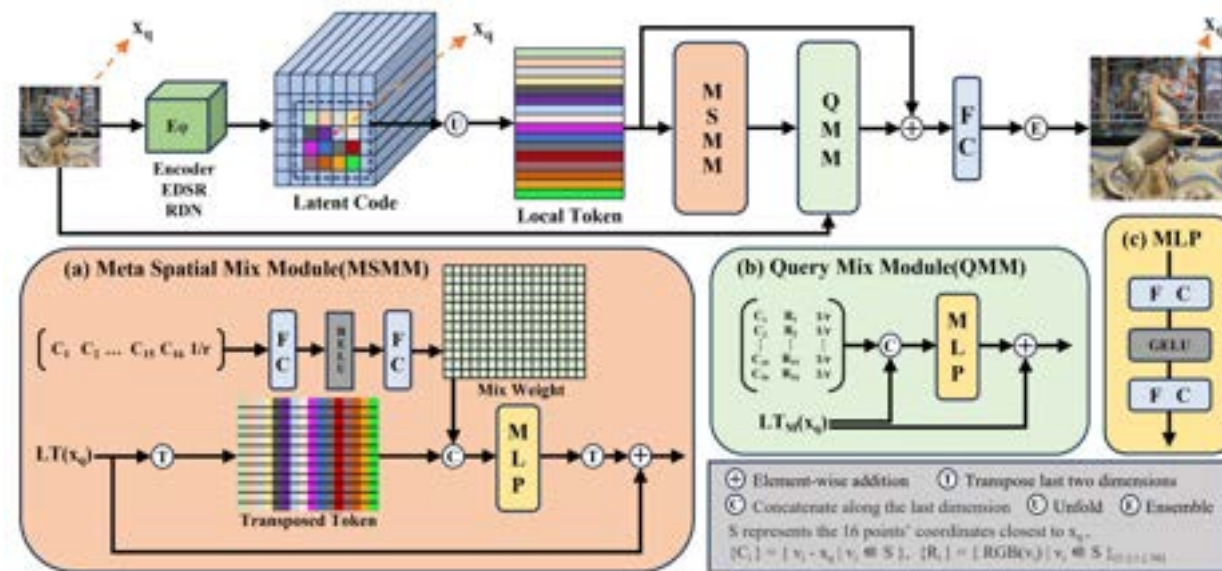


(c) A simplified pipeline of the proposed method SinSR (distill only). It directly learns the deterministic mapping between  $x_T$  and  $x_0$ , therefore the inference process can be further compressed into only one step ( $T=1$ ).

## Continuous Optical Zooming: A Benchmark for Arbitrary-Scale Image Super-Resolution in Real World

- 概要:** 任意スケール画像超解像で使用されている既存データセットは、実世界の劣化に非対応、もしくは固定倍率であるため課題があった。そこで、実世界向けの任意スケール画像超解像データセットであるCOZデータセットを提案。加えて、モデルのロバスト性を高めるため、MLP-mixerとメタラーニングに基づくLocal Mix Implicit network (LMI)を提案。
- 手法:** (COZデータセット) 撮影はモーターとベルトを利用して自動で焦点距離を変更する撮影システムを利用。ズームによる撮影時の輝度および解像度の変動に対処するため輝度調整後に位置合わせを行う二段階のSIFTアルゴリズムを採用。(LMI)MSMMでは各座標を用いて複数のlatent codeの重みを算出し結合。QMMでは元画像のRGB値や座標情報をクエリとしてMSMMの出力した特徴量に埋め込む。
- 感想:** 下記の表のように既存データセットとの違いを明確に示しているのが良いと感じた。LMIのEncoderとしてEDSRもしくはRDNを使っているが、それが妥当なのかは不明。

Dataset	Conference	Real-World	Arbitrary-Scale	Zoom
DIV2K [1]	CVPRW 2017	✗	✓	-
RealSR [4]	ICCV 2019	✓	✗	Manual
City100 [6]	CVPR 2019	✓	✗	Manual
SR-RAW [33]	CVPR 2019	✓	✗	Manual
DRealSR [30]	ECCV 2020	✓	✗	Manual
Ours (COZ)	-	✓	✓	Automatic



## SportsHHI: A Dataset for Human-Human Interaction Detection in Sports Videos

- ❑ **概要:** 新しいビデオベースの視覚関係検出タスク:「ビデオ中の人間同士の相互作用検出」を提案, 本提案を実現するために「SportsHHIデータセット」を構築, 構築したデータセットの使用により, 複数人が存在する複雑なビデオにおける人間同士の相互作用検出のパフォーマンスが大幅に向上
- ❑ **手法:** Youtubeからバスケットボールとバレーボールのビデオクリップを収集, 34種類の相互作用カテゴリに分類, VIAツールで118,075個の人体境界ボックスと50,649個の相互作用インスタンスをアノテーション. オブジェクト検出器でビデオフレーム内の人を検出, 検出結果に基づき, 相互作用提案を生成, 提案特徴を抽出. 抽出した特徴に基づき, 相互作用を分類.
- ❑ **感想:** 試合動画から, 特定選手のスキルとパフォーマンスを定量的に評価できるシステムの先行研究になりそう. 局後の検討にも有用



Figure 1. Comparison between previous video visual relation detection datasets and our SportsHHI. In the upper row,

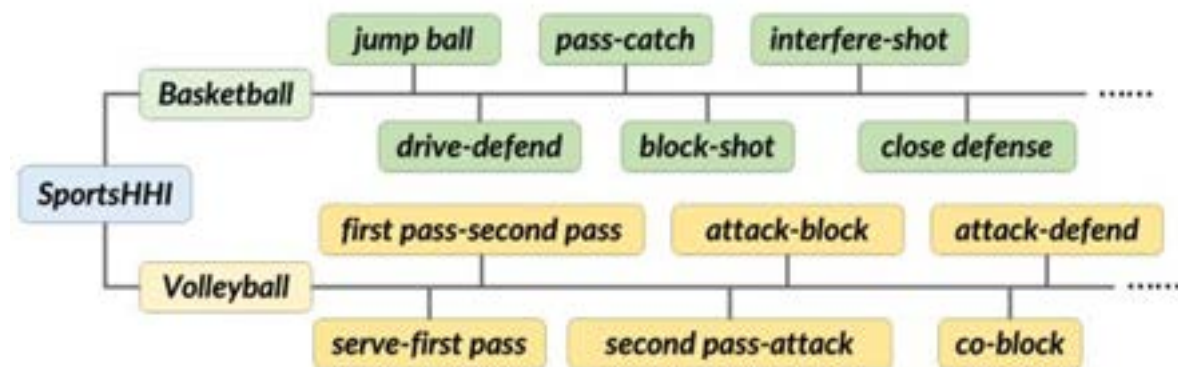
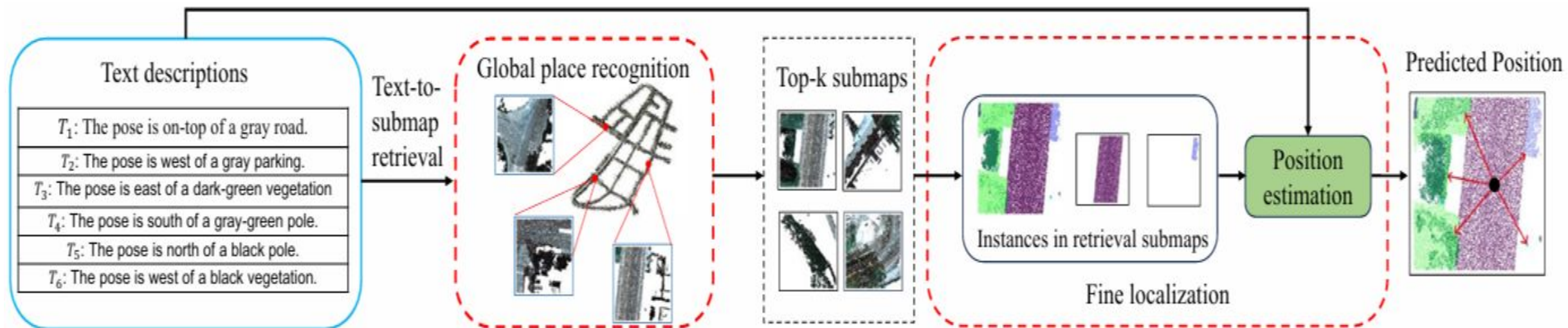


Figure 3. **Interaction classes hierarchy.** There are 34 interaction classes of high-level semantics in total in SportsHHI. 16 for basketball and 18 for volleyball.

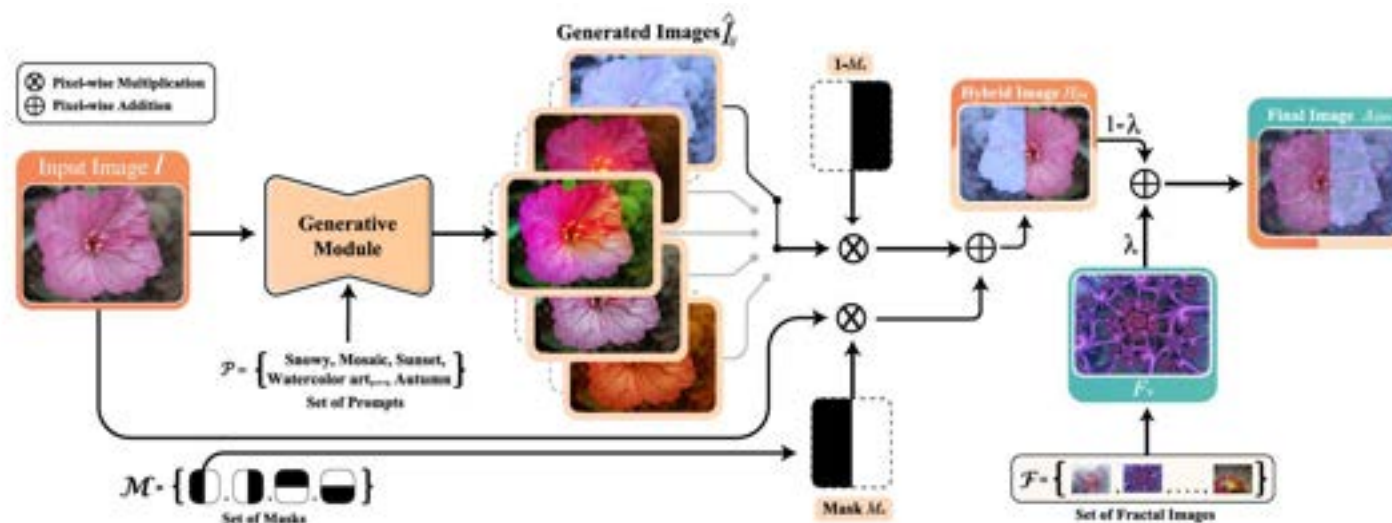
## Text2Loc: 3D Point Cloud Localization from Natural Language

- ❑ **概要:** 自然言語の記述に基づいて3D点群上での位置を特定する手法。周囲を表す点群と、位置を記述したテキストクエリが提供されると、Text2Locはその記述された位置の地図内での最も確率の高い場所を特定する。ベースラインを最大2倍上回り、テキストクエリを5m以下の精度で位置特定可能。
- ❑ **新規性:** 場所の文記述内および文間のコンテキストの詳細を階層的に表現する新しい注意ベースの方法を提案。最終的な位置特定段階でテキストインスタンスマッチャーの使用を完全に廃止した初めての存在。
- ❑ **手法:** テキストベースの位置記述が与えられると、まずターゲット位置を含む可能性のある粗い候補位置「サブマップ」のセットを特定(事前に構築されたサブマップのデータベースから上位k個の最も近いサブマップを取得)。次に、取得したサブマップの中心座標を、設計したマッチングフリーの位置推定モジュールを使用して精緻化し、ターゲット位置の精度を向上。
- ❑ **感想:** 現状のサブセットはデータベースから入手しているため、そこを物体検出のRPNのように提案ベースにできれば、新規性として新たに提案できるのでは？



## DiffuseMix: Label-Preserving Data Augmentation with Diffusion Models

- Diffusionのimg2imgで画像の雰囲気(色, 季節, 画風など)を変化
  - 元画像にマスク(半々)
  - Mixing Set (Fractals)を重ねる



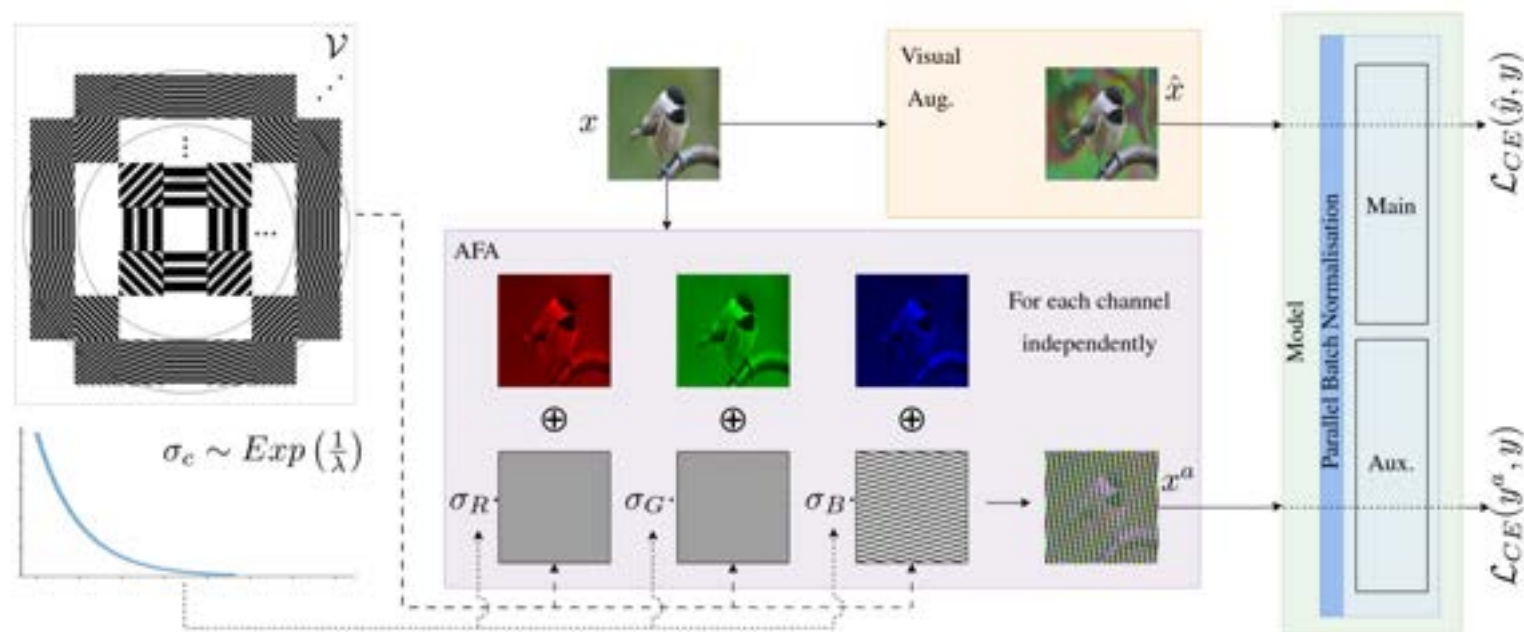
arXiv: <https://arxiv.org/abs/2405.14881>

GitHub: <https://github.com/khawar-islam/diffuseMix>

Fourier-basis functions to bridge augmentation gap:

Rethinking frequency augmentation in image classification

- 各チャンネルにフーリエ基底関数に基づく縞模様を合成したものを学習に使用
- 画像の見た目を変化させるデータ拡張(PRIME, AugMix, etc)と組み合わせることで、さらに性能を上げられる

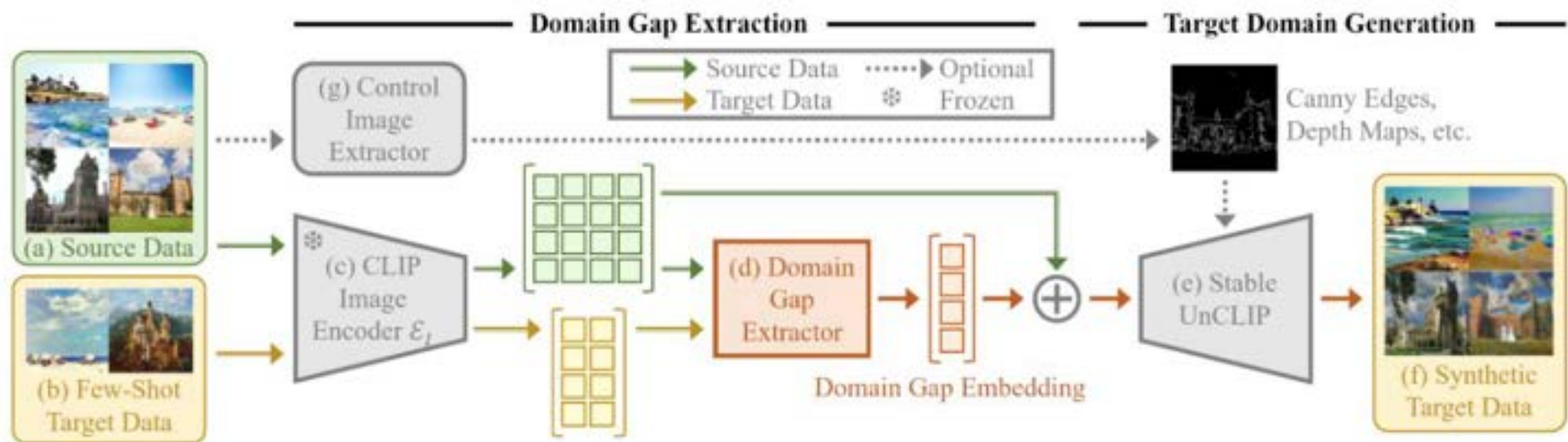


arXiv: <https://arxiv.org/abs/2403.01944>

GitHub: <https://github.com/nis-research/afa-augment>

## Domain Gap Embeddings for Generative Dataset Augmentation

- ❑ 対象のデータ数が不足している場合, 画像生成でデータ拡張したい
- ❑ 大規模な別のデータセットと対象データセットのドメインギャップを抽出
- ❑ 大規模なデータセットの画像をソースとして画像生成するとき, 対象のデータセットとのドメインギャップを考慮して生成することで, 少量データセットのドメインで多様なデータセットを構築できる





## Distilling Vision-language Models on Millions of Video

- ❑ **目的:** ImageベースのVLMを動画に適用し, 動画内の動的シーンや因果関係を理解して, 高品質な擬似キャプションを生成する
- ❑ **課題:** 人力のラベリングコスト, 画像ベースのモデルをそのまま用いると時間的情報を捉えられない
- ❑ **提案:** Visual Adaptation, Language Adaptationを2段階で行い, 動画にAdapt  
キャプションは, 静的な外観, 一般的な動作, 詳細な身体動作など複数の粒度で生成される

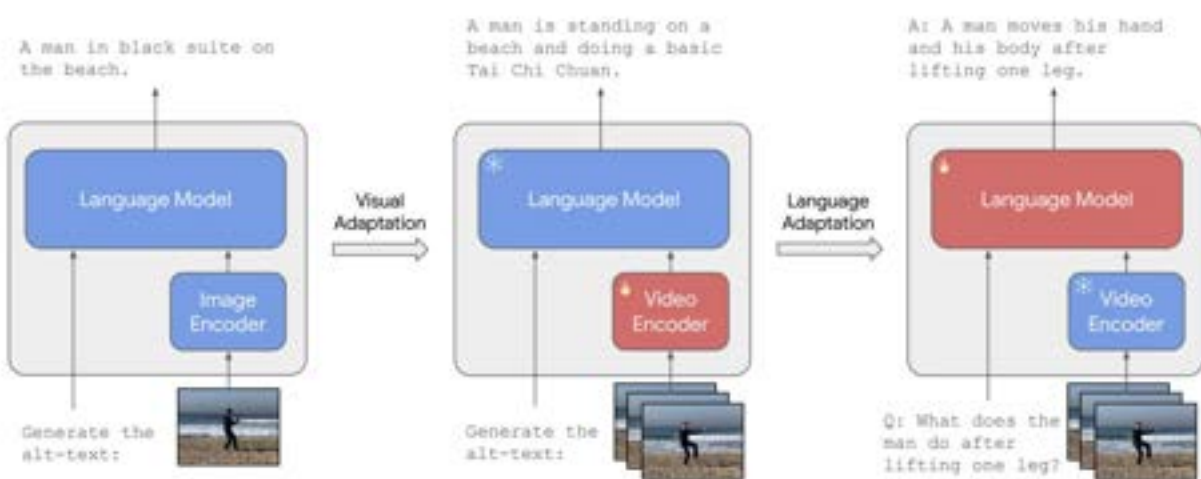
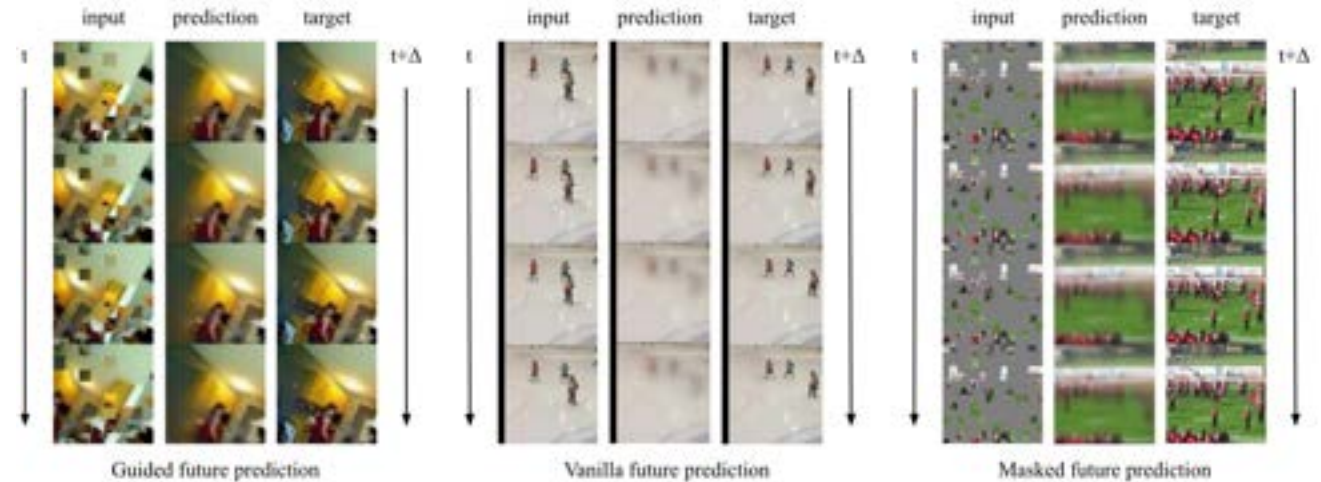
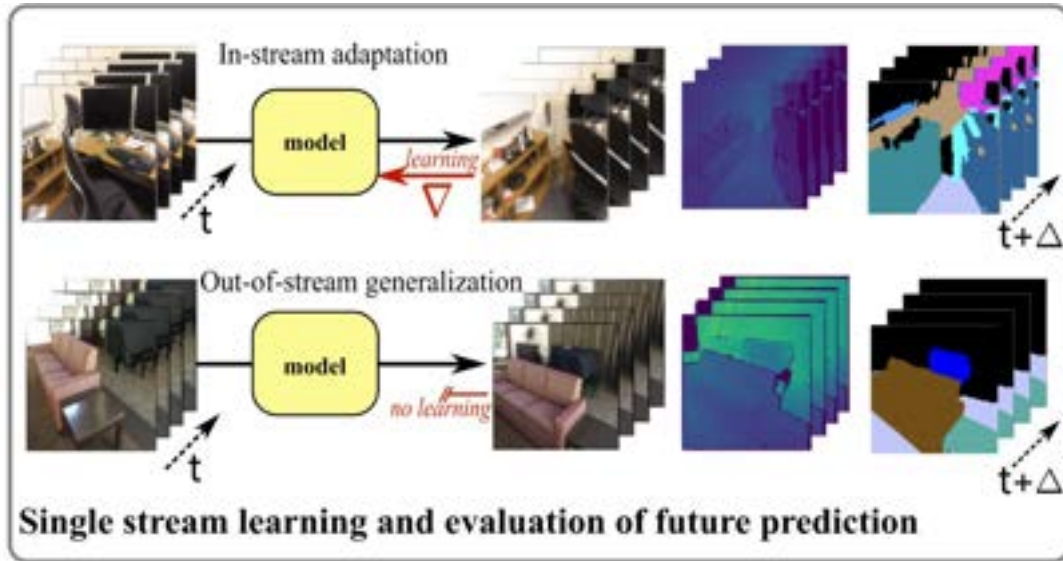


Figure 4. An example of video captions by PaLI-3 before and after video-specific adaptation. We show the keyframes on top for illustration purposes and the generated captions in the following blocks. Different details in text are highlighted. Best viewed in color.

## Learning from One Continuous Video Stream

- ❑ **目的:** 単一の連続ビデオストリームからのオンライン学習, 人間の実際の学習のようにaugmentationやミニバッチなどなしでの学習
- ❑ **提案:** 全てのタスクのターゲット出力をRGB空間にマッピングし, モデルや損失の切り替えなしにタスクを切り替え可能に独立同分布(IID)設定で学習された大規模なバッチを使用した未来予測事前学習タスクを導入



## MovieChat: From Dense Token to Sparse Memory for Long Video Understanding

- ❑ **目的:** 長期的な時間相関を保持しながら, 長時間の動画理解を行う
- ❑ **課題:** 長い時間の動画を扱う際の計算量, ビデオフレームを保存する際のメモリコスト, 長期的な時間相関の保持
- ❑ **提案:** スライディングウィンドウ方式で抽出した特徴量を順次保存し, 類似度の高い隣接特徴量は適宜マージすることでメモリ使用量の削減と長期的なビデオ特徴の表現を保持する

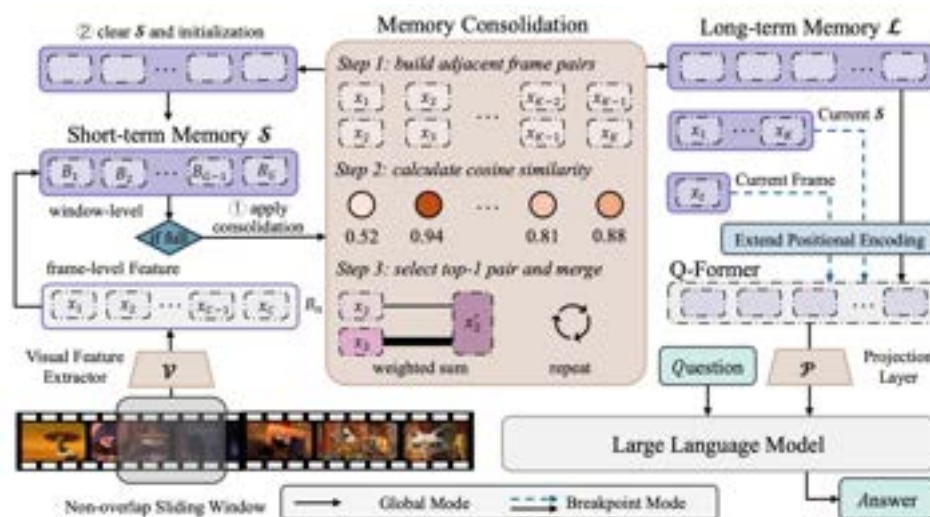


Figure 2. **Illustration of MovieChat.** MovieChat extracts video features with a sliding window and represents them in token form, which are then sequentially fed into the short-term memory frame by frame. When the fixed-length short-term memory reaches its preset limit, the earliest tokens are popped and consolidated into the long-term memory. MovieChat incorporates two distinct inference modes: the global mode, which exclusively utilizes the long-term memory, and the breakpoint mode, which additionally incorporates the current short-term memory as part of the video representation. The breakpoint mode allows for understanding the video at a specific moment in time. After passing through a projection layer, the video representation is inputted into a large language model for interaction with the user.

