CVPR 2025 Report

Hirokatsu Kataoka, Yoshihiro Fukuhara,

Ryousuke Yamada, Daichi Otsuka, Rintaro Yanagi, Oishi Deb,
Kazuya Nishimura, Moeri Okuda, Yuto Matsuo, Ren Ohkubo, Yue Qiu, Noritake Kodama,
Gido Kato, Kenzo Yamabuki, Joe Hasei, Ryuichi Nakahara, Yukinori Yamamoto, Sho
Okazaki, Kohsuke Ide, Yuiga Wada,
Daichi Yashima, Shinichi Mae, Hinako Mitsuoka, Maika Takada,
Jianyuan Wang

LIMIT.Lab / cvpaper.challenge / Visual Geometry Group (VGG)

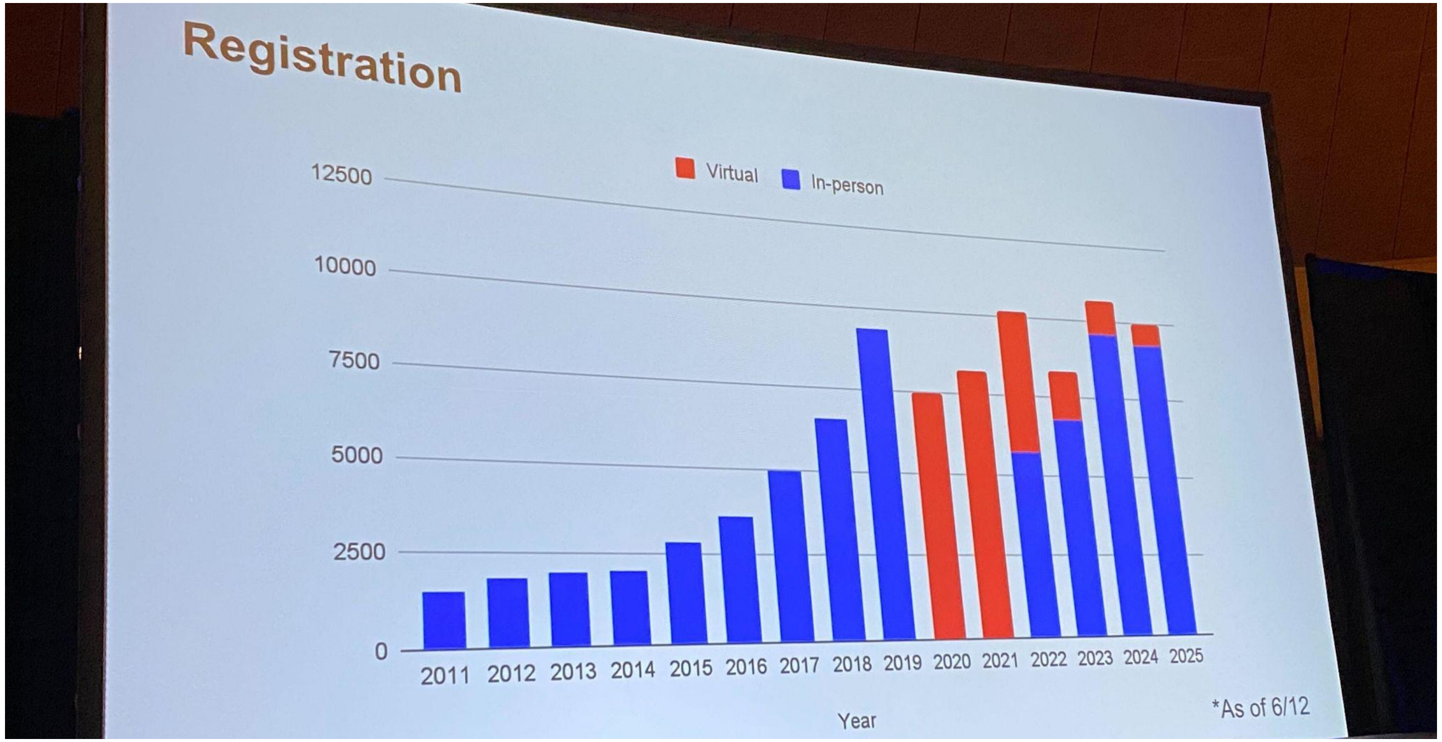
Meta Insights into Trends and Tendencies in CVPR 2025

- What kind of research is currently trending?
- What are overseas researchers working on?
- We have compiled the “trends” and “insights.”

CVPR25: Meta Insights into Trends and Tendencies (1/179)

From opening slide

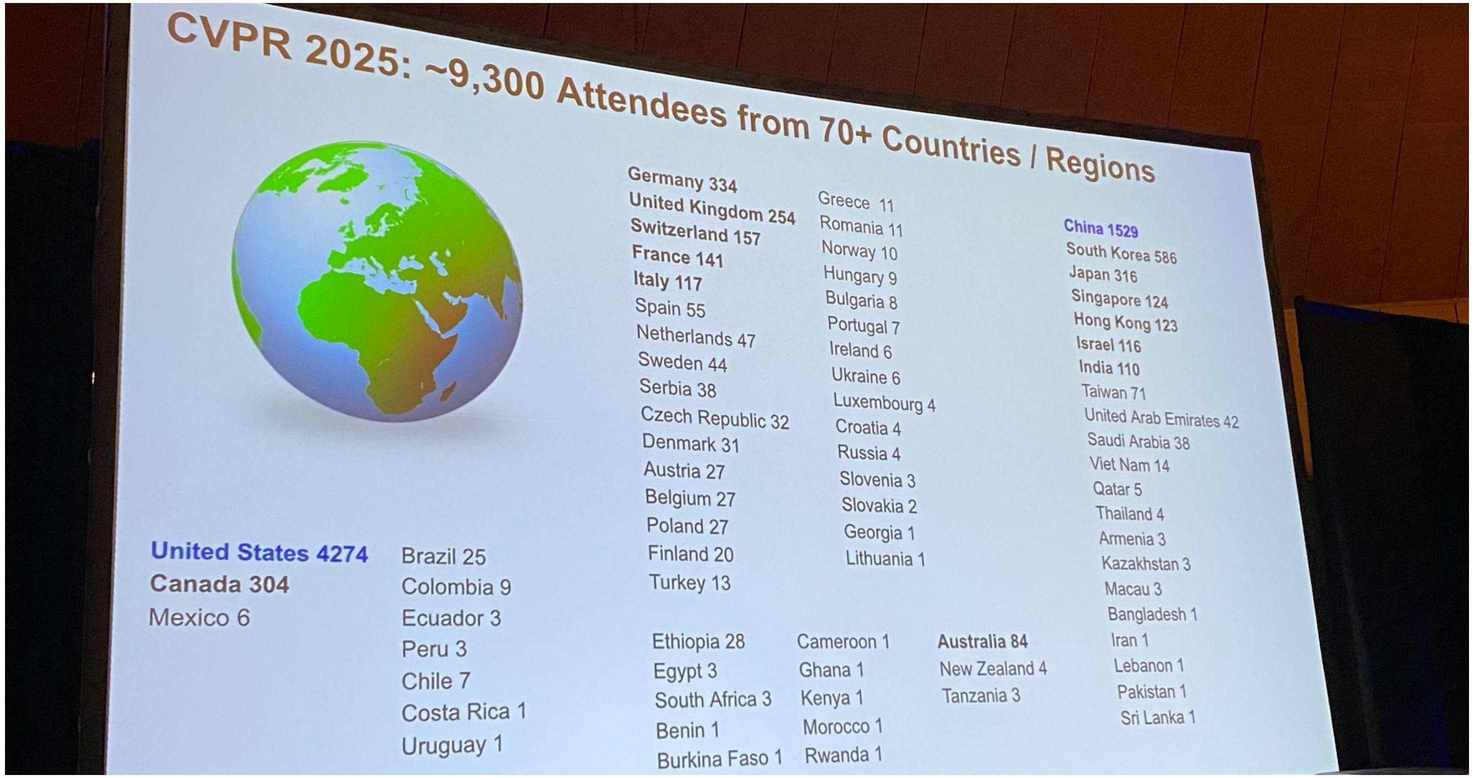
- The attendees slightly decreased this year
  - By comparing to the CVPR 2024 attendees
- Virtual attendance is being allowed



CVPR25: Meta Insights into Trends and Tendencies (2/179)

From opening slide

- ❑ Listed at each continent
  - ❑ US: 4,274, China: 1,529, South Korea: 586, Germany: 334, Japan: 316, Canada: 304, United Kingdom: 254
  - ❑ Totally 70+ countries/regions



CVPR 2025: ~9,300 Attendees from 70+ Countries / Regions

United States 4274
Canada 304
Mexico 6

Brazil 25
Colombia 9
Ecuador 3
Peru 3
Chile 7
Costa Rica 1
Uruguay 1

Germany 334
United Kingdom 254
Switzerland 157
France 141
Italy 117
Spain 55
Netherlands 47
Sweden 44
Serbia 38
Czech Republic 32
Denmark 31
Austria 27
Belgium 27
Poland 27
Finland 20
Turkey 13

Greece 11
Romania 11
Norway 10
Hungary 9
Bulgaria 8
Portugal 7
Ireland 6
Ukraine 6
Luxembourg 4
Croatia 4
Russia 4
Slovenia 3
Slovakia 2
Georgia 1
Lithuania 1

Ethiopia 28
Egypt 3
South Africa 3
Benin 1
Burkina Faso 1

Cameroon 1
Ghana 1
Kenya 1
Morocco 1
Rwanda 1

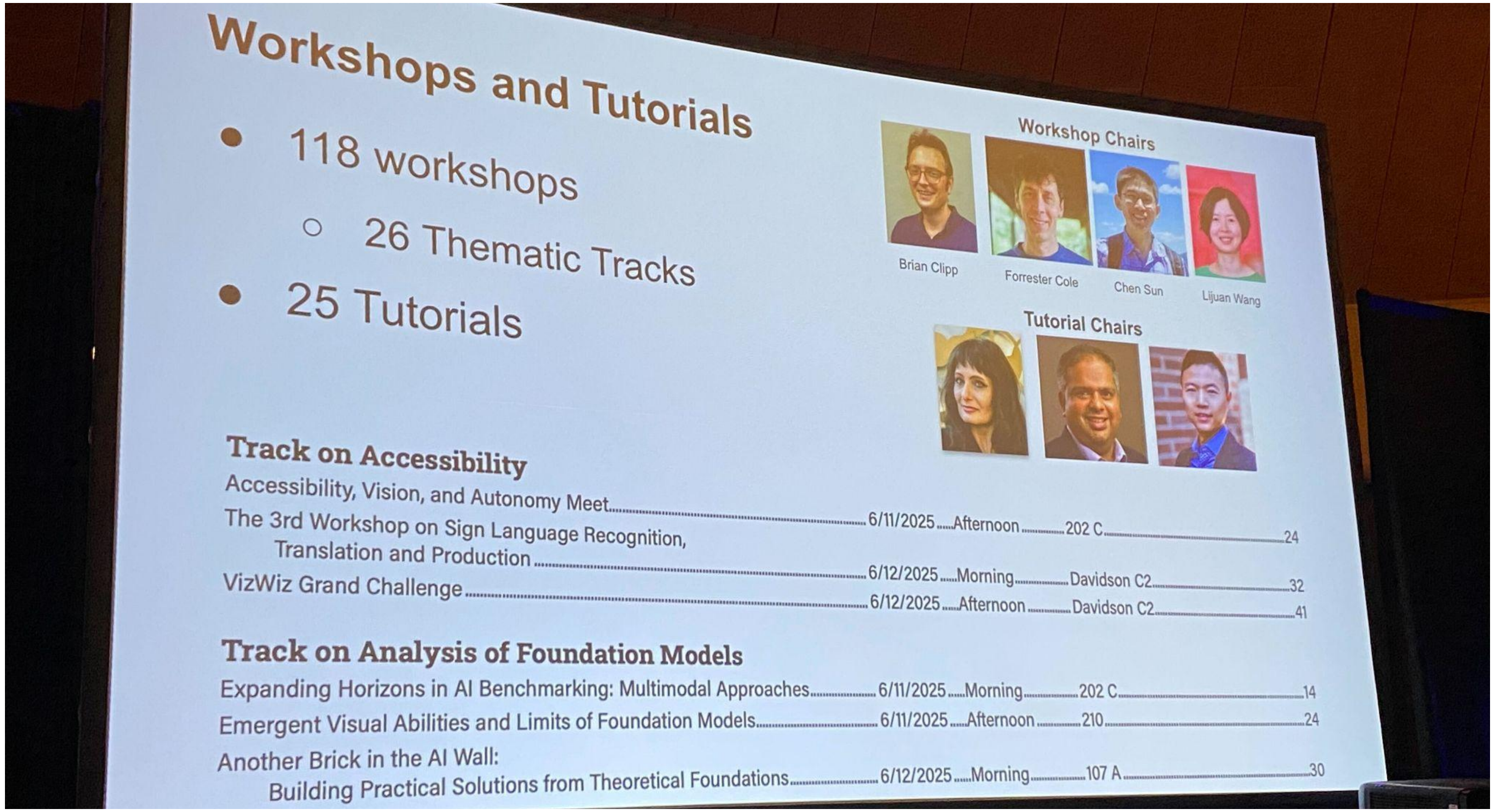
Australia 84
New Zealand 4
Tanzania 3

China 1529
South Korea 586
Japan 316
Singapore 124
Hong Kong 123
Israel 116
India 110
Taiwan 71
United Arab Emirates 42
Saudi Arabia 38
Viet Nam 14
Qatar 5
Thailand 4
Armenia 3
Kazakhstan 3
Macau 3
Bangladesh 1
Iran 1
Lebanon 1
Pakistan 1
Sri Lanka 1

CVPR25: Meta Insights into Trends and Tendencies (3/179)

From opening slide

- The #workshops slightly decreased from the previous year
  - 2024: 123 vs. 2025: 118 workshops
- The 118 ws have been divided into 26 tracks



Workshops and Tutorials

- 118 workshops
  - 26 Thematic Tracks
- 25 Tutorials

Workshop Chairs

Brian Clipp
Forrester Cole
Chen Sun
Lijuan Wang

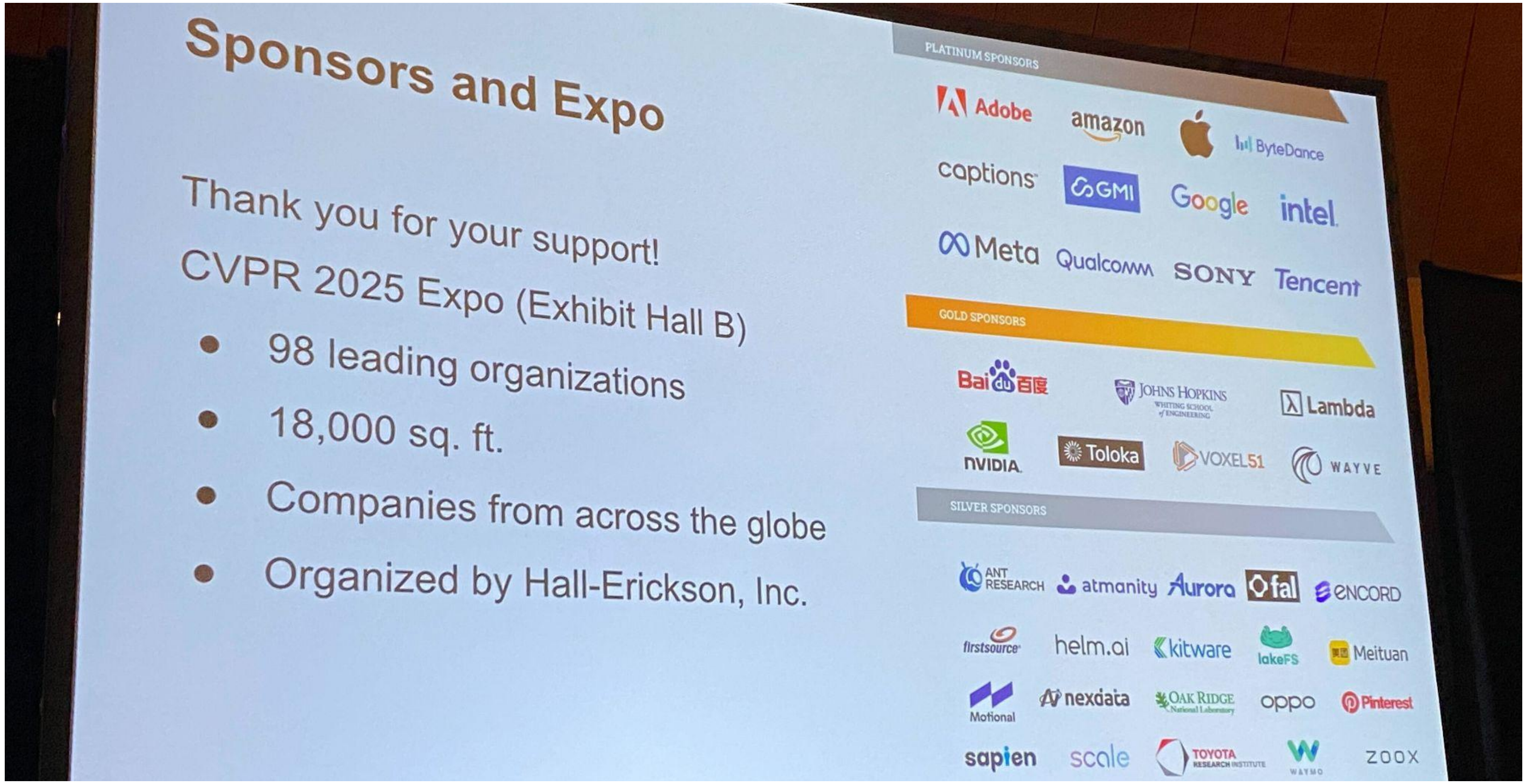
Tutorial Chairs

Track on Accessibility				
Accessibility, Vision, and Autonomy Meet	6/11/2025	Afternoon	202 C	24
The 3rd Workshop on Sign Language Recognition, Translation and Production	6/12/2025	Morning	Davidson C2	32
VizWiz Grand Challenge	6/12/2025	Afternoon	Davidson C2	41
Track on Analysis of Foundation Models				
Expanding Horizons in AI Benchmarking: Multimodal Approaches	6/11/2025	Morning	202 C	14
Emergent Visual Abilities and Limits of Foundation Models	6/11/2025	Afternoon	210	24
Another Brick in the AI Wall: Building Practical Solutions from Theoretical Foundations	6/12/2025	Morning	107 A	30

CVPR25: Meta Insights into Trends and Tendencies (4/179)

From opening slide

- You can see the leading companies in CVPR from the list!
- Thanks to the companies, the CVPR community has been accelerated



Sponsors and Expo

Thank you for your support!

CVPR 2025 Expo (Exhibit Hall B)

- 98 leading organizations
- 18,000 sq. ft.
- Companies from across the globe
- Organized by Hall-Erickson, Inc.

PLATINUM SPONSORS

Adobe, amazon, Apple, ByteDance, captions, GMI, Google, intel, Meta, Qualcomm, SONY, Tencent

GOLD SPONSORS

Baidu 百度, JOHNS HOPKINS WHITING SCHOOL of ENGINEERING, Lambda, NVIDIA, Toloka, VOXEL51, WAYVE

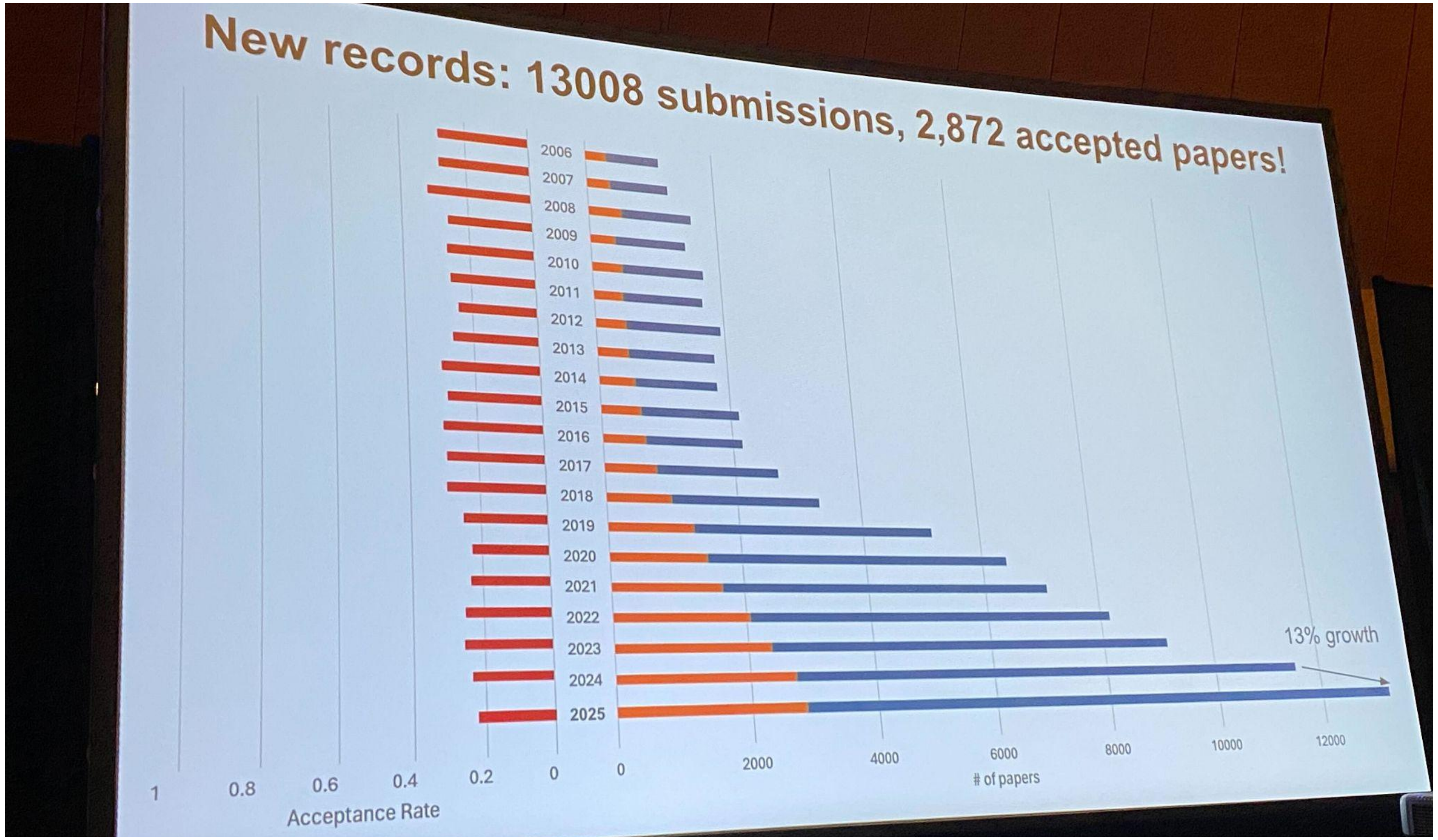
SILVER SPONSORS

ANT RESEARCH, atmanity, Aurora, fal, ENCORD, firstsource, helm.ai, kitware, lakeFS, Meituan, Motional, nexdata, OAK RIDGE National Laboratory, oppo, Pinterest, sapien, scale, TOYOTA RESEARCH INSTITUTE, WAYMO, zoox

CVPR25: Meta Insights into Trends and Tendencies (5/179)