## MVBench: A Comprehensive Multi-modal Video Understanding Benchmark

- ❑ **目的:** 動的なビデオタスクにおける時間的理解を評価するベンチマークの作成
- ❑ **課題:** 既存のビデオベンチマークは基本的なビデオタスクや特定のドメインに限定されており, 手動アノテーションの負担も大きい
- ❑ **提案:** 新ベンチマークの設計とそのベンチマークで高性能を実現するVLLMの提案  
静的な画像タスクを動画タスクに変換(“Is the man on the stage?” → “What direction is the man moving?”), 自動アノテーション生成を行うことでアノテーションコストを削減

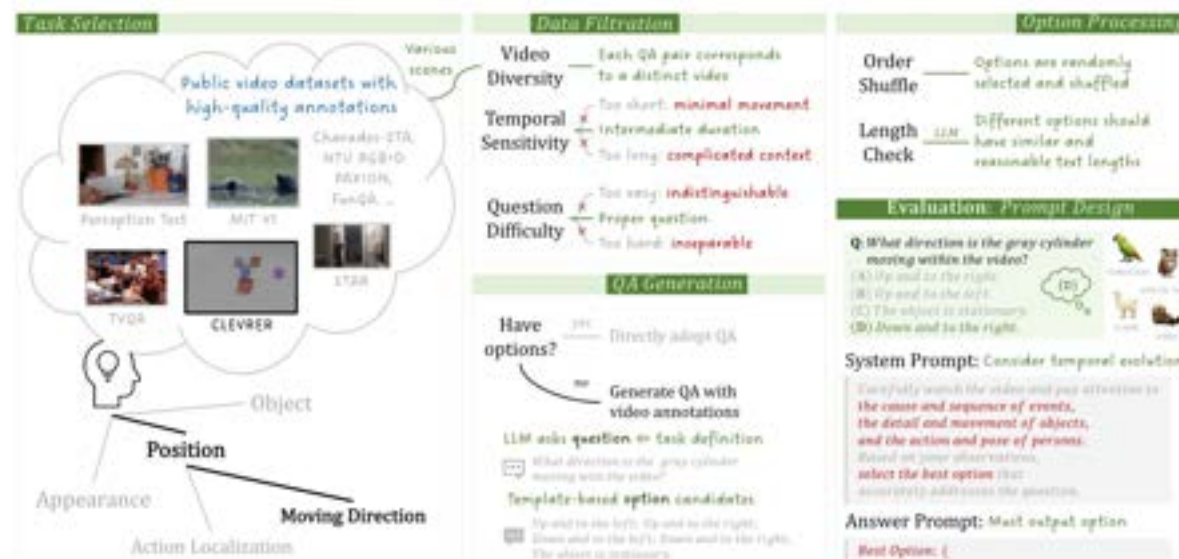
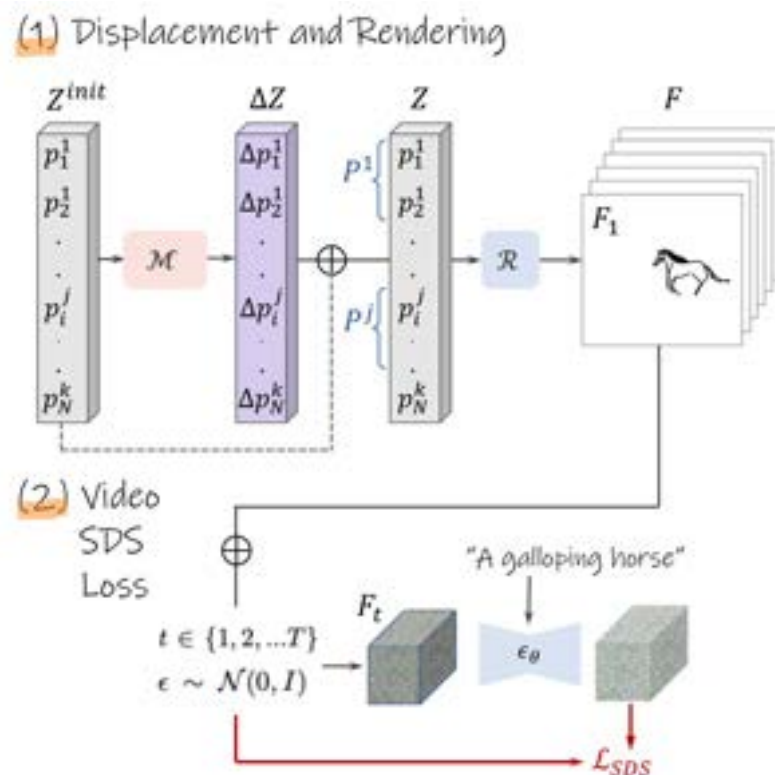


Figure 2. Generation pipeline of MVBench. Within public annotations, data is carefully filtered and relevant multiple-choice QAs are auto-generated. The effective system prompt and efficient answer prompt are employed to guide MLLMs toward precise outputs.

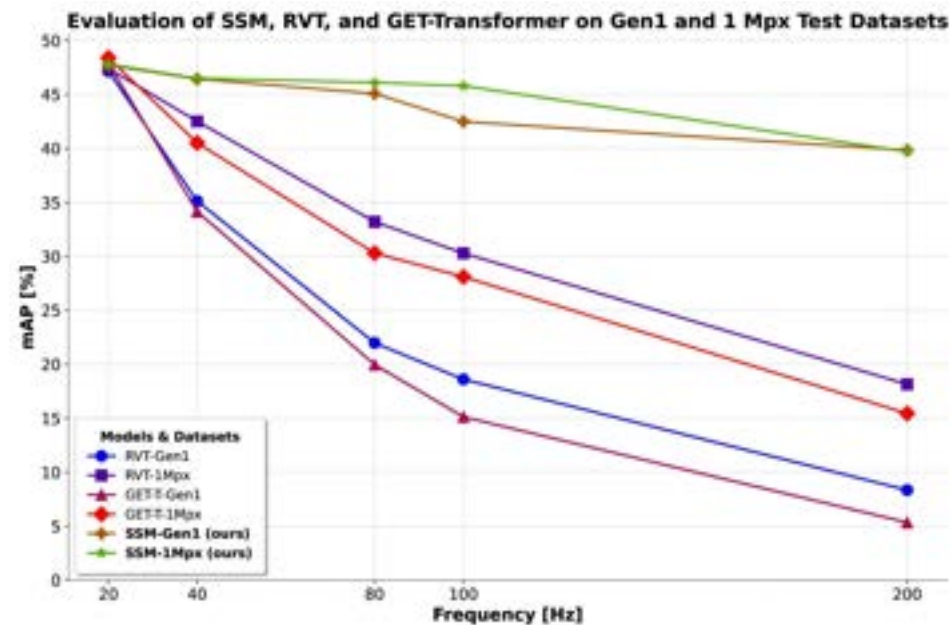
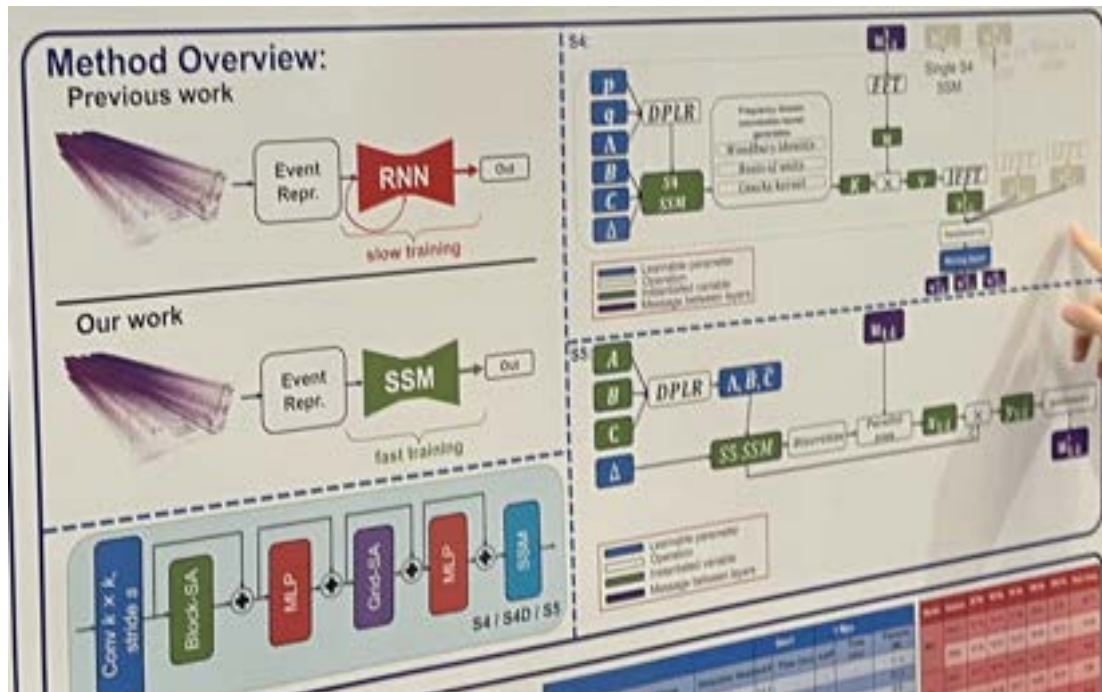
## Breathing Life Into Sketches Using Text-to-Video Priors

- ❑ **目的:** スケッチアニメーションの作成自動化
- ❑ **課題:** 参照動作やアノテーションを必要とした
- ❑ **提案:** 事前学習されたtext-to-videoのモデルを用いて, 動きの事前知識からアニメーション化の学習を進める(SDS lossを用いて 図中のMを学習)→追加のアノテーションや参照動作を必要としない



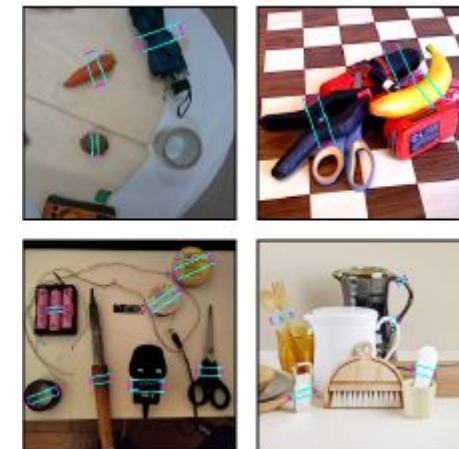
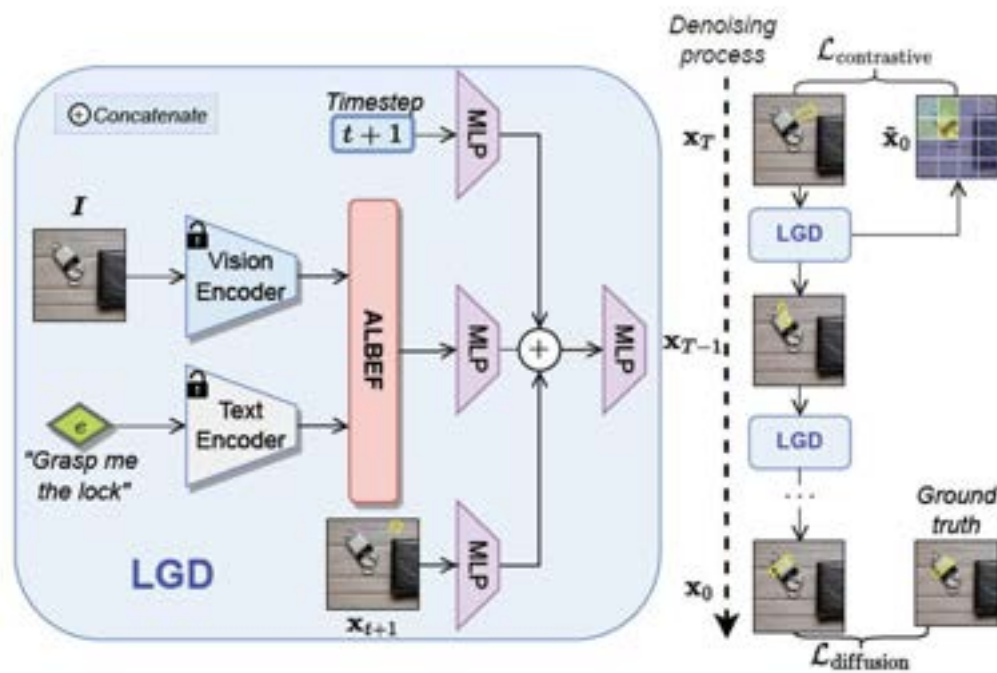
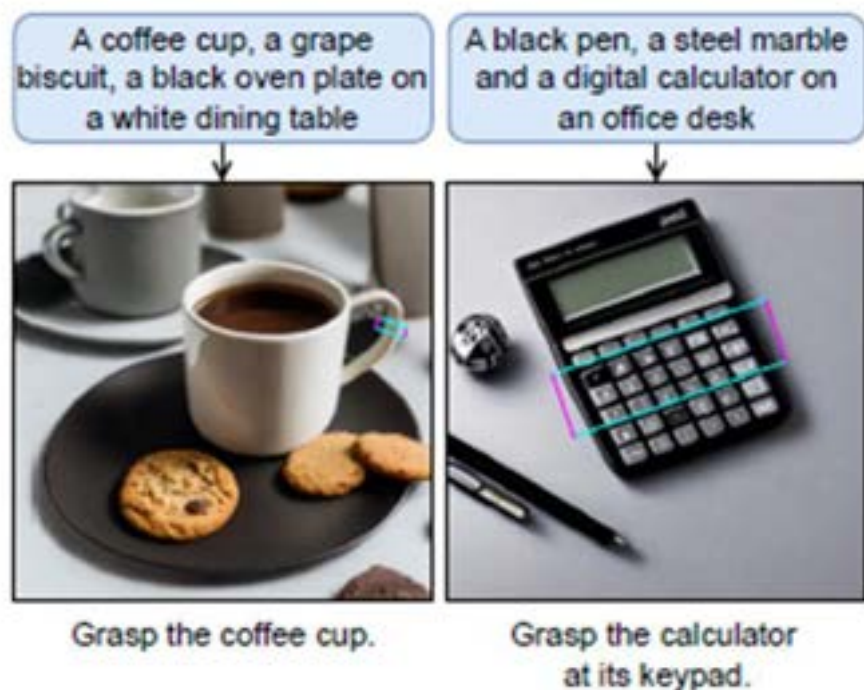
## State Space Models for Event Cameras

- ❑ 概要: イベントカメラを処理する既存のモデルはRNNやTransformerを用いるが, これらは周波数に関する汎化性能が低く, 学習速度も遅い.
- ❑ 新規性: SSM (状態空間モデル)を用いてイベントカメラを処理する手法を提案. SSMはタイムスケールに関する学習可能なパラメタを有するため, 学習時とは異なる周波数においても適切に汎化できると考えられる. 本研究では, SSMとしてS4およびS5を使用し, 周波数に関する高い汎化性能, ならびに学習過程における33%の高速化を達成した.



## Language-driven Grasp Detection

- ❑ **概要:** ①10M規模の把持データセット(GraspAnything++)を発表、②条件付き誘導拡散モデルに基づく言語ドリブンな把持検出。
- ❑ **手法:** 言語指示から正確に把持ポーズを検出するためのDenoisingによって学習。対照学習的な構造を持っていることが重要
- ❑ **感想:** GraspAnything (ICRA2024)では、単純な計算式によって粗くサンプリングした成果発表を行い、今回からは、言語に対応づけた検出モデルを構築している。物理的な評価がされていないが、プロジェクトページの充実から今後の展望だろう。従来型の把持データモデルに比べると、3Dモデルを利用していない点や言語を介したOpen-Vocabularyなデータセットである点で、ユーザビリティの高さが評価できる。トレンドを適切に把握している点も 🍌



## 動的なGaussian Splattingに関する研究を3つピックアップ

### □ 傾向

- 動的シーンの高品質な再構成とレンダリングへの応用が進んでいる
- 3Dガウシアンをベースとしつつ、ニューラルネットワークによる変形フィールドを組み合わせることで、効率的かつ高精度な表現を実現
- オンライン学習やストリーミングへの対応など、実用性を重視した手法の開発が進められている
- 単眼カメラからの入力への対応や、変形フィールドの効率化など、限られた入力からの高品質な再構成を目指す傾向がある

### □ 課題

- 大きな動きや複雑な形状変化のモデル化には、より柔軟で表現力の高い変形場が必要
- 初期フレームや点群の品質に依存するため、不完全な入力へのロバスト性を高める必要がある
- 遠景などの再構成が難しい領域での品質向上が求められる
- オンライン学習やストリーミングにおいて、限られた計算時間内で高品質な結果を得るためのさらなる効率化が必要
- 大規模な動的シーンや長時間の動画への対応には、メモリ効率やスケーラビリティの向上が求められる



## Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction

### 概要

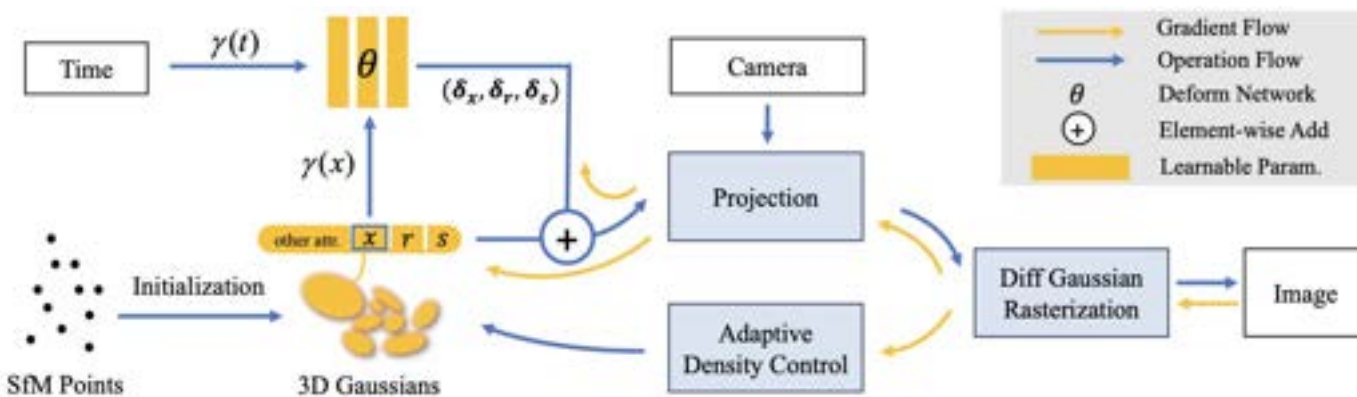
- 単眼動的シーンの高品質再構成・レンダリングのための変形可能な3Dガウシアンプラットフォームを提案
- 3DガウシアンとMLPベースの変形場を組み合わせることで、動的シーン再構成を実現
- Annealing Smooth Training機構により、実データセットの不正確なポーズによる時間方向のジッターを低減

### 新規性

- 従来の動的NeRF手法は暗示的(implicit)表現に依存し、細部の獲得が困難で、リアルタイムレンダリングも難しかった
- 点ベースレンダリングの3D-GSを動的シーンに拡張し、標準空間での3Dガウシアン学習とMLPによる変形場の推定を組み合わせることで高品質かつ高速なレンダリングを実現
- Annealing Smooth Trainingにより、計算量を増やさずに時間方向の滑らかさを担保しつつ動的な詳細も保持

### リミテーション

- 評価は中程度の動きを含むシーンが主体で、細かな表情変化などへの対応は今後の課題



## 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering

### 概要

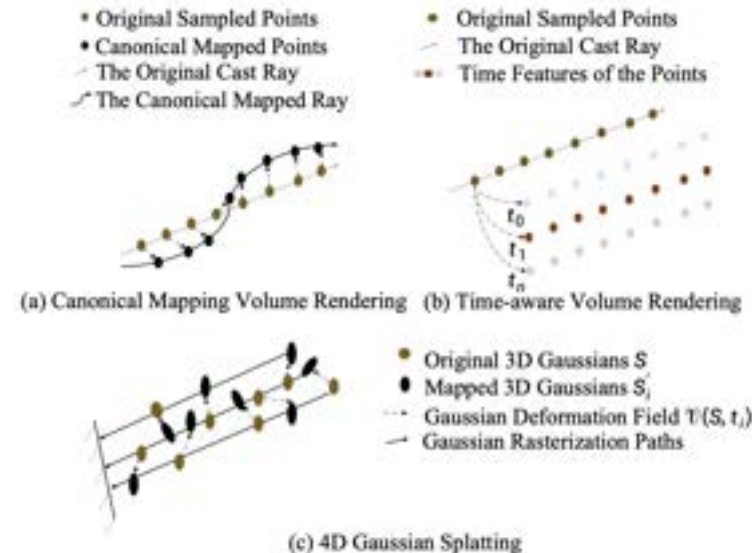
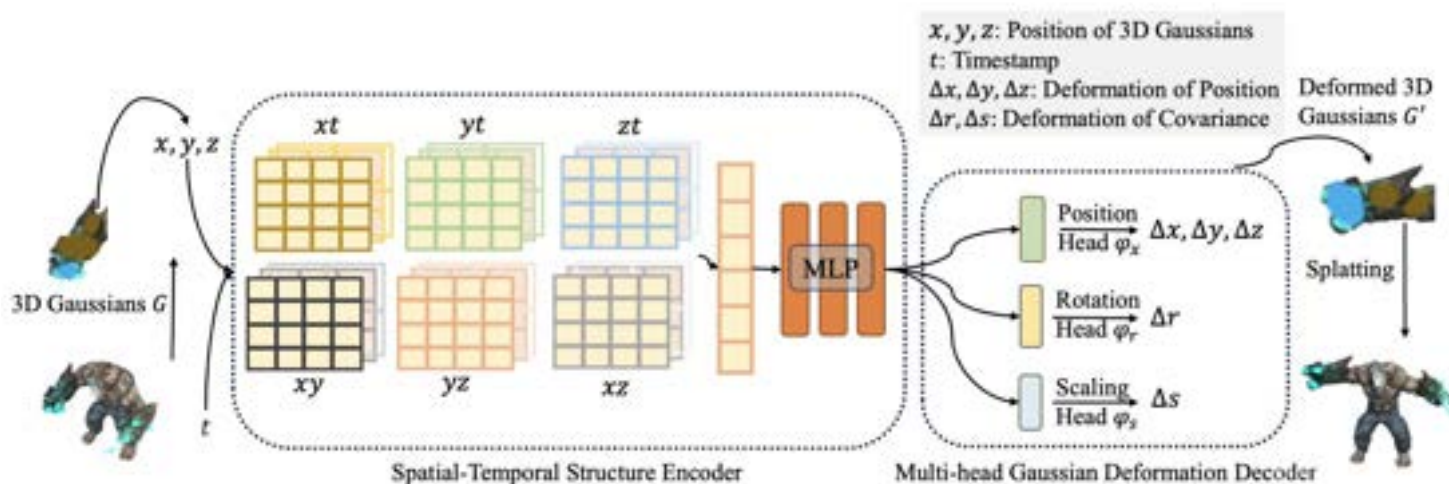
- 動的シーンのリアルタイムレンダリングを実現する4D Gaussian Splatting (4D-GS)手法を提案
- 3DガウシアンとMLPベースの4D変形場を組み合わせることで、動的シーン表現を実現
- Spatial-Temporal Structure Encoderにより、近接する3Dガウシアン間の変形を考慮した変形の学習が可能に

### 新規性

- 従来の動的NeRF手法は暗示的表現に依存し、レンダリング速度が遅い問題があった
- 3D-GSを4Dに拡張し、コンパクトな変形場を導入することで、学習とレンダリングの高速化を実現
- マルチ解像度のHexPlane表現により、3DGS間の空間的・時間的関係をエンコード

### リミテーション

- 単眼動的シーンでは入力データが疎で、複雑なシーンで過学習が生じやすい
- 大きな動きを含む多視点データセットでは、短時間での収束が難しい
- 単眼設定での大きな動きや劇的なシーン変化のモデリングは困難な場合がある



## 3DStream: On-the-Fly Training of 3D Gaussians for Efficient Streaming of Photo-Realistic Free-Viewpoint Videos

### □ 概要

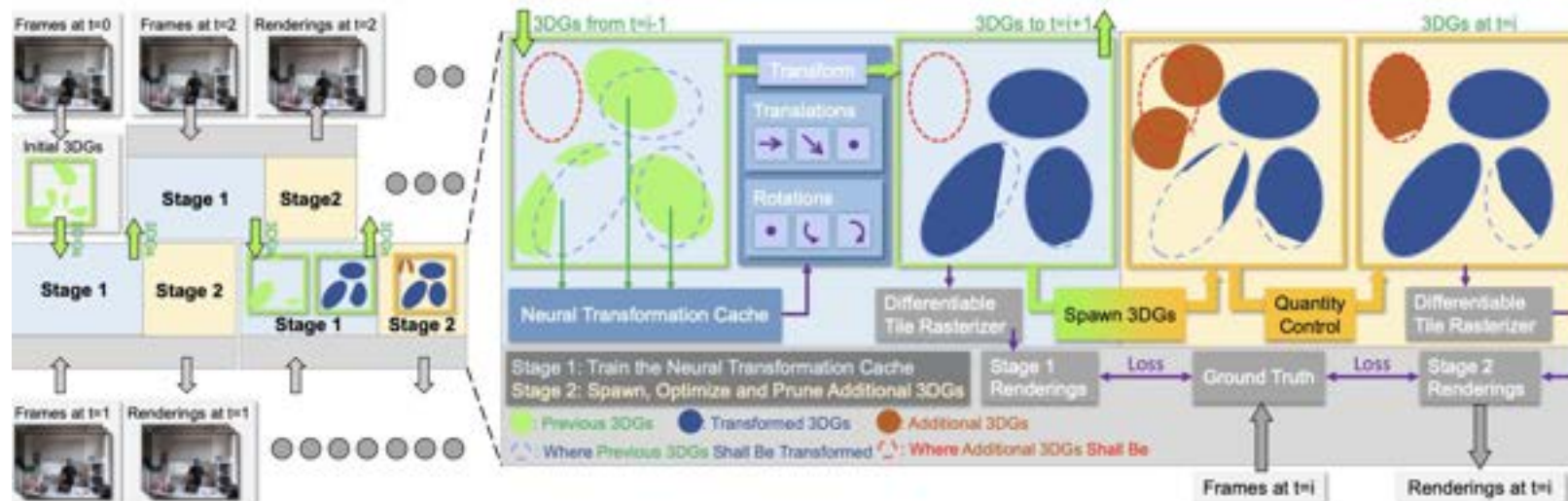
- 多視点動画からの自由視点動画ストリーミングのための、オンザフライな3D Gaussian Splatting (3DStream)手法を提案
- Neural Transformation Cache (NTC) で3DGSの変換をコンパクトにモデル化することで、高速な逐次再構成とリアルタイムレンダリングを実現
- Adaptive 3DG Additionにより、動的シーンの新出オブジェクトにも対応

### □ 新規性

- 完全な動画シーケンスを必要とするオフライン学習ではなく、動画ストリームに対するオンザフライ学習を実現
- NTCにより3DGSの変換を効率的にモデル化し、フレーム毎の追加の3DGSと合わせて用いることで、限られたストレージで動的シーンの変化に対応
- 既存のFVV構築手法に比べ、モデルサイズと画質を維持しつつ、学習速度とレンダリング速度で優位性

### □ リミテーション

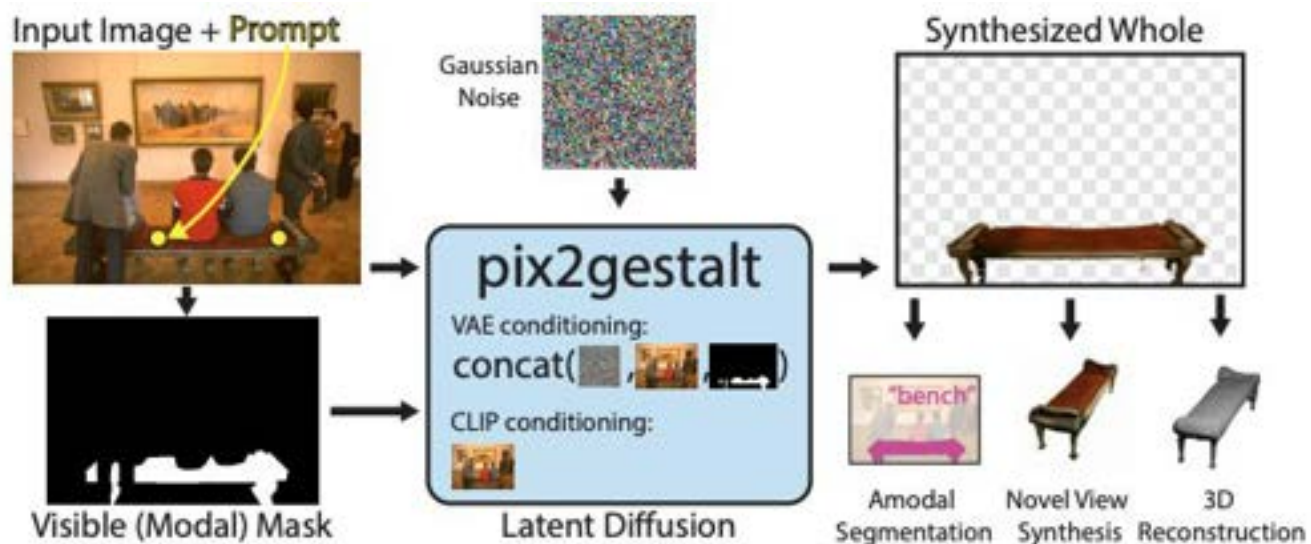
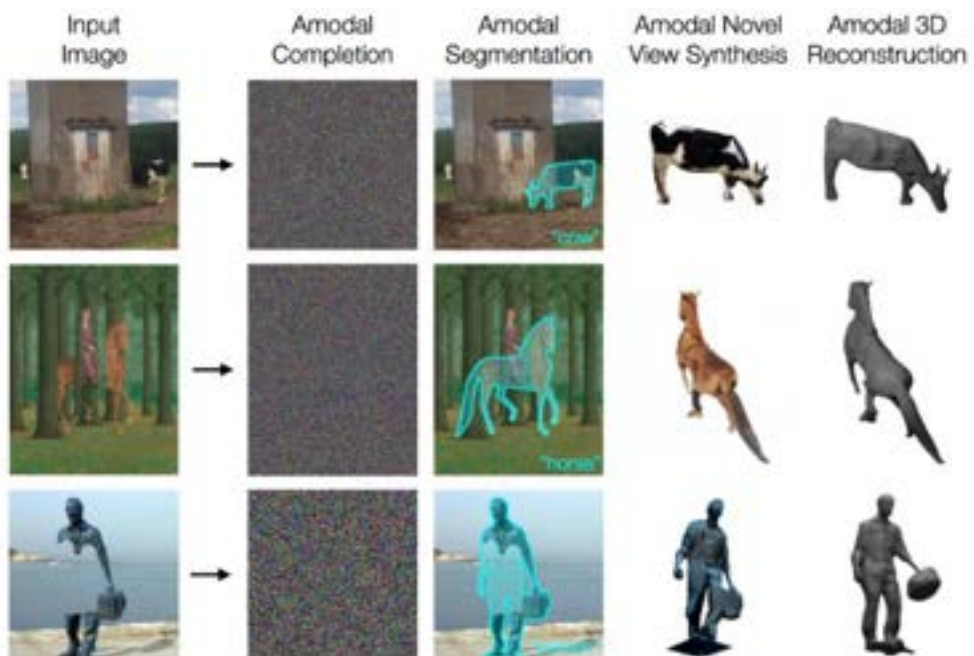
- 初期フレームでの3DGSの品質に大きく依存。遠景などCOLMAPでの再構成が難しい領域では品質低下の可能性
- オンザフライ学習のため学習反復数が限られ、Stage 1での大きな動きやStage 2での複雑な新出オブジェクトのモデル化が制約される(以下の図を参照)



<https://arxiv.org/pdf/2403.01444>

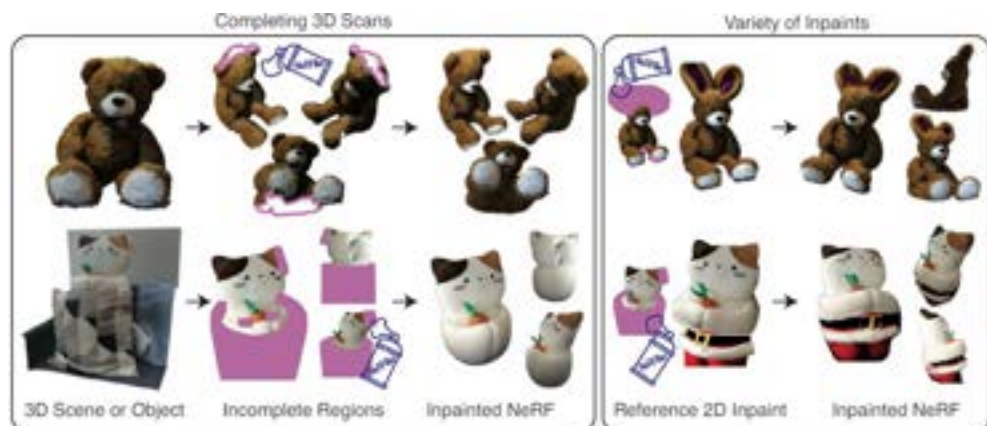
## pix2gestalt: Amodal Segmentation by Synthesizing Wholes

- ❑ **概要:** Stable Diffusionを用いて部分的な観測から物体の形と外観(新しい視点画像生成)を生成することで、Zero-shot でamodal segmentation手法を提案
- ❑ **新規性:** 大規模データで学習済みStable Diffusionは、すでにImplicitlyでamodal representationを学習しているという仮説のもと、Zero-shot で様々な遮蔽を含むAmodal Segmentationに対応可能なフレームワークを提案。
- ❑ **手法:** 合成データセットで学習。条件付き拡散モデルを使用し、部分的なRGB画像とポイントのプロンプトから物体の全体像を生成する。
- ❑ **感想:** 遮蔽を含む画像から物体の新しい視点を生成することが素晴らしい。著者欄から見ると、おそらくZero-1-to-3 [ICCV2023]の発展論文です。(ただし、遮蔽は含まれていません)

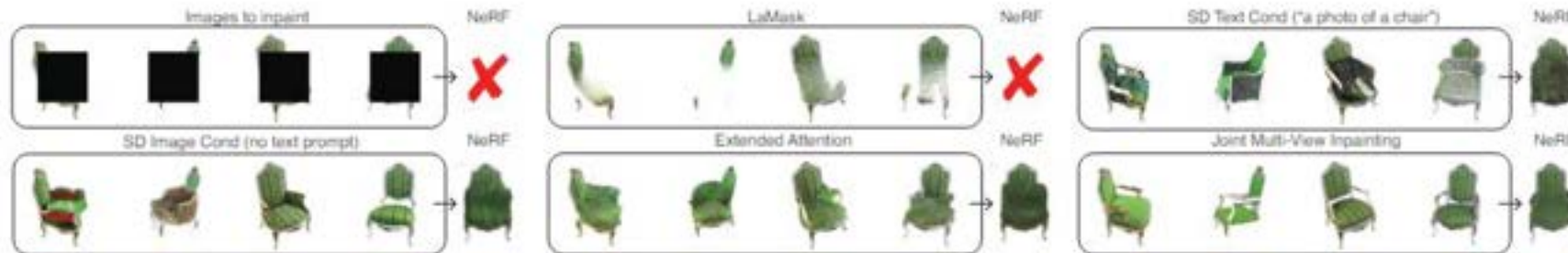


## NeRFiller: Completing Scenes via Generative 3D Inpainting

- ❑ **目的:** 3Dインペインティングによりセンサにより収集された3Dデータの欠損を補完
- ❑ **課題:** センサより取得された3Dデータには欠損が多く存在
- ❑ **提案:** 2Dインペインティング拡散モデルを利用することで3Dインペインティングを実現
- ❑ **新規性:** 4つの画像を2×2のグリッドにタイリングすることで、独立に塗りつぶすよりも一貫性のある塗りつぶしができるという、2Dインペインティング拡散モデルのユニークな特性を発見

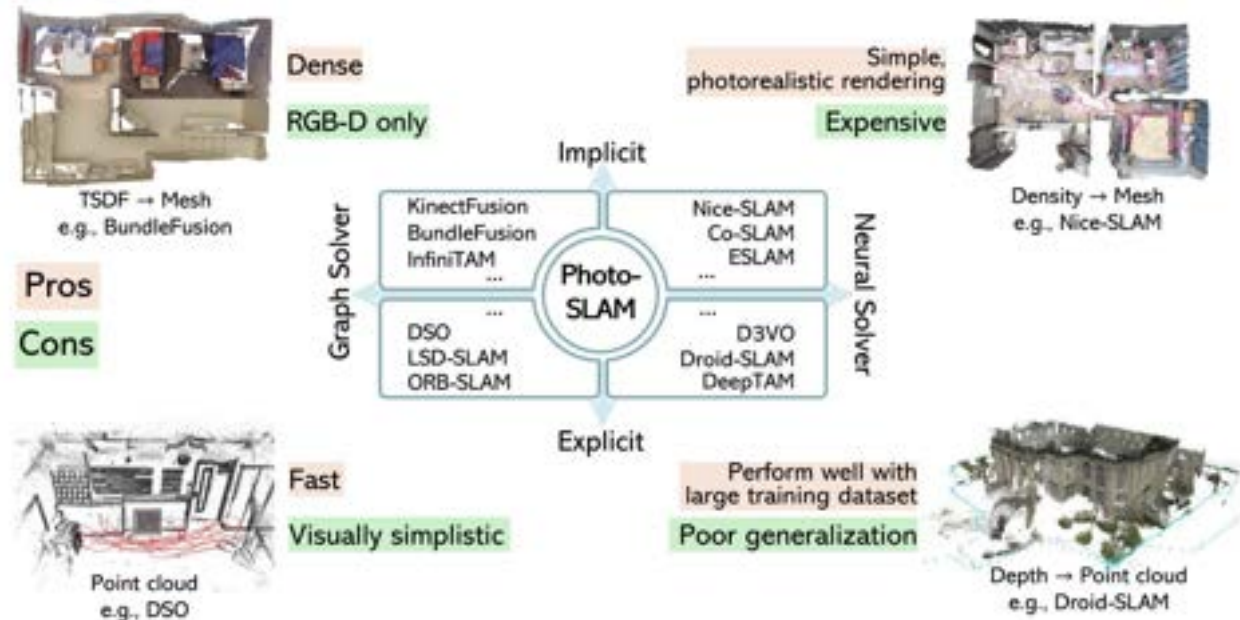


	PSNR ↑	SSIM ↑	LPIPS ↓	MUSIQ ↑	Corrs ↑
Masked NeRF	14.71	0.78	0.26	3.71	675
LaMask	27.39	0.90	0.05	3.76	643
SD Image Cond	22.03	0.86	0.11	3.68	665
Inpaint + DU	26.60	0.89	0.08	3.76	660
Ours w/o depth	28.41	0.92	0.06	3.72	682
Ours	28.28	0.91	0.06	3.73	696



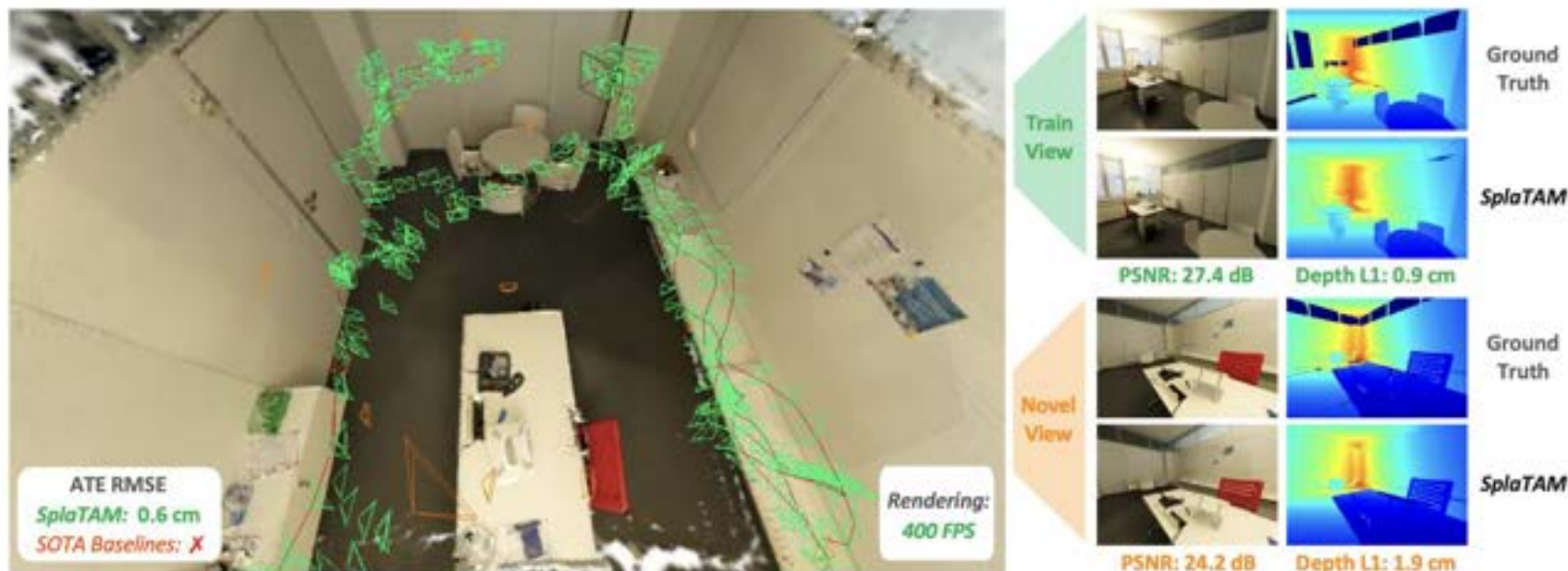
## Photo-SLAM: Real-time Simultaneous Localization and Photorealistic Mapping for Monocular, Stereo, and RGB-D Cameras

- ❑ **概要:** 幾何学的特徴に基づくハイパープリミティブを用いたSLAMを提案
- ❑ **手法:** 取り扱う解像度のピラミッド構造を構築し、明示的に得られた点群からそれぞれ高密度化したハイパープリミティブを生成する。ガウシアンピラミッドベースの学習により各レベルを漸進的に学習することで、フォトリアリスティックなマッピング性能を向上。
- ❑ **新規性:** 従来のNeRFベースのSLAMは暗黙的な表現に依存するためリソースの消費が莫大。提案手法ではハイパープリミティブを介して明示的な幾何学特徴による軽量な地図と高品質なテクスチャを含む地図を接続した



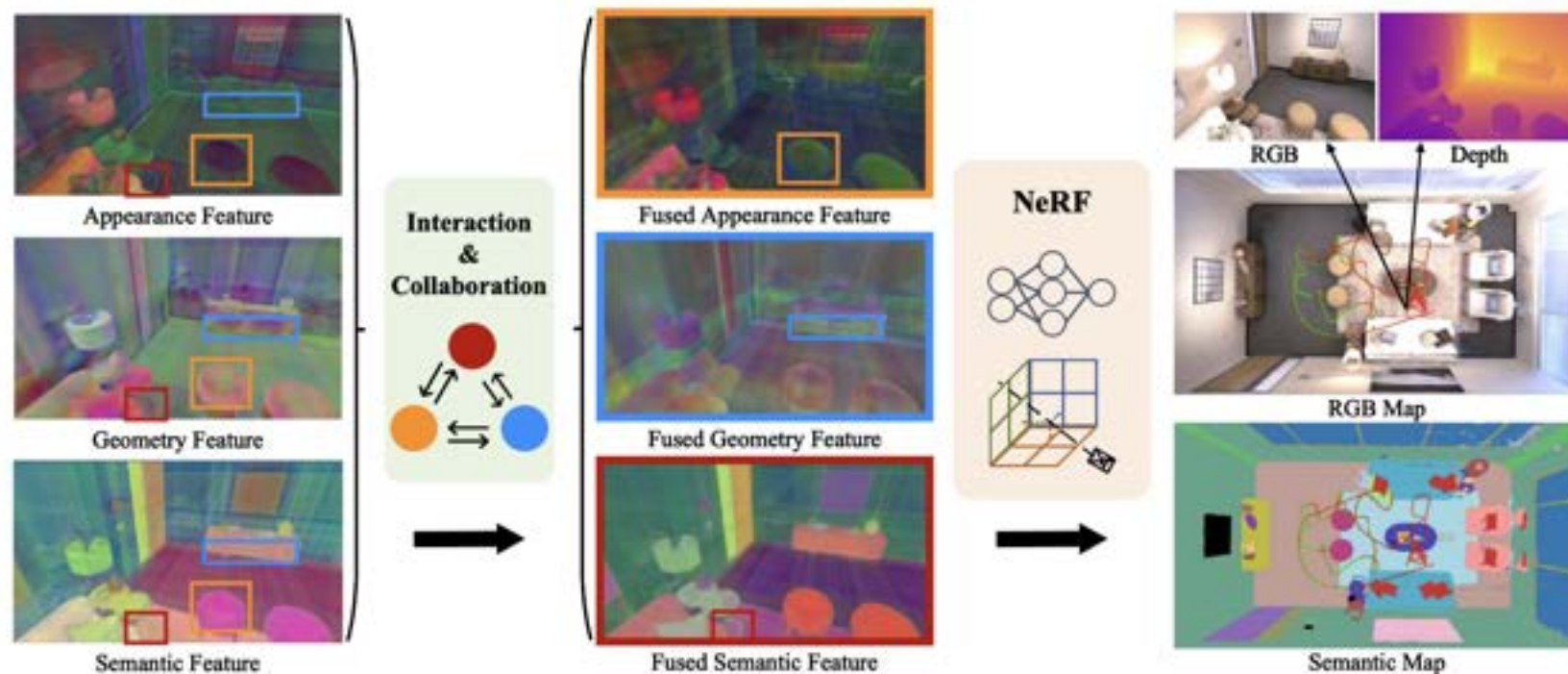
## SplaTAM: Splat, Track & Map 3D Gaussians for Dense RGB-D SLAM

- ❑ **概要:** 3Dガウシアンを活用し、単一のポーズなしRGB-Dカメラからの高忠実度の再構成を実現
- ❑ **手法:** ガウス表現に合わせたシンプルなオンライントラッキングとマッピングシステムを導入
- ❑ **新規性:** 明示的な3Dガウス表現を利用することで、高速なレンダリングと高密度な最適化、以前にマッピングされた領域かどうかの迅速な判定、より多くのガウシアンを追加することによる構造化されたマップ拡張



## SNI-SLAM: Semantic Neural Implicit SLAM

- **概要:** NeRFに基づく高密度RGB-DセマンティックSLAMシステムを提案
- **手法:** 多レベルの意味理解を可能にする階層的意味表現を導入し、シーンのトップダウン構造化セマンティックを記述。クロスアテンションを通して、外観、形状、意味的特徴を統合する。多階層の特徴量から意味的、RGB、TSDF値をデコーダで融合する
- **新規性:** クロスアテンションに基づいて、幾何学的特徴、外観特徴、意味的特徴を統合することで、異なる特徴間の相互強化を実現。また、有向な相関を扱う新しいデコーダを導入することで、単一の属性に欠陥がある場合でも頑健性を保つことができる





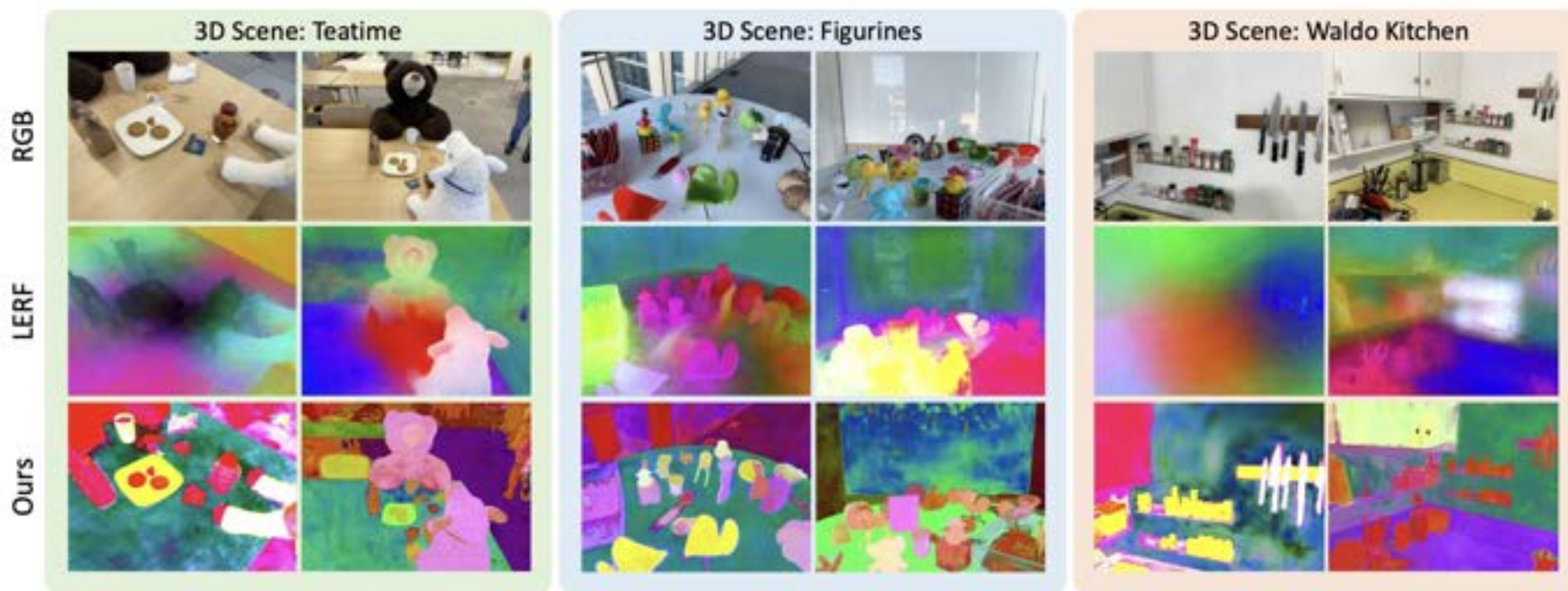
## GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models

- ❑ **概要:** Gaussian Splattingを介して2Dと3Dの拡散モデルの橋渡しを行うText-to-3D手法を提案案
- ❑ **手法:** 3D拡散モデルを用いてプロンプトから3D点群を初期化し、2D拡散モデルを利用して高精細化する
- ❑ **新規性:** 2D拡散モデルは3Dの一貫性に、3D拡散モデルは利用可能なデータ量の制限に課題があった。本手法ではGaussianSplattingを介して2つのモデルを接続することでメリットを両立



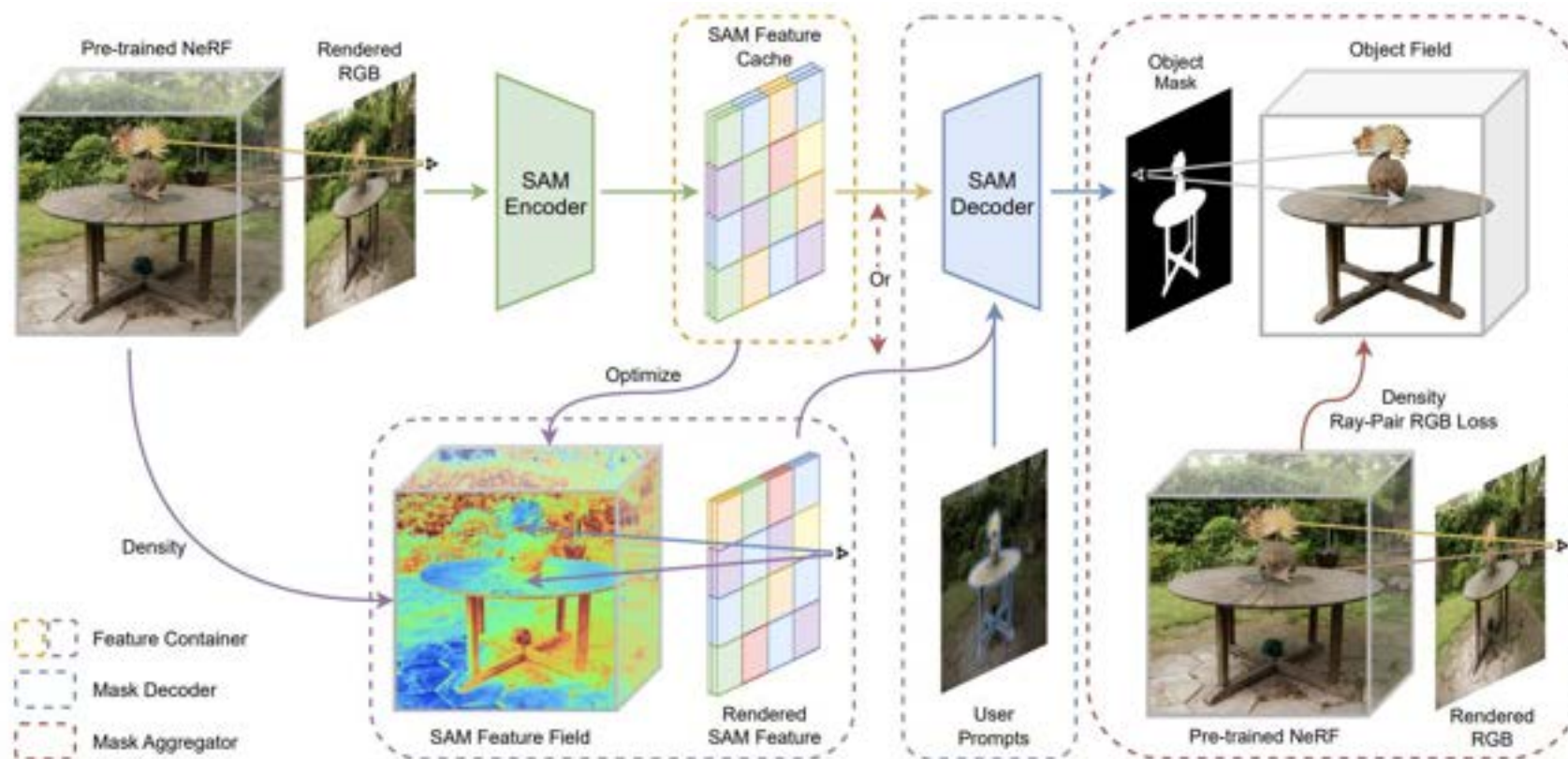
## LangSplat: 3D Language Gaussian Splatting