From opening slide

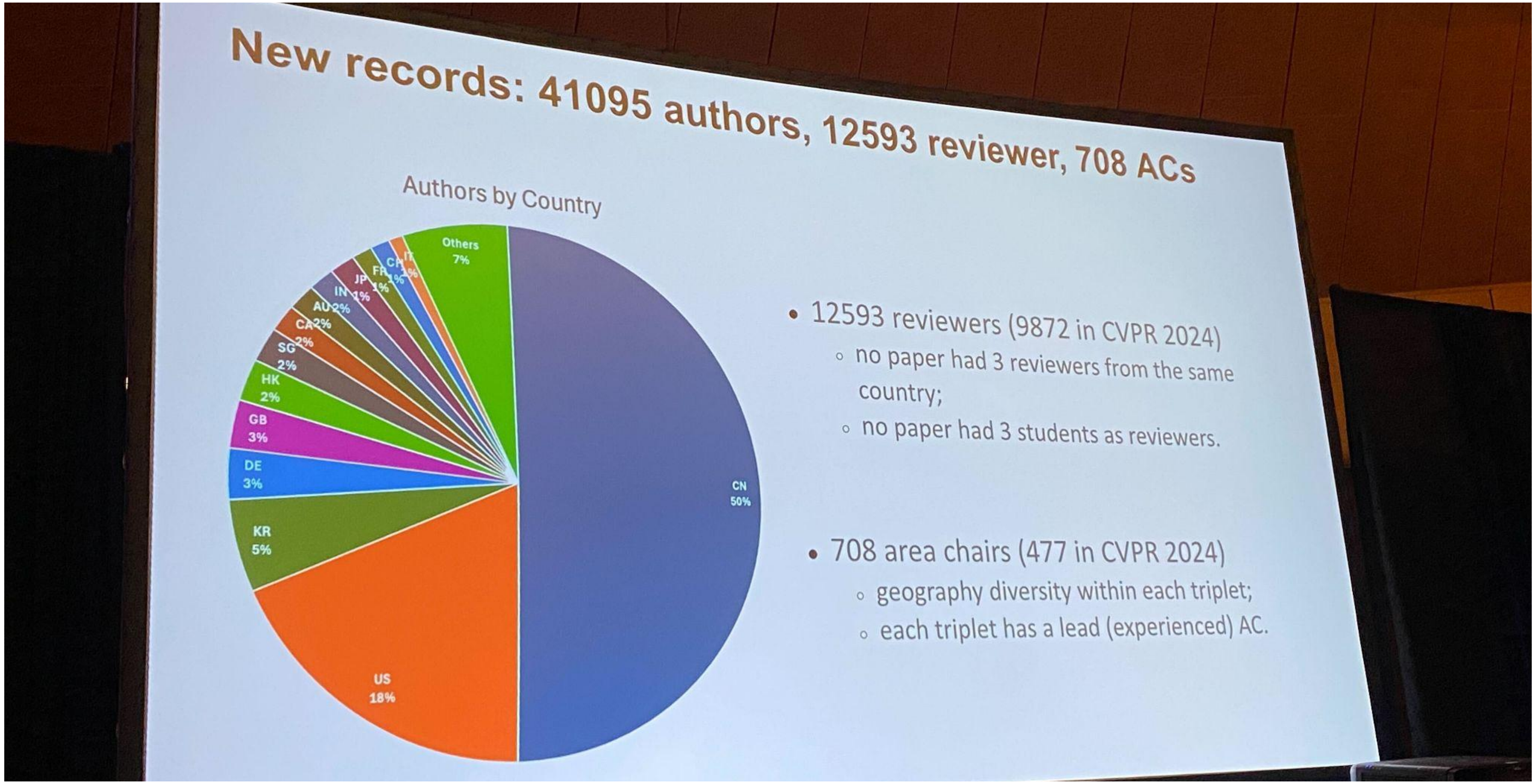
- ❏ #Submissions – 2024: 11,532 vs. 2025: 13,008 (12.8% increased)
- ❏ Acceptance rate – 2024: 23.55% vs. 2025: 22.08% (1.47 pt decreased)



CVPR25: Meta Insights into Trends and Tendencies (6/179)

From opening slide

- ❏ Rapidly growing CVPR community!
  - ❏ in terms of #authors, #reviewers, and #area chairs



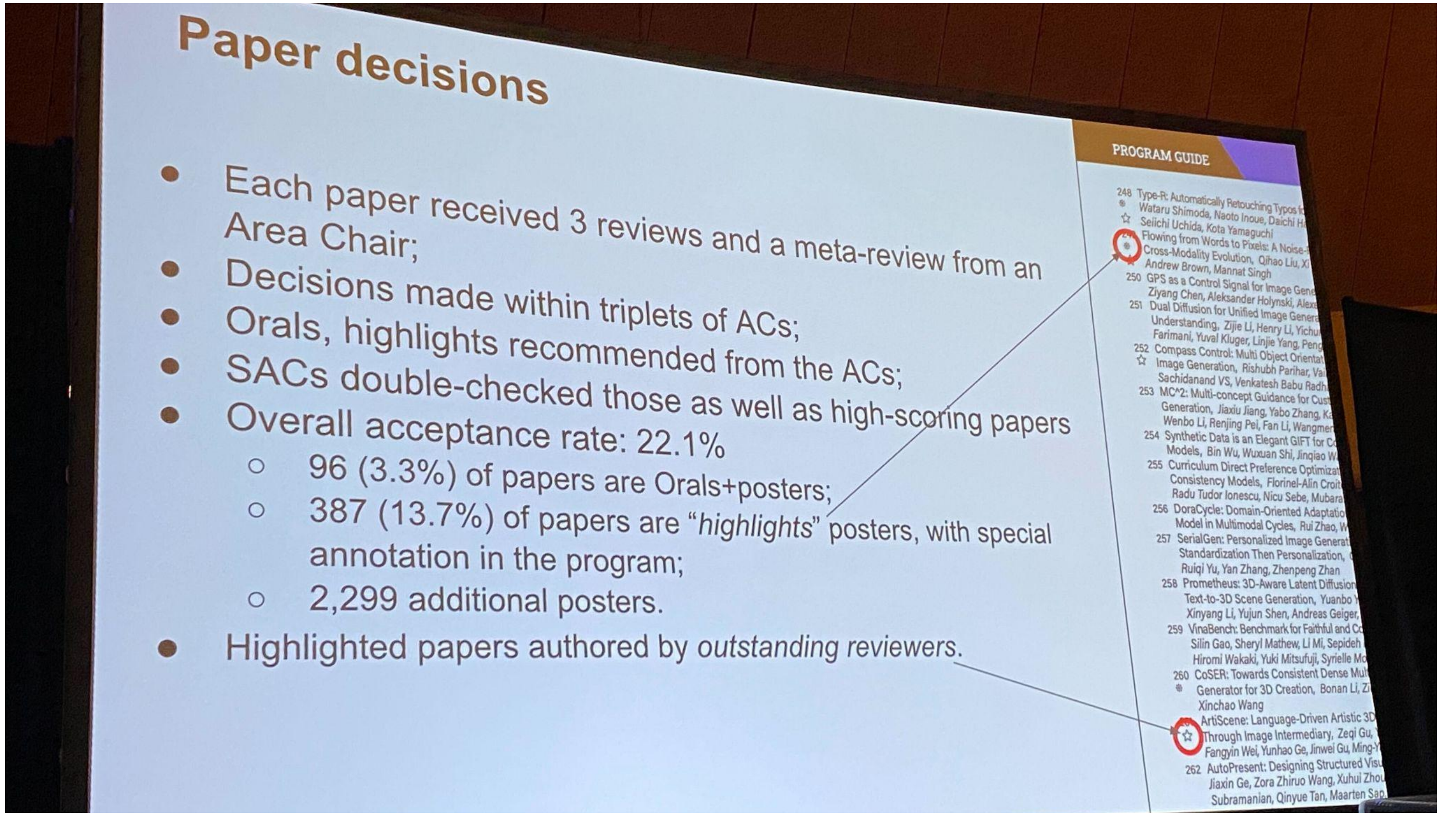
New records: 41095 authors, 12593 reviewer, 708 ACs

- 12593 reviewers (9872 in CVPR 2024)
  - no paper had 3 reviewers from the same country;
  - no paper had 3 students as reviewers.
- 708 area chairs (477 in CVPR 2024)
  - geography diversity within each triplet;
  - each triplet has a lead (experienced) AC.

CVPR25: Meta Insights into Trends and Tendencies (7/179)

From opening slide

- ❏ Newly induced CVPR paper from reviewer contribution
  - ❏ A paper included outstanding reviewer is highlighted
  - ❏ Review contributions may influence accepted papers!



Paper decisions

- Each paper received 3 reviews and a meta-review from an Area Chair;
- Decisions made within triplets of ACs;
- Orals, highlights recommended from the ACs;
- SACs double-checked those as well as high-scoring papers
- Overall acceptance rate: 22.1%
  - 96 (3.3%) of papers are Orals+posters;
  - 387 (13.7%) of papers are "*highlights*" posters, with special annotation in the program;
  - 2,299 additional posters.
- Highlighted papers authored by *outstanding reviewers.*

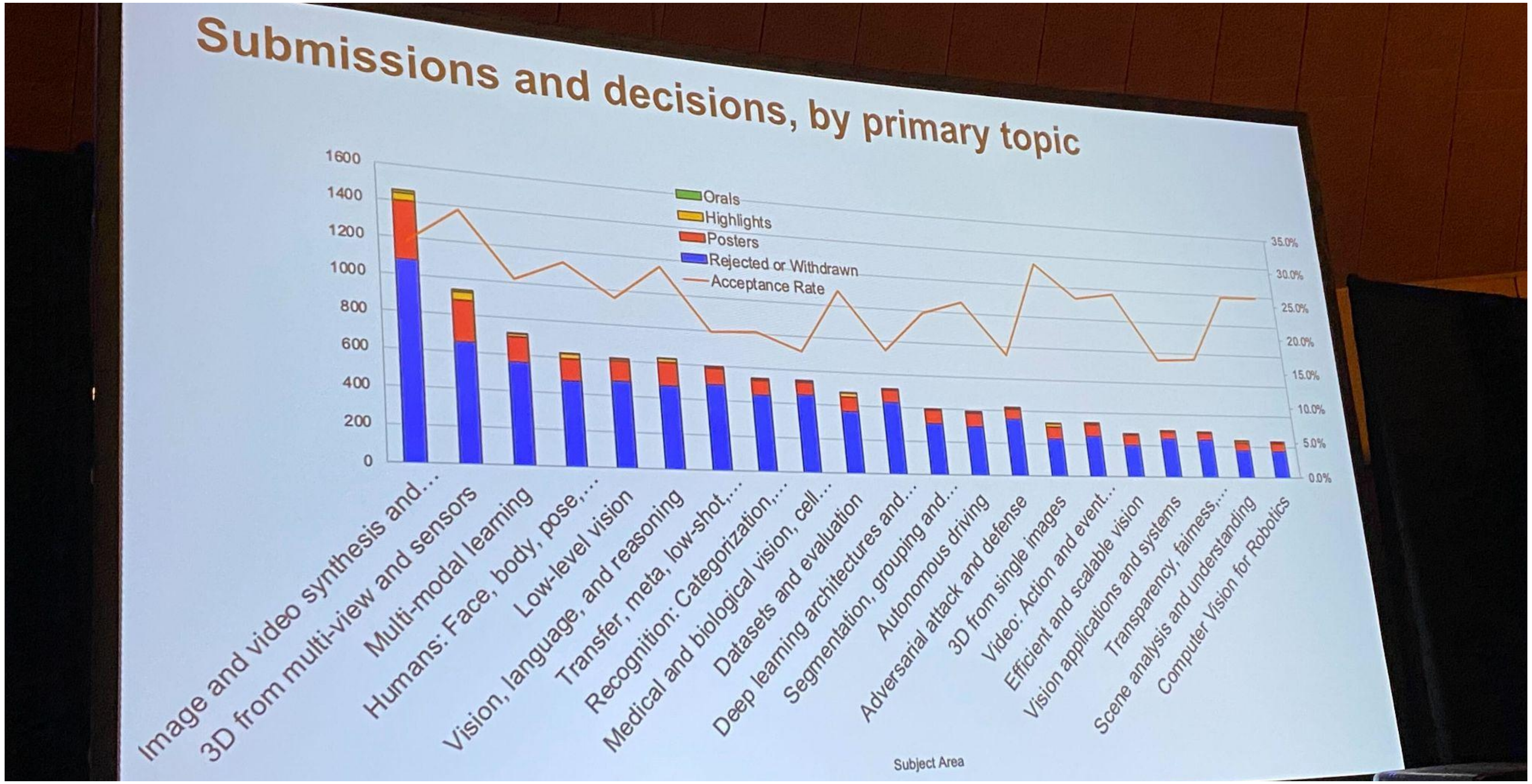
PROGRAM GUIDE

248 Type-R: Automatically Retouching Typos fo
Wataru Shimoda, Naoto Inoue, Daichi H
☆ Seiichi Uchida, Kota Yamaguchi
Flowing from Words to Pixels: A Noise-
Cross-Modality Evolution, Qihao Liu, Xi
Andrew Brown, Mannat Singh
250 GPS as a Control Signal for Image Gene
Ziyang Chen, Aleksander Holynski, Alexe
251 Dual Diffusion for Unified Image Genera
Understanding, Zijie Li, Henry Li, Yichu
Farimani, Yuval Kluger, Linjie Yang, Peng
252 Compass Control: Multi Object Orientat
☆ Image Generation, Rishubh Parihar, Vai
Sachidanand VS, Venkatesh Babu Radh
253 MC^2: Multi-concept Guidance for Cust
Generation, Jiaxiu Jiang, Yabo Zhang, Ka
Wenbo Li, Renjing Pei, Fan Li, Wangmen
254 Synthetic Data is an Elegant GIFT for Co
Models, Bin Wu, Wuxuan Shi, Jinqiao W
255 Curriculum Direct Preference Optimizat
Consistency Models, Florinel-Alin Croit
Radu Tudor Ionescu, Nicu Sebe, Mubara
256 DoraCycle: Domain-Oriented Adaptatio
Model in Multimodal Cycles, Rui Zhao, W
257 SerialGen: Personalized Image Generat
Standardization Then Personalization,
Ruiqi Yu, Yan Zhang, Zhenpeng Zhan
258 Prometheus: 3D-Aware Latent Diffusion
Text-to-3D Scene Generation, Yuanbo Y
Xinyang Li, Yujun Shen, Andreas Geiger,
259 VinaBench: Benchmark for Faithful and Co
Silin Gao, Sheryl Mathew, Li Mi, Sepideh
Hiromi Wakaki, Yuki Mitsufuji, Syrielle Mo
260 CoSER: Towards Consistent Dense Mul
Generator for 3D Creation, Bonan Li, Zi
Xinchao Wang
ArtiScene: Language-Driven Artistic 3D
☆ Through Image Intermediary, Zeqi Gu,
Fangyin Wei, Yunhao Ge, Jinwei Gu, Ming-Y
262 AutoPresent: Designing Structured Visu
Jiaxin Ge, Zora Zhiruo Wang, Xuhui Zhou
Subramanian, Qinyue Tan, Maarten Sap

CVPR25: Meta Insights into Trends and Tendencies (8/179)

From opening slide

- Listed the recent trends in CVPR
  - What are the next trends?

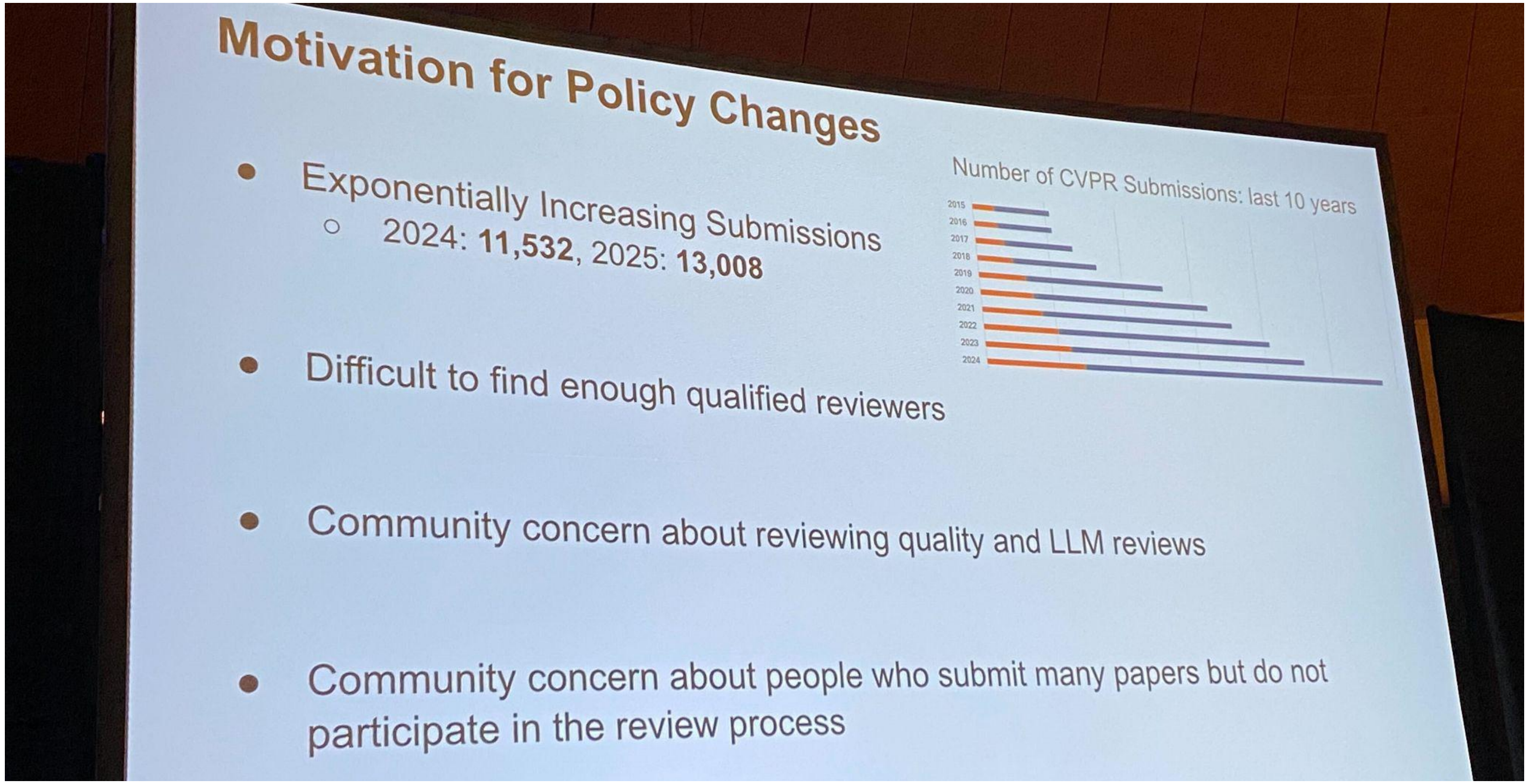


LIMIT.LAB
https://limitlab.xyz/

CVPR25: Meta Insights into Trends and Tendencies (9/179)

From opening slide

- The program committee is struggling to maintain the review quality
  - Some decisions for reviewing process
  - Qualified reviewers / reviewer education should be required



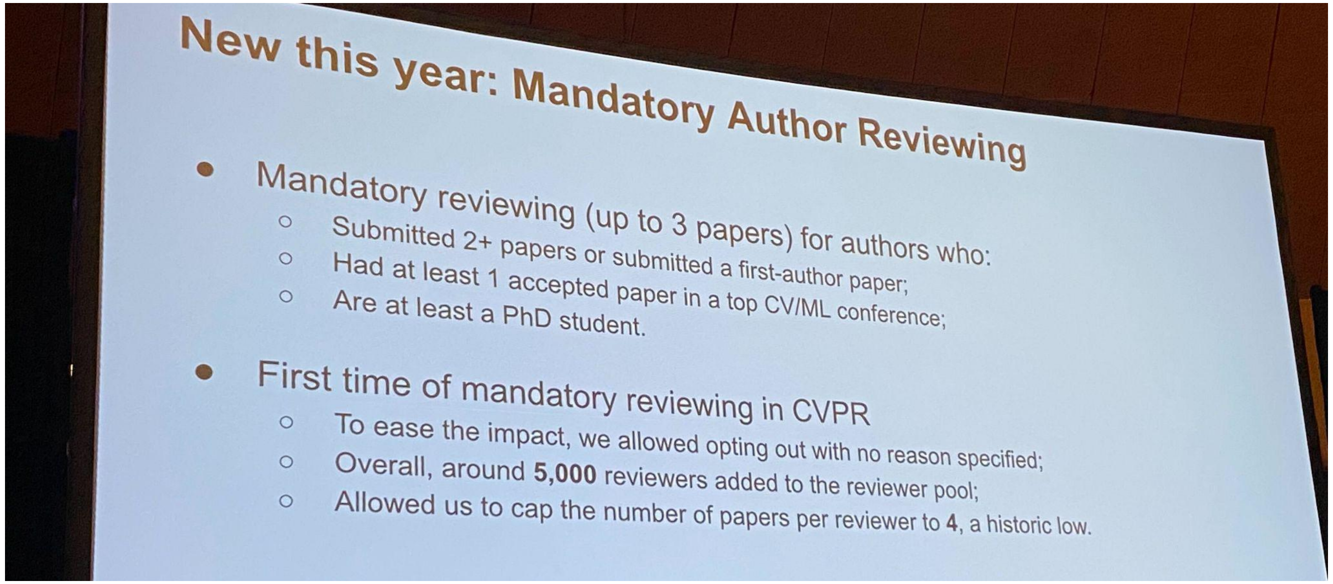
Motivation for Policy Changes

- Exponentially Increasing Submissions
  - 2024: **11,532**, 2025: **13,008**
- Difficult to find enough qualified reviewers
- Community concern about reviewing quality and LLM reviews
- Community concern about people who submit many papers but do not participate in the review process

CVPR25: Meta Insights into Trends and Tendencies (10/179)

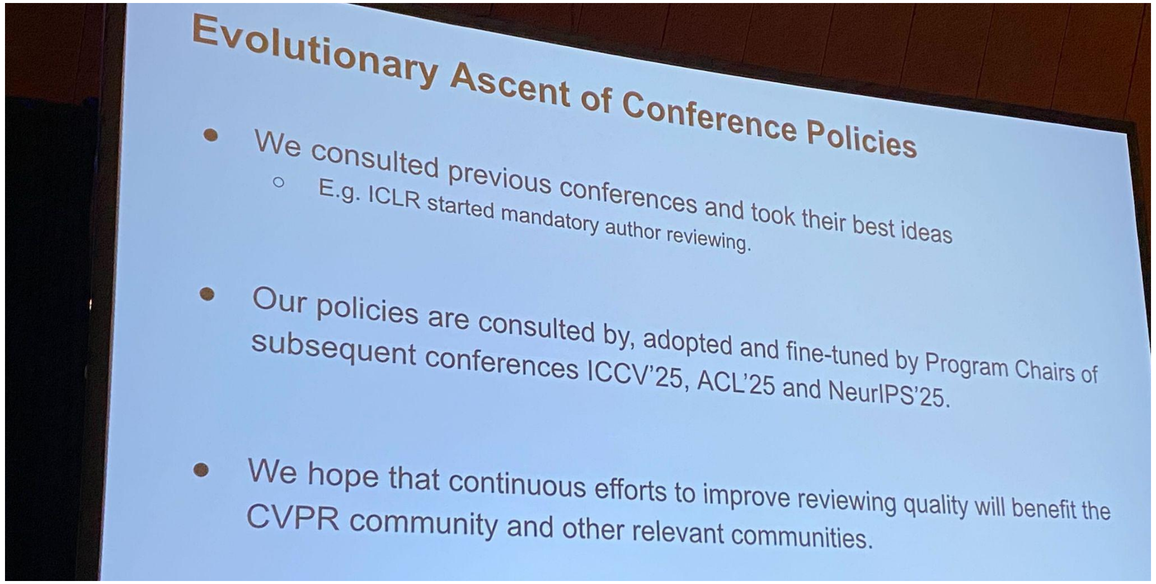
From opening slide

- 1st author / multi-paper submitter must review others
  - It’s now mandatory
  - Only voluntary review is difficult



New this year: Mandatory Author Reviewing

- Mandatory reviewing (up to 3 papers) for authors who:
  - Submitted 2+ papers or submitted a first-author paper;
  - Had at least 1 accepted paper in a top CV/ML conference;
  - Are at least a PhD student.
- First time of mandatory reviewing in CVPR
  - To ease the impact, we allowed opting out with no reason specified;
  - Overall, around **5,000** reviewers added to the reviewer pool;
  - Allowed us to cap the number of papers per reviewer to **4**, a historic low.



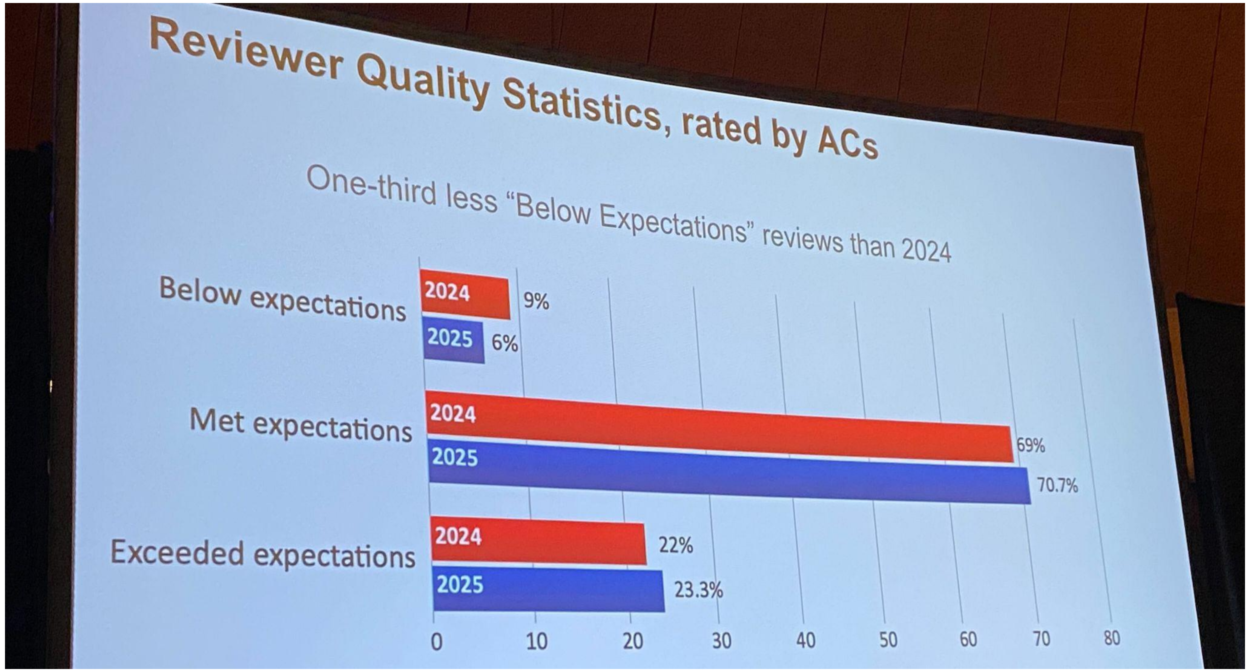
Evolutionary Ascent of Conference Policies

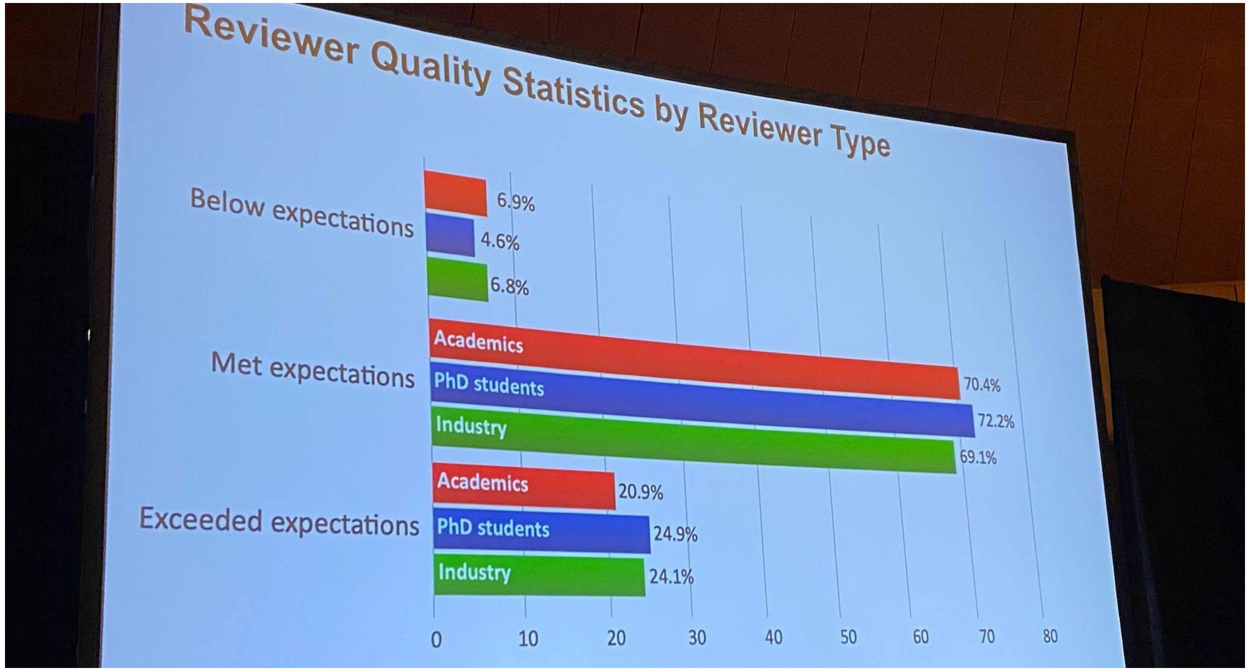
- We consulted previous conferences and took their best ideas
  - E.g. ICLR started mandatory author reviewing.
- Our policies are consulted by, adopted and fine-tuned by Program Chairs of subsequent conferences ICCV’25, ACL’25 and NeurIPS’25.
- We hope that continuous efforts to improve reviewing quality will benefit the CVPR community and other relevant communities.

CVPR25: Meta Insights into Trends and Tendencies (11/179)

From opening slide

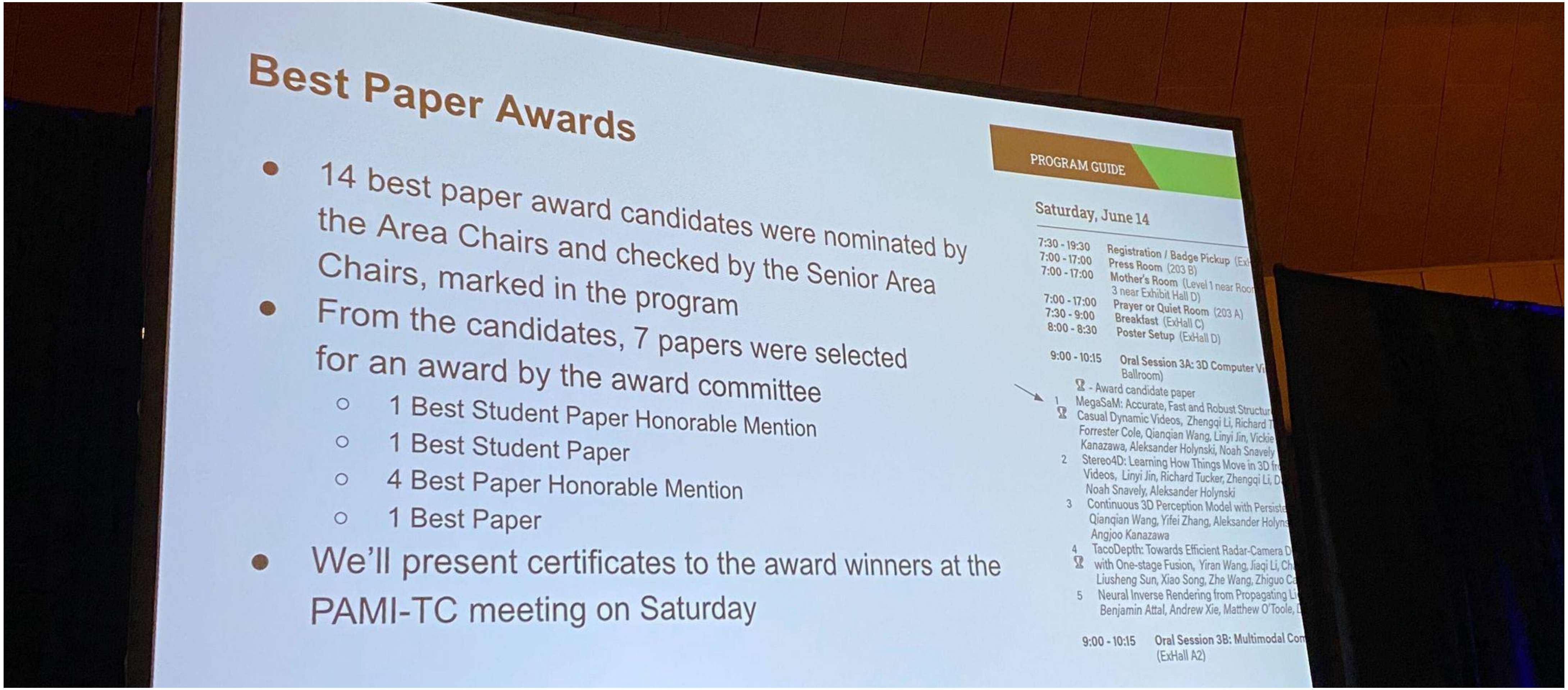
- According to the area chairs,
  - Relatively better reviews have been gathered in CVPR 2025
  - A Ph.D. student can be a good reviewer!





CVPR25: Meta Insights into Trends and Tendencies (12/179)

From award ceremony



Best Paper Awards

- 14 best paper award candidates were nominated by the Area Chairs and checked by the Senior Area Chairs, marked in the program
- From the candidates, 7 papers were selected for an award by the award committee
  - 1 Best Student Paper Honorable Mention
  - 1 Best Student Paper
  - 4 Best Paper Honorable Mention
  - 1 Best Paper
- We'll present certificates to the award winners at the PAMI-TC meeting on Saturday

PROGRAM GUIDE

Saturday, June 14

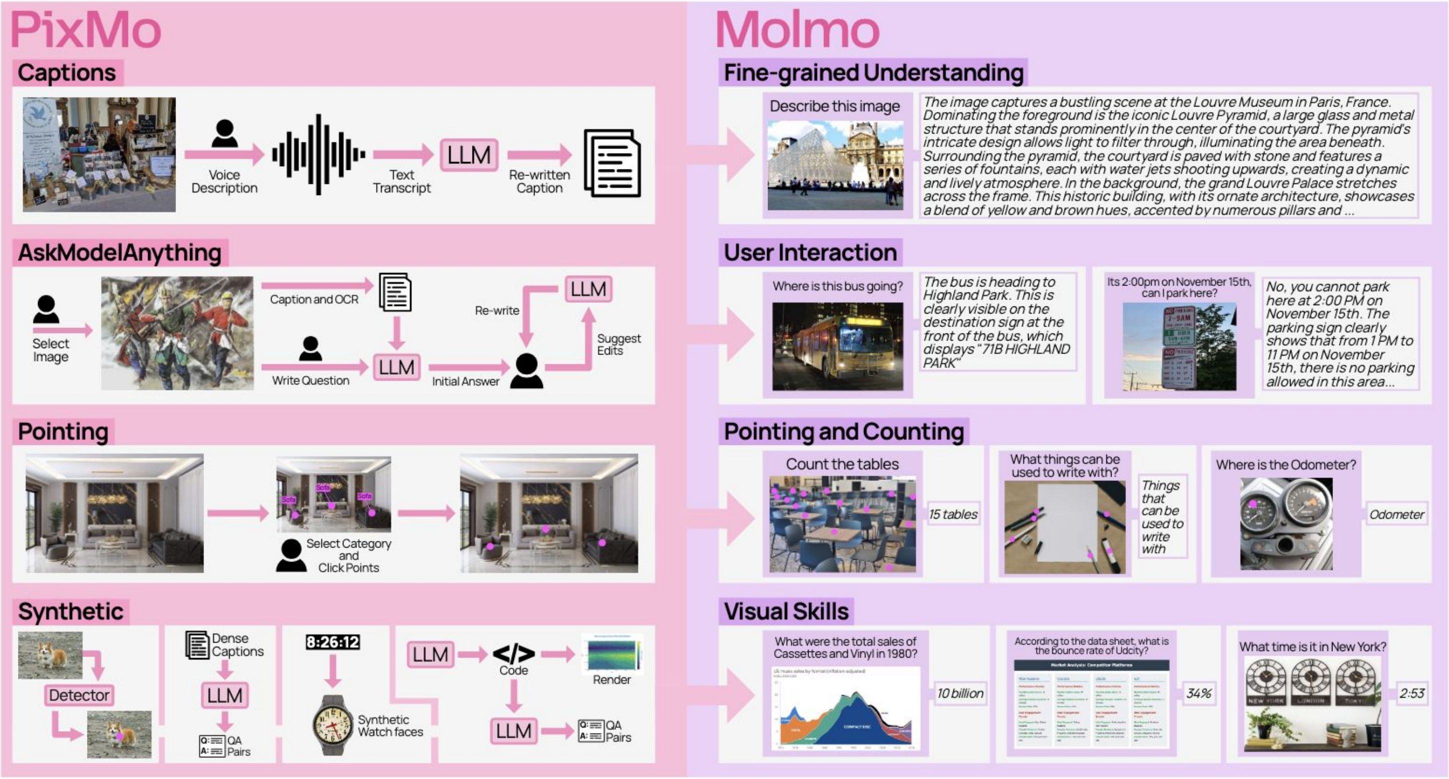
7:30 - 19:30 Registration / Badge Pickup (Ex
7:00 - 17:00 Press Room (203 B)
7:00 - 17:00 Mother's Room (Level 1 near Roo
3 near Exhibit Hall D)
7:00 - 17:00 Prayer or Quiet Room (203 A)
7:30 - 9:00 Breakfast (ExHall C)
8:00 - 8:30 Poster Setup (ExHall D)

9:00 - 10:15 Oral Session 3A: 3D Computer Vi
Ballroom)
🏆 - Award candidate paper
1 🏆 MegaSaM: Accurate, Fast and Robust Structur
Casual Dynamic Videos, Zhengqi Li, Richard T
Forrester Cole, Qianqian Wang, Linyi Jin, Vickie
Kanazawa, Aleksander Holynski, Noah Snavely
2 Stereo4D: Learning How Things Move in 3D fr
Videos, Linyi Jin, Richard Tucker, Zhengqi Li, D
Noah Snavely, Aleksander Holynski
3 Continuous 3D Perception Model with Persiste
Qianqian Wang, Yifei Zhang, Aleksander Holyns
Angjoo Kanazawa
4 🏆 TacoDepth: Towards Efficient Radar-Camera D
with One-stage Fusion, Yiran Wang, Jiaqi Li, Ch
Liusheng Sun, Xiao Song, Zhe Wang, Zhiguo Ca
5 Neural Inverse Rendering from Propagating Li
Benjamin Attal, Andrew Xie, Matthew O'Toole, D

9:00 - 10:15 Oral Session 3B: Multimodal Com
(ExHall A2)

Summary of best paper candidates 1: Molmo and PixMo: Open Weights and Open Data for State-of- the-Art Vision-Language Models

- Molmo is a fully open VLM family built without using proprietary data or models, addressing the lack of foundational knowledge in open VLM development.
- Trained on the newly collected PixMo datasets, Molmo's 72B model outperforms major proprietary models like Gemini 1.5 Pro and ranks just behind GPT-4o.



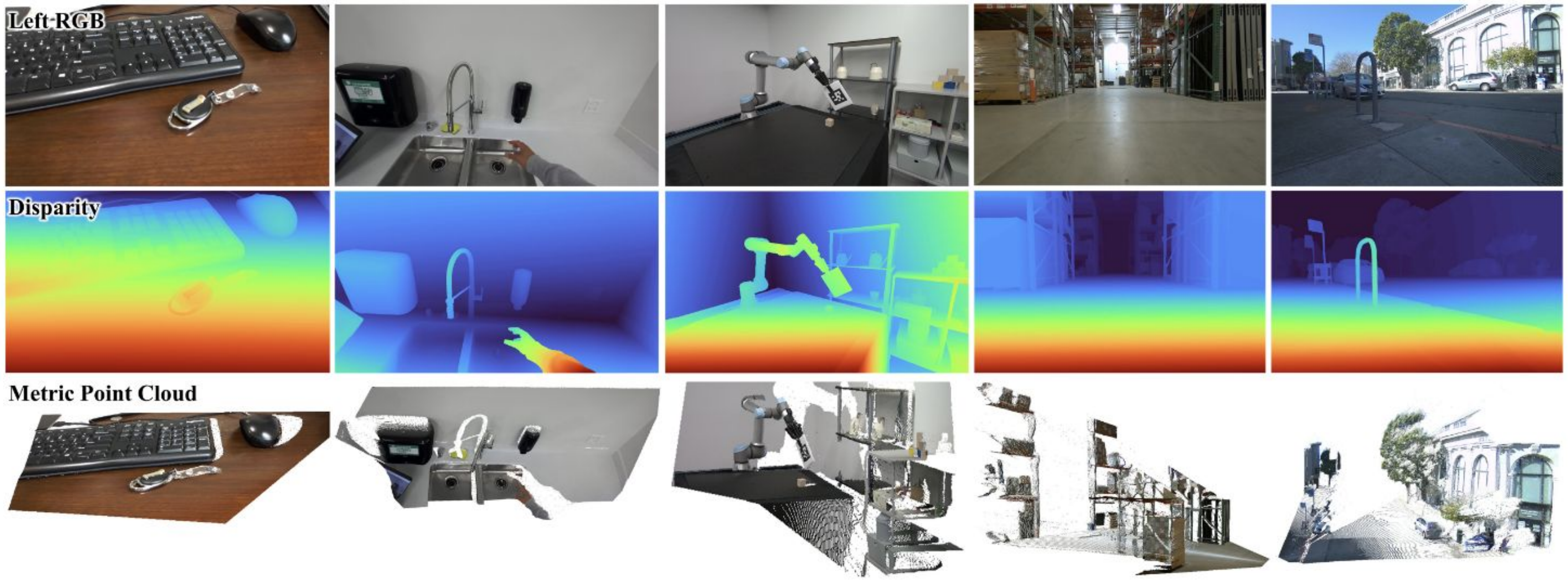
The image captures a bustling scene at the Louvre Museum in Paris, France. Dominating the foreground is the iconic Louvre Pyramid, a large glass and metal structure that stands prominently in the center of the courtyard. The pyramid's intricate design allows light to filter through, illuminating the area beneath. Surrounding the pyramid, the courtyard is paved with stone and features a series of fountains, each with water jets shooting upwards, creating a dynamic and lively atmosphere. In the background, the grand Louvre Palace stretches across the frame. This historic building, with its ornate architecture, showcases a blend of yellow and brown hues, accented by numerous pillars and ...

The bus is heading to Highland Park. This is clearly visible on the destination sign at the front of the bus, which displays "71B HIGHLAND PARK"

No, you cannot park here at 2:00 PM on November 15th. The parking sign clearly shows that from 1 PM to 11 PM on November 15th, there is no parking allowed in this area...

Summary of best paper candidates 2: Foundation Stereo: Zero-shot Stereo Matching

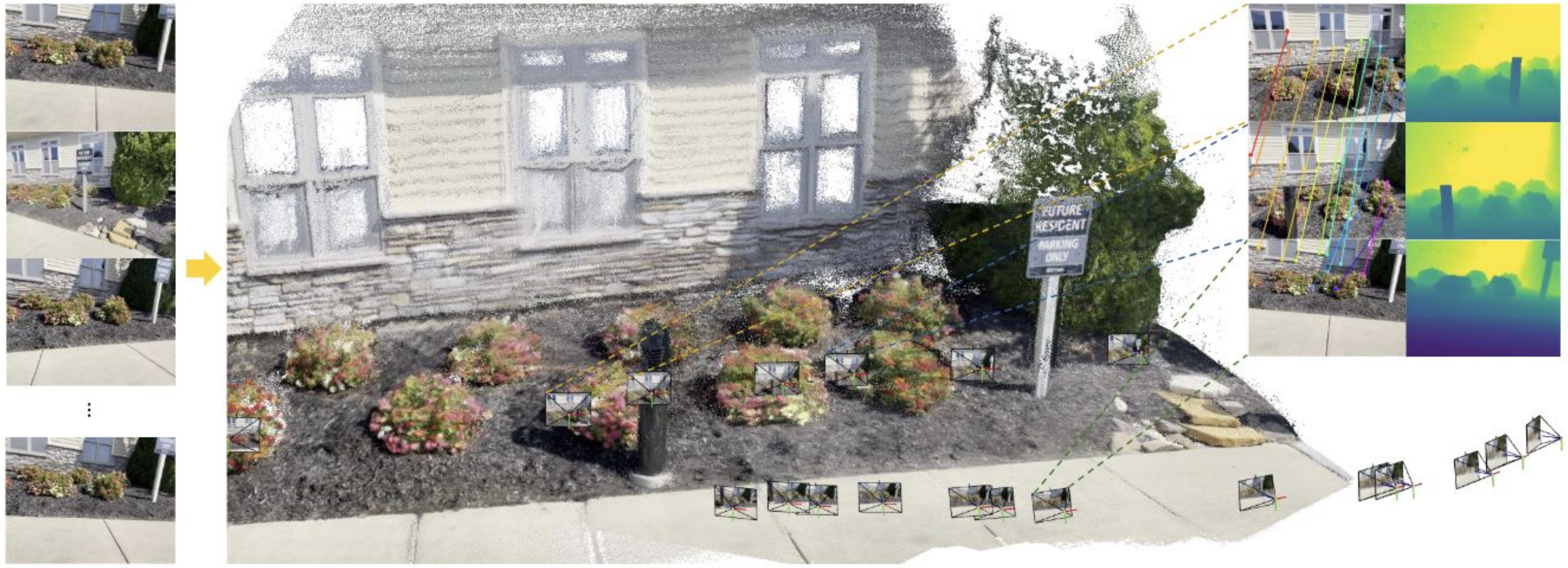
- Foundation Stereo is a foundation model for stereo depth estimation, achieving strong zero-shot generalization without target-domain fine-tuning
- It is trained on a large-scale synthetic dataset and enhanced by self-curation, monocular prior adaptation, and long-range context reasoning for cross-domain robustness



CVPR25: Meta Insights into Trends and Tendencies (15/179)

Summary of best paper candidates 3: VGGT: Visual Geometry Grounded Transformer

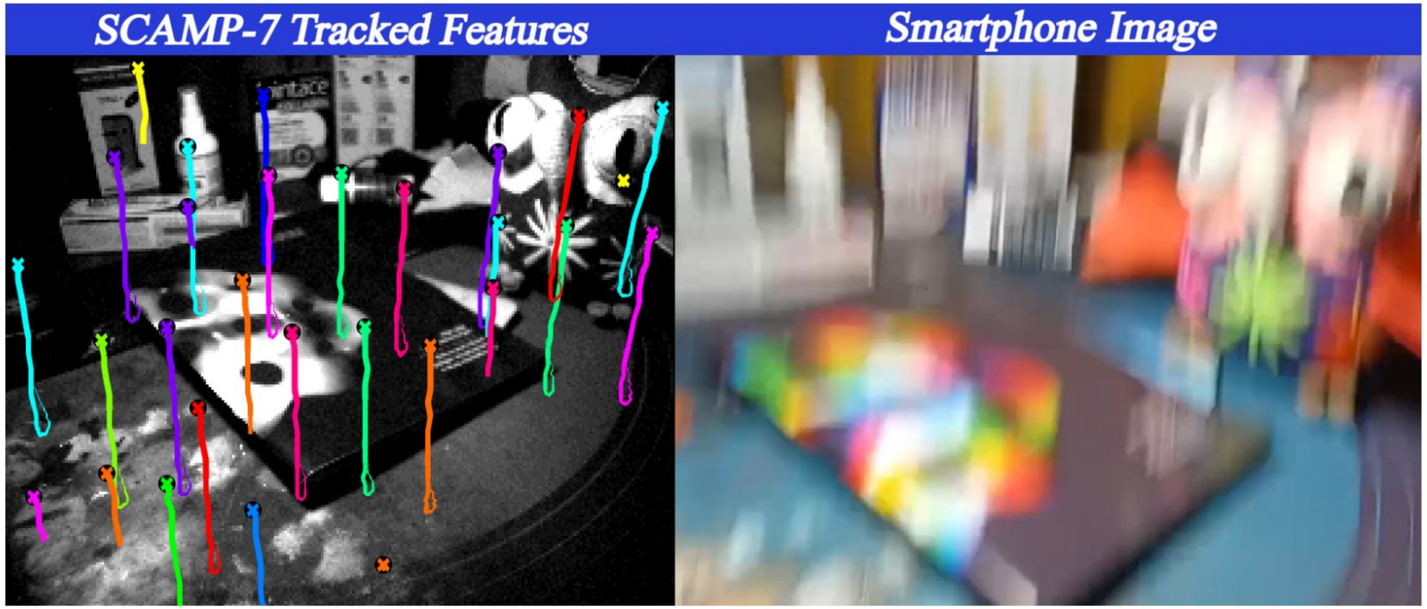
- VGGT is a feed-forward network that predicts full 3D scene attributes from one or more views, including depth, camera parameters, and 3D point tracks.
- It achieves state-of-the-art performance across multiple 3D tasks with high speed and efficiency, and boosts downstream tasks when used as a feature backbone.



CVPR25: Meta Insights into Trends and Tendencies (16/179)

Summary of best paper candidates 4: Descriptor-In-Pixel: Point-Feature Tracking For Pixel Processor Arrays

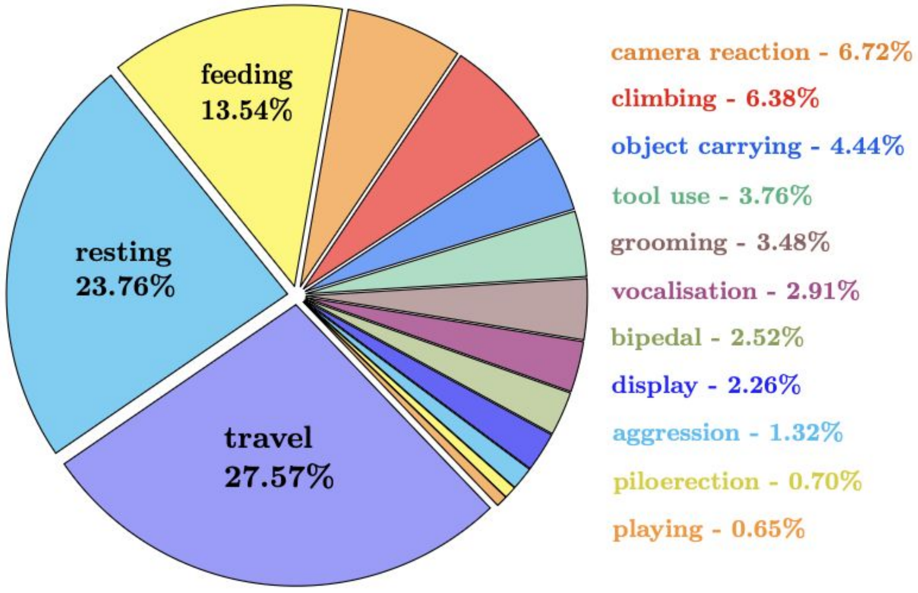
- This paper introduces a fully in-pixel method for point-feature detection and tracking on PPA vision sensors, eliminating the need to output raw image data.
- Using a Descriptor-In-Pixel paradigm, the system runs at over 3000 FPS on the SCAMP-7 prototype, achieving over 1000× data reduction while maintaining reliable tracking under fast motion.