- ❑ **概要:** 3D空間内で正確かつ効率的なopen-vocabulary queryingを可能にする3D言語場を構築
- ❑ **手法:** シーンごとの言語オートエンコーダを学習し、次にシーン固有の潜在空間上でCLIP言語特徴を学習することで、明示的なモデリングによって課されるメモリの大幅な要求を軽減
- ❑ **新規性:** 従来、3D空間場でオブジェクトの境界を明確にするのは困難であった。SAMを用いて階層的セマンティクスを学習することで、点の曖昧性の問題を解決



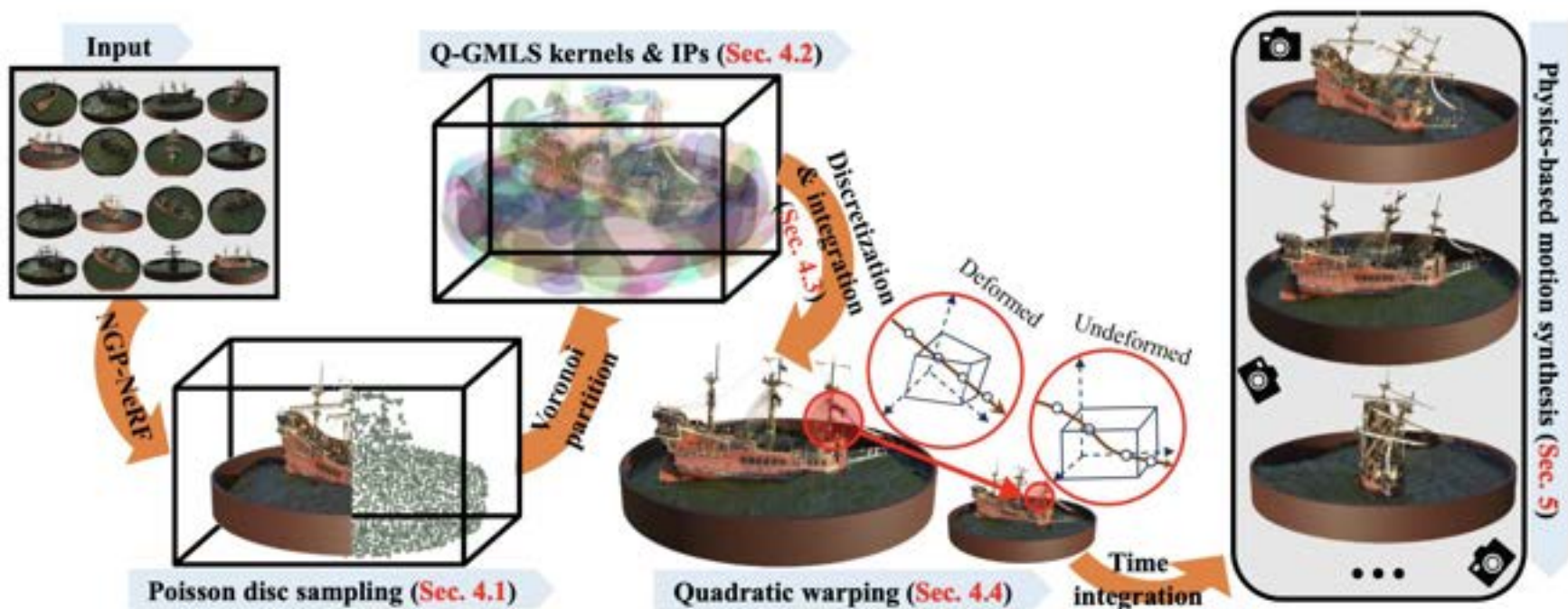
## SANeRF-HQ: Segment Anything for NeRF in High Quality

- ❑ **概要:** 与えられたシーン内のあらゆるターゲットオブジェクトの高品質な3Dセグメンテーションを実現
- ❑ **手法:** SAMを利用してオープンワールドのオブジェクトセグメンテーションを行い、NeRFを利用して異なる視点からの情報を集約
- ❑ **新規性:** 密度場とRGB類似度を採用してセグメンテーション境界の精度を向上させることで、従来難しかった複雑なシナリオでも物体を正確かつ一貫してセグメンテーション可能



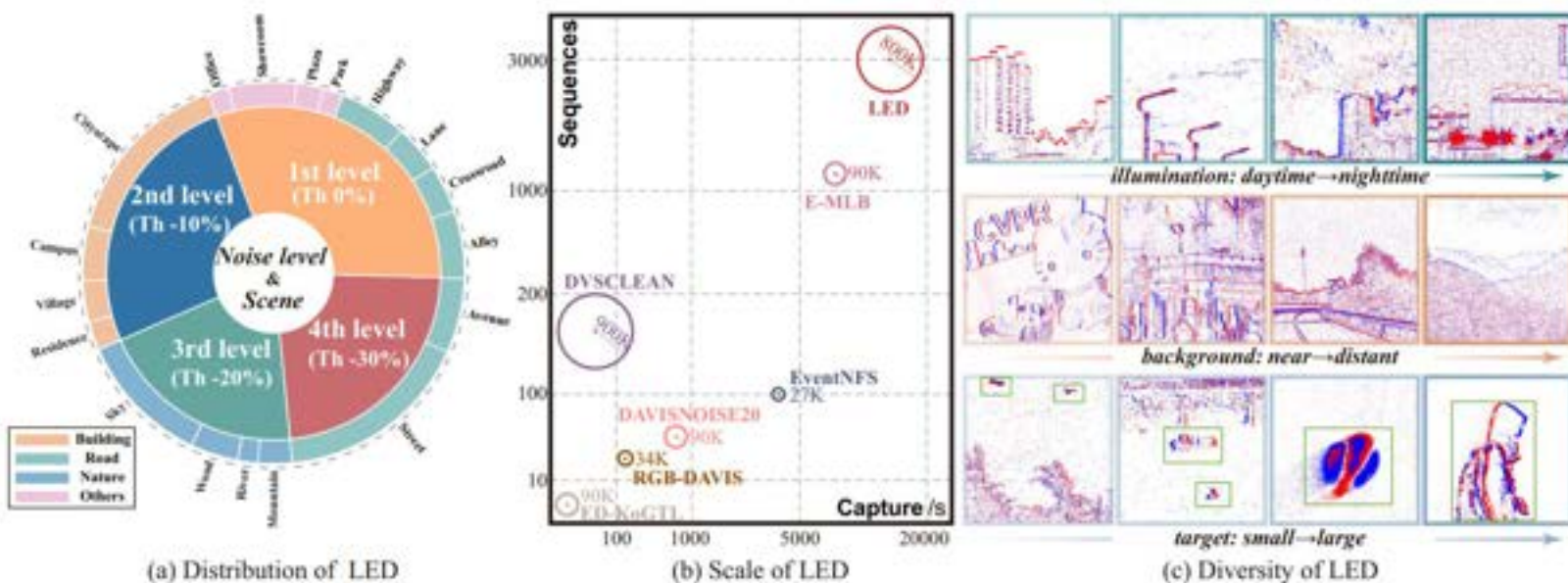
## PIE-NeRF: Physics-based Interactive Elastodynamics with NeRF

- ❑ **概要:** 物理ベースのシミュレーションをNeRFとシームレスに統合し、実世界の物体の高品質な弾性力学変形を生成
- ❑ **手法:** 適応的なポアソンディスクサンプリングにより、学習済みのNeRF(実装場はiNGP)を照会し、 $n$ 個のQ-GMLSカーネルにスパース化する。積分点はQ-GMLSカーネルの中心を含むモデル上に配置。カーネルでの離散化とIPでの数値積分により、物理学に基づいた運動を合成
- ❑ **新規性:** メッシュレスに変形を離散化するため、四面体メッシュやボクセルグリッドのような中間的な補助形状プロキシの必要性を排除できる



## LED: A Large-scale Real-world Paired Dataset for Event Camera Denoising

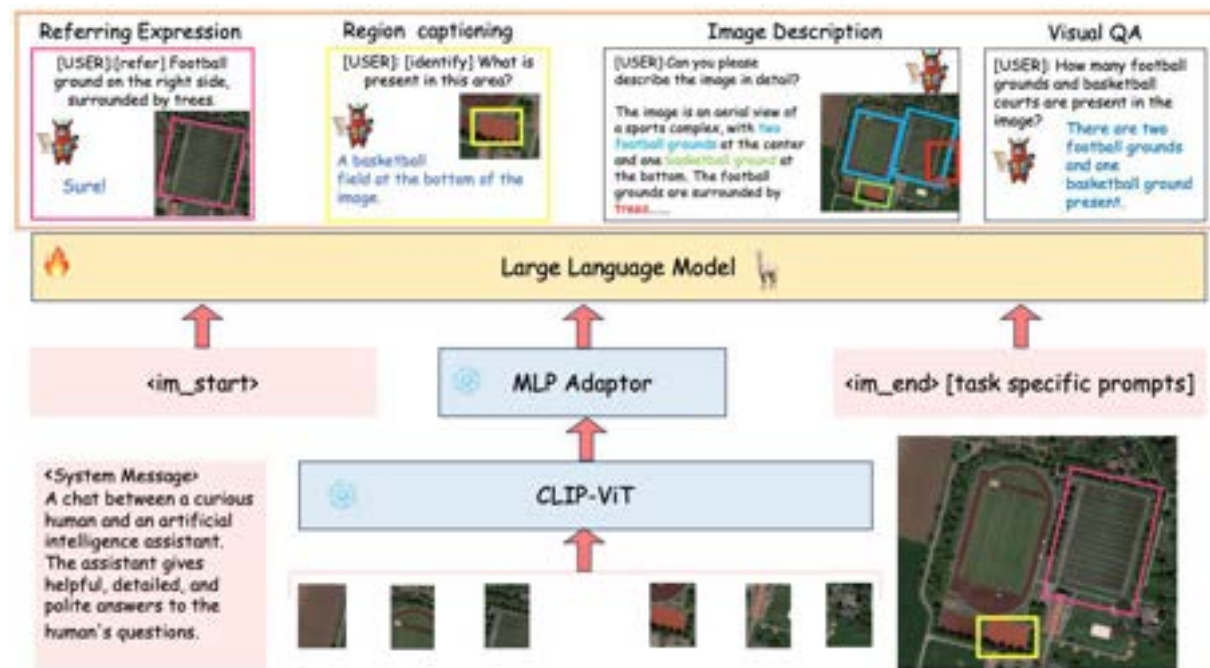
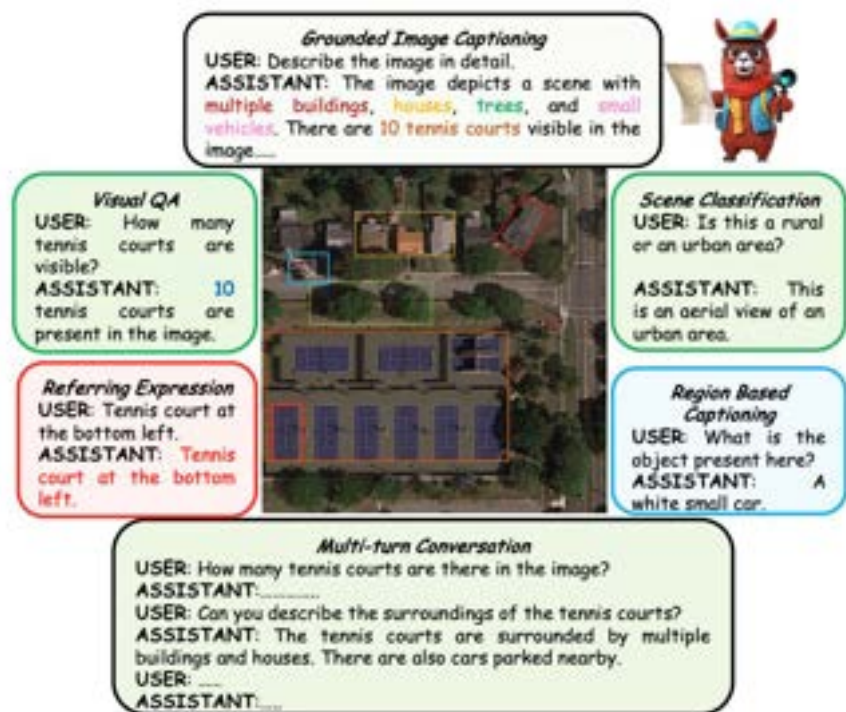
- イベントカメラはノイズの干渉を受けやすい一方で理想的な環境下で研究, 検証される
- 本研究ではより実世界に則した, イベントデノイジングデータセット(LED)を提案
- 同時にノイズ除去フレームワークとしてDEDを提案
- 新しいペアデータのベースラインとなる



<https://github.com/Yee-Sing/led>

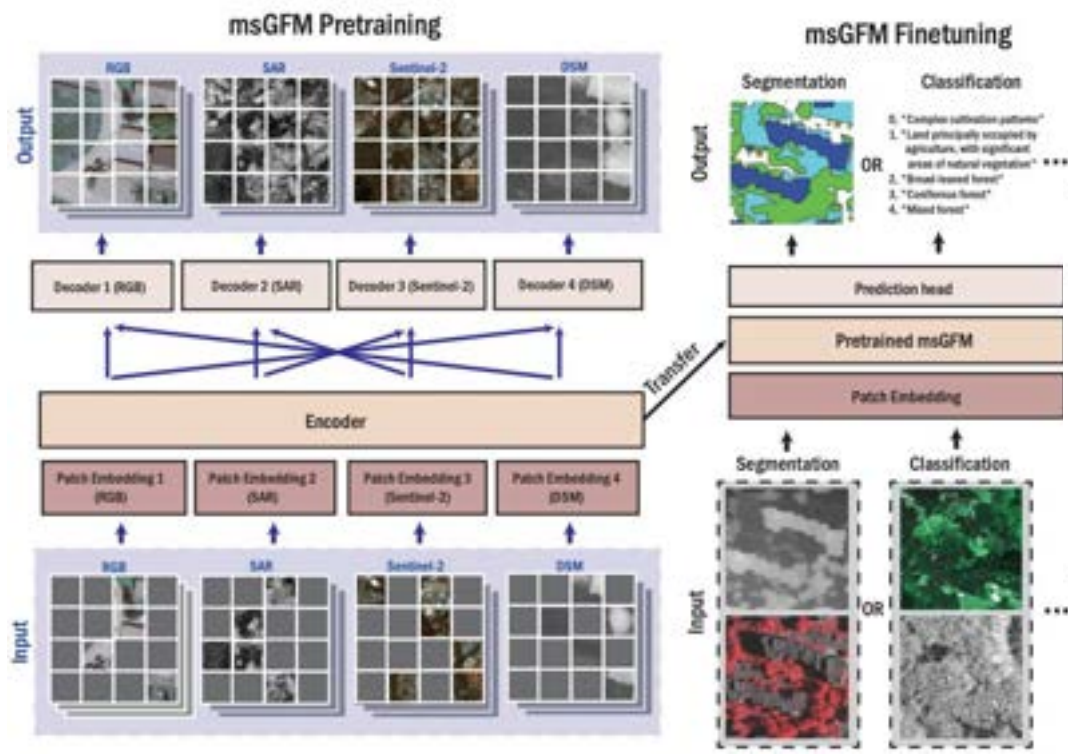
## GeoChat : Grounded Large Vision-Language Model for Remote Sensing (RS)

- ❑ **概要:** 一般的な大規模視覚言語モデル(VLM)では高解像度を扱う際にRS画像特有のドメインを理解できなかつたり、大域的な情報を考慮することができない。そこで、RS画像ドメインに特化したVLMデータセットおよびモデルの構築方法を提案
- ❑ **手法:** 画像のエンコードにはCLIP-ViT(Vision Transformer)を使用、MLP Adaptorを通して言語特徴量に変化しLLM(Vicuna-v1.5)をベースにRSタスク向けにLoRAで調整。タスク固有のトークンを使用している。
- ❑ **感想:** RSではRemote CLIPが提案されているため視覚的エンコーダーを単純に差し替えるだけでも性能は向上しそう。今回はCVタスクを用いており、単純に雲除去などのRS特有のタスクを加えてみても面白そう。また、どの程度季節性が考慮されているか分からなく、(雪が積もった時みたいな)ロバストが気になる。取り組み自体は面白い。



## Bridging Remote Sensors with Multisensor Geospatial Foundation Models

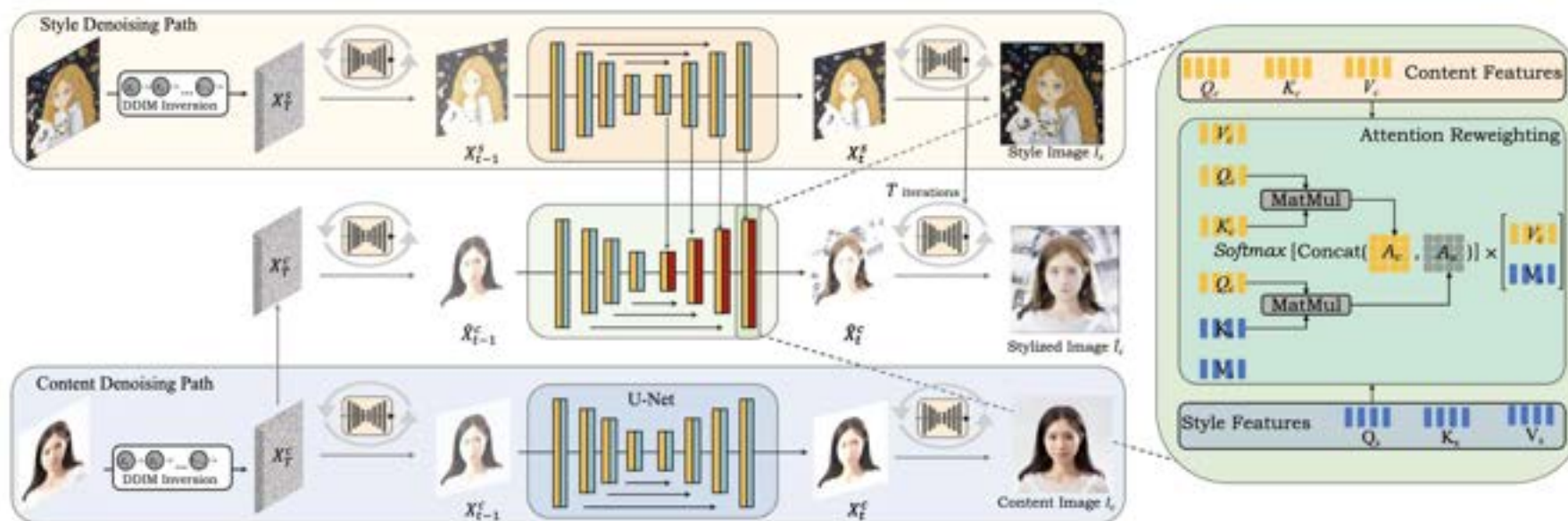
- 概要:** 複数のモダリティを活用できるマルチセンサー地理空間事前学習モデルを開発した。また、本論文ではマルチセンサーを教師なしで学習したモデルが対応するセンサー間の表現をどのように表現できるのか？マルチセンサーから取得したデータを事前学習することで下流タスクで性能が向上するのかを検証。
- 手法:** 各センサごとに異なる埋め込みレイヤーを用意し、マスキングされたデータをEncoderに入力、マスクの予測は複数のセンサーを使ってマスクの値を予測する事前学習を行い、そのEncoderを解きたいタスクへ再学習を行う。
- 感想:** 衛星データは位置情報が保有しているので異なるセンサーから取得された同じ位置のデータの取得が比較的容易で今回ラベル無しなので、Updateされたデータセットが1年以内に提案されそう。と思ったら比較手法のGFMの著者だった。



Methods	10% BEN mAP (↑)	100% BEN mAP (↑)	SEN12MS-CR			SpaceNet		Vaihingen mIOU (↑)
			MAE (↓)	SAM (↓)	SSIM (↑)	PSNR (↑)	SSIM (↑)	
SeCo [38]	82.6	87.8	-	-	-	-	-	68.9
SatMAE [15]	82.1	-	-	-	-	22.742	0.621	70.6
MoCoV2 [8]	-	89.3	-	-	-	-	-	-
DINO-MC [62]	84.2	88.6	-	-	-	-	-	-
GFM [39]	86.3	-	-	-	-	22.599	0.638	75.2
Random	82.6	86.2	0.048	14.78	0.572	21.825	0.594	67.0
IN-22k [36]	85.7	89.5	-	-	-	21.655	0.612	74.7
msGFM	<b>87.5</b>	<b>92.9</b>	<b>0.026</b>	<b>4.87</b>	<b>0.842</b>	<b>22.850</b>	<b>0.668</b>	<b>75.8</b>

## Z\*: Zero-shot Style Transfer via Attention Reweighting

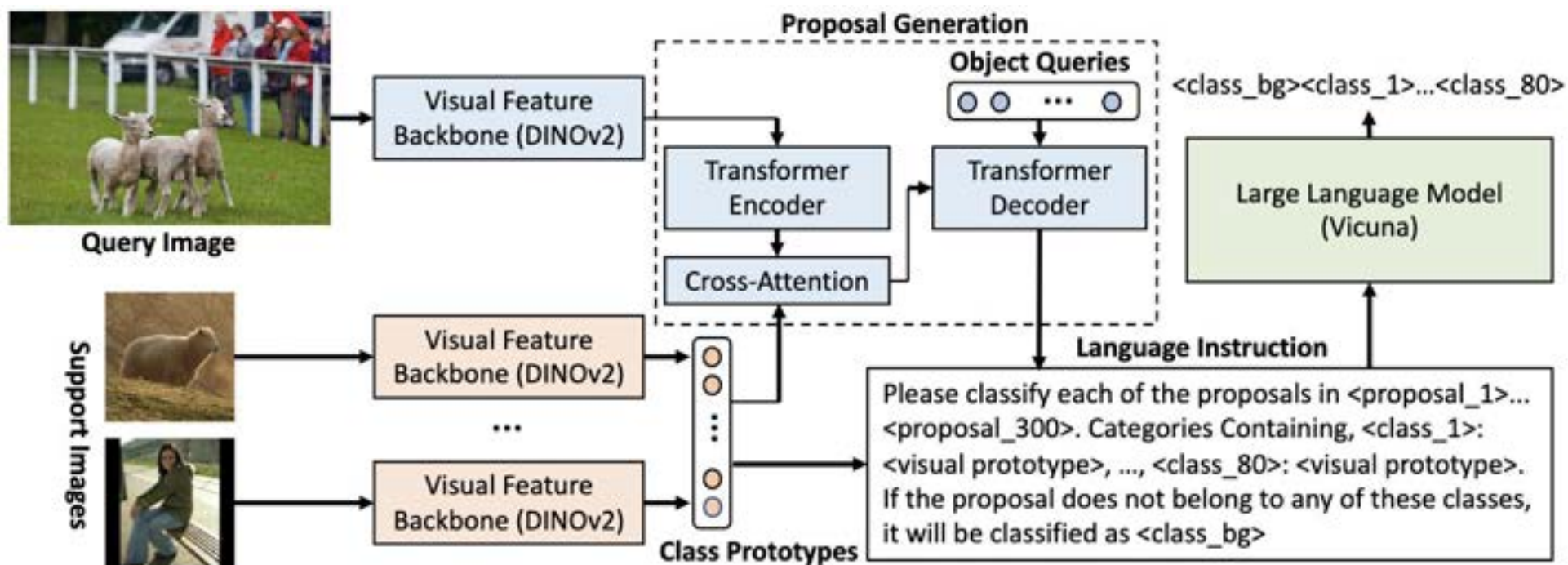
- 概要:** Diffusion Modelを用いたZero-Shot Style Transferを提案。これまで取り組まれてきた、Text-to-ImageのDiffusion Modelを用いたスタイル変換ではなく従来通りContent画像及びStyle画像を使用したStyle変換を目指したもの。
- 手法:** タイムステップを通してプロンプト(Stable Diffusionでいうところのテキスト)の埋め込みが変化しないことに対処するため、DDIM Inversionを用いてContent画像とStyle画像の各タイムステップの特徴量を取得するDual-Path schemeを提案。Attention Reweighting内ではDual-Path schemeで得たContent特徴量をクエリとすることで、Content画像の構造を保ったままStyle特徴量からStyle情報を取り入れる。
- 感想:** 全体としてはよくStyle変換できている。しかし、人の絵画をStyle画像及び人の画像をContent画像にした時、Style画像のContentに誘導されて髪型が大きく変化している場合がある。よって、Style画像からContent特徴を抑制する方法を提案できると更によくなりそう。





## Few-Shot Object Detection with Foundation Models

- ❑ **概要:** 基盤モデルをFew-Shot Object Detectionに活用した研究。特徴量の抽出と検出されたオブジェクトの分類に基盤モデルを使用することで従来よりも高い精度を達成した。
- ❑ **手法:** Query画像とSupport画像を重みを固定した同一のDINOv2に入力。Deformable DETRのEncoderに入力されたQuery画像の特徴量とSupport画像の特徴量をCross-Attentionに入力。その後、Decoderに入力しBounding Boxを得る(Object Queriesはランダムに初期化されたシーケンス)。また、LLMのin-context learningを用いてBounding Box内のObjectのクラスを分類する。
- ❑ **感想:** DINOv2の重みは固定な点とDINOv2のモデルサイズが小さくてもそこそこのパフォーマンスが出るの点が魅力的。