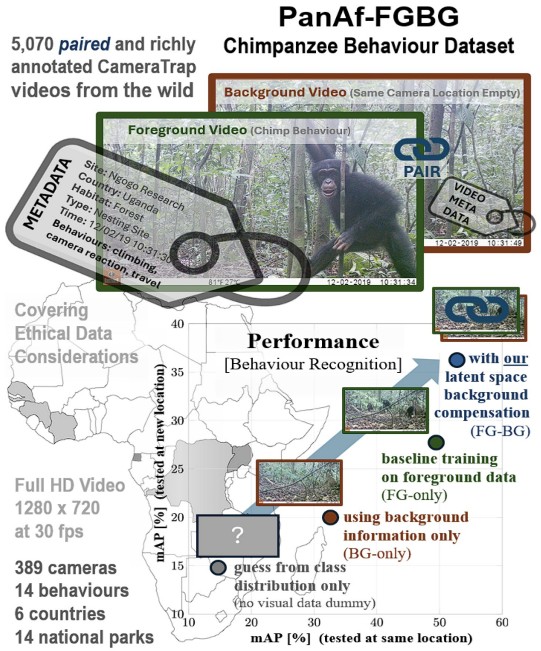


CVPR25: Meta Insights into Trends and Tendencies (17/179)

Summary of best paper candidates 5: The PanAf-FGBG Dataset: Understanding the Impact of Backgrounds in Wildlife Behaviour Recognition

- PanAf-FGBG is a novel dataset of wild chimpanzee behaviours paired with background videos from the same camera locations, enabling direct evaluation of background impact on behaviour recognition.
- It supports in- and out-of-distribution testing and introduces a normalization technique that significantly improves model generalization.





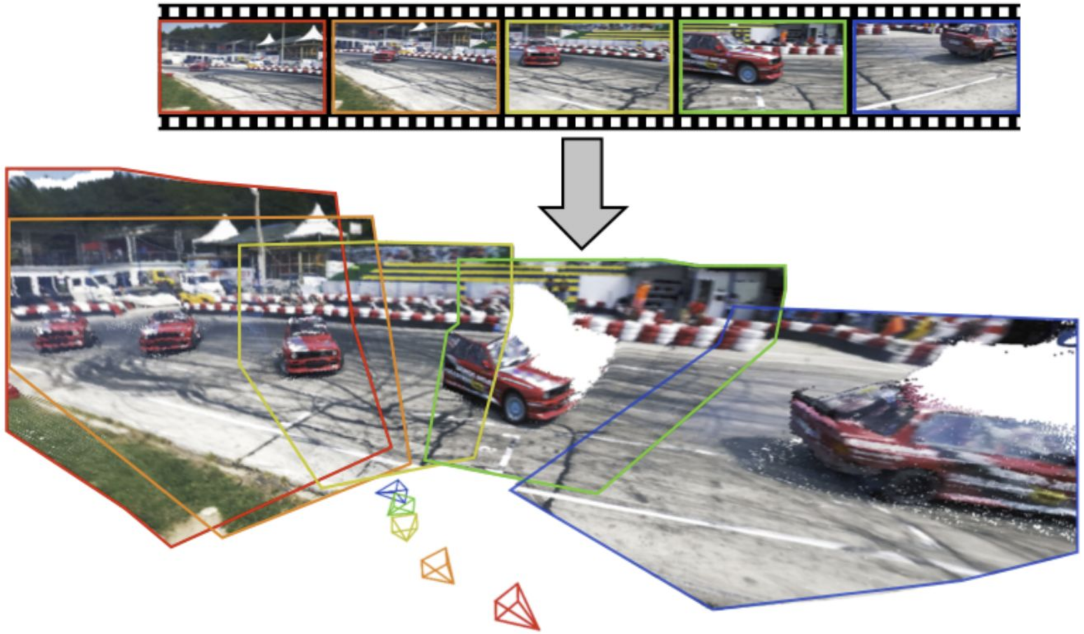
PanAf-FGBG
Chimpanzee Behaviour Dataset
5,070 paired and richly annotated CameraTrap videos from the wild
Background Video (Same Camera Location Empty)
Foreground Video (Chimp Behaviour)
PAIR
METADATA
Site: Ngogo Research
Country: Uganda
Habitat: Forest
Type: Nesting Site
Time: 12/02/19 10:31:30
Behaviours: climbing, camera reaction, travel
VIDEO META DATA
Covering Ethical Data Considerations
Full HD Video 1280 x 720 at 30 fps
389 cameras
14 behaviours
6 countries
14 national parks
Performance
[Behaviour Recognition]
with our latent space background compensation (FG-BG)
baseline training on foreground data (FG-only)
using background information only (BG-only)
guess from class distribution only (no visual data dummy)
?
mAP [%] (tested at new location)
mAP [%] (tested at same location)
10
15
20
25
30
35
40
10
20
30
40
50
60

CVPR25: Meta Insights into Trends and Tendencies (18/179)

Summary of best paper candidates 6: MegaSaM: Accurate, Fast and Robust Structure and

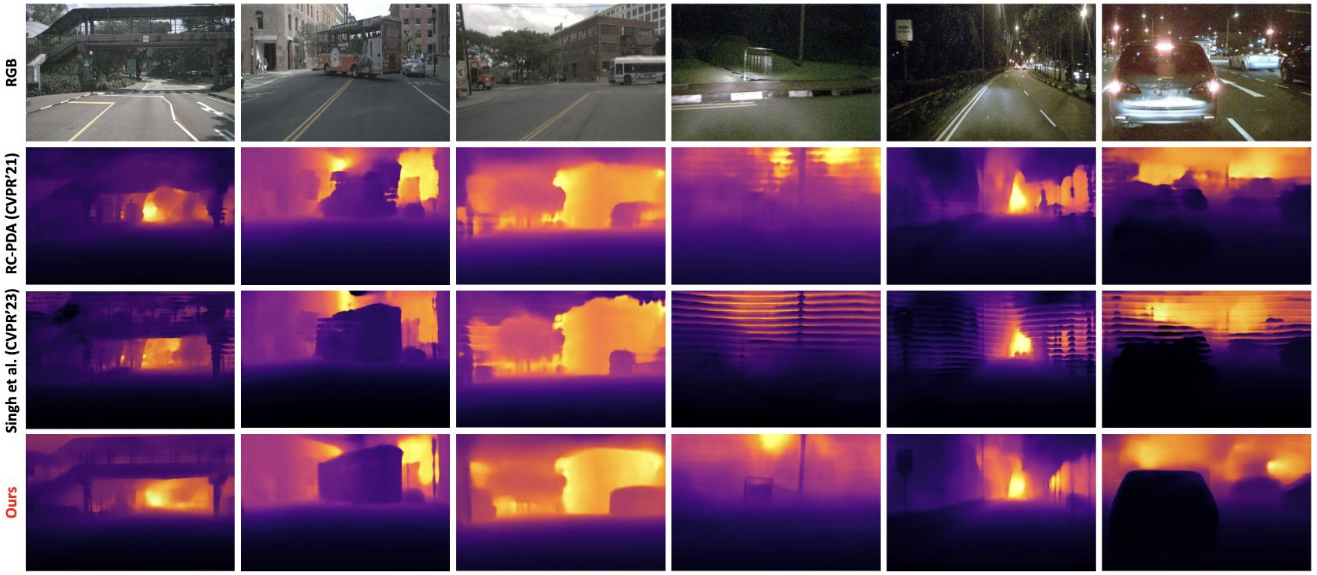
Motion from Casual Dynamic Videos

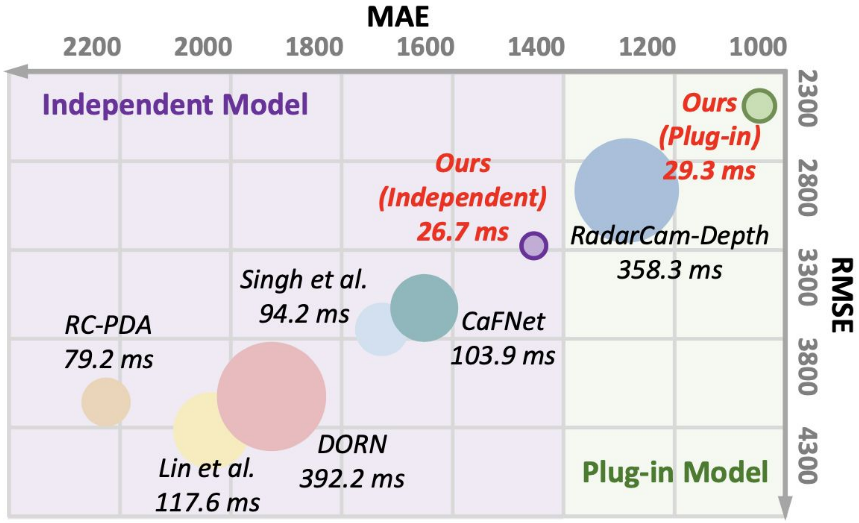
- This work presents a deep visual SLAM system that accurately estimates camera poses and depth maps from monocular videos of dynamic scenes with minimal parallax.
- With tailored training and inference schemes, the system outperforms existing methods in both accuracy and speed on real and synthetic datasets.



Summary of best paper candidates 7: TacoDepth: Towards Efficient Radar-Camera Depth Estimation with One-stage Fusion

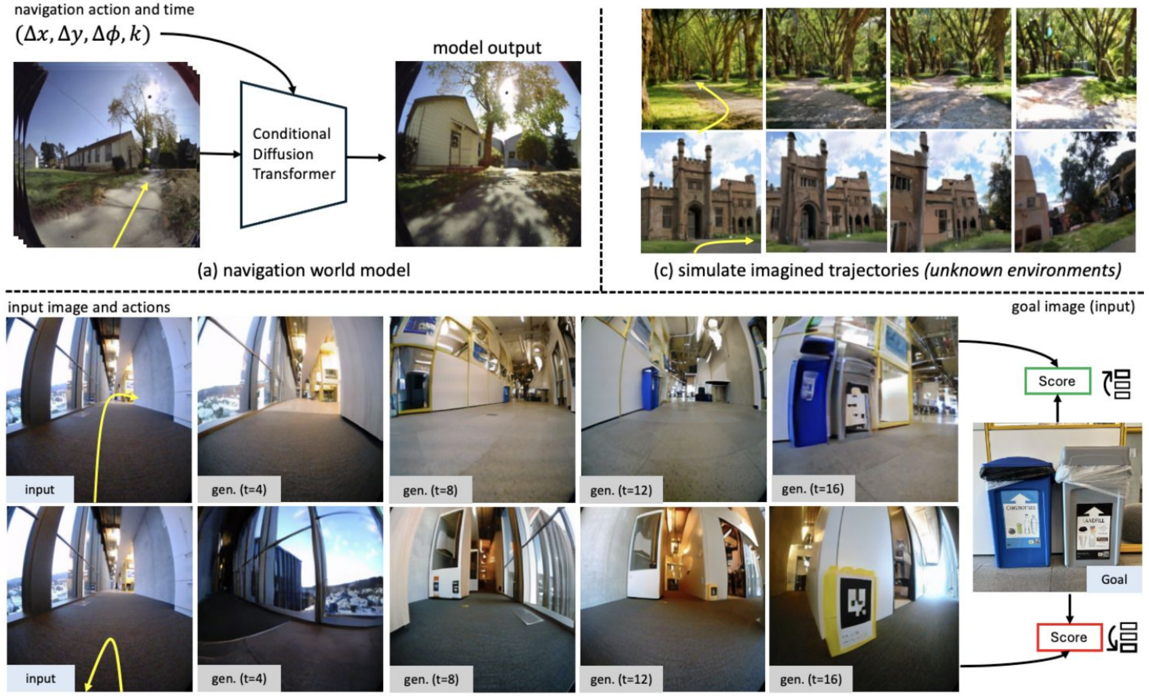
- TacoDepth is a one-stage Radar-Camera depth estimation model that improves efficiency and robustness without relying on intermediate depth results.
- It achieves 12.8% better accuracy and 91.8% faster processing than previous methods, offering a strong balance between speed and performance for real-time applications.





Summary of best paper candidates 8: Navigation World Models

- NWM is a controllable video generation model that predicts future views for navigation using a Conditional Diffusion Transformer trained on egocentric videos.
- It enables dynamic, constraint-aware planning and can simulate trajectories even in unfamiliar environments from a single image.



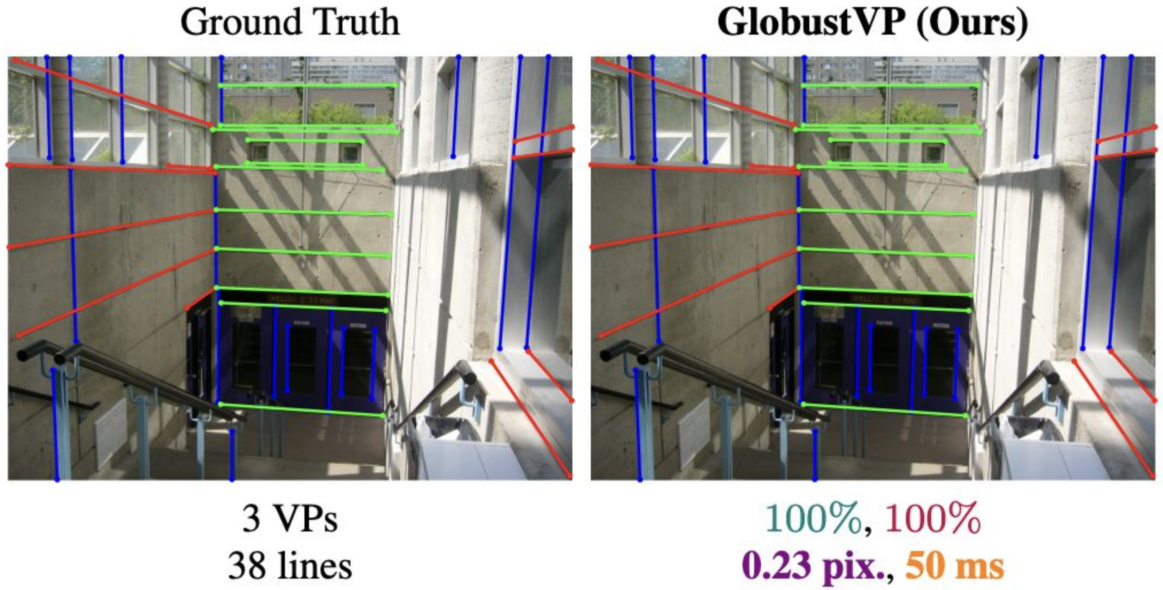
(a) navigation world model

(c) simulate imagined trajectories *(unknown environments)*

(b) evaluate trajectories for **navigation planning** by synthesizing videos (*known environments*)

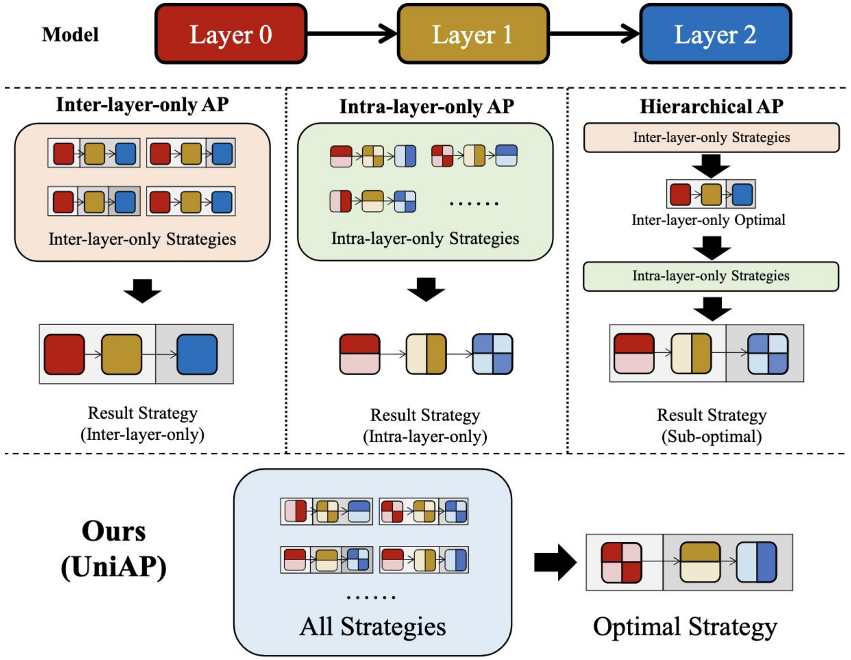
Summary of best paper candidates 9: Convex Relaxation for Robust Vanishing Point Estimation in Manhattan World

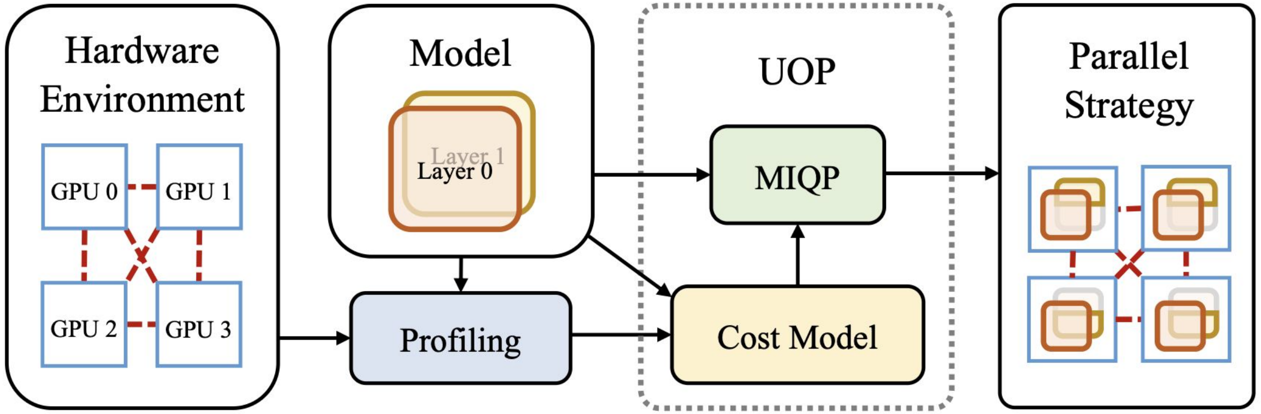
- GlobustVP introduces a convex relaxation approach for vanishing point detection in Manhattan worlds, combining soft line-VP associations with semidefinite programming.
- It achieves global optimality with high efficiency and robustness, outperforming prior methods on both synthetic and real-world datasets.



Summary of best paper candidates 10: UniAP: Unifying Inter- and Intra-Layer Automatic Parallelism by Mixed Integer Quadratic Programming

- UniAP is the first automatic parallelism method that jointly optimizes inter- and intra-layer strategies using mixed integer quadratic programming.
- It achieves up to 3.80× higher throughput and 107× faster optimization than previous state-of-the-art methods on Transformer-based models.

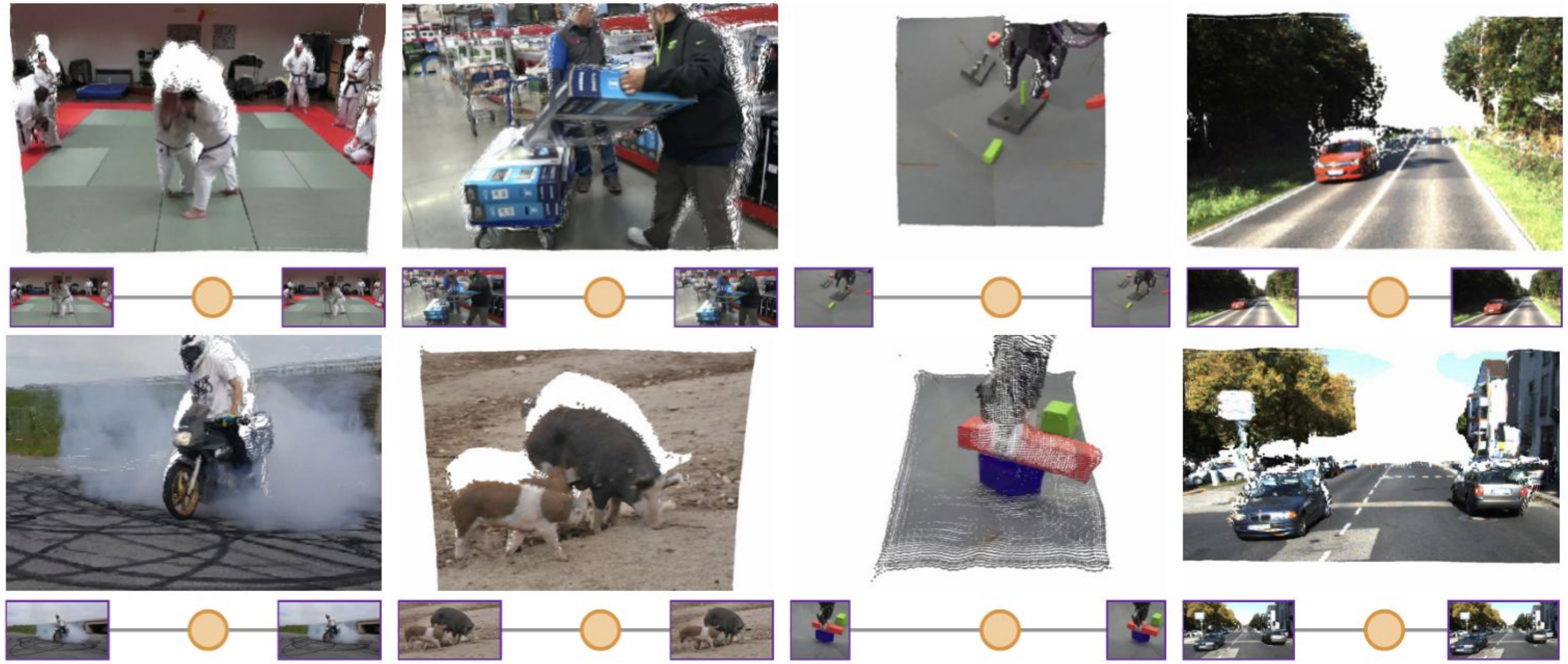




CVPR25: Meta Insights into Trends and Tendencies (23/179)

Summary of best paper candidates 11: Zero-Shot Monocular Scene Flow Estimation in the Wild

- This work introduces a generalizable scene flow model that jointly estimates geometry and motion, addressing key limitations in current approaches.
- It leverages a diverse synthetic dataset of 1M samples and achieves strong zero-shot performance on real-world videos and robotic scenes.

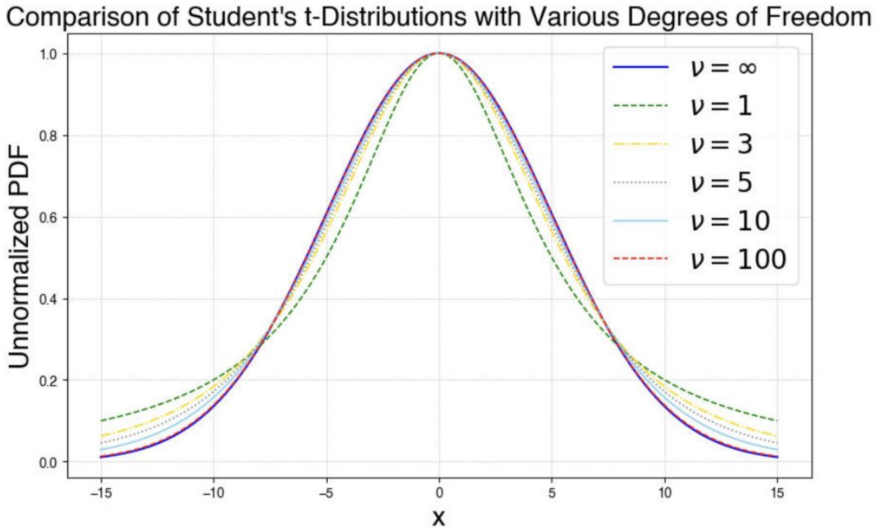


CVPR25: Meta Insights into Trends and Tendencies (24/179)

Summary of best paper candidates 12: 3D Student Splatting and Scooping

- This paper introduces Student Splatting and Scooping (SSS), a new mixture model using Student’s t distributions with both positive and negative densities for novel view synthesis.
- SSS improves quality and parameter efficiency over standard 3D Gaussian Splatting, reducing components by up to 82% while maintaining or surpassing rendering performance.

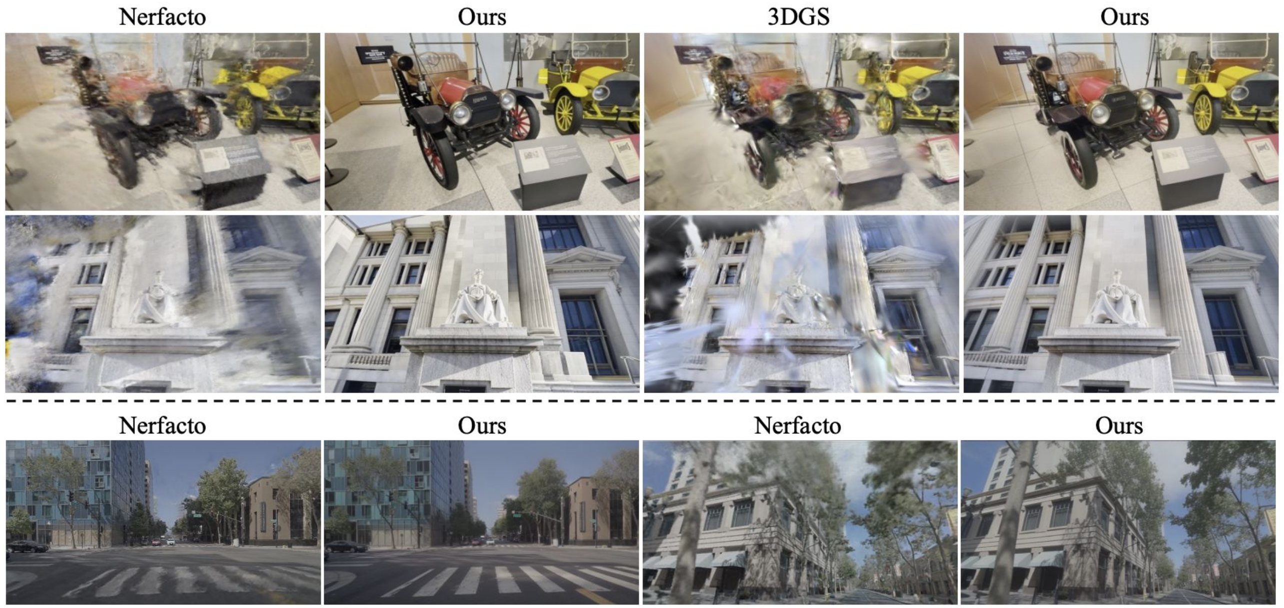
(a) Bonsai from Mip-NeRF 360 [2]
(b) Gardern from Mip-NeRF 360 [2]
(c) Kitchen from Mip-NeRF 360 [2]
(d) Truck from Tanks & Temples [16]
(e) Playroom from Deep Blending [10]



CVPR25: Meta Insights into Trends and Tendencies (25/179)

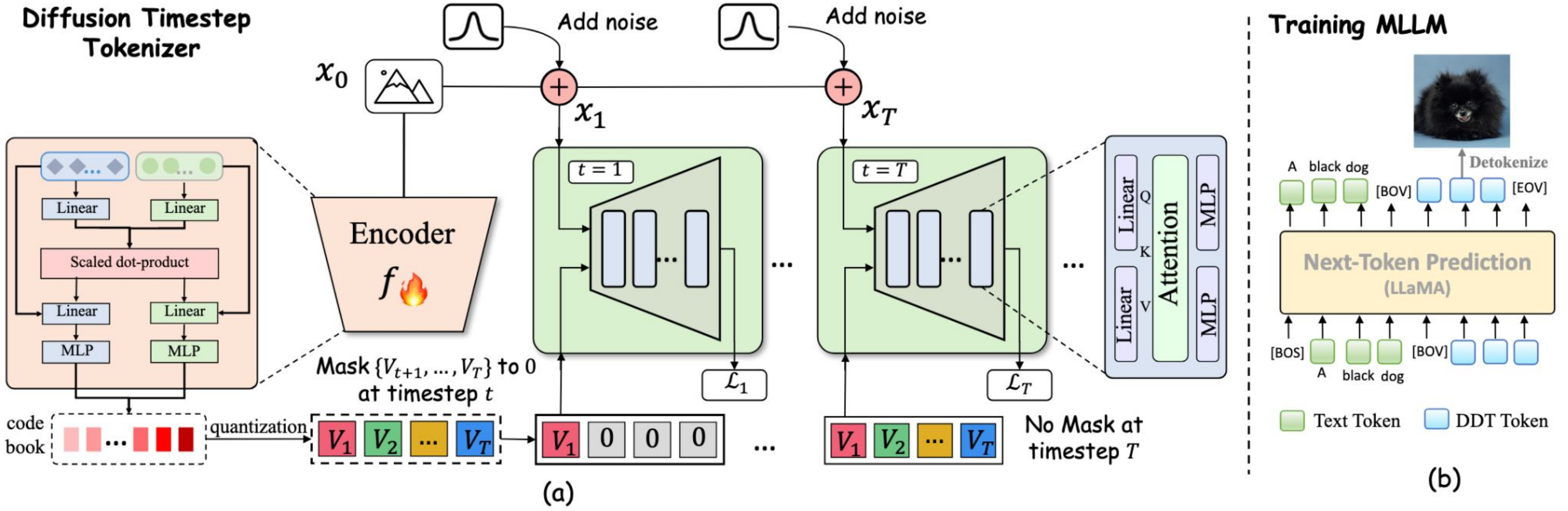
Summary of best paper candidates 13: DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models

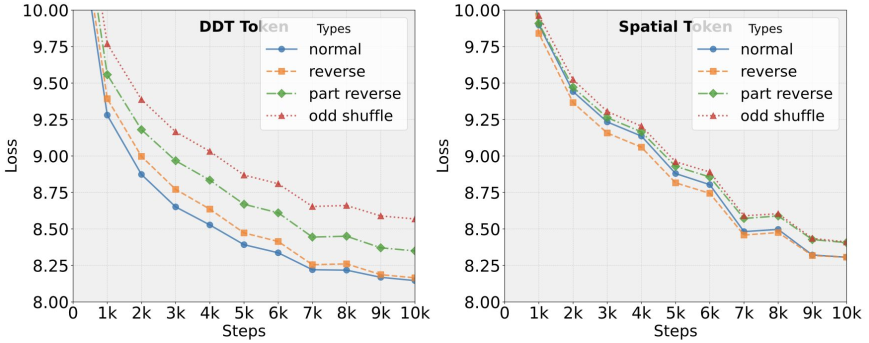
- DIFIX3D+ is a novel pipeline that enhances 3D reconstruction and novel-view synthesis using a single-step diffusion model called DIFIX to remove artifacts.
- Compatible with both NeRF and 3DGS, it improves underconstrained regions and achieves 2× better FID scores while preserving 3D consistency.



Summary of best paper candidates 14: Generative Multimodal Pretraining with Discrete Diffusion Time-step Tokens

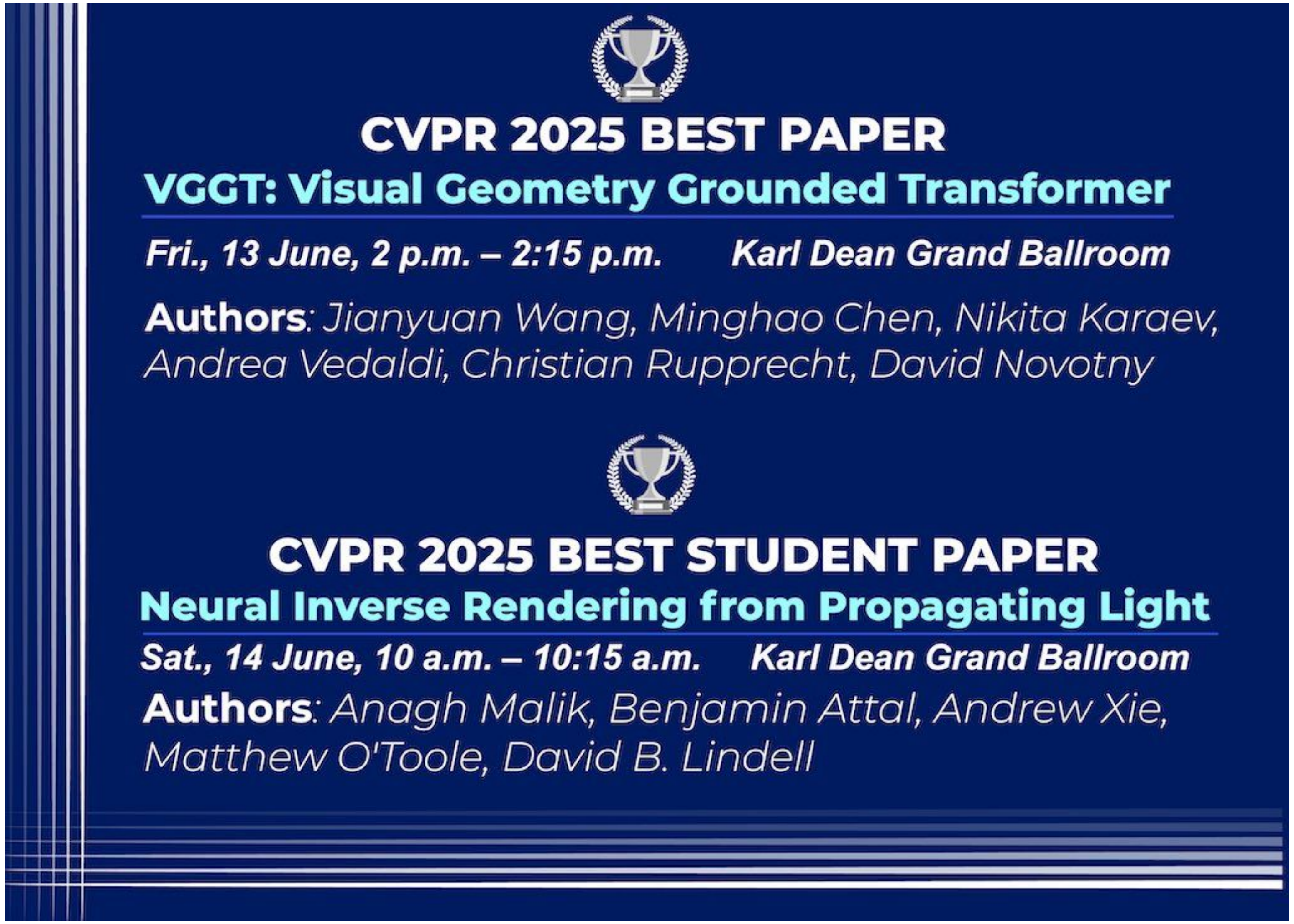
- This work introduces a unified framework that combines LLMs and diffusion models using recursive visual tokens derived from diffusion timesteps.
- The approach enables effective multimodal comprehension and generation, outperforming existing MLLMs across both tasks.





CVPR25: Meta Insights into Trends and Tendencies (27/179)

From award ceremony



CVPR 2025 BEST PAPER

VGGT: Visual Geometry Grounded Transformer

Fri., 13 June, 2 p.m. – 2:15 p.m. ***Karl Dean Grand Ballroom***

Authors: *Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, David Novotny*

CVPR 2025 BEST STUDENT PAPER

Neural Inverse Rendering from Propagating Light

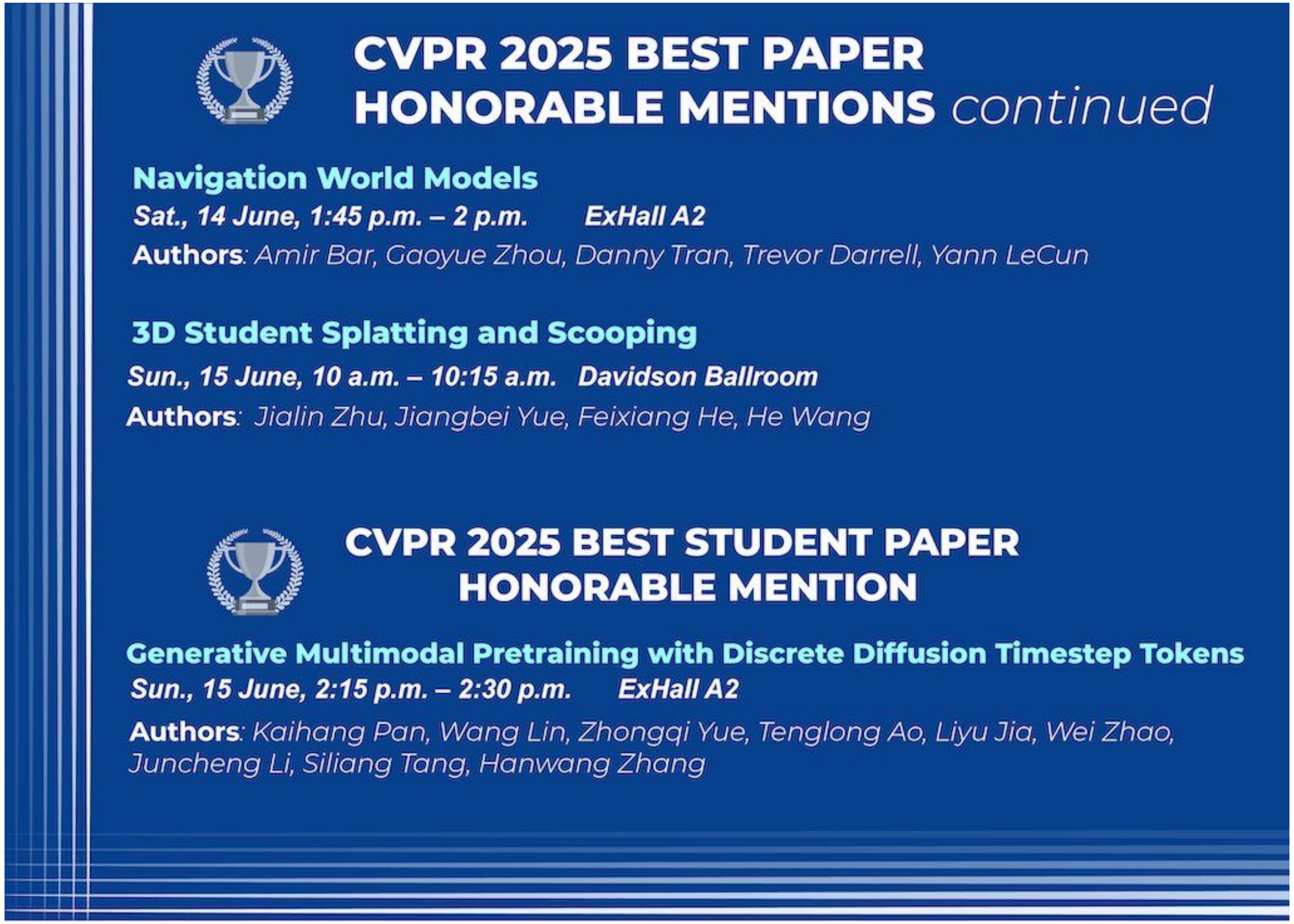
Sat., 14 June, 10 a.m. – 10:15 a.m. ***Karl Dean Grand Ballroom***

Authors: *Anagh Malik, Benjamin Attal, Andrew Xie, Matthew O'Toole, David B. Lindell*

Reference: CVPR X website

CVPR25: Meta Insights into Trends and Tendencies (28/179)

From award ceremony



CVPR 2025 BEST PAPER HONORABLE MENTIONS *continued*

Navigation World Models

Sat., 14 June, 1:45 p.m. – 2 p.m. ***ExHall A2***

Authors: *Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, Yann LeCun*

3D Student Splatting and Scooping

Sun., 15 June, 10 a.m. – 10:15 a.m. ***Davidson Ballroom***

Authors: *Jialin Zhu, Jiangbei Yue, Feixiang He, He Wang*

CVPR 2025 BEST STUDENT PAPER HONORABLE MENTION

Generative Multimodal Pretraining with Discrete Diffusion Timestep Tokens

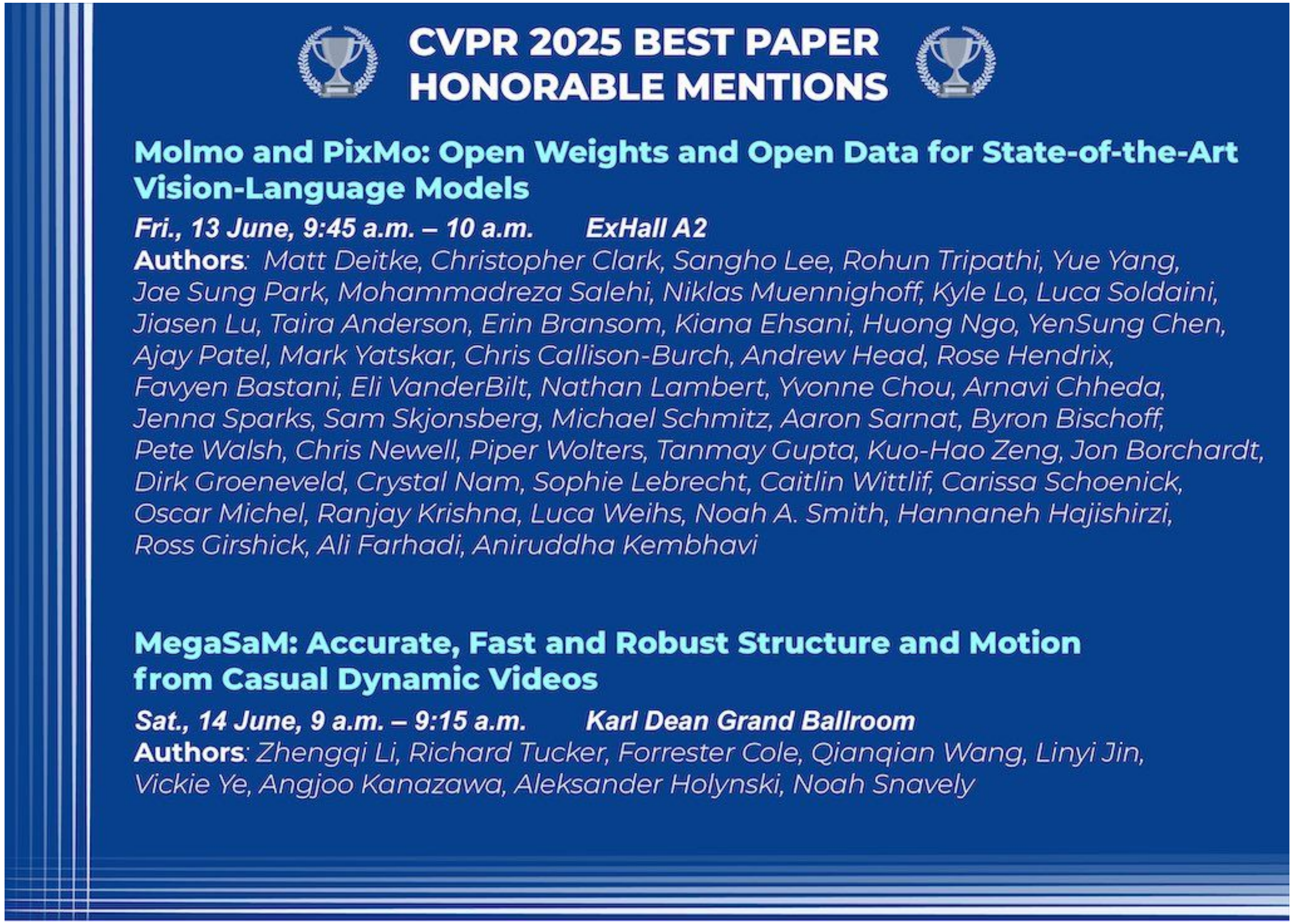
Sun., 15 June, 2:15 p.m. – 2:30 p.m. ***ExHall A2***

Authors: *Kaihang Pan, Wang Lin, Zhongqi Yue, Tenglong Ao, Liyu Jia, Wei Zhao, Juncheng Li, Siliang Tang, Hanwang Zhang*

Reference: CVPR X website

CVPR25: Meta Insights into Trends and Tendencies (29/179)

From award ceremony



CVPR 2025 BEST PAPER HONORABLE MENTIONS

Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models

Fri., 13 June, 9:45 a.m. – 10 a.m. ***ExHall A2***

Authors: *Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, Aniruddha Kembhavi*

MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos

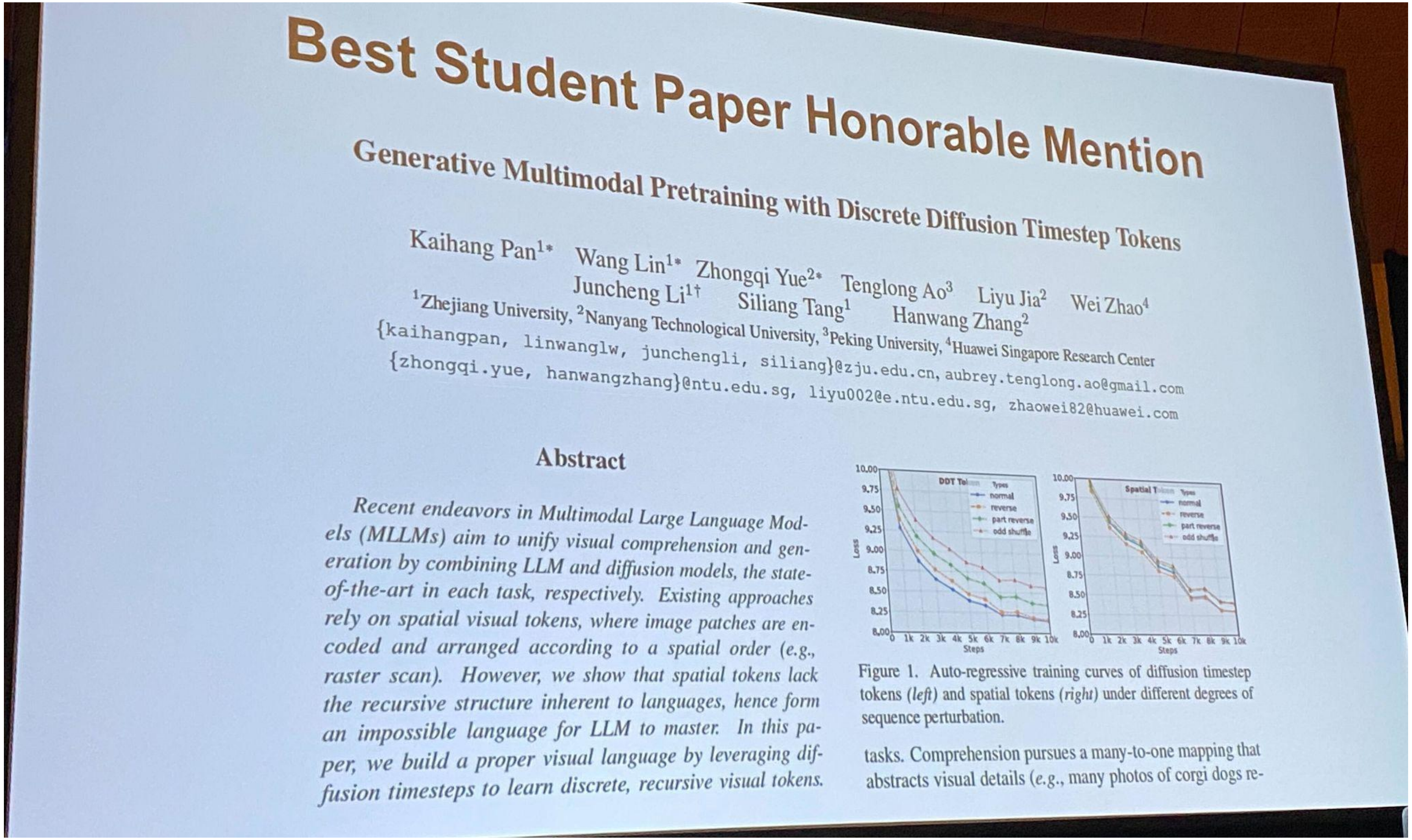
Sat., 14 June, 9 a.m. – 9:15 a.m. ***Karl Dean Grand Ballroom***

Authors: *Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, Noah Snavely*

Reference: CVPR X website

CVPR25: Meta Insights into Trends and Tendencies (30/179)

From award ceremony



Best Student Paper Honorable Mention

Generative Multimodal Pretraining with Discrete Diffusion Timestep Tokens

Kaihang Pan[1]* Wang Lin[1]* Zhongqi Yue[2]* Tenglong Ao[3] Liyu Jia[2] Wei Zhao[4]
Juncheng Li[1]† Siliang Tang[1] Hanwang Zhang[2]
[1]Zhejiang University, [2]Nanyang Technological University, [3]Peking University, [4]Huawei Singapore Research Center
{kaihangpan, linwanglw, junchengli, siliang}@zju.edu.cn, aubrey.tenglong.ao@gmail.com
{zhongqi.yue, hanwangzhang}@ntu.edu.sg, liyu002@e.ntu.edu.sg, zhaowei82@huawei.com

Abstract

Recent endeavors in Multimodal Large Language Models (MLLMs) aim to unify visual comprehension and generation by combining LLM and diffusion models, the state-of-the-art in each task, respectively. Existing approaches rely on spatial visual tokens, where image patches are encoded and arranged according to a spatial order (e.g., raster scan). However, we show that spatial tokens lack the recursive structure inherent to languages, hence form an impossible language for LLM to master. In this paper, we build a proper visual language by leveraging diffusion timesteps to learn discrete, recursive visual tokens.