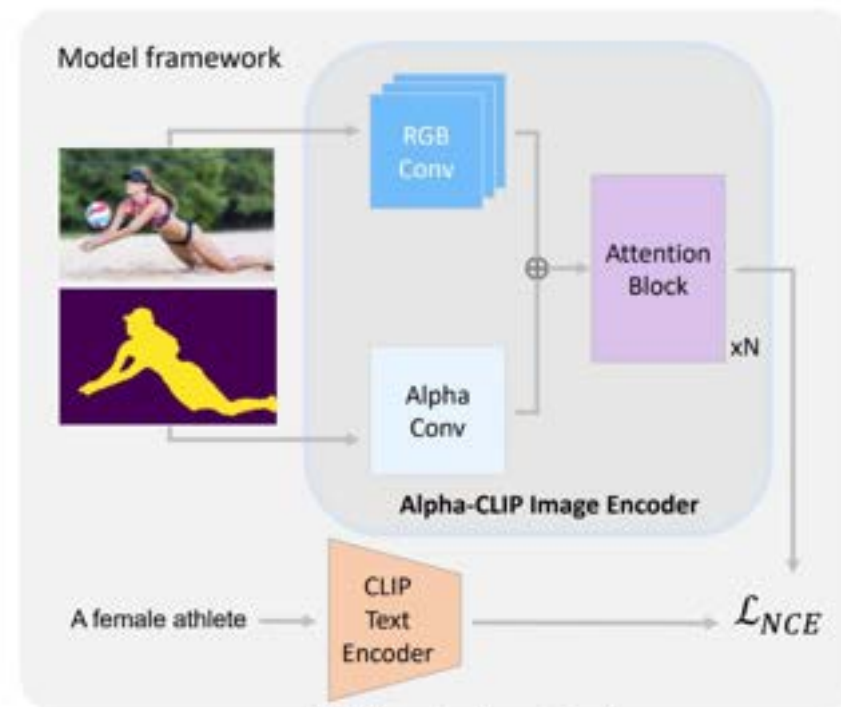


## Alpha-CLIP: A CLIP Model Focusing on Wherever You Want

- **概要:** CLIPは、特定の領域に焦点を当てるのが難しい。画像の特定領域に焦点を当てるためにアルファチャンネルを追加し、画像とテキストのペアを学習することで、画像認識精度を維持しつつ、詳細な領域認識を実現。
- **手法:** 従来のRGBチャンネルに加えて、アルファチャンネルを追加し、特定領域の情報を強調して学習。
- **感想:** CLIP学習において粒度の細かい学習を行う工夫が昨今取り組まれている中、この手法は非常にシンプルかつ有効に見える。



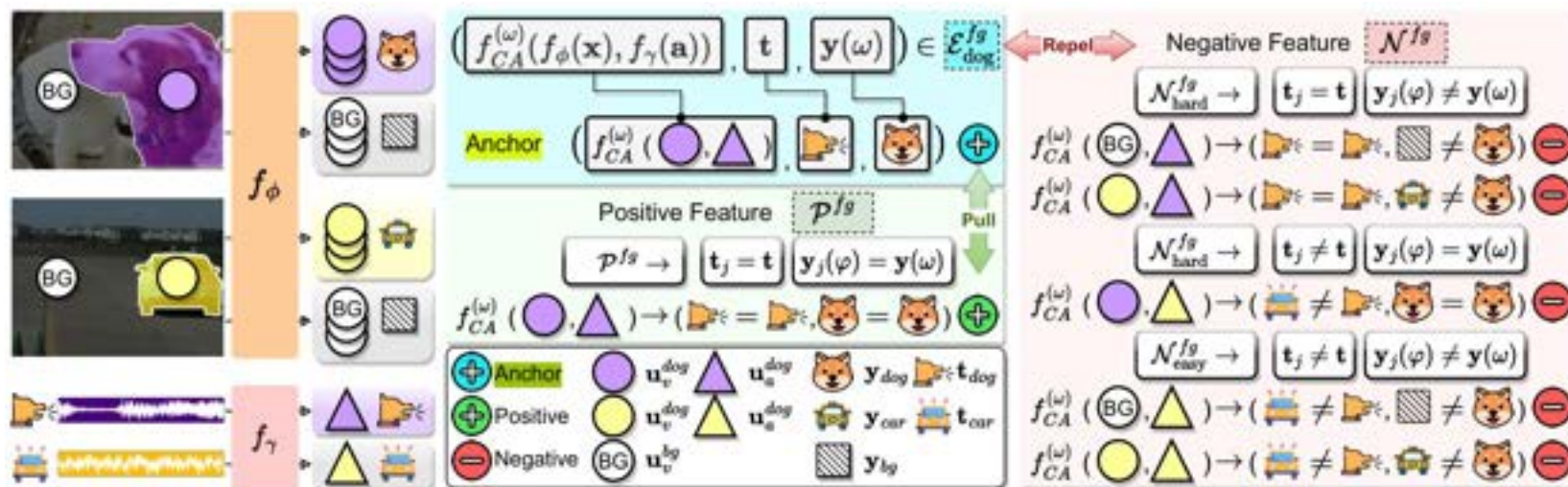
(a) Data generation pipeline



(b) Fine-tuning pipeline

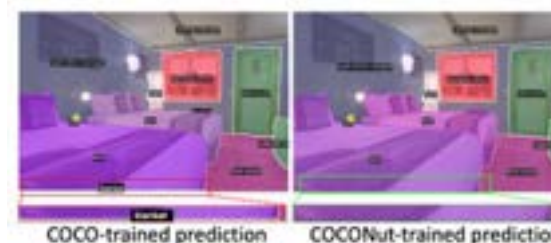
## Unraveling Instance Associations: A Closer Look for Audio-Visual Segmentation

- 概要:** 一つのシーンの中で特定のオブジェクトが音源であると仮定してしまう”commonsense”(常識的な)バイアスを解消。1.COCOデータセットの画像とVGGSoundのデータセットをペアリングしたVisual Post-Production(VPO)ベンチマークを構築。2.VPOベンチマークでaudio-visualの対照学習を行うことで、ポジティブとネガティブな特徴表現を区別する。
- 手法:** 1.VPOベンチマークはCOCO画像から21のセグメンテーションマスクを、そしてVGGSoundから10秒にトリミングした音声をインスタンスレベルに基づきランダムにマッチさせる。2.audio-visualデータセットからセグメントとオーディオデータを取り出し、セグメントとオーディオの組み合わせによりポジティブ、ネガティブaudio-visual特徴を作成する。損失関数によりポジティブ特徴は近づき、ネガティブ特徴は遠ざかるように学習を行う。
- 感想:** 従来のバイアスを解消するベンチマークを提案することで、マルチモダリティの学習の制約コストをより削減することが可能だと感じた。



## COCONut: Modernizing COCO Segmentation

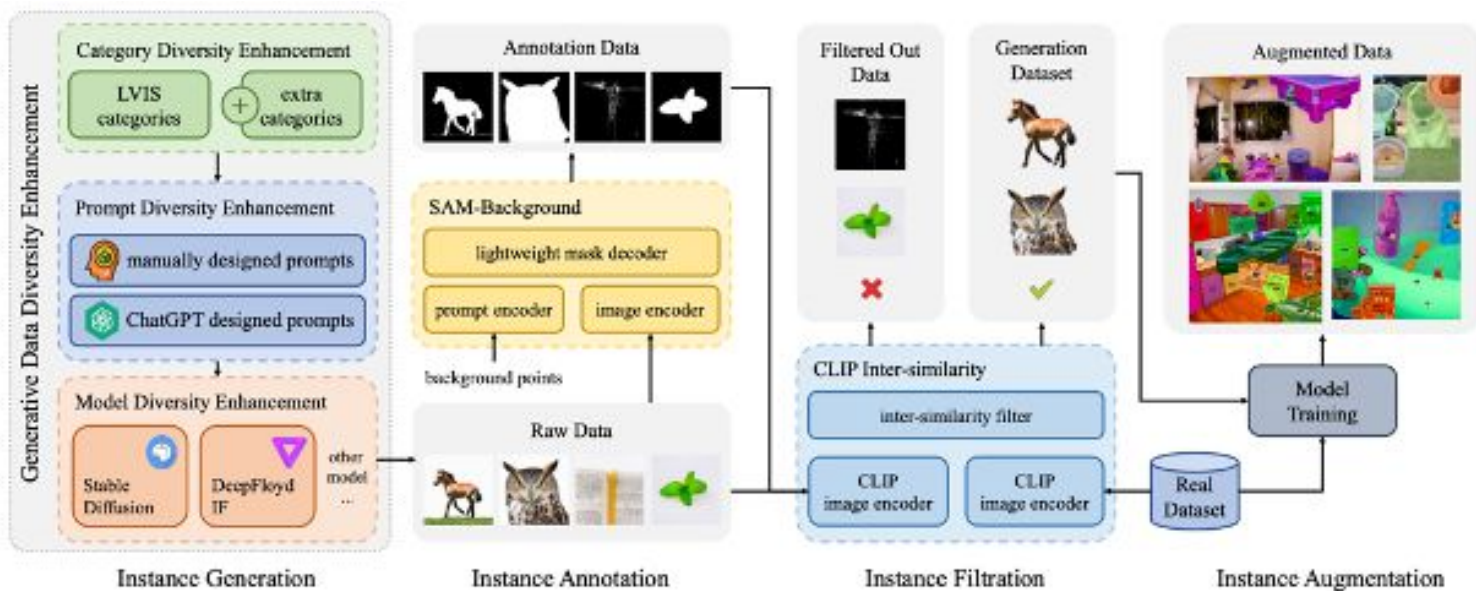
- 概要:** 物体検出やセグメンテーションモデルを用いたアノテーションパイプラインを用いてCOCOの改善を目指し、より高品質なアノテーションとデータセットの拡張を特徴とした、383,000枚の画像と5.18百万のパノプティクスマスクを含んだCOCONutを提案。
- 手法:** 物体検出やセグメンテーションモデルにて推定されたmaskやbboxに対して人間が確、修正するプロセスを複数回行うことで高精度なアノテーションを実施。作成したCOCONutにてモデルを学習させ、新しいラベルデータにおいて高いスコアを確認。
- 感想:** COCOのアノテーション精度については以前から問題になっていたためとても貢献度の高い取り組みだと思う。SAMとCOCONutで学習したモデルの比較がどうなるか検証したい



backbone	training set	COCO-val			relabelled COCO-val			COCONut-val		
		PQ	$AP^{mask}$	mIoU	PQ	$AP^{mask}$	mIoU	PQ	$AP^{mask}$	mIoU
ResNet50	COCO	53.3	39.6	61.7	55.1	40.6	63.9	53.1	37.1	62.5
	COCONut-S	51.7	37.5	59.4	58.9	44.4	64.4	56.7	41.2	63.6
	COCONut-B	53.4	39.3	62.6	60.2	45.2	65.7	58.1	42.9	64.7
	COCONut-L	54.1	40.2	63.1	60.7	45.8	66.1	60.7	44.8	68.3
ConvNeXt-L	COCO	57.9	45.0	66.9	60.4	46.4	69.9	58.3	44.1	66.4
	COCONut-S	55.9	41.9	66.1	64.4	50.8	71.4	59.4	45.7	67.8
	COCONut-B	57.8	44.8	66.6	64.9	51.2	71.8	61.3	46.5	69.5
	COCONut-L	58.1	45.3	67.3	65.1	51.4	71.9	62.7	47.6	70.6

## DiverGen: Improving Instance Segmentation by Learning Wider Data Distribution with More Diverse Generative Data

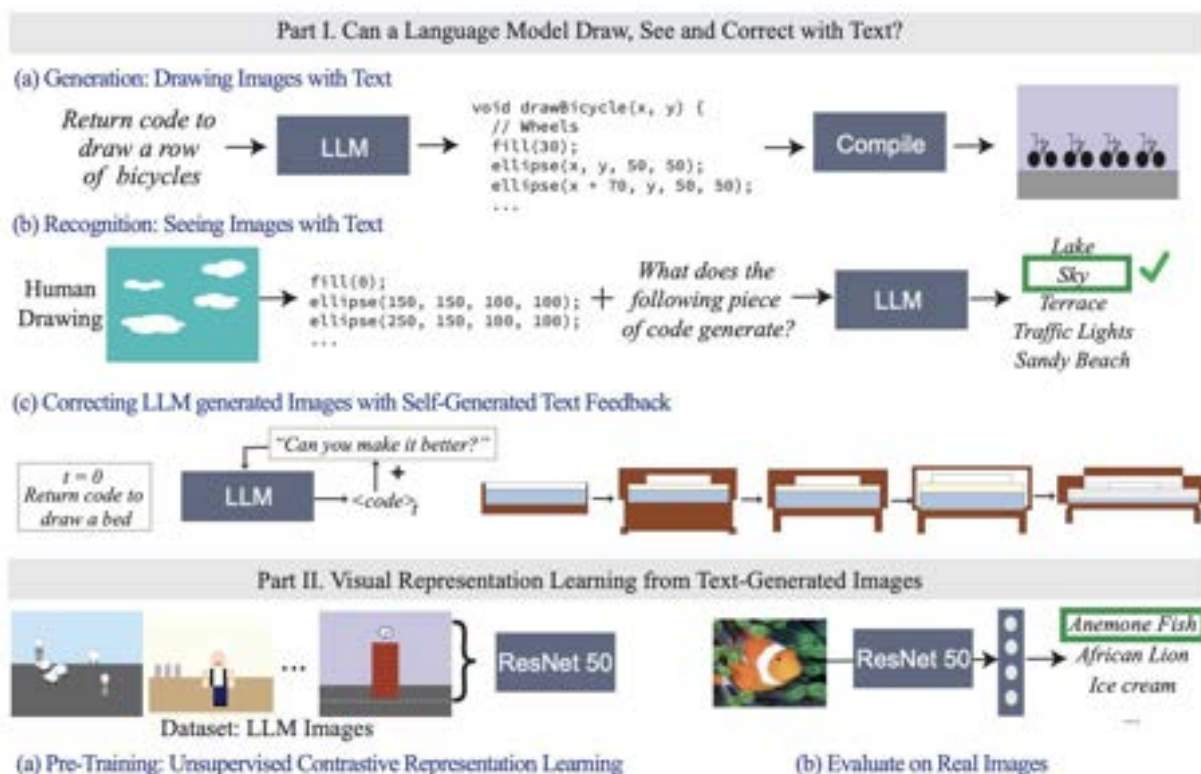
- 概要:** 生成モデルで生成したデータと実画像データとの分布の不一致の観点から、実画像データの分布を拡大させるデータ多様性拡張手法を提案。この手法によりLVISを用いたインスタンスセグメンテーションにおいて従来のデータ拡張手法よりも高いスコアを確認
- 手法:** カテゴリの多様性はImageNet-1kのカテゴリを、プロンプトの多様性はChatGPTを用いて各プロンプトが異なり、1画像に1オブジェクトになるよう設定、生成モデルの多様性は2種類のモデルを用いた。アノテーション生成はSAMを使用
- 感想:** 生成モデルを用いてデータセットの多様性を拡張するのは良い手法だと思った。一方でアノテーションにはSAMを用いているため、end-to-endでインスタンスセグメンテーションデータを生成できるようになると良い



Method	Backbone	$AP^{box}$	$AP^{mask}$	$AP_r^{box}$	$AP_r^{mask}$
Copy-Paste [7]	EfficientNet-B7	41.6	38.1	-	32.1
Tan et al. [26]	ResNeSt-269	-	41.5	-	30.0
Detic [36]	Swin-B	46.9	41.7	45.9	41.7
CenterNet2 [35]	Swin-L	47.5	42.3	41.4	36.8
X-Paste [34]	Swin-L	50.1	44.4	48.2	43.3
<b>DiverGen (Ours)</b>	Swin-L	<b>51.2</b> (+1.1)	<b>45.5</b> (+1.1)	<b>50.1</b> (+1.9)	<b>45.8</b> (+2.5)

## A Vision Check-up for Language Models

- 概要:** LLMがどのようにVisionを捉えているのかを様々な方法で調査した論文。LLMが生成した画像でPre-trainingの高い効果が確認できた。
- 手法:** (a) LLMにコードを書かせ指示物体を線画させる。(b) 人が書いた絵をコード化し、LLMに分類問題として解かせる。(c) LLMで生成した画像でモデルを訓練し、実画像で評価。
- 新規性:** LLMが視覚情報をどれほど持つのか、文字としてだけではなく視覚的に表すことができるのかを調査した点。抽象的な概念をうまく実験に落とし込んでいる。



	Pre-training Dataset	I-1k	I-100
		Linear	5-NN
Random Init.	None	4.36	4.28
Real	Places	55.59	57.04
Procedural	Dead-leaves	20.00	12.76
	FractalDB-1k	23.86	17.24
	StyleGAN O	38.12	33.00
	S-21k	44.83	43.24
	LLMs (w/o GPT-4)	33.60	22.42
	LLMs (w/ GPT-4)	36.16	27.44
	LLMs (w/o GPT-4) + S-21k	45.79	<b>43.40</b>
	LLMs (w/ GPT-4) + S-21k	<b>46.03</b>	43.36

## Person-in-WiFi 3D: End-to-End Multi-Person 3D Pose Estimation with Wi-Fi

- **概要:** WiFiを使用した複数人物3次元姿勢推定モデルの提案
- **手法:** CSIデータに対して、subcarrierをまとめてトークン化。アンテナとレシーバーの関係性は位置情報を含むので位置一時間Embeddingを追加。100のqueryを追加してDecoderを通し、ハンガリアンマッチングロスなどを使って複数人物姿勢推定
- **新規性:** 概要の内容+WiFi信号を用いたDETRベースのTransformerで姿勢推定を初めて実現

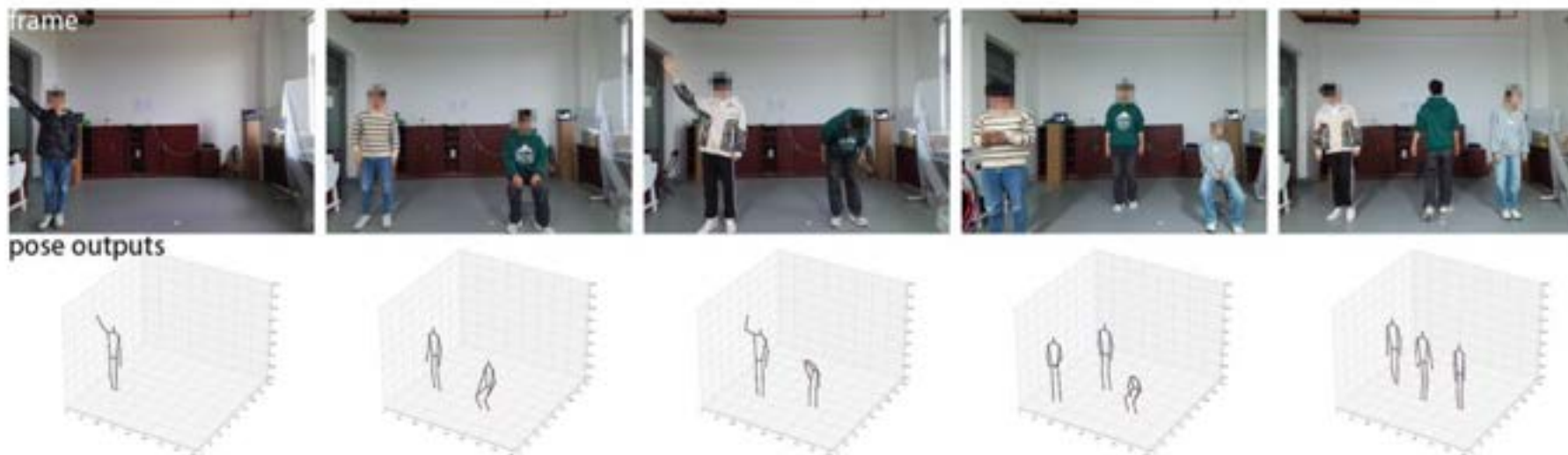
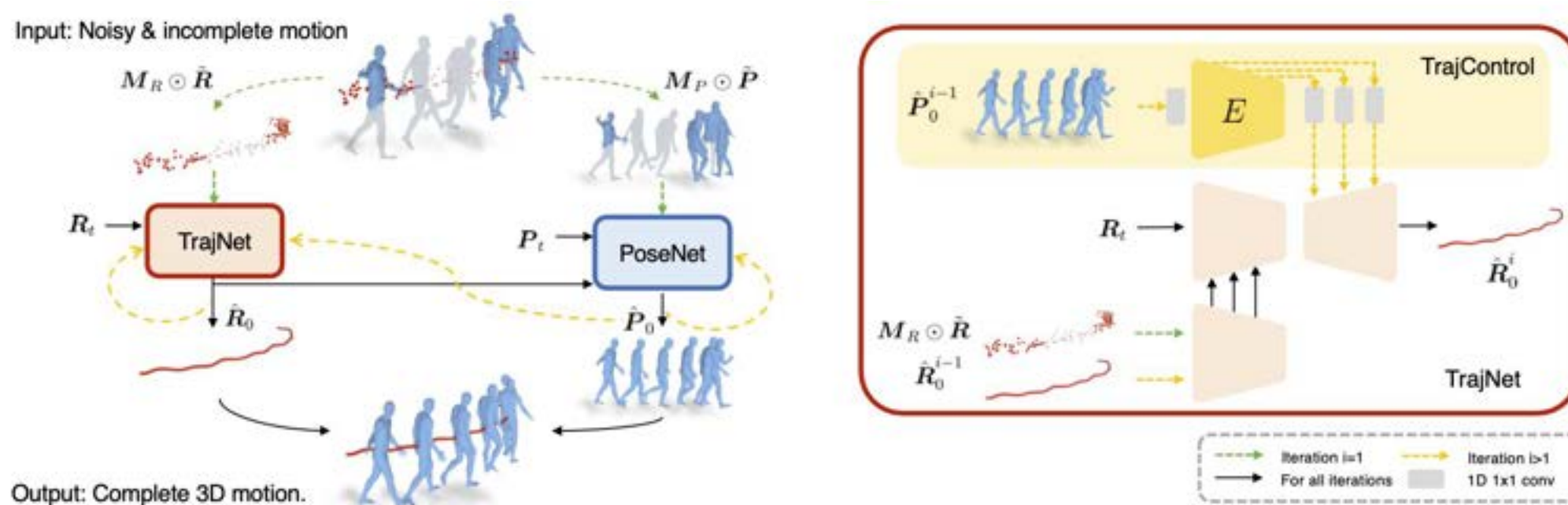


Figure 1. This paper presents Person-in-WiFi 3D, the first multi-person 3D pose estimation system with Wi-Fi signals.

## RoHM: Robust Human Motion Reconstruction via Diffusion

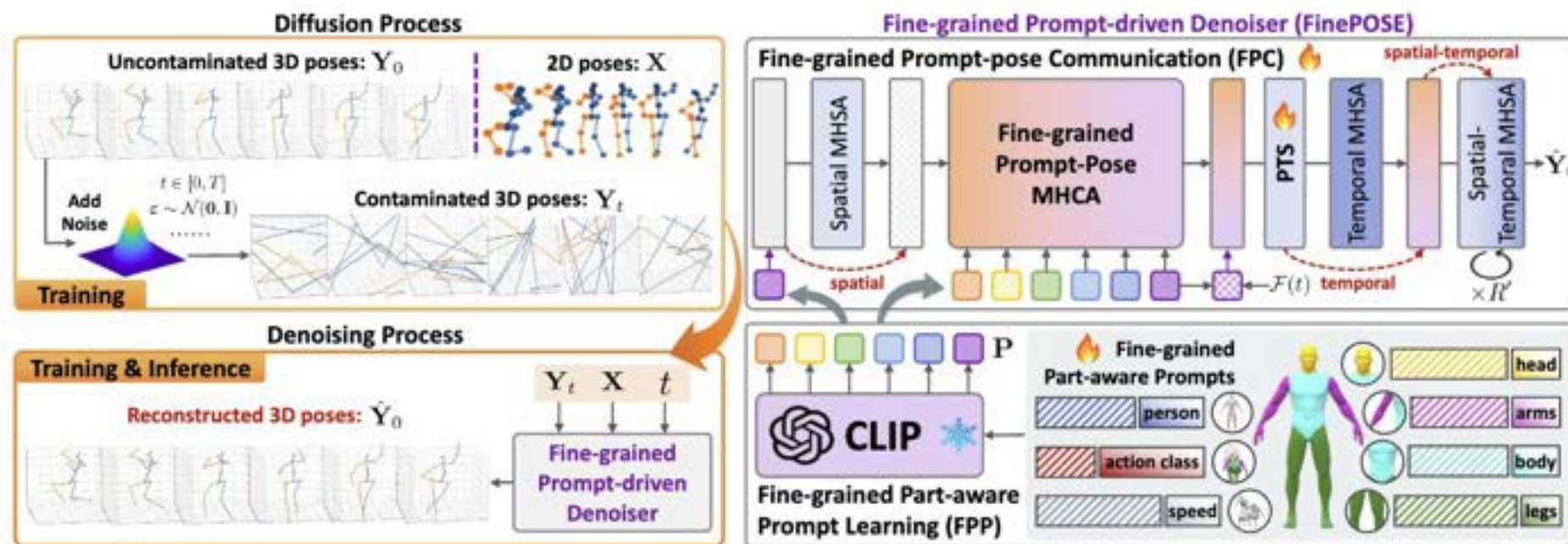
- ❑ **概要:** 拡散モデルを活用し、遮蔽やノイズに頑健な3次元人物姿勢推定モデルを提案
- ❑ **手法:** マスクされたRootの動きと人物姿勢に対してTrajNetとPoseNetによりDenoising。得られたPoseを条件としてRootを更にアップデート(TrajControl)
- ❑ **新規性:** 大局的/局所的な人の動きをそれぞれマスク+拡散モデルでモデル化。遮蔽によって下半身が見えないような状況であっても正確に3次元姿勢を推定可能。また推論時のガイダンスによってPlausibilityの高い姿勢出力を実現





## FinePOSE: Fine-Grained Prompt-Driven 3D Human Pose Estimation via Diffusion Models

- **概要:** CLIPのテキスト特徴とプロンプト最適化を利用した3次元人物姿勢推定
- **手法:** 関節点位置に関してデノイズングする際に、2D関節点座標と時間、人物姿勢に関わる学習可能プロンプトをconcatして活用。
- **新規性:** 大規模に事前学習されているCLIPのtext embeddingを利用することによって人間の知識を活用。また部位毎のプロンプトを用意し高い粒度で情報を抽出