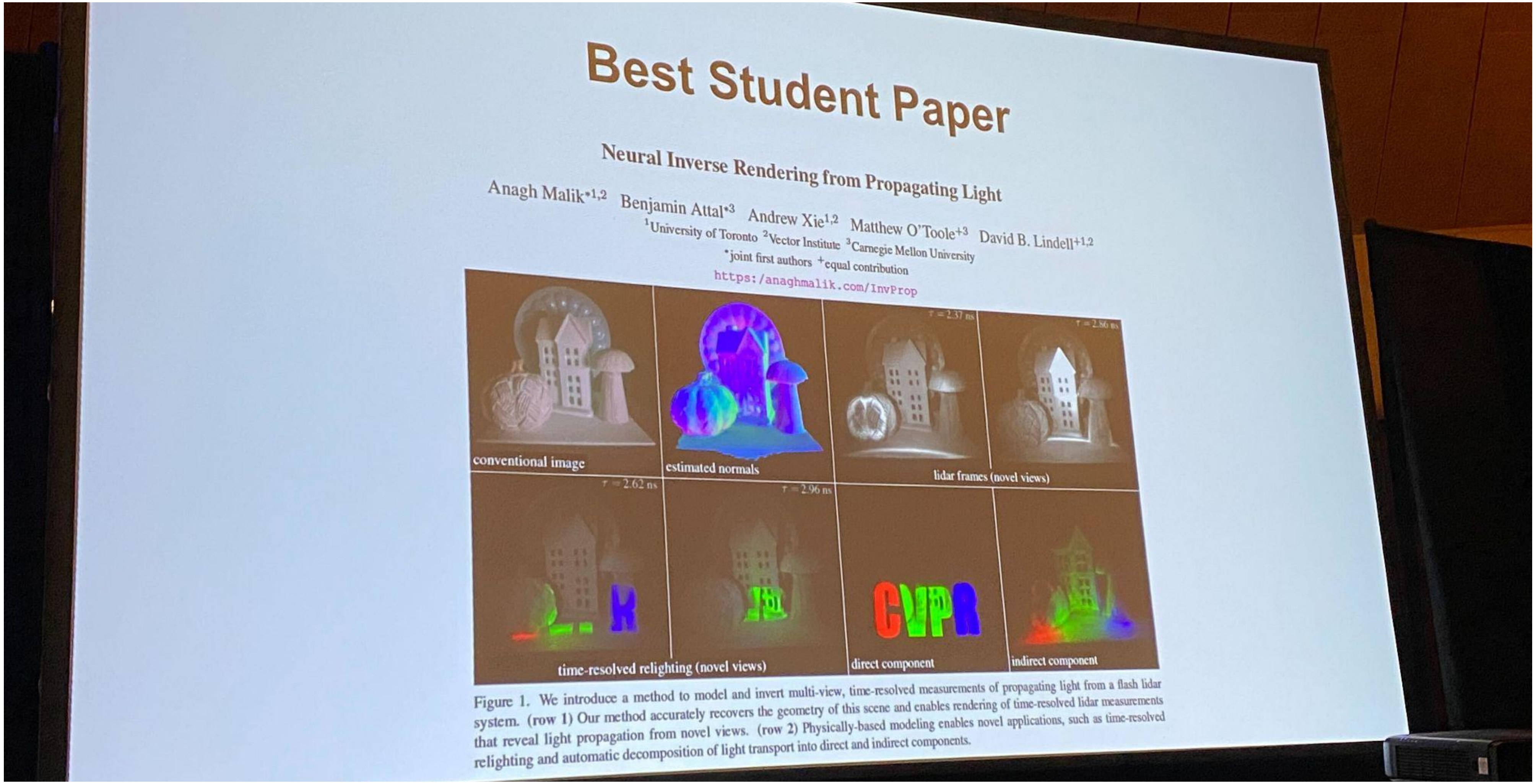
Figure 1. Auto-regressive training curves of diffusion timestep tokens (*left*) and spatial tokens (*right*) under different degrees of sequence perturbation.

tasks. Comprehension pursues a many-to-one mapping that abstracts visual details (*e.g.*, many photos of corgi dogs re-

LIMIT.LAB
https://limitlab.xyz/

CVPR25: Meta Insights into Trends and Tendencies (31/179)

From award ceremony



Best Student Paper

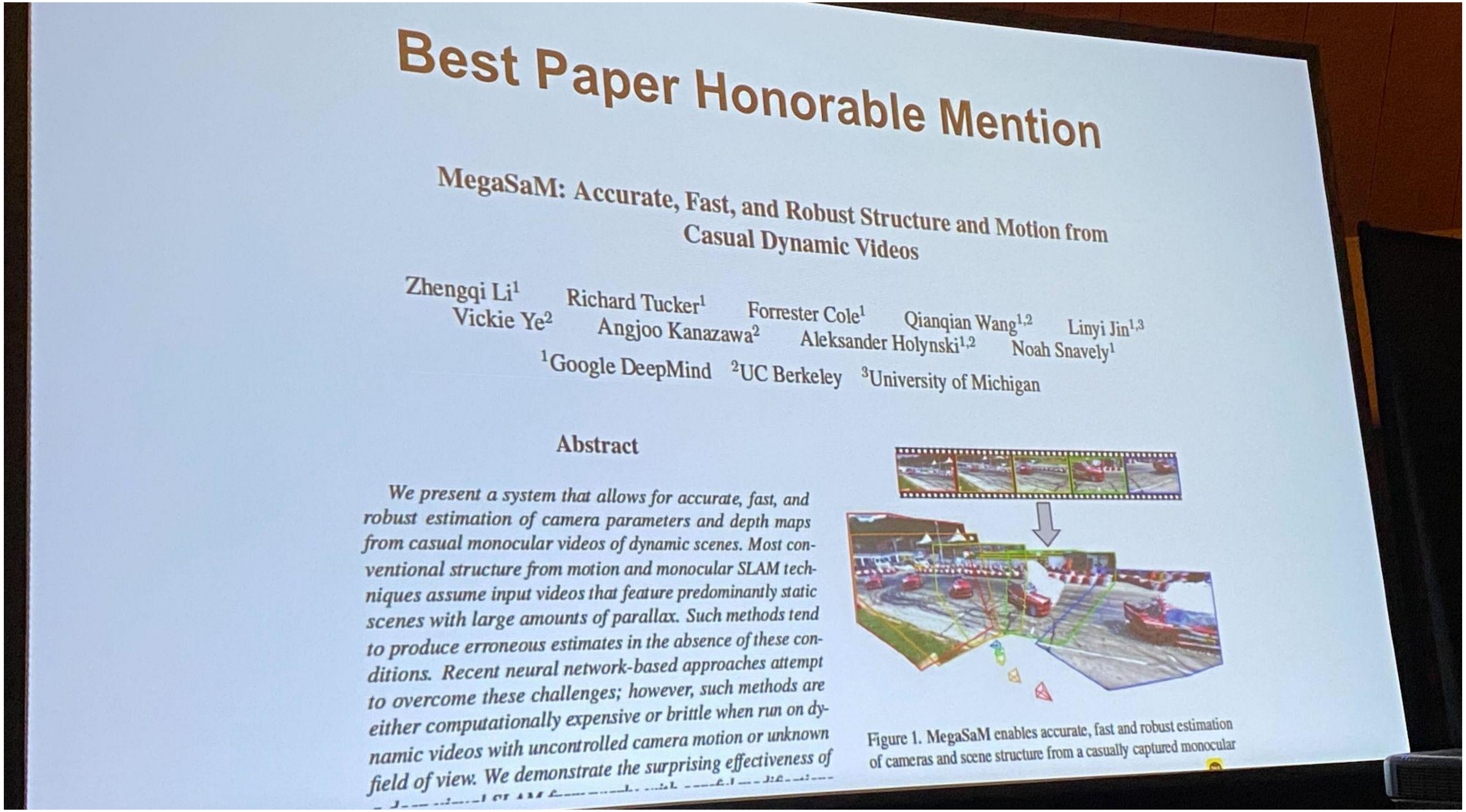
Neural Inverse Rendering from Propagating Light

Anagh Malik*[1,2] Benjamin Attal*[3] Andrew Xie[1,2] Matthew O'Toole+[3] David B. Lindell+[1,2]
[1]University of Toronto [2]Vector Institute [3]Carnegie Mellon University
*joint first authors +equal contribution
https://anaghmalik.com/InvProp

Figure 1. We introduce a method to model and invert multi-view, time-resolved measurements of propagating light from a flash lidar system. (row 1) Our method accurately recovers the geometry of this scene and enables rendering of time-resolved lidar measurements that reveal light propagation from novel views. (row 2) Physically-based modeling enables novel applications, such as time-resolved relighting and automatic decomposition of light transport into direct and indirect components.

CVPR25: Meta Insights into Trends and Tendencies (32/179)

From award ceremony



Best Paper Honorable Mention

MegaSaM: Accurate, Fast, and Robust Structure and Motion from Casual Dynamic Videos

Zhengqi Li[1] Richard Tucker[1] Forrester Cole[1] Qianqian Wang[1,2] Linyi Jin[1,3]
Vickie Ye[2] Angjoo Kanazawa[2] Aleksander Holynski[1,2] Noah Snavely[1]
[1]Google DeepMind [2]UC Berkeley [3]University of Michigan

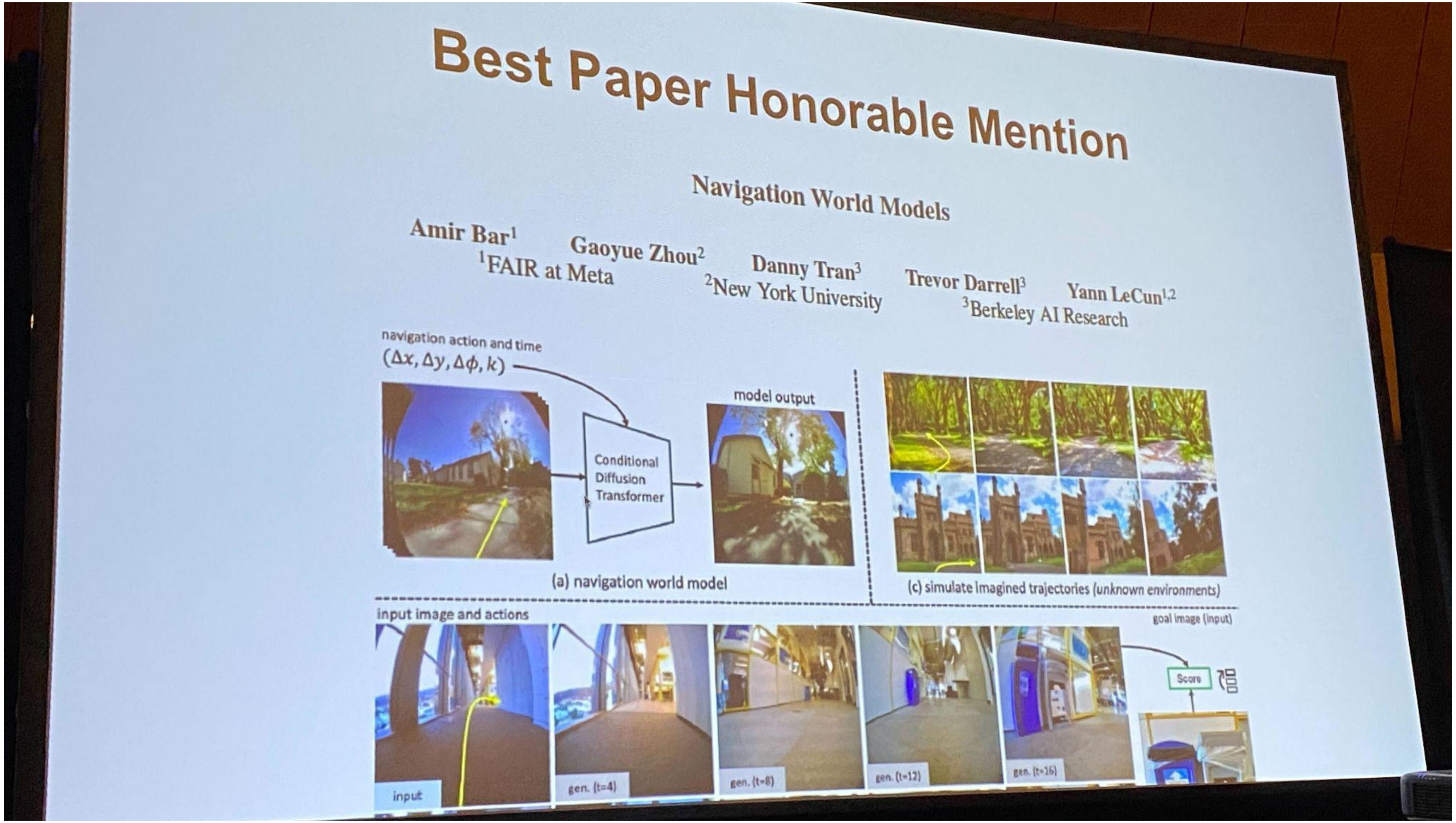
Abstract

We present a system that allows for accurate, fast, and robust estimation of camera parameters and depth maps from casual monocular videos of dynamic scenes. Most conventional structure from motion and monocular SLAM techniques assume input videos that feature predominantly static scenes with large amounts of parallax. Such methods tend to produce erroneous estimates in the absence of these conditions. Recent neural network-based approaches attempt to overcome these challenges; however, such methods are either computationally expensive or brittle when run on dynamic videos with uncontrolled camera motion or unknown field of view. We demonstrate the surprising effectiveness of

Figure 1. MegaSaM enables accurate, fast and robust estimation of cameras and scene structure from a casually captured monocular

CVPR25: Meta Insights into Trends and Tendencies (33/179)

From award ceremony



Best Paper Honorable Mention

Navigation World Models

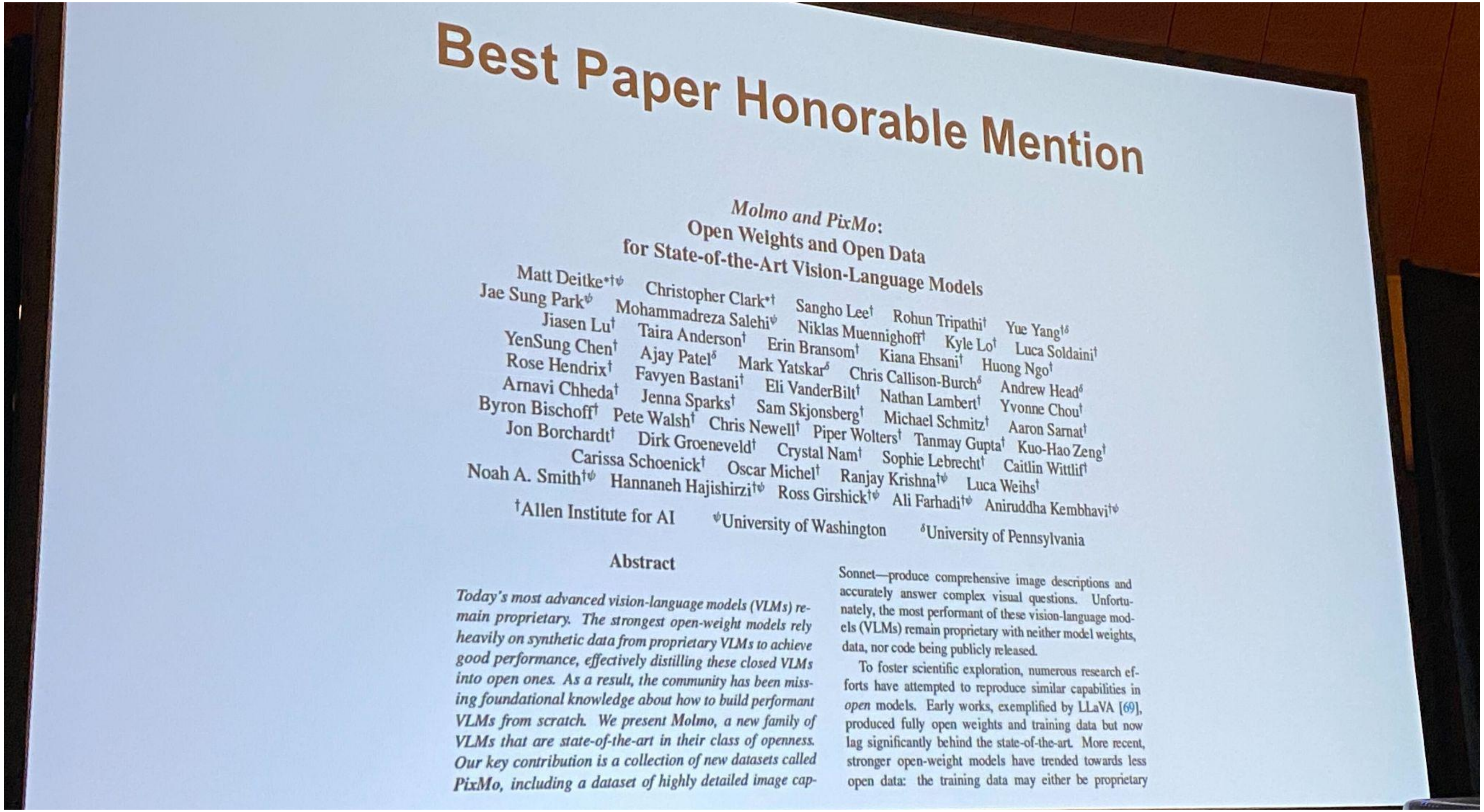
Amir Bar[1] Gaoyue Zhou[2] Danny Tran[3] Trevor Darrell[3] Yann LeCun[1,2]

[1]FAIR at Meta [2]New York University [3]Berkeley AI Research

navigation action and time
$(\Delta x, \Delta y, \Delta \phi, k)$
Conditional Diffusion Transformer
model output
(a) navigation world model
(c) simulate imagined trajectories (unknown environments)
input image and actions
goal image (input)
Score
input
gen. (t=4)
gen. (t=8)
gen. (t=12)
gen. (t=16)

CVPR25: Meta Insights into Trends and Tendencies (34/179)

From award ceremony



Best Paper Honorable Mention

Molmo and PixMo:
Open Weights and Open Data
for State-of-the-Art Vision-Language Models

Matt Deitke[*†ψ] Christopher Clark[*†] Jae Sung Park[ψ] Mohammadreza Salehi[ψ] Sangho Lee[†] Rohun Tripathi[†] Yue Yang[†δ] Jiasen Lu[†] Taira Anderson[†] Niklas Muennighoff[†] Kyle Lo[†] Luca Soldaini[†] YenSung Chen[†] Ajay Patel[δ] Erin Bransom[†] Kiana Ehsani[†] Huong Ngo[†] Rose Hendrix[†] Mark Yatskar[δ] Chris Callison-Burch[δ] Andrew Head[δ] Arnavi Chheda[†] Favyen Bastani[†] Eli VanderBilt[†] Nathan Lambert[†] Yvonne Chou[†] Byron Bischoff[†] Jenna Sparks[†] Sam Skjonsberg[†] Michael Schmitz[†] Aaron Sarnat[†] Jon Borchardt[†] Pete Walsh[†] Chris Newell[†] Piper Wolters[†] Tanmay Gupta[†] Kuo-Hao Zeng[†] Dirk Groeneveld[†] Crystal Nam[†] Sophie Lebrecht[†] Caitlin Wittlif[†] Carissa Schoenick[†] Oscar Michel[†] Ranjay Krishna[†ψ] Luca Weihs[†] Noah A. Smith[†ψ] Hannaneh Hajishirzi[†ψ] Ross Girshick[†ψ] Ali Farhadi[†ψ] Aniruddha Kembhavi[†ψ]

[†]Allen Institute for AI [ψ]University of Washington [δ]University of Pennsylvania

Abstract

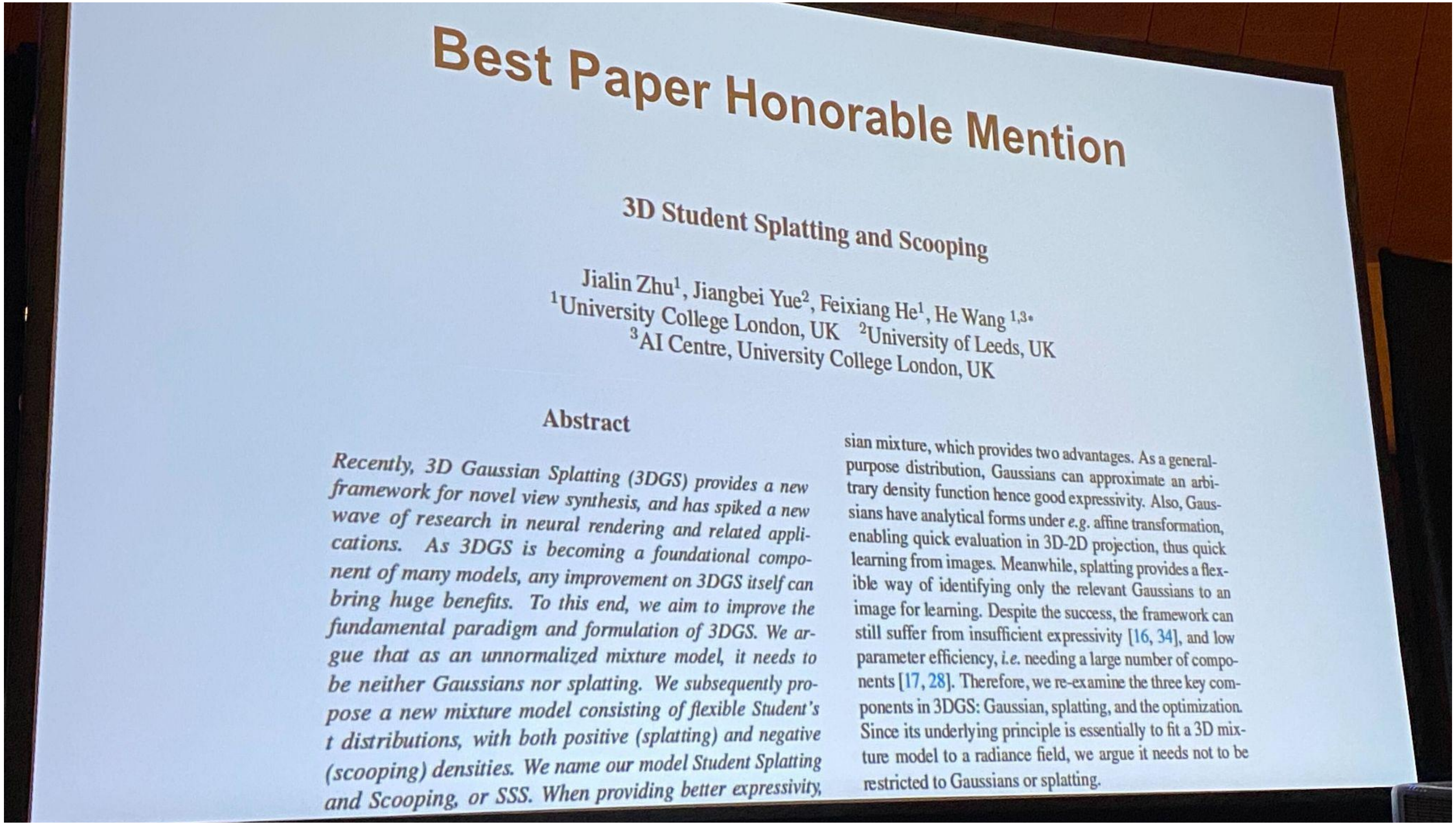
Today's most advanced vision-language models (VLMs) remain proprietary. The strongest open-weight models rely heavily on synthetic data from proprietary VLMs to achieve good performance, effectively distilling these closed VLMs into open ones. As a result, the community has been missing foundational knowledge about how to build performant VLMs from scratch. We present Molmo, a new family of VLMs that are state-of-the-art in their class of openness. Our key contribution is a collection of new datasets called PixMo, including a dataset of highly detailed image cap-

Sonnet—produce comprehensive image descriptions and accurately answer complex visual questions. Unfortunately, the most performant of these vision-language models (VLMs) remain proprietary with neither model weights, data, nor code being publicly released.