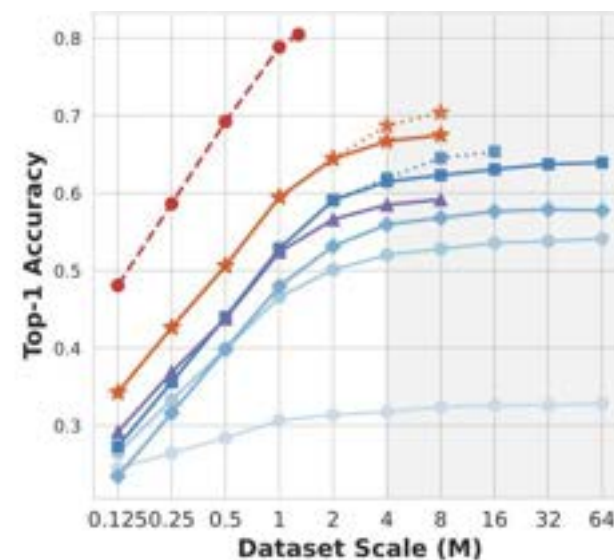
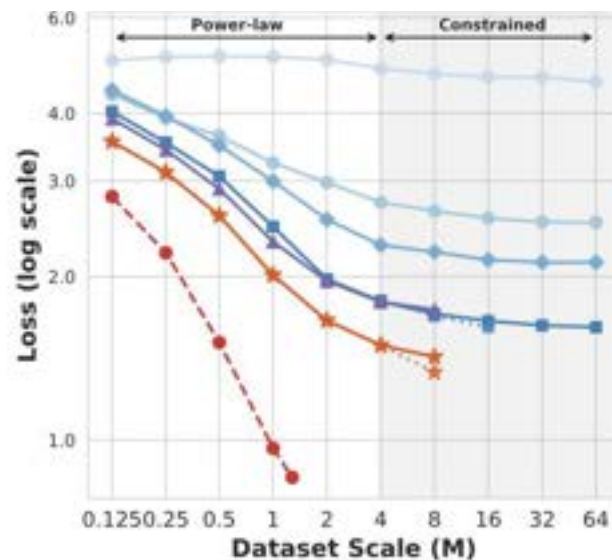
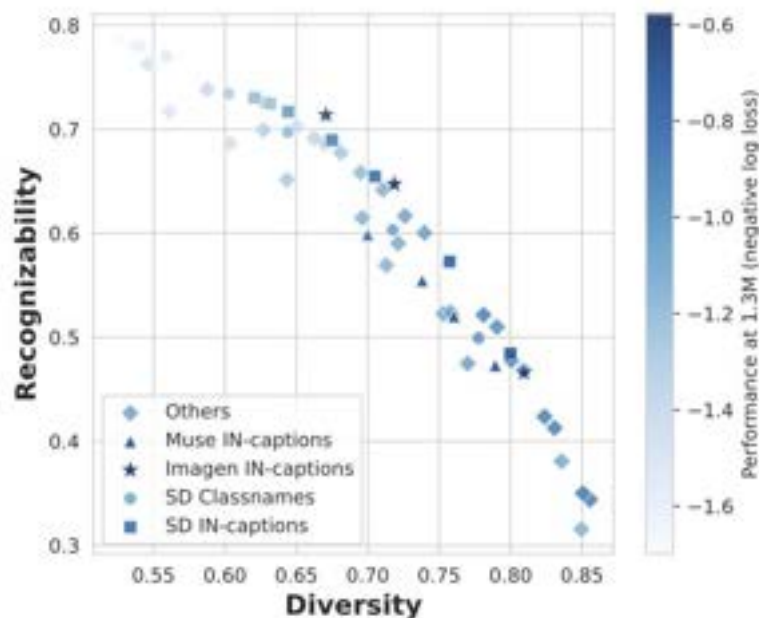
## ImageNet-D: Benchmarking Neural Network Robustness on Diffusion Synthetic Object

- ❑ **概要:** 生成画像から構成される、分布外頑健性評価用データセットImageNet-Dを提案
- ❑ **手法:** Background, Texture, Materialを指定したキャプションを用いて画像生成。その後代表的な4つのモデルが一致して不正解だったものを選出
- ❑ **新規性:** 頑健性評価データセット(ImageNet-Cなど)はルールベースの画像編集などを使用しておりリアルさがない。作成にもコストがかかる。そこでリアルかつ自動生成可能な頑健性評価用データセットを構築。既存の画像識別モデルやVQAモデルに対して16~40%程度の精度低下を起こした。



## Scaling Laws of Synthetic Images for Model Training ... for Now

- **概要:** 人工画像データを用いた事前学習のスケーリング則に関する検証
- **手法:** 生成データを用いた時の事前学習に関して、①モデルの種類②textガイダンスの強さ③textプロンプトを変化させスケーリング則を検証。また得られた画像の多様性、クラスに関する正確さと事後学習の精度の関係を検証。
- **実験結果:** (i) 生成データによる事前学習は現実データと比べてスケーリング則が弱い(中央図)。(ii) データの多様性が大事(左下図) (iii) スケーリングの程度にクラス間で差がある(生成の精度が悪そう) (iv) ドメインギャップがあると大規模人工データ>リアルデータ



Legend for the graphs:

- SD Classname, CFG 7.5 (light blue circle)
- SD CLIP template, CFG 2.0 (light blue diamond)
- MUSE IN-Captions (dark blue triangle)
- Real (red dashed line with circle)
- SD Classname, CFG 2.0 (light blue square)
- SD IN-Captions, CFG 2.0 (dark blue square)
- Imagen IN-Captions (orange star)

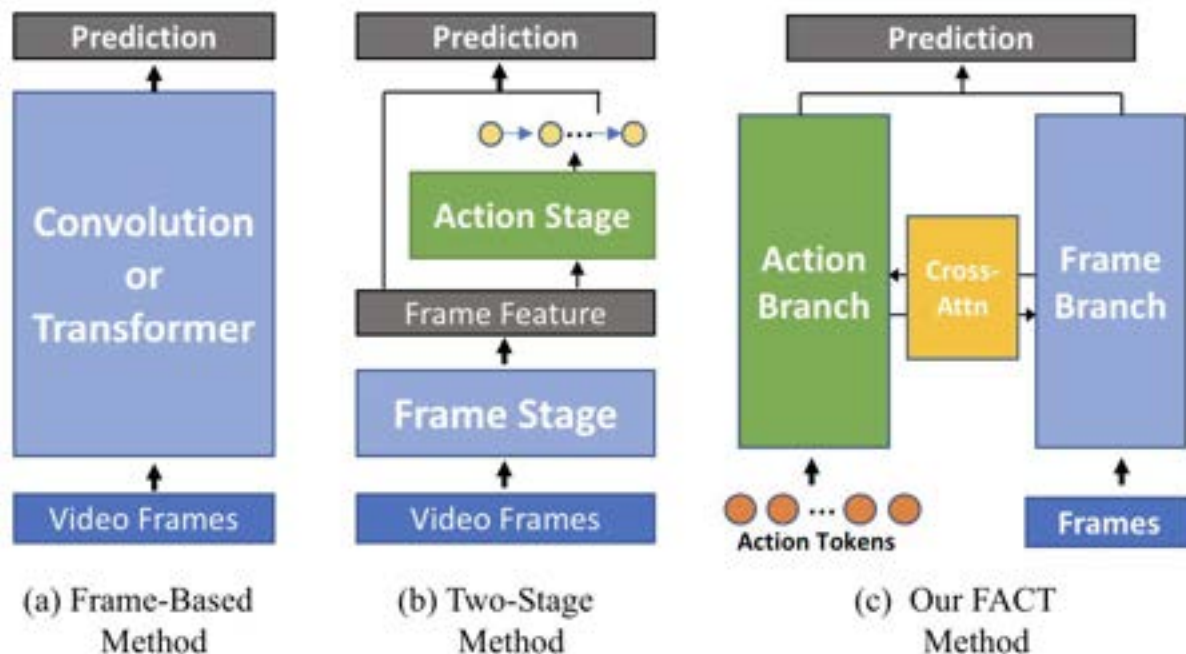
## Learning Vision from Models Rivals Learning Vision from Data

- **概要:** LLMとStable Diffusionによって生成された画像データのみを用いた事前学習
- **手法:** in-context learningにより背景や位置情報を含んだキャプションをLLMにより生成。同一キャプションを使用した画像を陽性として対照学習。MIMも追加で使用
- **新規性:** 画像生成のためのキャプションをLLMを用いて生成。LIPやDINO v2に匹敵する精度を下流タスクで記録。



## FACT: Frame-Action Cross-Attention Temporal Modeling for Efficient Action Segmentation

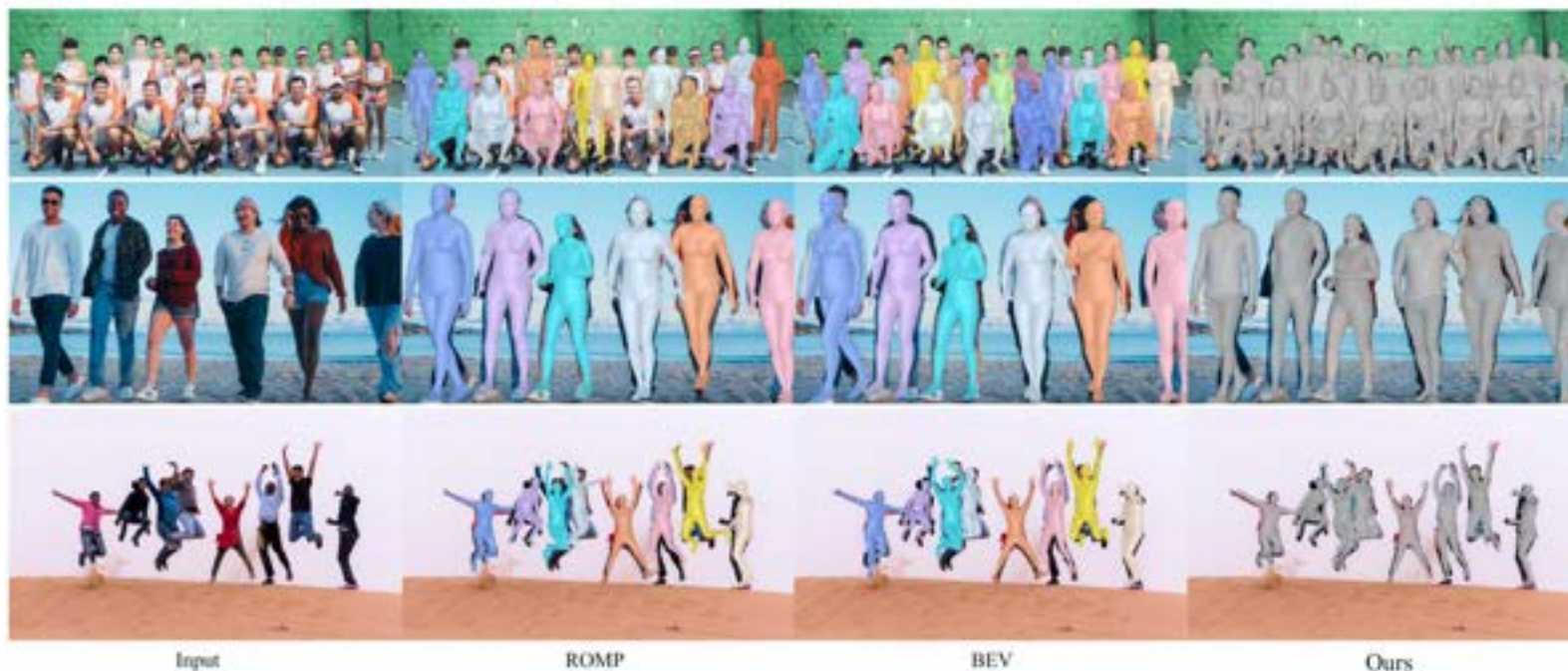
- 概要:** アクションとフレーム間の相補的な情報を学習させる手法。
- 手法:** 従来通り特徴量抽出→時系列解析を行なった後、解析結果とフレーム特徴をcross-attentionする工程を特徴量抽出のブランチと時系列解析のブランチで交互に行う。
- 新規性:** アクションレベルとフレームレベルの二つの視点で特徴量を洗練させることで時系列の解析能力を向上させた。



	Note	Breakfast			GTEA				
		F1@[10, 25, 50]	Edit	Acc	F1@[10, 25, 50]	Edit	Acc		
ED-TCN [30]	F <sub>1</sub>	-	-	-	72.2	69.3	56.0	64.0	-
TDRN [32]	F <sub>1</sub>	-	-	-	79.2	74.4	62.7	74.1	70.1
SSA-GAN [21]	F <sub>2</sub>	-	-	-	43.3	80.6	79.1	74.2	76.0
Bridge-Prompt [36]	F <sub>3</sub>	-	-	-	94.1	92.0	83.0	91.6	81.2
MSTCN [16]		52.6	48.1	37.9	61.7	63.3	87.5	85.4	74.6
MSTCN++ [37]		64.1	58.6	45.9	64.9	67.6	88.8	85.7	76.0
MuCon [56]	P	73.2	66.1	48.4	76.3	62.8	-	-	-
C2F-TCN [55]	A	72.2	68.7	57.6	69.6	76.0	84.3	81.8	72.6
ASRF[24]	C <sub>1</sub>	74.3	68.9	56.1	72.4	67.6	89.4	87.8	79.8
HASR [1]		74.7	69.5	57.0	71.9	69.4	90.9	88.6	76.4
ASFormer[61]		76.0	70.6	57.4	75.0	73.5	90.1	88.8	79.2
DTL[59]	C <sub>2</sub>	78.8	74.5	62.9	77.7	75.8	-	-	-
MVGA[4]	A	75.6	72.1	59.7	76.8	74.2	91.3	90.0	79.3
TCTr [3]		76.6	71.1	58.5	76.1	77.5	91.3	90.1	80.0
UVAST[7]	P	76.9	71.5	58.0	77.1	69.7	92.7	91.3	81.0
RTK [25]		76.9	72.4	60.5	76.1	73.3	91.2	90.6	83.4
LTCContext [5]		77.6	72.6	60.1	77.0	74.2	-	-	-
DiffAct[39]	C <sub>1</sub>	80.3	75.9	64.6	78.4	75.1	92.5	91.5	84.7
<b>FACT</b>		<b>81.4</b>	<b>76.5</b>	<b>66.2</b>	<b>79.7</b>	<b>76.2</b>	<b>93.5</b>	<b>92.1</b>	<b>84.1</b>

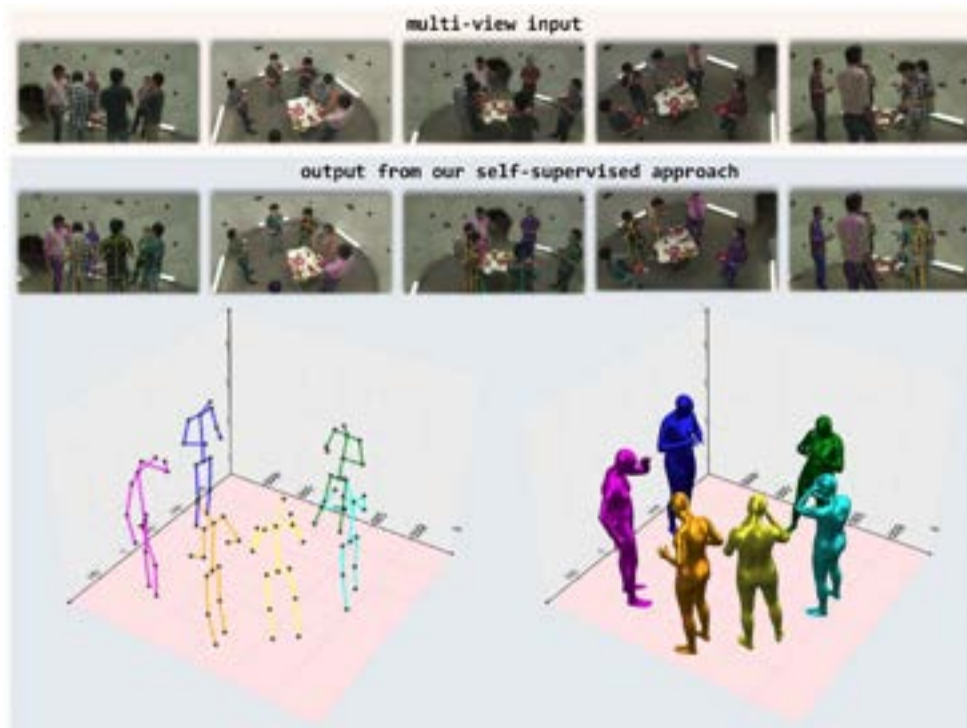
## AiOS: All-in-One-Stage Expressive Human Pose and Shape Estimation

- **概要:** 一つのモデルで画像から3次元のメッシュ復元を行う手法.
- **手法:** 特徴量抽出→BBOXを検出するステージ→関節点推定を行うステージ→検出された関節点のrefinement&メッシュ復元するステージの順に行う. 特徴量抽出以外の前ステージはDETRをベースとしており, ステージ間でトークンを橋渡しする.
- **新規性:** 関節点推定を行うステージで人物それぞれのマスクを生成することで重なっている人物間における関節の誤認識を抑制.



## SelfPose3d: Self-Supervised Multi-Person Multi-View 3d Pose Estimation

- ❑ **概要:** 完全な自己教師あり学習のみで多視点の3次元の多人数の姿勢推定を行う手法
- ❑ **手法:** 多視点の2D画像から3D上の各人物の中心位置を推定. 中心位置と各視点における2D姿勢推定の結果から3D姿勢推定を行い, その3D姿勢推定結果を各視点の2D画像に投影し, 元の2D姿勢推定結果との間で損失を計算する. また, オクルージョンが発生しているかどうかを別途2DCNNを用いて計算し損失の比率を自動的に決定させる.
- ❑ **新規性:** 完全な自己教師あり学習のみで多視点の3次元の多人数の姿勢推定を行う. 多視点特有の情報をうまく活用することで精度を向上させている.



## DiffusionRegPose: Enhancing Multi-Person Pose Estimation using a Diffusion-Based End-to-End Regression Approach

- **概要:** 拡散モデルベースでend-to-endで画像から多人数の2D関節点を推定する手法
- **手法:** 人物検出器の出力のBBOXとdiffusionベースの姿勢推定器の出力をcross-attentionさせ、その出力を得る。学習時は、DiffusionPoseと同様にGTの関節点に徐々にノイズを付与するモジュールとそれをデノイズさせるモジュールを用意し、ノイズに対する除去性能を獲得させる。推論時はデノイズさせるモジュールのみを利用し徐々にデノイズを行う。
- **新規性:** 人物検出とポーズのノイズ除去のそれぞれの過程で相互に情報を活用することで、人物検出の精度の向上とノイズ除去性能の向上の両方を達成している。





## Improving Spectral Snapshot Reconstruction with Spectral-Spatial Rectification

- スナップショット分光法で取得したHSI(Hyperspectral Image)を復元する際の空間データの劣化を緩和するための手法SSR(Spectral-Spatial Rectification)を提案。実データおよびシミュレーションデータセットを対象に復元能力が改善することが実証された
  - 従来手法(the spectra-wise transformer)では、帯域全体をトークンとみなすことでスペクトル分布の局所的差異が無視される
  - スナップショット分光から取得したデータではスペクトル密度の低い空間スペクトル回復が困難な場合がある
- 特徴量を空間的に多数のウィンドウに分割するWSSA(Window-based Spectra-wise Self-Attention)と大きいカーネルの畳み込み、スペクトル重みの学習により空間の劣化を緩和するARB(spAtial Rectification Block)を新たに提案し、HSI復元の際の劣化を緩和
- WSSAとARBを加えることでHSI復元性能が向上することのほか、画質指標のPSNRとSSIMによる評価にて、SOTA手法よりも高い復元性能を示した

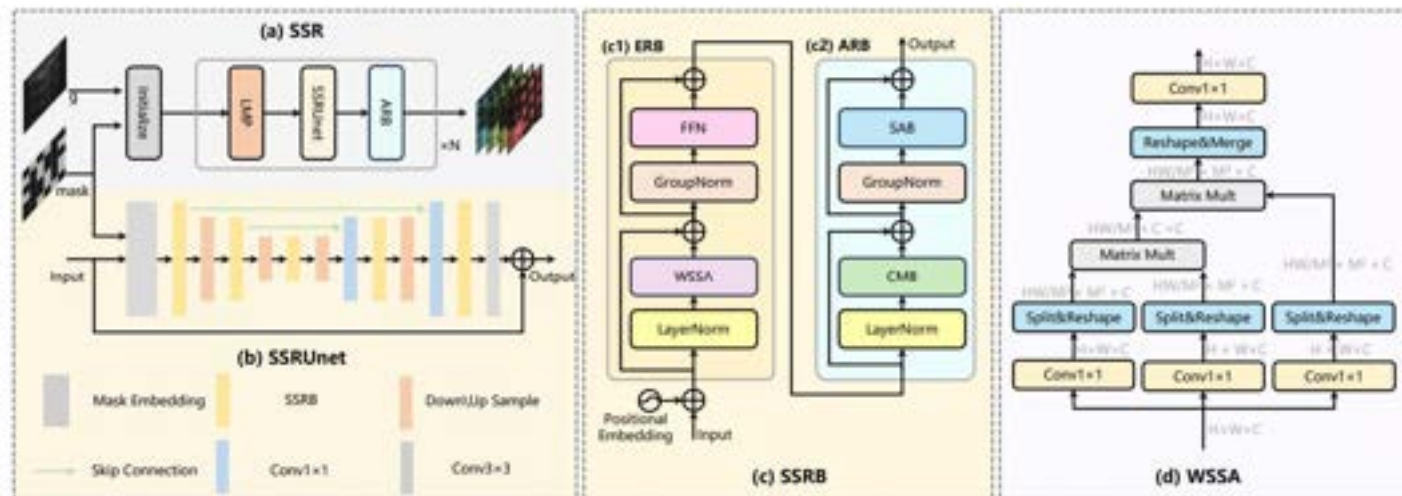


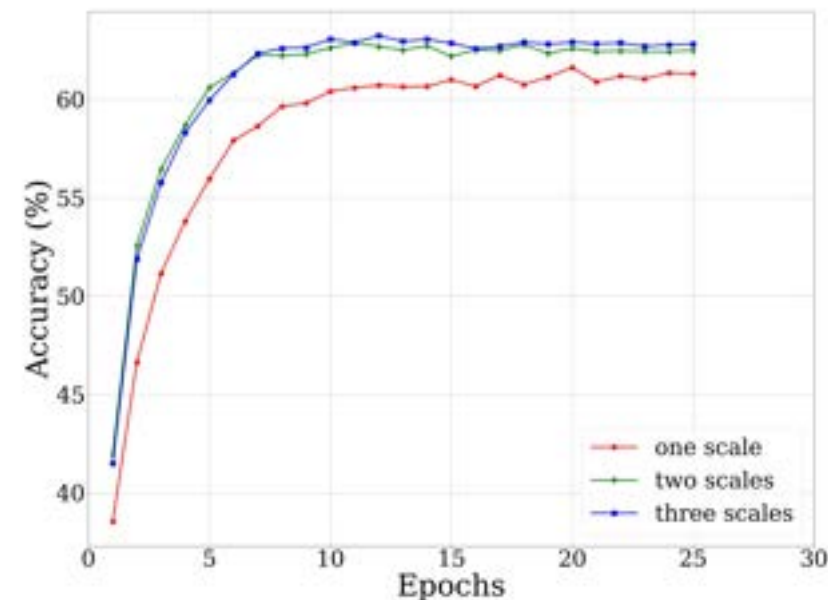
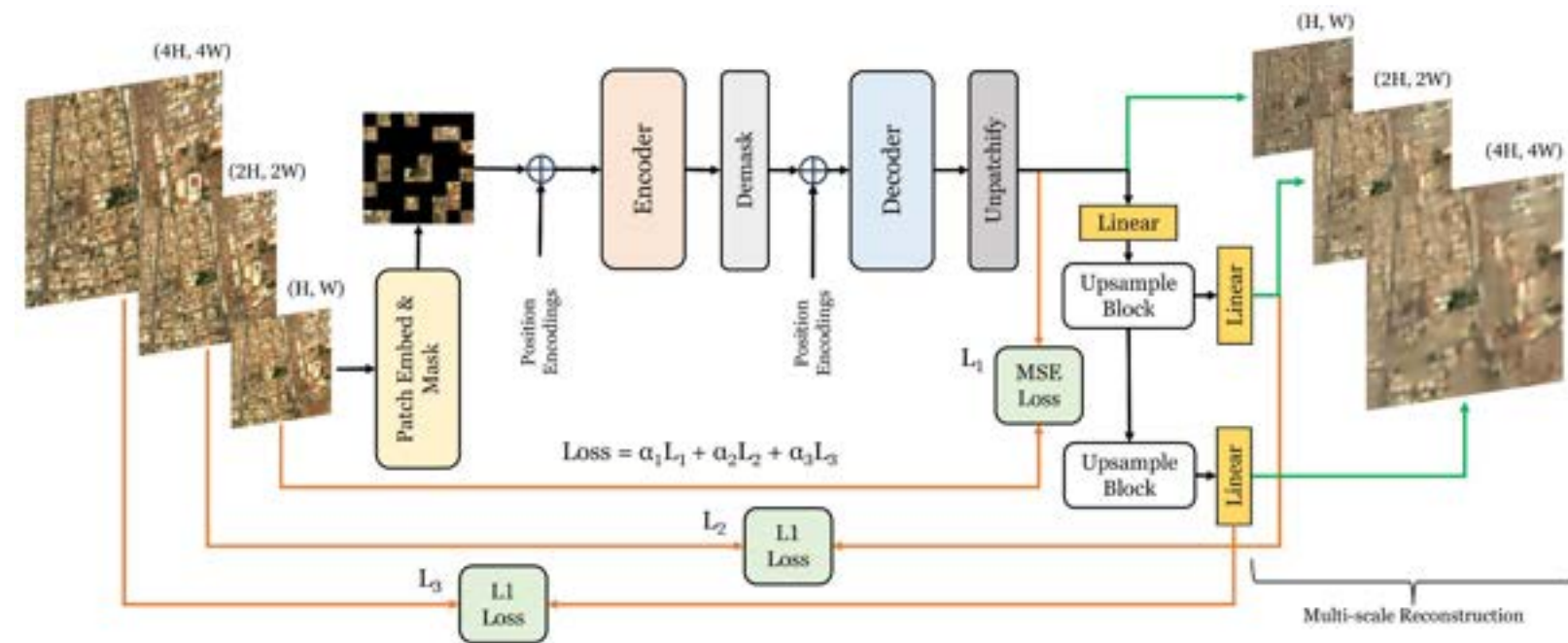
Figure 2. (a)-(c) The overall architecture of SSR, SSRNet, and SSRB. LMP is achieved via a formula. (d) Details of WSSA.

Table 2. Ablation study of WSSA and ARB.