To foster scientific exploration, numerous research efforts have attempted to reproduce similar capabilities in *open* models. Early works, exemplified by LLaVA [69], produced fully open weights and training data but now lag significantly behind the state-of-the-art. More recent, stronger open-weight models have trended towards less open data: the training data may either be proprietary

CVPR25: Meta Insights into Trends and Tendencies (35/179)

From award ceremony



Best Paper Honorable Mention

3D Student Splatting and Scooping

Jialin Zhu[1], Jiangbei Yue[2], Feixiang He[1], He Wang [1,3]
[1]University College London, UK [2]University of Leeds, UK
[3]AI Centre, University College London, UK

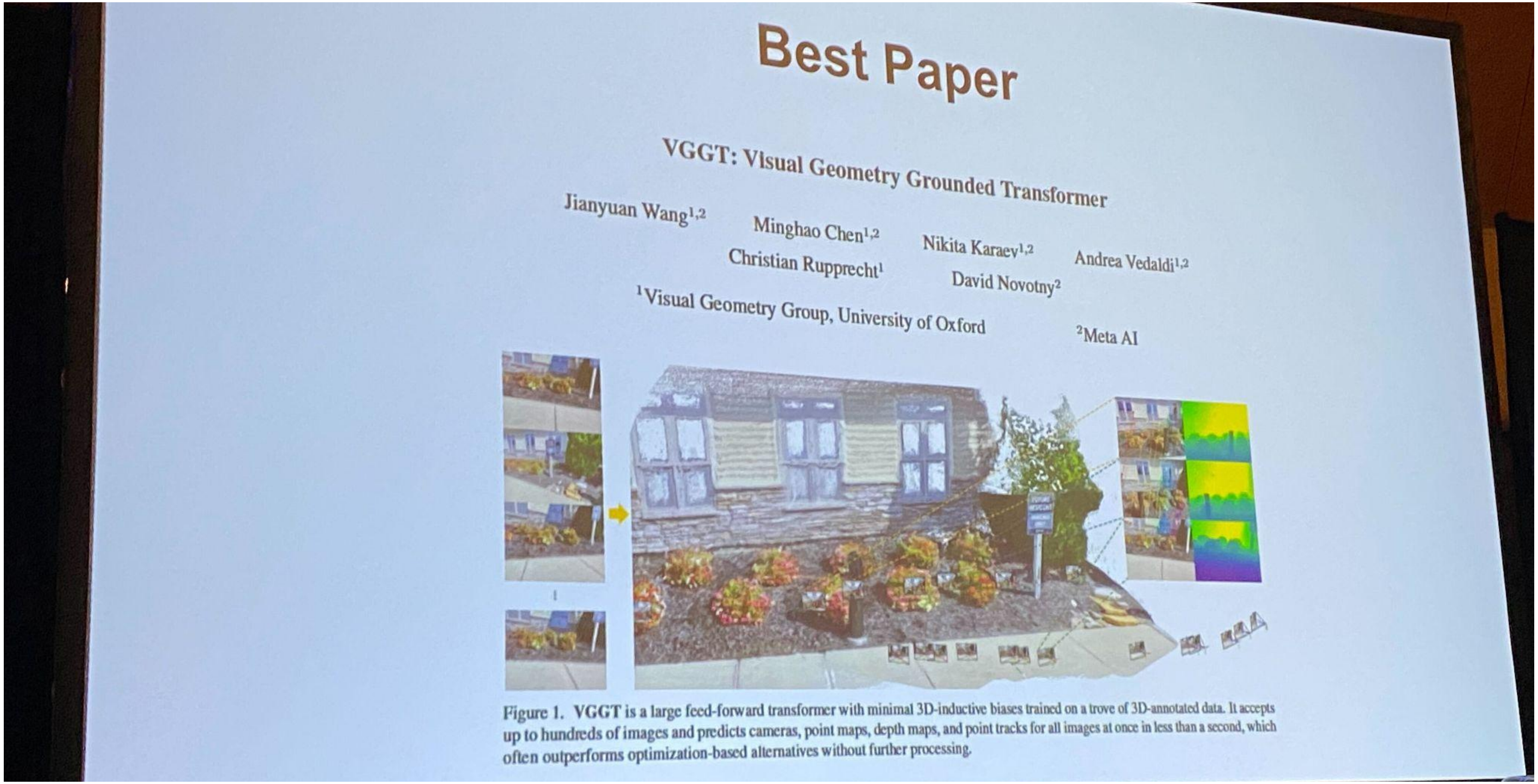
Abstract

Recently, 3D Gaussian Splatting (3DGS) provides a new framework for novel view synthesis, and has spiked a new wave of research in neural rendering and related applications. As 3DGS is becoming a foundational component of many models, any improvement on 3DGS itself can bring huge benefits. To this end, we aim to improve the fundamental paradigm and formulation of 3DGS. We argue that as an unnormalized mixture model, it needs to be neither Gaussians nor splatting. We subsequently propose a new mixture model consisting of flexible Student's t distributions, with both positive (splatting) and negative (scooping) densities. We name our model Student Splatting and Scooping, or SSS. When providing better expressivity,

sian mixture, which provides two advantages. As a general-purpose distribution, Gaussians can approximate an arbitrary density function hence good expressivity. Also, Gaussians have analytical forms under *e.g.* affine transformation, enabling quick evaluation in 3D-2D projection, thus quick learning from images. Meanwhile, splatting provides a flexible way of identifying only the relevant Gaussians to an image for learning. Despite the success, the framework can still suffer from insufficient expressivity [16, 34], and low parameter efficiency, *i.e.* needing a large number of components [17, 28]. Therefore, we re-examine the three key components in 3DGS: Gaussian, splatting, and the optimization. Since its underlying principle is essentially to fit a 3D mixture model to a radiance field, we argue it needs not to be restricted to Gaussians or splatting.

CVPR25: Meta Insights into Trends and Tendencies (36/179)

From award ceremony



Best Paper

VGGT: Visual Geometry Grounded Transformer

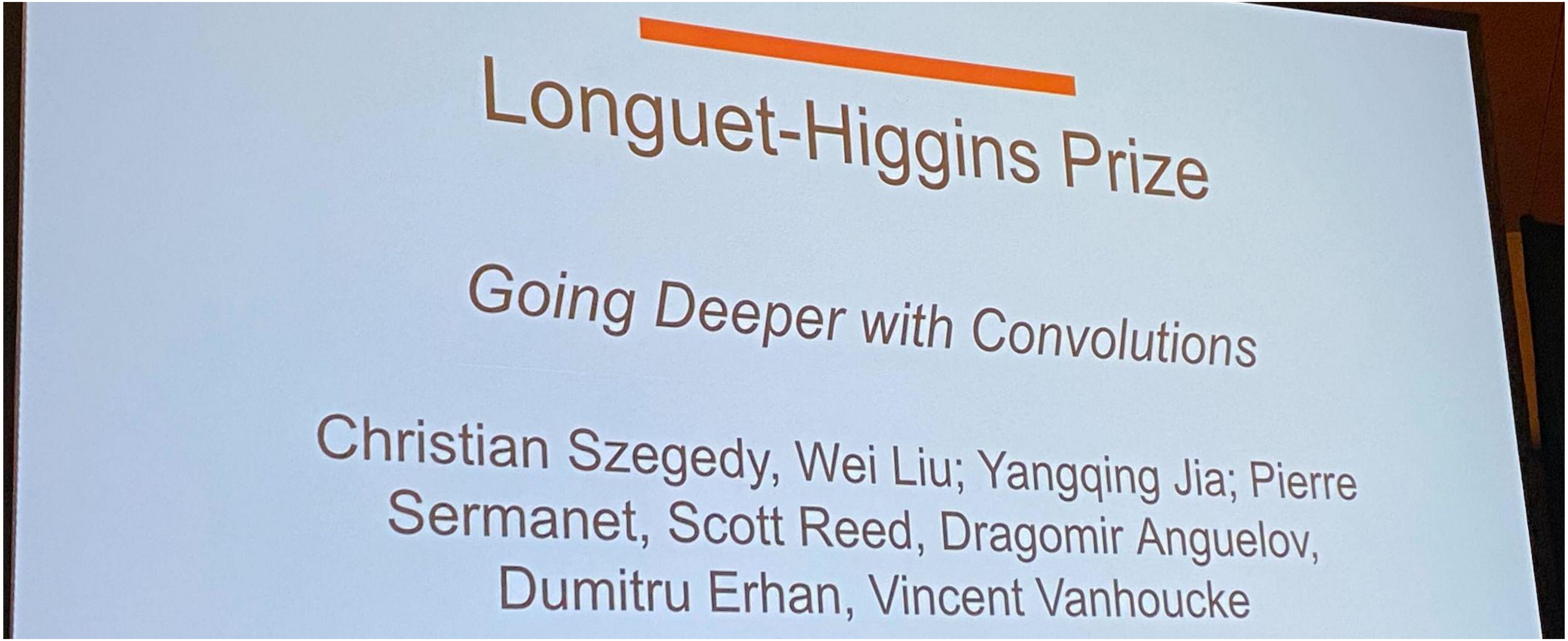
Jianyuan Wang[1,2] Minghao Chen[1,2] Nikita Karaev[1,2] Andrea Vedaldi[1,2]
Christian Rupprecht[1] David Novotny[2]

[1]Visual Geometry Group, University of Oxford [2]Meta AI

Figure 1. **VGGT** is a large feed-forward transformer with minimal 3D-inductive biases trained on a trove of 3D-annotated data. It accepts up to hundreds of images and predicts cameras, point maps, depth maps, and point tracks for all images at once in less than a second, which often outperforms optimization-based alternatives without further processing.

CVPR25: Meta Insights into Trends and Tendencies (37/179)

From award ceremony



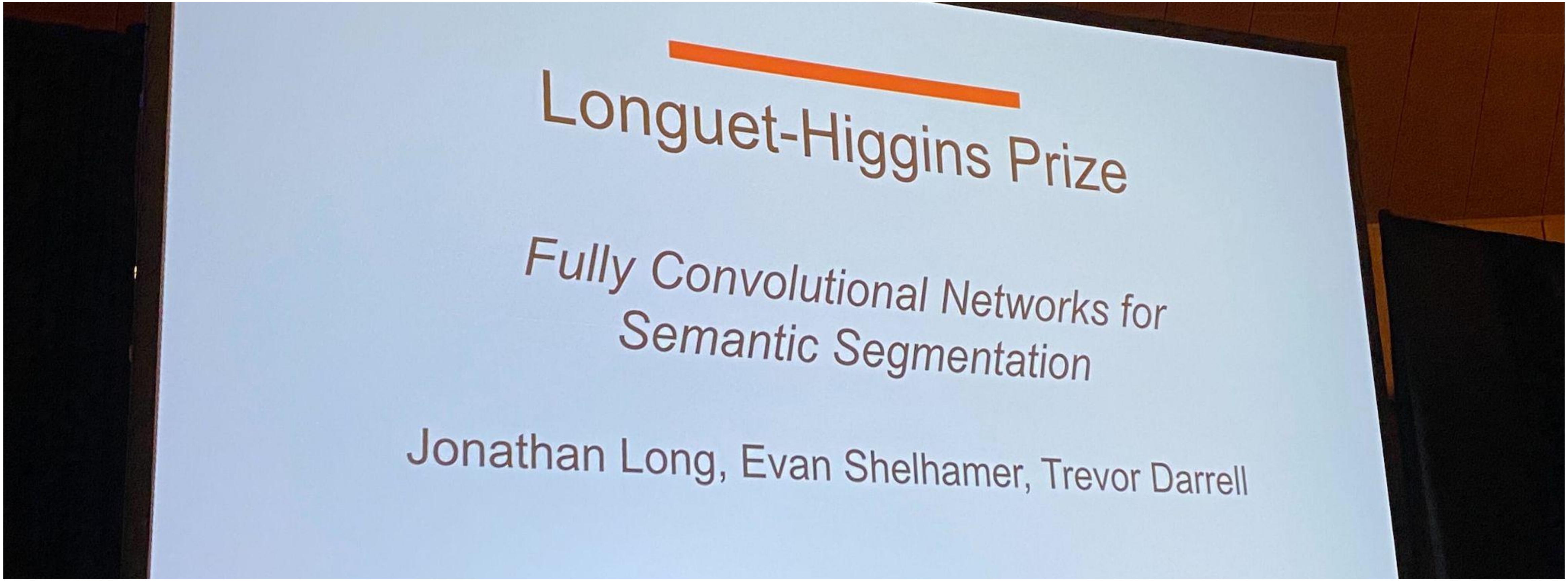
Longuet-Higgins Prize

Going Deeper with Convolutions

Christian Szegedy, Wei Liu; Yangqing Jia; Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke

CVPR25: Meta Insights into Trends and Tendencies (38/179)

From award ceremony



Longuet-Higgins Prize

Fully Convolutional Networks for Semantic Segmentation

Jonathan Long, Evan Shelhamer, Trevor Darrell

CVPR25: Meta Insights into Trends and Tendencies (39/179)

From award ceremony

Young Researcher Award
(Honorable Mention)

Ishan Misra

LIMIT.LAB
https://limitlab.xyz/

CVPR25: Meta Insights into Trends and Tendencies (40/179)

From award ceremony

Young Researcher Award

Hao Su

Saining Xie

LIMIT.LAB
https://limitlab.xyz/

CVPR25: Meta Insights into Trends and Tendencies (41/179)

From award ceremony

Thomas Huang Memorial Prize

Kristen Grauman

CVPR25: Meta Insights into Trends and Tendencies (42/179)

From award ceremony

IEEE Awards

IEEE TECHNICAL FIELD AWARDS

ANIMA ANANDKUMAR

For contributions to AI, including tensor methods and neural operators with applications to scientific domains.

2025 IEEE KIYO TOMIYASU AWARD RECIPIENT

SPONSORED BY The late Dr. Kiyo Tomiyasu

#IEEEAwards

CVPR25: Meta Insights into Trends and Tendencies (43/179)

Unofficial best paper prediction in LIMIT.Lab

❑ This initiative was organized by a group of volunteers to coincide with the release of the Best Paper Award Candidates for CVPR 2025.

❑ Independent from the official review process, it aims to predict which paper will receive the award, based on our own perspectives.

❑ Beyond simple prediction, the goal is to deepen our understanding of the field by carefully reading the nominated papers and evaluating their novelty, impact, and future potential.

❑ Through discussion and exchanging opinions, the initiative also helps broaden our perspectives and refine our criteria for assessing research.

CVPR25: Meta Insights into Trends and Tendencies (44/179)

Unofficial best paper prediction in LIMIT.Lab

- Our Award Selection Process
  - We initiated a call on Slack to form the (unofficial) CVPR 2025 Award Committee and created a dedicated Slack channel for coordination.
  - For every paper, at least one committee member read it thoroughly and created a summary. All members reviewed these summaries to ensure they had at least a basic understanding of every paper before forming their judgments.
  - Each member shared their individual award list, which was then discussed and consolidated through committee deliberation.
  - After forming the award list, the selected papers were re-reviewed before making the final decision.
- Selection Criteria
  - The evaluation primarily follows CVPR regulations, focusing on the level of contribution to the field and the potential to significantly impact its future.
  - Of course, factors such as author prominence or existing citation count are excluded from consideration.

CVPR25: Meta Insights into Trends and Tendencies (45/179)

Unofficial best paper prediction in LIMIT.Lab

- ❑ Our selected papers for each award
  - ❑ Total: 9 papers (referring the CVPR 2024 list; 10 awarded papers in CVPR 2024)
  - ❑ Best Student Paper Honorable Mention（2）
  - ❑ Best Student Paper Award（2）
  - ❑ Best Paper Honorable Mention（3）
  - ❑ Best Paper Award（2）
  - ❑ Note that the award candidate list（https://cvpr.thecvf.com/virtual/2025/events/AwardCandidates2025）and the awardees are slightly different

Unofficial best paper prediction in LIMIT.Lab

- ❑ Our selected papers for each award
  - ❑ Total: 9 papers from the CVPR 2024 list (10 awarded papers in CVPR 2024)
  - ❑ Best Student Paper Honorable Mention（2）
    - ❑ DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models
    - ❑ 3D Student Splatting and Scooping
  - ❑ Best Student Paper Award（2）
    - ❑ Zero-Shot Monocular Scene Flow Estimation in the Wild
    - ❑ Convex Relaxation for Robust Vanishing Point Estimation in Manhattan World
  - ❑ Best Paper Honorable Mention（3）
    - ❑ Navigation World Models
    - ❑ MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos
    - ❑ Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models
  - ❑ Best Paper Award（2）
    - ❑ Generative Multimodal Pretraining with Discrete Diffusion Time-step Tokens
    - ❑ VGGT: Visual Geometry Grounded Transformer

CVPR25: Meta Insights into Trends and Tendencies (47/179)

Unofficial best paper prediction in LIMIT.Lab

- ❑ Our selected papers for each award and reality
  - ❑ Total: 9 → 7 papers from the CVPR 2024 list (10 awarded papers in CVPR 2024)
  - ❑ Best Student Paper Honorable Mention (2→ 1)
    - ❑ ❌DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models
    - ❑ ✅ 3D Student Splatting and Scooping → But, it's Best Paper Honorable Mention
  - ❑ Best Student Paper Award (2→ 1)
    - ❑ ❌Zero-Shot Monocular Scene Flow Estimation in the Wild
    - ❑ ❌Convex Relaxation for Robust Vanishing Point Estimation in Manhattan World
  - ❑ Best Paper Honorable Mention (3→ 4)
    - ❑ ✅ Navigation World Models
    - ❑ ✅ MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos
    - ❑ ✅ Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models
  - ❑ Best Paper Award (2→ 1)
    - ❑ ✅ Generative Multimodal Pretraining with Discrete Diffusion Time-step Tokens → But, it's Best Student Paper Honorable Mention
    - ❑ ✅ VGGT: Visual Geometry Grounded Transformer

CVPR25: Meta Insights into Trends and Tendencies (48/179)

Meta-level insights and BP selected discussion points

- Even before reading the papers, the titles alone were fairly accurate, 6 out of the 7, picked matched well. This shows that it's important to craft titles that clearly convey the contribution and impact to some extent.
- You may not win an award even if you aim for it, but you'll never reach it if you don't aim at all!
- VGGT is the result of years of accumulated effort, including extensive method exploration and parameter tuning, as well as close collaboration within the lab, combining components like VGGSfM and CoTracker.

CVPR25: Meta Insights into Trends and Tendencies (49/179)

Meta-level insights and BP selected discussion points

- ❏ 3D reconstruction was once again a major highlight this year.
- ❏ 3D Reconstruction → Faster and higher-quality results!
- ❏ VLMs (Vision-Language Models) → Open-source momentum and improved controllability over images!
- ❏ Robotics → Groundbreaking applications of world models!

CVPR25: Meta Insights into Trends and Tendencies (50/179)

Meta-level insights and BP selected discussion points

- Reading the CVPR Best Paper candidate studies felt like diving into a collection of science fiction stories, each one filled with a sense of wonder and possibility.
- These works weren’t just about incremental technical improvements; they felt like presentations of futuristic tools. They made me think, “If something this amazing is possible, I’d want to build something like this.” Each paper sparked real inspiration.
- What made them even more compelling was how thoroughly the authors validated their ideas through meticulous experimentation.
- Just like good science fiction is grounded in a sense of realism, these carefully designed experiments gave the research a tangible, exciting credibility.

CVPR25: Meta Insights into Trends and Tendencies (51/179)

Backstory of the Best Paper Award – VGGT (Thank you, Jianyuan Wang, for sharing the story!)

- ❑ VGGSfM in the beginning
  - ❑ Was too complex and difficult to get started with
  - ❑ Follows the classical SfM pipeline: images – correspondences – camera poses & points – bundle adjustment – refined camera poses & points
- ❑ VGGT authors also noticed DUSt3R
  - ❑ Reveals the power of large-scale data-driven paradigm
  - ❑ Went for VGGT – Transformer feedforward only, everyone can use it
- ❑ VGG (Lab) has proposed CoTracker for an accurate point tracker

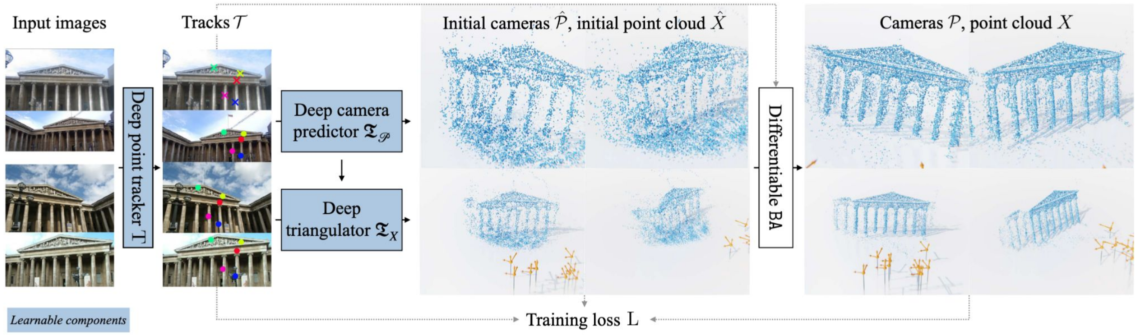

Input images
Tracks $\mathcal{T}$
Initial cameras $\hat{\mathcal{P}}$, initial point cloud $\hat{X}$
Cameras $\mathcal{P}$, point cloud X
Deep point tracker $\mathfrak{T}$
Deep camera predictor $\mathfrak{T}_{\mathcal{P}}$
Deep triangulator $\mathfrak{T}_{X}$
Differentiable BA
Training loss L
Learnable components

Figure 2. **Overview of VGGSfM.** Our method extracts 2D tracks from input images, reconstructs cameras using image and track features, initializes a point cloud based on these tracks and camera parameters, and applies a bundle adjustment layer for reconstruction refinement. The whole framework is fully differentiable and designed for end-to-end training.

J. Wang et al. "VGGSfM: Visual Geometry Grounded Deep Structure From Motion," CVPR 2024. [Link]

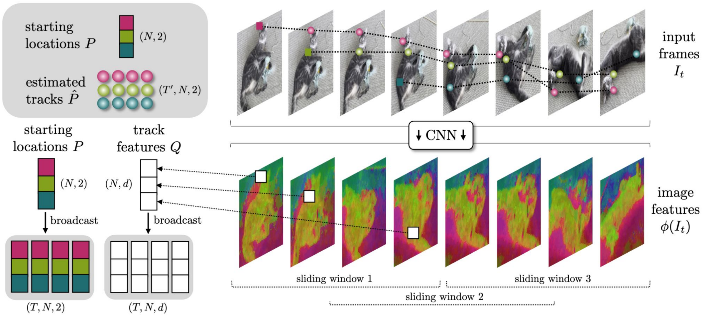


Fig. 3: CoTracker architecture. We compute convolutional features $\phi(I_t)$ for every frame and process them with sliding windows. To initialize track features Q, we bilinearly sample from $\phi(I_t)$ with starting point locations P. Locations P also serve to initialize estimated tracks $\hat{P}$. See Fig. 4 for a visualization of one sliding window.

N. Karaev et al. "CoTracker: It is Better to Track Together," ECCV 2024. [Link]

CVPR25: Meta Insights into Trends and Tendencies (52/179)

Backstory of the Best Paper Award – VGGT (Thank you, Jianyuan Wang, for sharing the story!)

- Flight from 🇬🇧 to 🇺🇸
  - Has been waiting for the passport
  - Got his passport just before the departure
  - Arrived Nashville at midnight before the award ceremony
- Award ceremony
  - No contact from the Program Chairs about the award, CVPR this year felt more like the Oscars
  - Best Paper Award was saved for last (this means it was 'all or nothing' after the honorable mention)
  - Comments from Jianyuan:

  I was extremely nervous. I still don't know how to fully describe the feeling.

  Even now, it all feels kind of unreal.

CVPR25: Meta Insights into Trends and Tendencies (53/179)

Backstory of the Best Paper Award – VGGT

- Perfect switching from VGGNet to VGGTransformer!
  - VGGNet (ICLR 2015)
    - Convolutional Neural Networks (CNN)
    - A backbone for image recognition
  - VGGTransformer / VGGT (CVPR 2025) – ten years after the VGGNet
    - Transformer
    - 3D reconstruction in a single Transformer
    - A backbone for 3D vision
  - Rough history of Visual Geometry Group (VGG)
    - Before 2010s: 3D geometry is the main focus, but there are recognition researches
    - After 2010s: Deep learning is the main focus, but there are 3D vision researches
    - VGGT: The strong cross point between 3D geometry & deep learning
    - After 2025?: Everyone is able to participate in 3D x deep learning?