Baseline-1	WSSA	ARB	PSNR	SSIM	Params (M)	FLOPs (G)
✓			36.60	0.956	1.06	14.43
✓	✓		37.76	0.965	1.30	18.82
✓	✓	✓	39.19	0.971	1.73	26.37

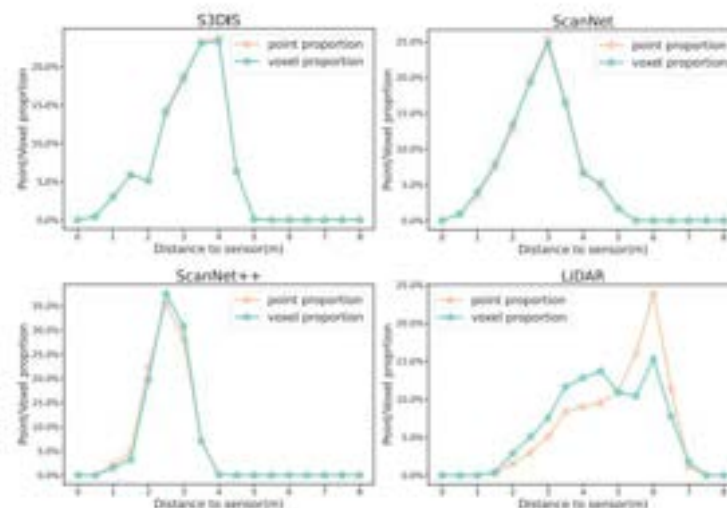
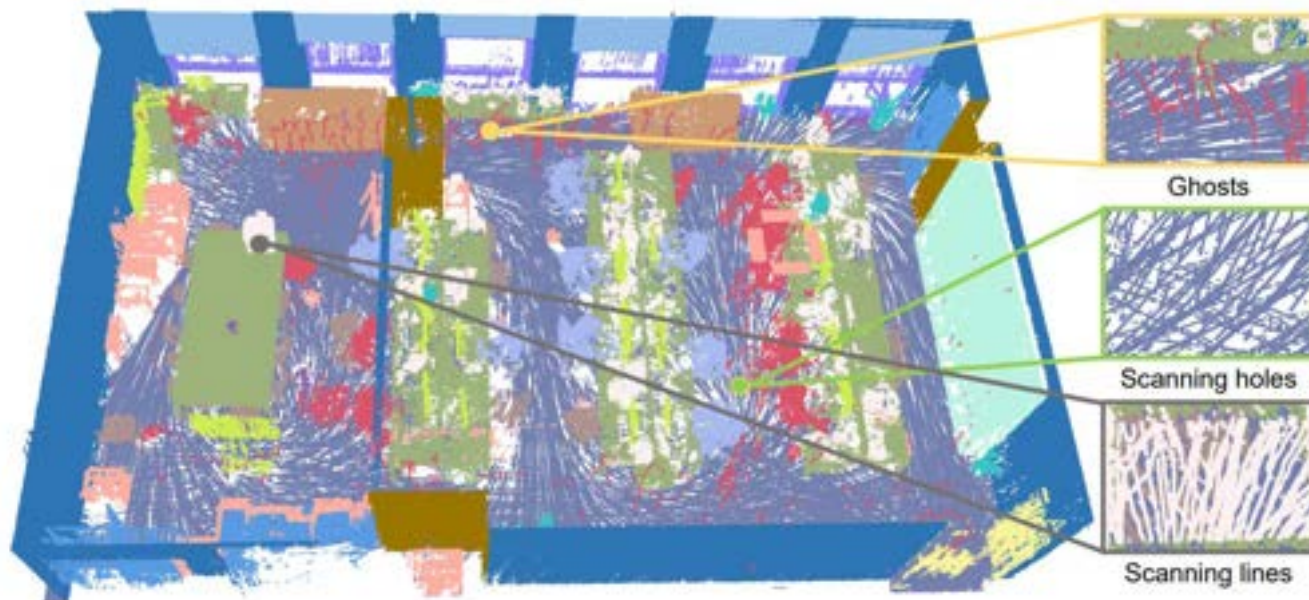
## Rethinking Transformers Pre-training for Multi-Spectral Satellite Imagery

- 概要:** マルチスケール性を有する衛星画像向けのTransformer事前学習手法を提案
- 手法:** 衛星画像はある程度解像度を高く観測できる可視光のセンサと、近赤外など解像度が低いセンサで観測される。RGBの場合は隣接する画素と画素の距離は数十cmオーダーだが、近赤外の場合は解像度が変わり数十 mオーダーになる。このようなマルチスケール性と多チャンネルを有する衛星画像に対する事前学習手法として、SatMAE++を提案した。ベースラインのSatMAEはMAEに対してRGB以外の近赤外などのスペクトル情報も扱えるように拡張した。SatMAE++では、さらにマルチスケール性を学習できるように拡張
- 新規性:** 入力画像を最大3つのスケールレベルで取り込む仕組みでマルチスケール性を学習させる



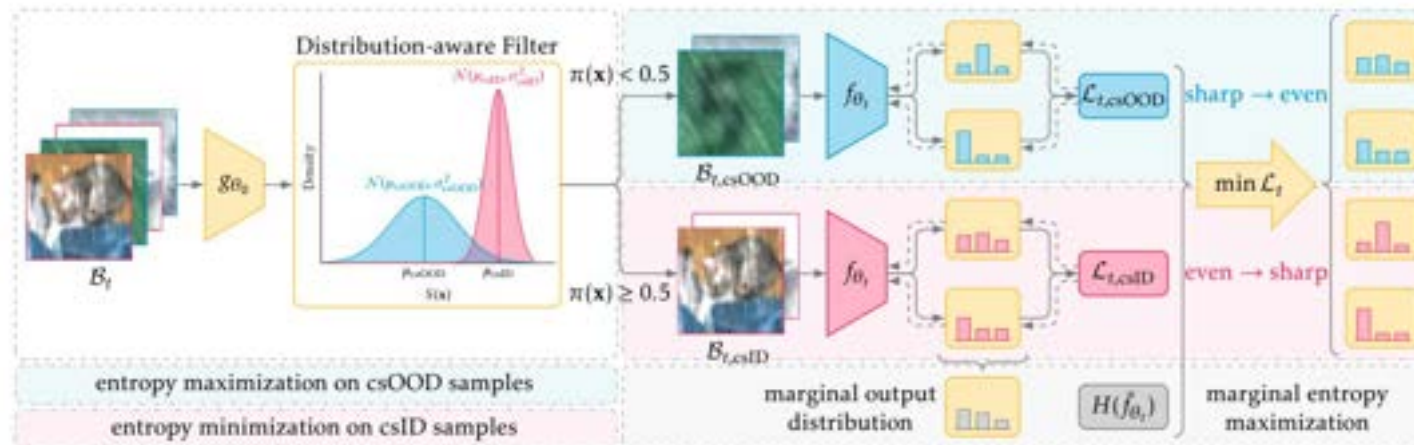
## LiDAR-Net: A Real-scanned 3D Point Cloud Dataset for Indoor Scenes

- ❑ **概要:** 点群のSem.Seg, Instance Seg. 物体検出向けのもちLiDARで屋内を撮影したデータセットを公開
- ❑ **手法:** 従来の屋内点群データセットはS3DISなどRGB-Dスキャナで観測されたものが多かった。しかし、屋内観測をする場合は地上設置型LiDARや手持ちのLiDARが使われることが多い。そこで、手持ちLiDARで観測した屋内点群に対してセグメンテーションなどを行うデータセットを構築して、より現実的な問題設定で点群深層学習手法の評価を行えるようにした。
- ❑ **新規性:** 従来のRGB-Dデータセットでは発生していないLiDAR特有のゴースト・スキヤンの穴・スキヤンラインが含まれているので、ユーザがLiDARで観測した状況に近い。



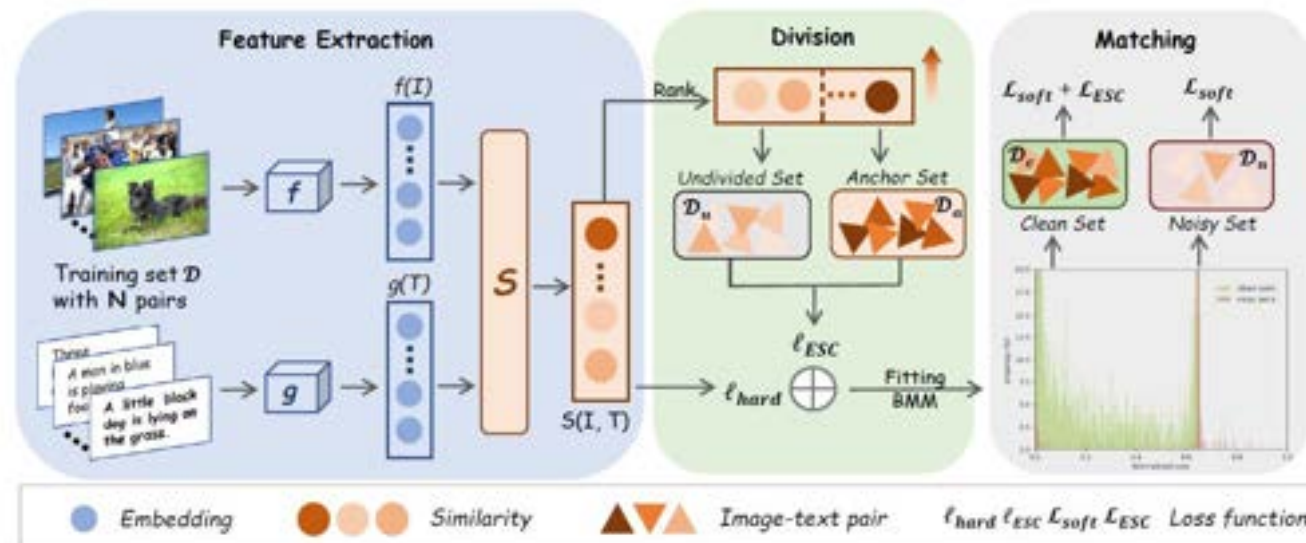
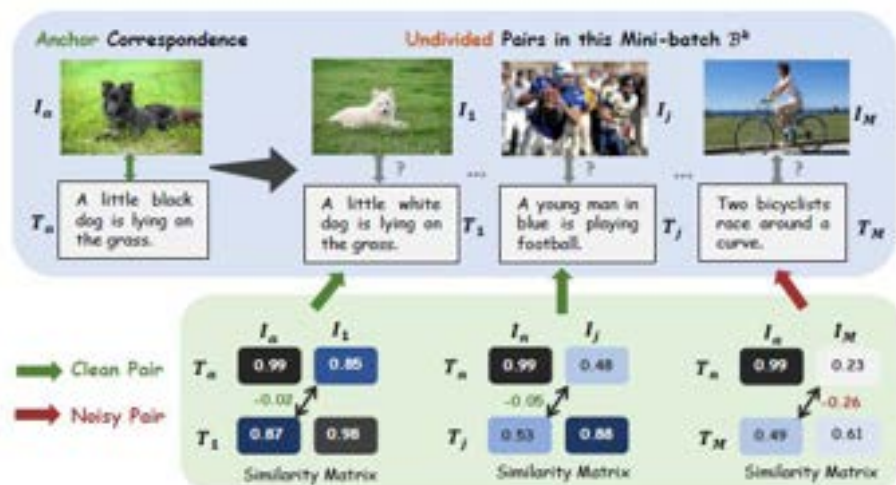
## Unified Entropy Optimization for Open-Set Test-Time Adaptation

- Open-Setに対しても有効な Test-Time Adaptation (TTA) の手法の提案
  - 既存のTTA手法はClosed-setには有効であるがOpen-setに対して性能の劣化が起こる
- TTAで広く用いられているエントロピー最小化は Open-set では上手くいかない
  - Open-set samples の存在がnormalization statisticsの推定を難しくする
  - 未知のサンプルに対するエントロピー最小化によってモデルの性能が悪化する
- ミニバッチごとにOODとIDに分ける処理を施し, OODサンプルに対してはエントロピー最大化, IDサンプルに対してはエントロピー最小化を用いて学習.



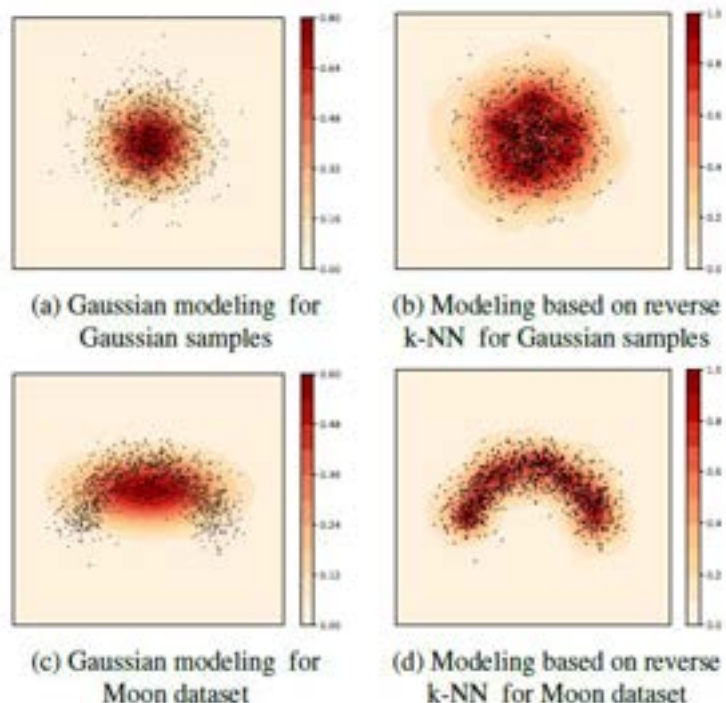
## Robust Noisy Correspondence Learning with Equivariant Similarity Consistency

- 概要:** マルチモーダルの学習に必要なデータに含まれるラベルノイズに対するロバスト性向上手法の提案。提案するEquivariant Similarity Consistency(ESC)によってクリーンラベルとノイズラベルの分離を可能とし、マルチモーダル検索の性能向上を達成
- 手法:** 画像とテキストの意味的变化は比例すると仮定。一部の適切な画像とテキストのペア画像を基にそれ以外のデータを判定。画像とテキストそれぞれを別のモデルで学習し、類似性を算出。その後2種類のlossを計算し、BMMに適用することでクリーンラベルとノイズラベルに分割
- 新規性:** 従来クロスモーダル検索で使用されていたtriplet lossではロバスト性が低下していたが、提案手法によりロバスト性を担保したまま、ノイズラベル対処が可能



## Learning with Structural Labels for Learning with Noisy Labels

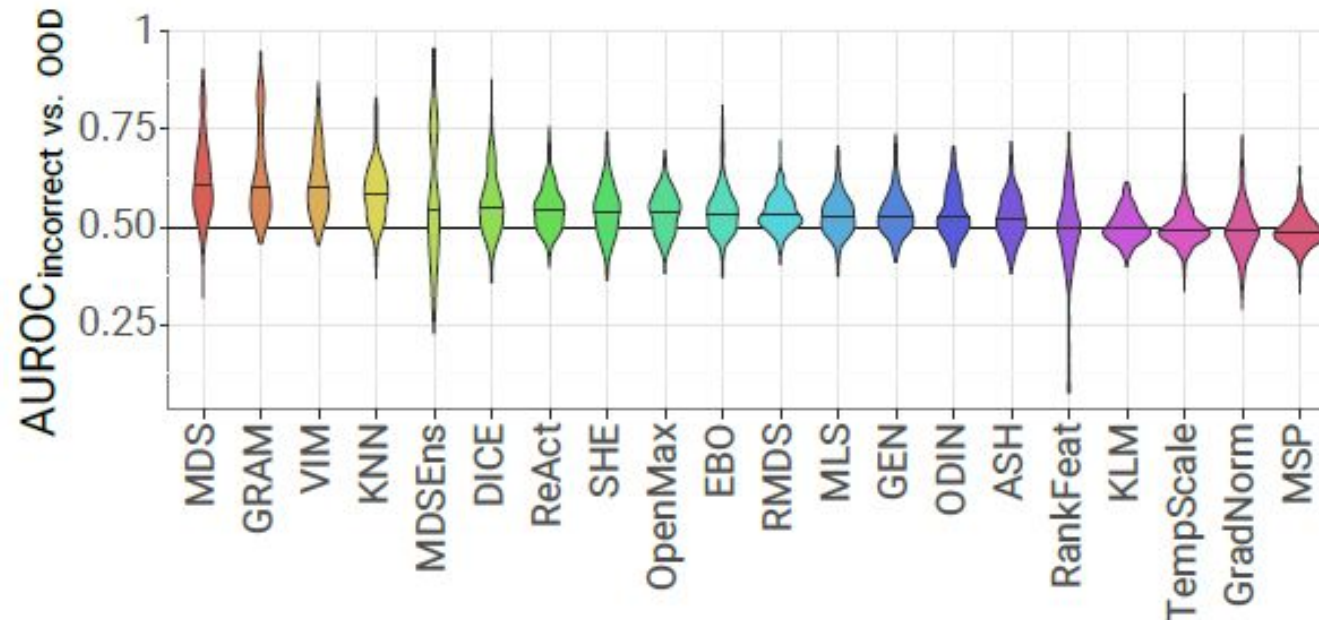
- 概要:** ラベルノイズを含んだ大規模データセットを学習し、性能を向上させる手法の提案。従来手法ではノイズラベルによる汎化性能の低下が問題であったが、構造化ラベルと呼ばれる特徴分布の追加やreverse k-NNアルゴリズムにより多様な特徴とラベルノイズの過適合を防止。従来のラベルノイズ手法に比べて性能向上や学習時間の効率化を実現
- 手法:** ソフトサンプルを定義し、各サンプルのクラスを予測。同じクラスを予測したサンプル同士で確率を計算し、確率に基づいてラベルを構造化ラベルを再定義。データ拡張を用いて構造化ラベルを学習
  - 各サンプルが同じクラスを予測したサンプルに向けて矢を放ち、受け取った数によって確率を近似。
- 新規性:** 手法によって効率的にデータの分布を示すことが可能(左図)。また構造化ラベルはデータセット全体にも効果的に影響を及ぼすことが判明



Method	WebVision		IISVRC2012	
	Top-1	Top-5	Top-1	Top-5
Co-teaching [19]	63.58	85.20	61.48	84.70
ELR+ [37]	77.78	91.68	70.29	89.76
NGC [65]	79.16	91.84	74.44	91.04
LongReMix [8]	78.92	92.32	-	-
RRL [32]	76.30	91.50	73.30	91.20
Sel-CL+ [33]	79.96	92.64	76.84	<b>93.04</b>
TCL [22]	79.10	92.30	75.40	92.40
DivideMix [31]	77.32	91.64	75.20	90.84
SSR [15]	80.92	92.80	75.76	91.76
<b>LSL (Ours)</b>	<b>81.40</b>	<b>93.00</b>	<b>77.00</b>	91.84

## A noisy elephant in the room: Is your out-of-distribution detector robust to label noise?

- 概要: Out-of-Distribution(OOD)における様々なシナリオでの大規模な実験
- 手法: CIFR10, CIFAR-100-Coarse, CIFAR-100-Fine, Clothing1Mなどのデータセット, ノイズの種類とラベル, アーキテクチャ、にわたり大規模な検証実験を実施
- 新規性: クラスラベルノイズの存在下で, OOD検出手法の性能は大幅に低下することが示された。



## Leveraging Cross-Modal Neighbor Representation for Improved CLIP Classification

- **概要:** CLIPの性能を向上させるために、クロスモーダルの隣接表現(CODER)を活用する方法を提案
- **手法:** 画像と生成テキストの埋め込みのコサイン類似度を用いて画像の隣接表現を構築し初期分類結果を再ランク付け
- **新規性:** 各データセットとモデルアーキテクチャにおいて一貫してCLIPのzero/few shot画像分類精度を向上

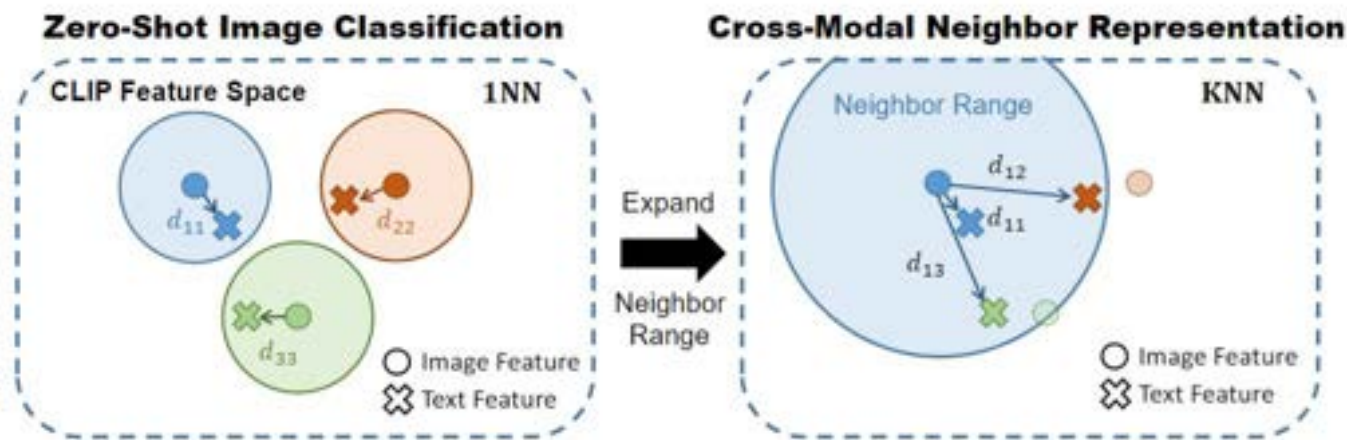


Figure 1. Illustration of image's Cross-modal Neighbor Representation (CODER). CLIP's powerful text-image matching capabilities endows it with a favorable cross-modal neighbor distance relation. And CLIP's Zero-Shot Image Classification process can be interpreted as using a 1NN algorithm to find the image's nearest text, with the text's class determining the image's predicted class. Inspired by this idea, we expand the image's neighbor range to leverage its distance to all texts for constructing the CODER. Here  $d_{ij}$  refers to the distance between the  $i$ -th image and the  $j$ -th text.

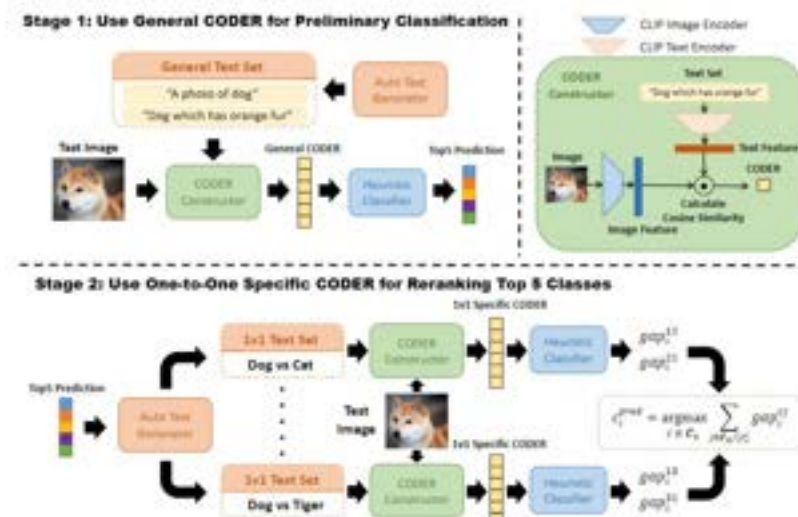


Figure 3. Illustration of two-stage zero-shot image classification process based on image's CODER. In the first stage, we use the Auto Text Generator to create a General Text Set, which contains general descriptions of classes. This set is utilized to construct the image's general CODER, and we use it for preliminary classification. In the second stage, we construct One-to-One Text Sets in pairs for the top five predicted classes of the preliminary classification results, focusing on attributes where the two specific classes differ most. We build one-to-one specific CODER for the image based on these One-to-One Text Sets and use heuristic classifier to get each class's classification score. Then we rerank the top five preliminary results based on the classification score gaps  $gap_j^{1,2}$  between classes. Here  $gap_j^{1,2}$  represents the difference obtained by subtracting the score of the class  $e$  from that of the class  $j$  for image  $x_i$ .



## A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions

- **概要:** CLIPスタイルのモデルが密なキャプション付き画像に対してどのように性能を発揮するかを評価
- **手法:** 7805枚の自然画像に対して、各画像に1000語以上の詳細なキャプションを人間がアノテートしたDensely Captioned Images (DCI) データセットを収集・ファインチューニング
- **新規性:** DCIデータセットで微調整されたモデルはCLIPベースラインを超える分類性能、BLIPやDACにも近いベンチマークを示した。

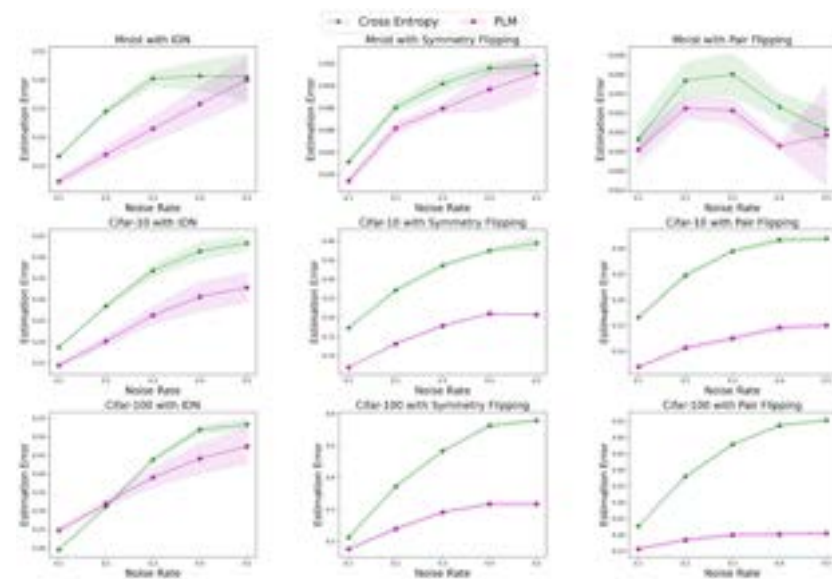
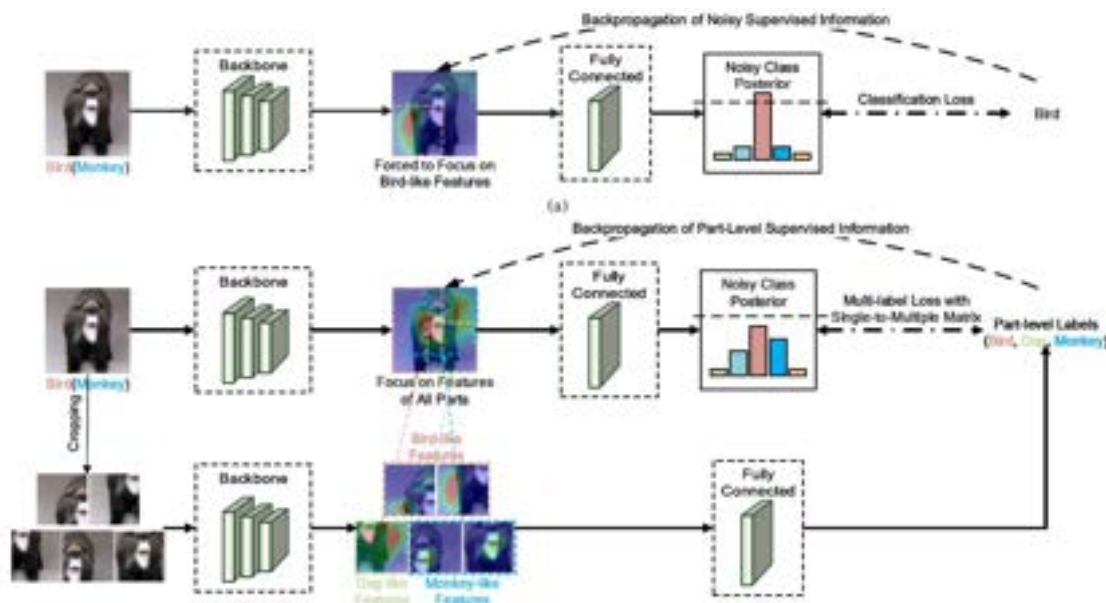


Model	ARO				VL-Checklist		
	VG-R	VG-A	COCO	Flicker	Object	Attribute	Relation
sDCI <sub>P1</sub>	76.23%	67.56%	88.58%	91.30%	80.71%	68.69%	70.12%
sDCI <sub>P1,NL0</sub>	57.34%	61.98%	39.36%	44.62%	88.37%	70.42%	61.28%
DAC <sub>LLM<sub>10,000</sub></sub>	61.53%	63.89%	46.28±1.5%	59.41±1.9%	66.90%	57.4%	56.96%
DAC <sub>LLM<sub>100,000</sub></sub>	61.0%	63.6%	48.2%	61.42%	66.87%	57.22%	57.18%
DAC <sub>LLM<sub>1,000,000</sub></sub>	60.1%	63.8%	50.2%	61.6%	66.54%	57.39%	56.77%
DAC <sub>LLM<sub>10,000,000</sub></sub>	<b>81.28%</b>	<b>73.91%</b>	<b>94.47%</b>	<b>95.68%</b>	87.30%	<b>77.27%</b>	86.41%
DAC <sub>SAM<sub>1,000,000</sub></sub>	77.16%	70.5%	91.22%	93.88%	<b>88.50%</b>	75.83%	<b>89.75%</b>
CLIP Baseline [27]	59.98%	63.18%	47.9%	60.2%	81.17%	67.67%	61.95%
BLIP2 [19]	41.16%	71.25%	13.57%	13.72%	84.14%	80.12%	70.72%
NegCLIP [42]	81%	71%	86%	91%	81.35%	72.24%	63.53%
SVLC [11]	80.61%	73.03%	84.73%	91.7%	85%	71.97%	68.95%

Table 3. sDCI fine-tuned CLIP performance against the ARO and VL-Checklist benchmark. We compare CLIP fine-tuned with sDCI against models fine-tuned using DAC captions. Since the DAC dataset contains 3M images whereas sDCI contains only 7805 images, we performed an ablation of the number of training images used in the DAC dataset. In this instance, DAC<sub>LLM<sub>10,000</sub></sub> refer to fine-tuning CLIP using only 10,000 images from DAC. We plot the mean across 5 different seeds and display the standard deviation when it is above 1% accuracy. We observe that training on sDCI lead to significant improvement in comparison to DAC for a comparable number of examples.

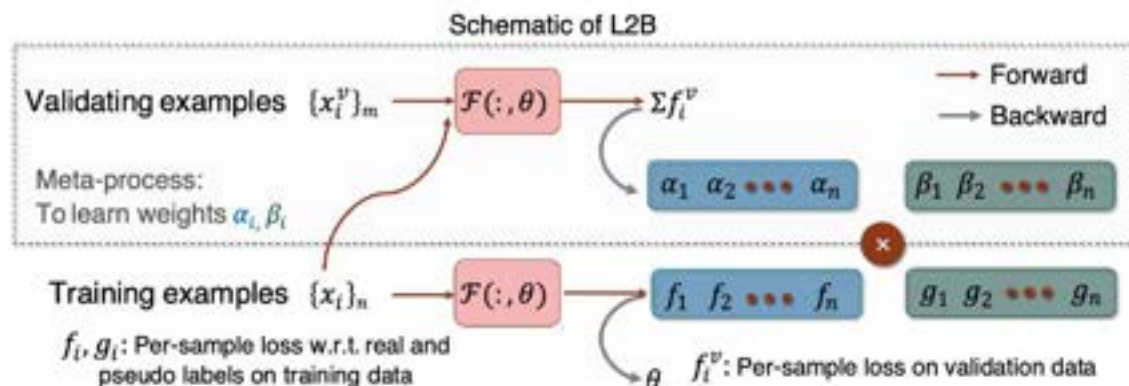
## Estimating Noisy Class Posterior with Part-level Labels for Noisy Label Learning

- 概要:** 部分レベルのラベルを用いてモデルが多様な情報を統合できるようにし、ノイズクラス事後確率をより正確に学習する方法”Part-Level Multi-labeling(PLM)”を提案
- 手法:** インスタンスをクロッピングして特徴が異なる部分に分割し、これらの部分に対応する部分ラベルを生成。これを用いてモデルを一貫性のある分類フレームワークに統合する
- 新規性:** 単一から複数への遷移行列を導入し、部分レベルのラベルを一貫性のある分類フレームワークに統合することで分類精度およびロバスト性が向上



## L2B: Learning to Bootstrap Robust Models for Combating Label Noise

- 概要:** 深層学習におけるノイズの多いラベルに過学習しやすい問題に対し、モデル自身の予測を利用してブートストラップを行う”Learning to Bootstrap(L2B)”を提案。L2Bは自然および医療画像タスクで優れた結果を出し、ノイズの多いラベルによる影響の軽減を実現
- 手法:** モデルが自分自身の予測を利用してブートストラップすることを可能にする手法で、メタラーニングを通じて実際の観測で得たラベルと生成されたラベル、および異なるサンプル間の重要度を動的に調整
- 新規性:** 固定された重みを使用せずにラベルやインスタンスを再重み付けすることで、従来のブートストラップ手法と異なる柔軟性を提供し、モデルのロバスト性が向上



Dataset	CIFAR-10				CIFAR-100				ISIC			
	20%	30%	40%	50%	20%	30%	40%	50%	20%	30%	40%	50%
Method/Noise ratio	20%	30%	40%	50%	20%	30%	40%	50%	20%	30%	40%	50%
Cross-Entropy (CE)	86.9	84.9	83.3	81.3	59.6	52.2	49.2	44.4	79.4	77.5	75.3	73.7
Bootstrap [37]	85.2	84.8	82.9	79.2	61.8	54.2	50.2	45.8	80.8	77.7	75.7	74.8
L2RW [38]	90.6	89.0	86.6	85.3	67.8	63.8	59.7	55.6	80.1	77.7	76.3	74.1
L2B (Ours)	92.2	90.7	89.9	88.5	71.8	69.5	67.3	64.5	81.1	80.2	78.6	76.8

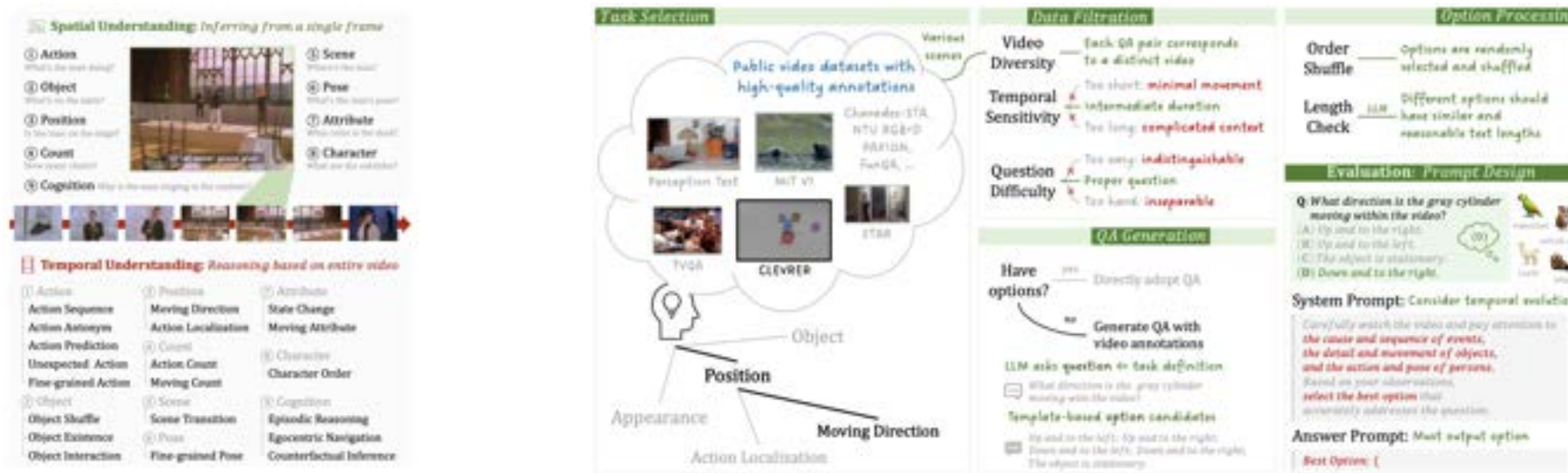
## Step Differences in Instructional Video

- **概要**: 同じ作業をしている2つの動画について, 対応する工程の作業手順の細かい差異を記述するという問題設定およびモデルを提案. 差異記述のデータはキャプション, 物体検出などを使いつつLLMで出力をまとめながら作業手順を記述し, LLMにより差異の記述を生成している.
- **手法**: 用意したデータを利用してLLMをinstruct tuningすることで差異記述モデルを構築.
- **新規性**: 作業手順の差異記述という新しい問題設定の提案とデータおよびベンチマークの構築.



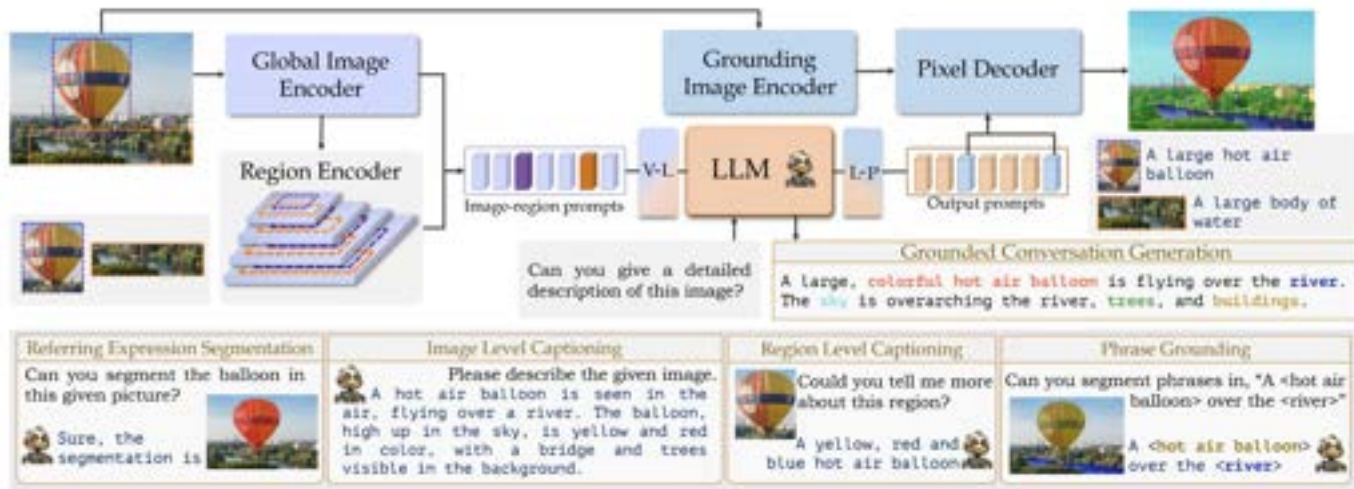
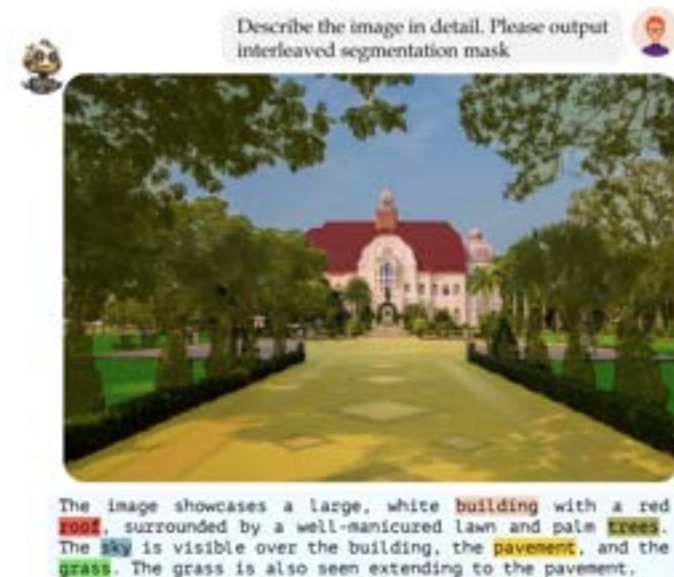
## MVBench: A Comprehensive Multi-modal Video Understanding Benchmark

- **概要**: マルチモーダル大規模言語モデル (MLLM) の時間的な理解も評価できるベンチマークを構築. 従来のベンチマークは静止画を用いた空間的な理解を主に評価しており時間的な理解については見過ごされていた. 提案するベンチマークは1枚の静止画情報だけでは解けないような20タスクをカバーしている. 加えて, ベースラインとしてVideoChat2というモデルを構築し時間的な理解の評価において従来のMLLMよりも高い精度を達成.
- **手法**: 既存の11動画データセットから短い動画を省く, 簡単すぎる・難しすぎる質問を省くなどしつつデータを収集し, 多肢選択問題に変換. 選択肢には正解の行動と反対の行動を入れるなどして時間的な理解を評価するよう問題を設計.
- **新規性**: 動画を用いたMLLM評価のためのベンチマークの構築.



## GLaMM: Pixel Grounding Large Multimodal Model

- **概要**: 画素単位での接地が可能なMLLMである **GLaMM** を提案  
 参照表現理解, キャプション生成などの下流タスクに応用可能  
 画像中の物体を画素単位で特定し自然な会話を生成するベンチマーク  
**Grounded Conversation Generation (GCG)** タスクを提案
- **手法**: **Grounding-anything Dataset (Grand)** の構築およびGrand上でGLaMMを学習
  - 810Mの領域に7.5Mのuniqueな概念を含む大規模データセット
  - 自動アノテーションパイプラインを使用して効率的に構築
- **新規性**: 大規模な視覚言語の接地を目的としたデータセットの自動構築手法の確立



Model	Validation Set					Test Set				
	M	C	AP50	mIoU	Recall	M	C	AP50	mIoU	Recall
BuboGPT [59]	17.2	3.6	19.1	54.0	29.4	17.1	3.5	17.3	54.1	27.0
Kosmos-2 [35]	16.1	27.6	17.1	55.6	28.3	15.8	27.2	17.2	56.8	29.0
LISA* [21]	13.0	33.9	25.2	62.0	36.3	12.9	32.2	24.8	61.7	35.5
GLaMM	15.2	43.1	28.9	65.8	39.6	14.6	37.9	27.2	64.6	38.0
GLaMM	16.2	47.2	30.8	66.3	41.8	15.8	43.5	29.2	65.6	40.8

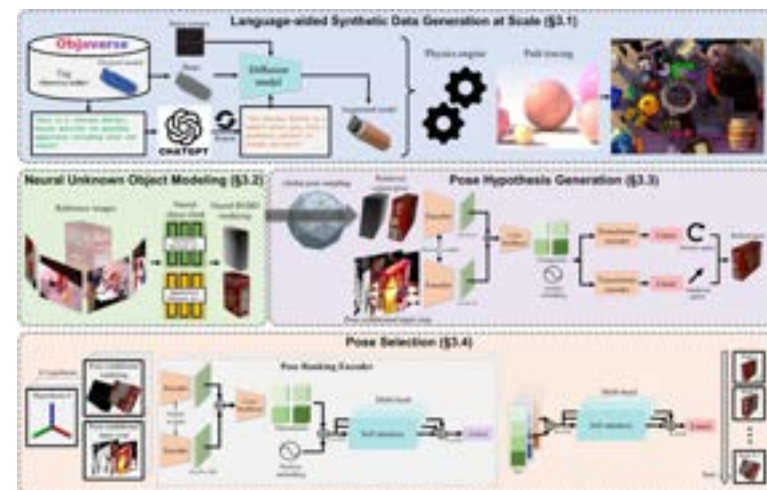
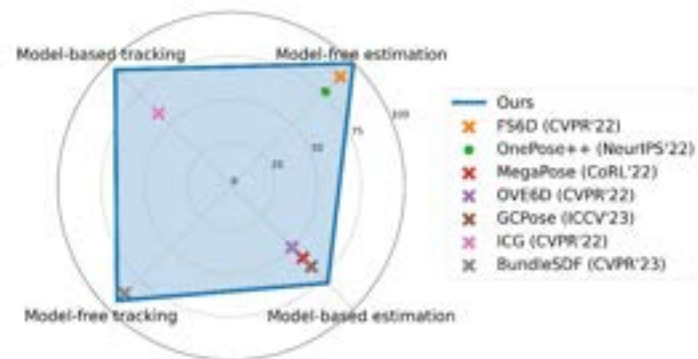
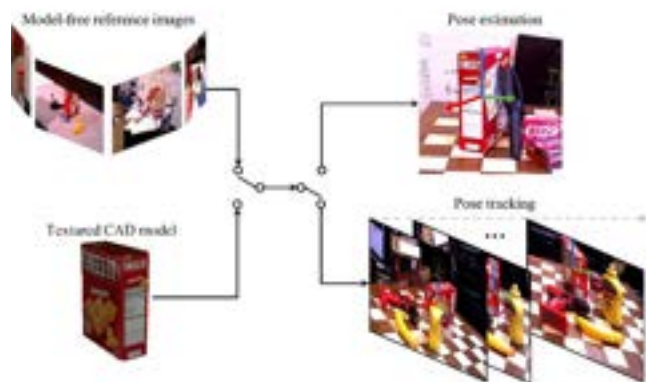
Table 3. Performance on GCG Task. Metrics include METEOR (M), CIDEr (C), AP50, mIoU, and Mask Recall. LISA\* denotes LISA adapted for GCG. GLaMM denotes training excluding 1K human annotated images. GLaMM shows better performance.

Method	refCOCO			refCOCO+			refCOCog	
	val	testA	testB	val	testA	testB	val(U)	test(U)
CRIS [47]	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
LAVT [50]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES [26]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder [63]	-	-	-	-	-	-	64.6	-
SEEM [64]	-	-	-	-	-	-	65.7	-
LISA-7B [21]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
GLaMM	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9

Table 4. Qualitative Assessment of GLaMM in Referring-Expression Segmentation. Performance across refCOCO, refCOCO+, and refCOCog in generating accurate segmentation masks based on text-based referring expressions surpasses that of closely related work, including LISA which is specifically designed for this task.

## FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects

- ❑ **概要:** 対象物体の6Dポーズ推定とトラッキングにおいて、**model-baseおよびmodel-freeの両方で使用できる** 手法を提案。現時点で未学習物体に対するModel-baseの6Dポーズ推定の検出手法でSoTA (BOP: Benchmark for 6D Object Pose EstimationのModel-based 6D localization of unseen objects)
- ❑ **新規性:** 従来の手法は、Model-base(推定対象物体の3Dモデルから推定する)もしくはModel-free(対象物体の複数の2D画像から推定する)な手法のどちらかにしか適応できなかった。
- ❑ **手法:** データセットの生成において、LLMとDiffusion Modelを用いた拡張を行った。chatGPTに、対象オブジェクトの取りうる外観について説明してほしいというプロンプトを入力し、出力を得る。この出力をDiffusion Modelに入力し、プロンプトに合ったテクスチャを生成する。またポーズ推定部ではまたポーズ推定部では、transformerのencoderを使用し、まずポーズの候補をいくつか推定し、推定したポーズのスコアをそれぞれで算出し、最もスコアが高いポーズをモデルの出力とする。



## Clustering for Protein Representation Learning

- **概要**: タンパク質における重要な構成要素(アミノ酸配列)を一次元的構造、三次元的構造から自動的にタンパク質の機能に重要な構成要素発見できるニューラルクラスタリングというフレームワークを提案。
- **新規性**: ニューラルクラスタリングフレームワークをタンパク質の表現学習に適用した。
- **感想**: GNNベースの手法と本手法の得意/不得意分野などの関係がどのようになっているのかが気になった。

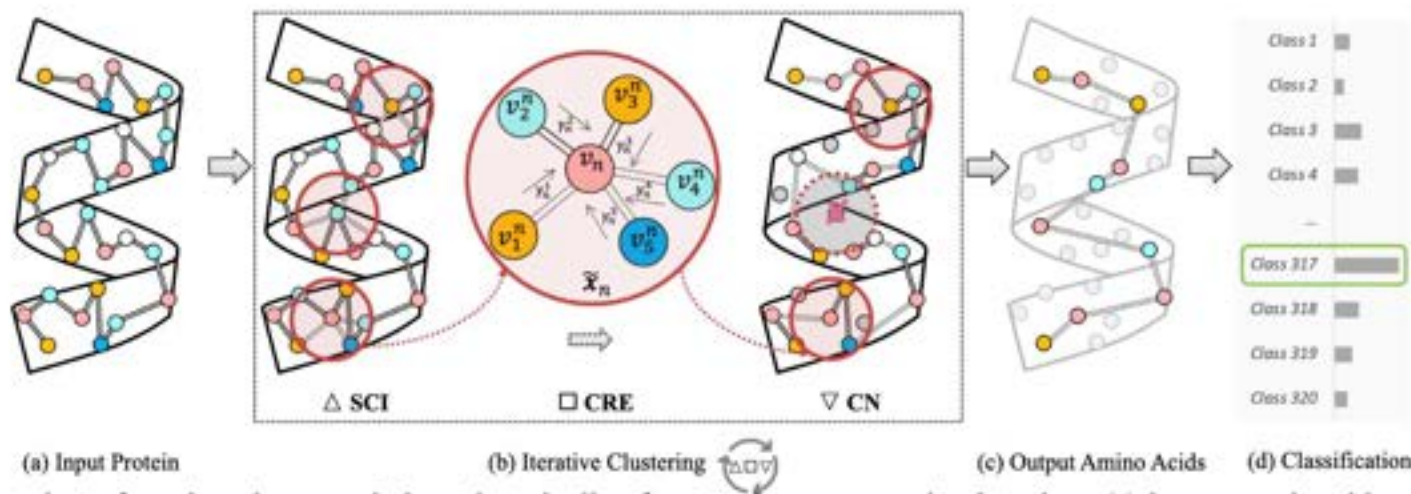
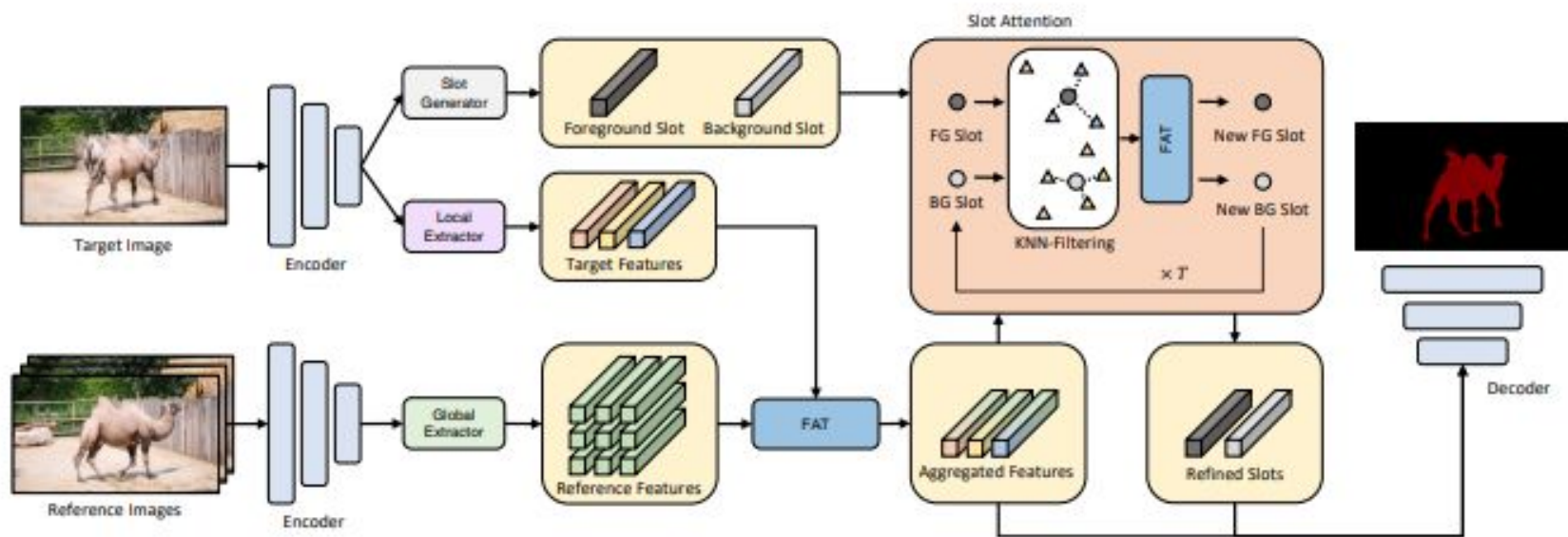


Figure 1. Overview of our iterative neural clustering pipeline for protein representation learning: (a) input protein with amino acids, (b) iterative clustering algorithm which repeatedly stacks three steps  $\Delta \square \nabla$ , (c) output can be seen as the critical amino acids of the protein, (d) output amino acids used for classification. The details of our iterative neural clustering method can be seen in §3.2.



## Guided Slot Attention for Unsupervised Video Object Segmentation

- ❑ 教師なしビジュアルセグメンテーション(VOS)に注目
  - ❑ ビデオ内のオブジェクトをフレームごとにセグメント化
  - ❑ 入力映像内のオブジェクトを探索し効果的に抽出する必要があり、半教師ありより困難
- ❑ 従来手法:動きの手がかりに過度に依存し色、テクスチャ、形状の構造情報を見落とす
- ❑ 本手法:k-最近傍フィルタリングと特徴集約変換を組み合わせることで前景と背景を分離



## 今後の展望

---

- 今後, CV分野はどうか？
- 我々としてはどうすれば良いか？

# 今後の展望(1/2)

---

## 「AI基盤モデル構築 → 適応学習」 全盛時代突入！？

- どのセッションにも必ず言及されていると思うくらいありふれている
  - CLIP/SAMなどはみんな使ってその上に構築しようという意気込み
- AI基盤モデル構築は特定研究機関に限定されている？
  - スクラッチでAI基盤モデル構築するという例は少数に限られる(今までもそう)
- 適応学習が面白い？
  - ロボットや医療などにも適応学習する例が多く見られる
  - 他にも, 基盤モデルによりはじめて実装される科学分野に関する発表が見られた

個別の面白い問題を見つけよう！

# 今後の展望(2/2)

## CVPR には参加者 ≡ 発表者になると良い？

- ❑ 日本ドメインからの参加者は約350人に対して日本人著者が入っているとされる論文数は60件前後
  - ❑ 参考1(CVPR 2024 Opening Remarks) <https://media.eventhosts.cc/Conferences/CVPR2024/OpeningRemarkSlides.pdf>
  - ❑ 参考2(ResearchPort) :<https://research-p.com/column/1889>
- ❑ みんなで投稿しよう！
  - ❑ CVPR来たら仲間を見つけて連携研究しよう
  - ❑ 研究したら論文を書こう
  - ❑ 論文書いたら何度も修正しよう
  - ❑ 修正したら論文投稿しよう
  - ❑ 以下, 繰り返して経験を蓄積しよう
  - ❑ 最後は日本からたくさんCVPRなどに採択されよう！

CV分野における日本のプレゼンスを上げよう！

※ワークショップ企画提案もどんどん投稿しましょう



ご質問・コメント等ありましたら, [cvpaper.challenge@gmail.com](mailto:cvpaper.challenge@gmail.com) / X @ CVPaperChalleng まで