LIMIT.LAB

CVPR25: Meta Insights into Trends and Tendencies (54/179)

Trends of Medical x CV research

a. Vision and language for medical domain

i. Medical VQA with noisy labels and diffusion [Guo+, CVPR 2025]

ii. VLMs alignment with CoOP [Koleilat+, CVPR 2025]

iii. VQA with visual reference [Chen+, CVPR 2025]

iv. VLM with soft label [Ko+, CVPR 2025]

v. Pre-training of VLM [Ziyang+, CVPR2025]

b. Semi-supervised learning for medical

i. The task has been focused on for a long time, but it is still prevalent this year.

ii. Is this evidence that the lack of labeled data has not been resolved, even with the foundation model in medical?

iii. Deal with annotation ambiguity [Kumari+, CVPR 2025]

iv. Depth guided segmentation [Li+, CVPR 2025]

v. Find overconfidence prediction of foundation model [Ma+, CVPR 2025]

vi. Unsupervised prompting for SAM by DPO-inspired loss [Konwer+, CVPR 2025]

vii. Uncertainty-aware consistency and contrastive [Assefa+, CVPR 2025]

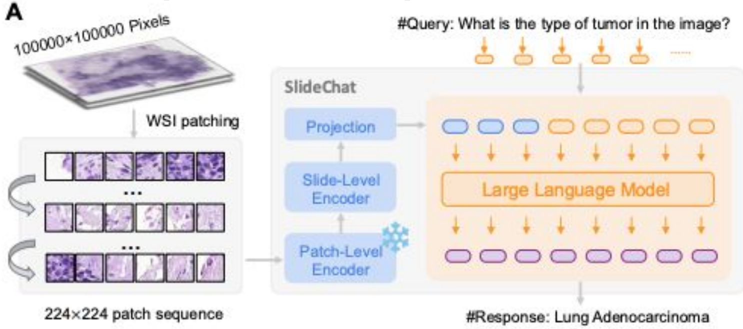
CVPR25: Meta Insights into Trends and Tendencies (55/179)

Trends of Pathology x CV research

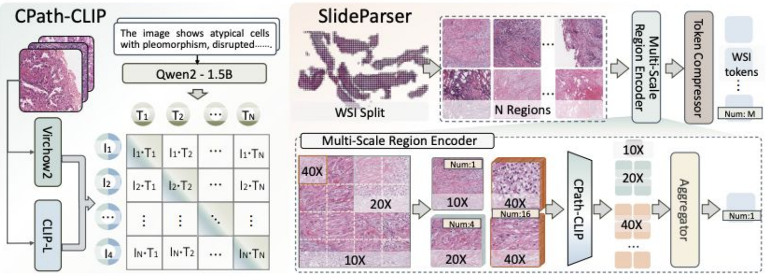
a. Vision and language for pathology

   i. Patch-level VLLM -> multi-resolution (including patch, slide-level) VLLM

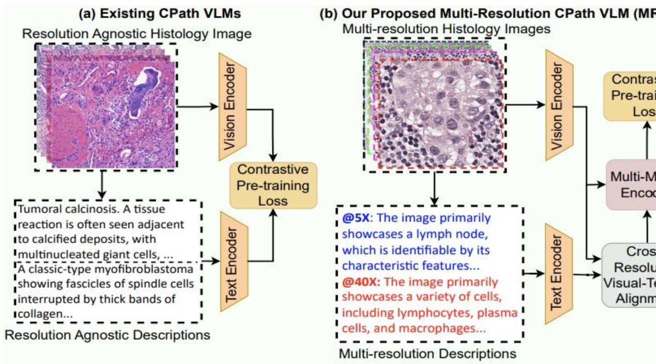
[Chen+, CVPR 2025]



[Sun+, CVPR 2025]



[Albastaki+, CVPR 2025]


(a) Existing CPath VLMs
Resolution Agnostic Histology Image
Vision Encoder
Contrastive Pre-training Loss
Text Encoder
Tumoral calcinosis. A tissue reaction is often seen adjacent to calcified deposits, with multinucleated giant cells, ...
A classic-type myofibroblastoma showing fascicles of spindle cells interrupted by thick bands of collagen...
Resolution Agnostic Descriptions
(b) Our Proposed Multi-Resolution CPath VLM (MR
Multi-resolution Histology Images
Vision Encoder
Text Encoder
@5X: The image primarily showcases a lymph node, which is identifiable by its characteristic features...
@40X: The image primarily showcases a variety of cells, including lymphocytes, plasma cells, and macrophages...
Multi-resolution Descriptions
Contras Pre-trai Los
Multi-M Encod
Cros Resolu Visual-Te Alignm

b. mamba-based model for whole slide image

   i. To effectively aggregate patch-level information for mamba

[Zhang+, CVPR 2025]

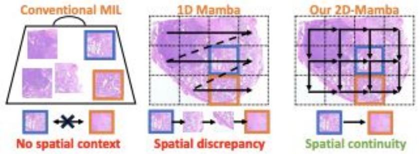
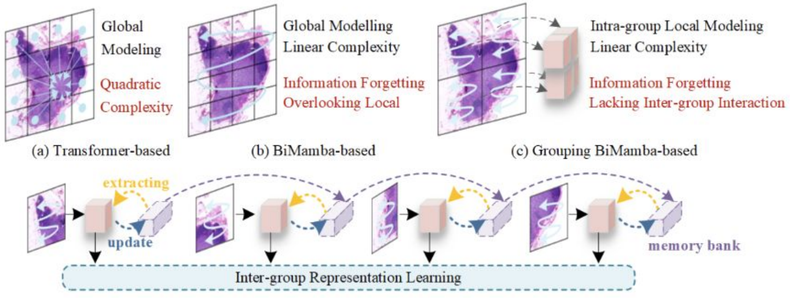


Figure 1. **Left: Conventional MIL Bagging** of patches adopts *no spatial context*. **Middle: 1D Mamba-based methods** flatten a WSI into a 1D sequence and lose the 2D structure. The adjacent blue and orange patches are far away in the sequence. We call this "*spatial discrepancy*". **Right: 2DMamba** processes a WSI in a 2D manner, preserving 2D structure. The blue and orange patches maintains adjacent in the sequence. We call it "*spatial continuity*".

[Zheng+, CVPR 2025]



(a) Transformer-based (b) BiMamba-based (c) Grouping BiMamba-based

(d) Memory BiMamba (Ours)

CVPR25: Meta Insights into Trends and Tendencies (56/179)

Direct Preference optimization (DPO) for Vision and language

Boosting for in-context leaning with SymDPO [Jia+, cvpr 2025]

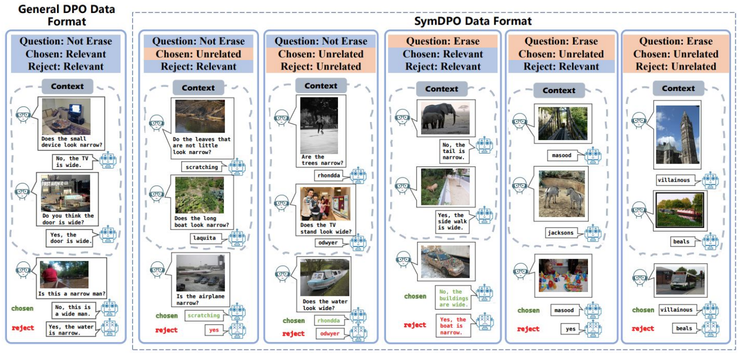


Figure 2. Comparison of General DPO and SymDPO Formats: General DPO relies solely on standard text for Questions, Answers, Chosen, and Rejected Answers, focusing on text-based training. In contrast, SymDPO replaces textual Answers with symbolized text to boost multimodal understanding, requiring models to interpret both visual and symbolized cues. This approach strengthens the model's ability to reason and decide in complex multimodal contexts.

Debiasing VLM with Noise-aware preference optimization [Zhang+, cvpr 2025]

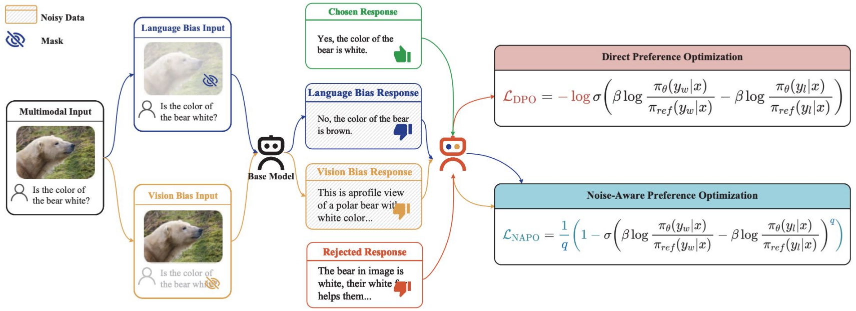

Noisy Data
Mask
Multimodal Input
Is the color of the bear white?
Language Bias Input
Vision Bias Input
Base Model
Chosen Response
Yes, the color of the bear is white.
Language Bias Response
No, the color of the bear is brown.
Vision Bias Response
This is aprofile view of a polar bear with white color...
Rejected Response
The bear in image is white, their white fur helps them...
Direct Preference Optimization
$\mathcal{L}_{\mathrm{DPO}} = -\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)$
Noise-Aware Preference Optimization
$\mathcal{L}_{\mathrm{NAPO}} = \frac{1}{q}\left(1-\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)^q\right)$

Figure 2. **Method details.** First, biased responses are constructed by using masking to guide the model toward over-relying on prompts and generating responses based on the base model. Next, NaPO is applied for noise-robust preference optimization to counteract noise in automatically constructed data, dynamically assessing data noise levels to calculate NaPO's noise robustness coefficient q (see Equation (12)). Here we assumed that the original data is of high quality, so DPO is used to train on it directly. Additional experiments were conducted with NaPO on the original data, and the results can be found in Appendix A.

Improving VLM with task preference optimization by introduce task specific token [Yan+, cvpr 2025]

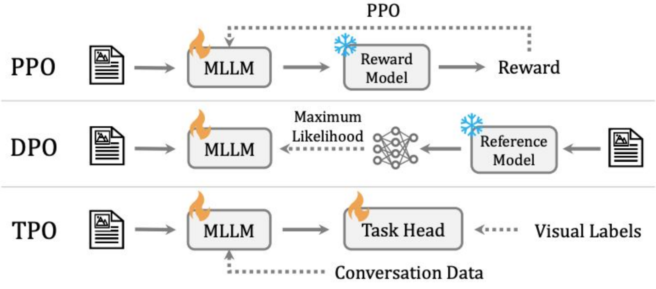


Figure 2. **Comparison of Learning Method.** A solid line indicates data flow, and a dotted line represents feedback. ❄ and 🔥 denote modules that are frozen and unfrozen.

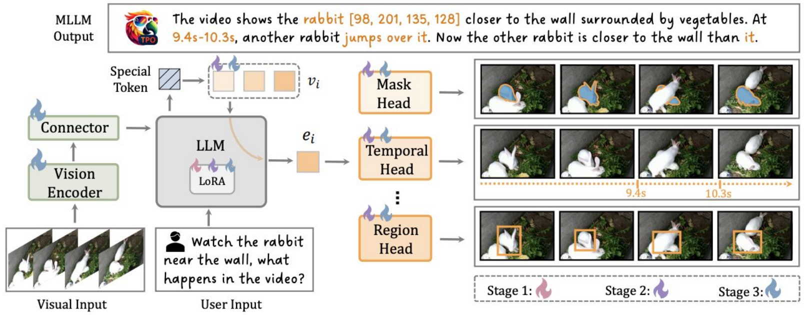

MLLM Output
The video shows the rabbit [98, 201, 135, 128] closer to the wall surrounded by vegetables. At 9.4s-10.3s, another rabbit jumps over it. Now the other rabbit is closer to the wall than it.
Special Token
v_i
e_i
Connector
Vision Encoder
LLM
LoRA
Mask Head
Temporal Head
Region Head
9.4s
10.3s
Watch the rabbit near the wall, what happens in the video?
Visual Input
User Input
Stage 1:
Stage 2:
Stage 3:

Figure 3. **Overall Pipeline of TPO.** The architecture of Task Preference Optimization (TPO) consists of four main components: (1) a vision encoder, (2) a connector, (3) a large language model, and (4) a series of visual task heads. Differently colored flame symbols indicate which components are unfrozen at various stages of the training process.

CVPR25: Meta Insights into Trends and Tendencies (57/179)

Direct Preference optimization (DPO) for diffusion model

Step-by-step preference optimization use PO for each step [Liang+, cvpr 2025]

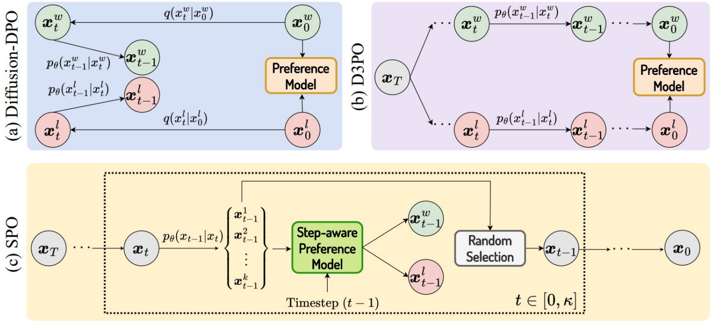


Figure 3. Comparing frameworks of SPO, Diffusion-DPO, and D3PO approaches. SPO does not adopt direct preference propagation as other DPO methods do. In SPO, a pool of samples are generated at each step, from which a proper win/lose pair is selected and used to fine-tune the diffusion model. Then, a single sample is randomly selected to initialize the next iteration.

DPO with curriculum learning [NA+, cvpr 2025] [Croitoru+, cvpr 2025]

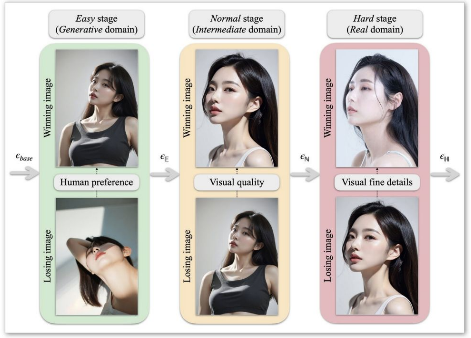
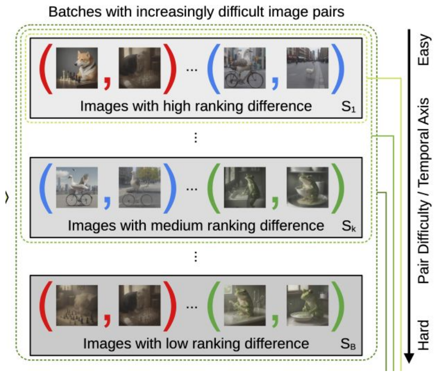


Figure 2. **Three-stage training of HG-DPO.** It progressively enhances the model's human image generation capabilities.



Batches with increasingly difficult image pairs
Images with high ranking difference
S₁
Images with medium ranking difference
Sₖ
Images with low ranking difference
S_B
Easy
Pair Difficulty / Temporal Axis
Hard

Calibrated Preference Optimization using multiple reward model [Lee+, cvpr 2025]

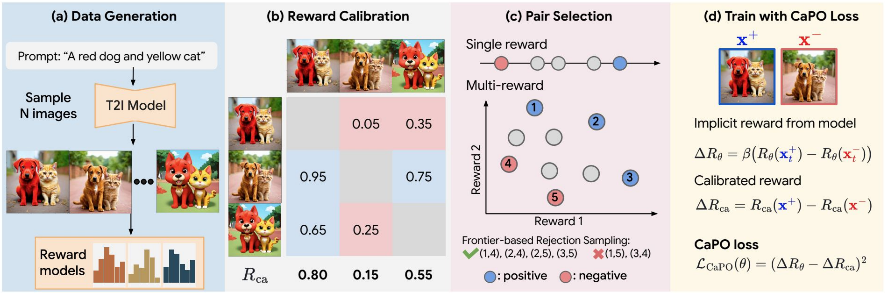


Figure 2. **Overview.** (a) We generate N images using pretrained T2I diffusion model using the prompt dataset, and infer the scores from reward models. (b) Then, we calibrate the rewards by making pairwise comparison between images. For each image, we compute the win-rates between other $N-1$ images using Eq. (2), and average them to obtain calibrated reward R_{ca} (see Sec. 4.2). (c) We select pair by choosing the best-of-N and worst-of-N when using single reward. For multi-reward, we use non-dominated sorting algorithm to select upper Pareto set as positives, and lower Pareto set as negatives. The accepted and rejected pairs are also listed using proposed rejection sampling method. (d) Lastly, during training, we select a pair from (c), and compute CaPO loss (*i.e.*, Eq. (8)), which perform regression task to match the difference in calibrated rewards (*i.e.*, ΔR_{ca} by the difference of implicit reward model (*i.e.*, ΔR_θ).

Inversion preference optimization with reparametrization DDIM [Lu+, cvpr 2025]

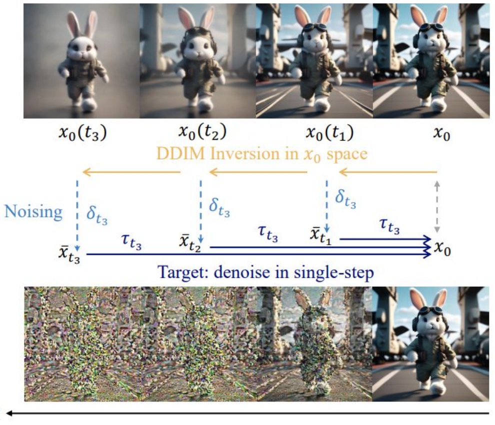


Figure 2. Illustration of Inversion for Preference Optimization.

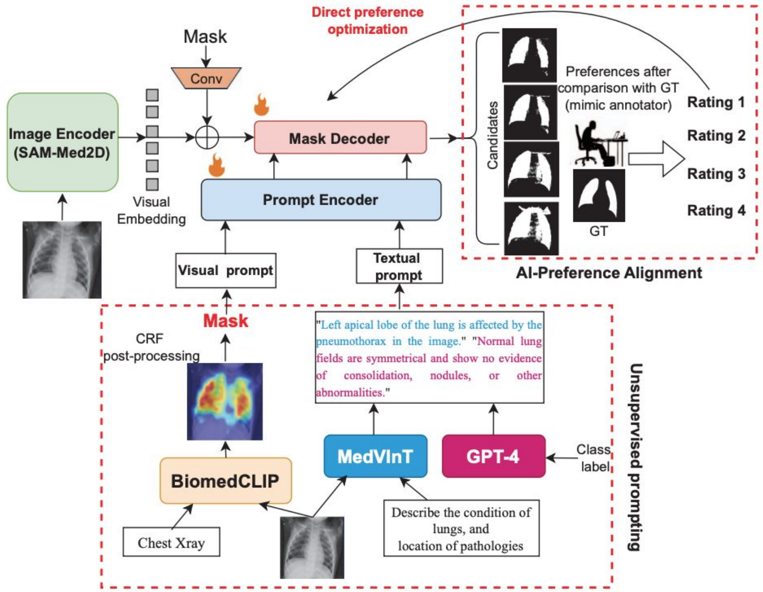
CVPR25: Meta Insights into Trends and Tendencies (58/179)

Direct Preference optimization (DPO) for downstream task

May DPO-based fine-tuning become the next trend?

Enhancing SAM with Efficient Prompting and Preference Optimization for Semi-supervised Medical Image Segmentation [Konwer+, cvpr 2025]

- Efficient unsupervised prompting strategy that enhances segmentation performance
- Step1: Add preferences ratings for candidates that is estimated by vision and language model.
- Step2: Train model with DPO



Direct preference optimization
Mask
Conv
Image Encoder (SAM-Med2D)
Visual Embedding
Mask Decoder
Prompt Encoder
Visual prompt
Textual prompt
Candidates
Preferences after comparison with GT (mimic annotator)
Rating 1
Rating 2
Rating 3
Rating 4
GT
AI-Preference Alignment
CRF post-processing
Mask
"Left apical lobe of the lung is affected by the pneumothorax in the image." "Normal lung fields are symmetrical and show no evidence of consolidation, nodules, or other abnormalities."
MedVInT
GPT-4
Class label
BiomedCLIP
Chest Xray
Describe the condition of lungs, and location of pathologies
Unsupervised prompting

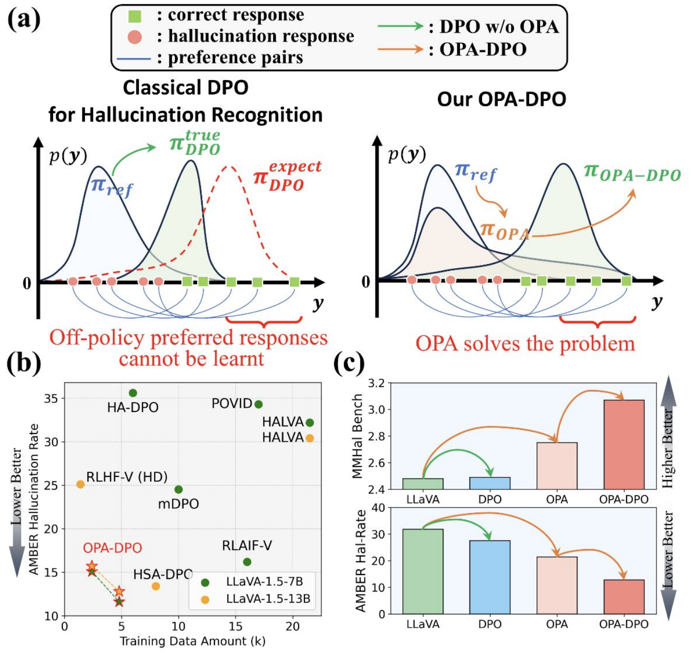
Figure 2. **Illustration of the proposed framework for semi-supervised segmentation:** Unsupervised geometric and text prompts, obtained from pretrained BiomedCLIP, MedVInT, and GPT-4 models, are fed into the prompt encoder for finetuning the framework on a small fraction of annotated data. In the next stage, we simulate a virtual annotation process that assigns ratings to the generated segmentation candidates, which are used to fine-tune the decoder. This stage handles unannotated data, as the model does not rely on ground truth for direct supervision but only for rating while simulating a human annotator's feedback.

We should tackle label-efficient dataset generation for DPO :

- Labor efficiency
  - Utilize foundation model?
- Ensure quality of data
  - Rethinking label-efficient techniques?
- Bias of data
  - Integrates multiple models?

Mitigating Hallucinations in Large Vision-Language Models via DPO: On-Policy Data Hold the Key

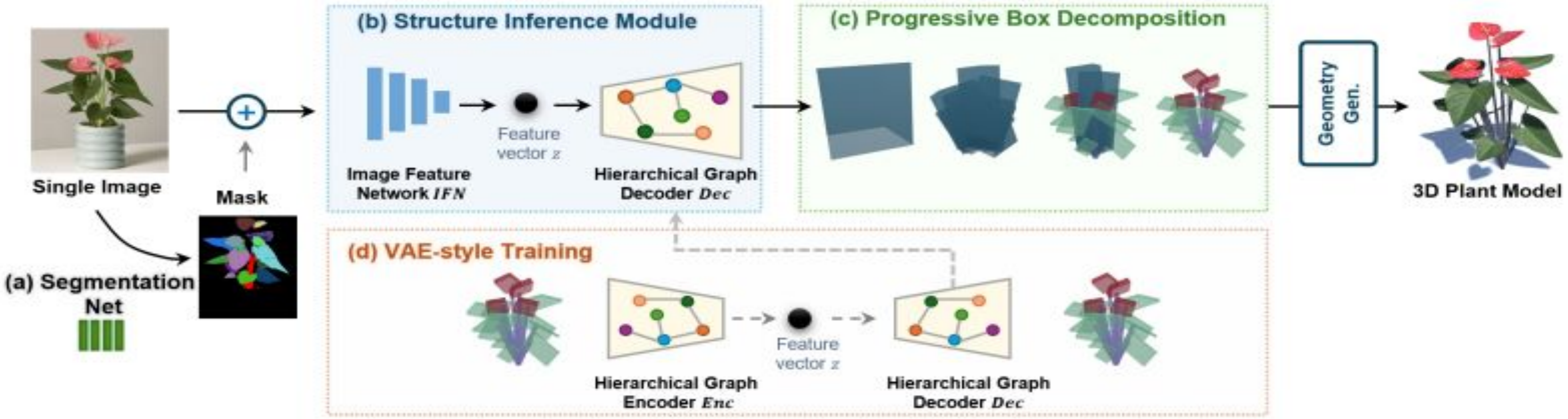
- ○ DPO cannot learn strictly off-policy preferred answers (reverse-KL → ∞)
- ○ Make expert-corrected responses on-policy before DPO
- ○ Overall Pipeline:
  - ■ 1. Collect & Rate
    - ● Base LVLM → response candidates
    - ● GPT-4V tags sentence-level hallucination severity / error type and minimally revises text (4.8k pairs)
  - ■ 2. On-Policy Alignment (OPA)
    - ● LoRA-SFT on original + revised answers → brings corrections into model's support.
  - ■ 3. OPA-DPO Fine-tune with three preference pairs
    - ● Language Correction (hallucination-weighted)
    - ● Image Focus (clean vs 30 %-masked image)
    - ● Anchored Preference (keep preferred probs from drifting)

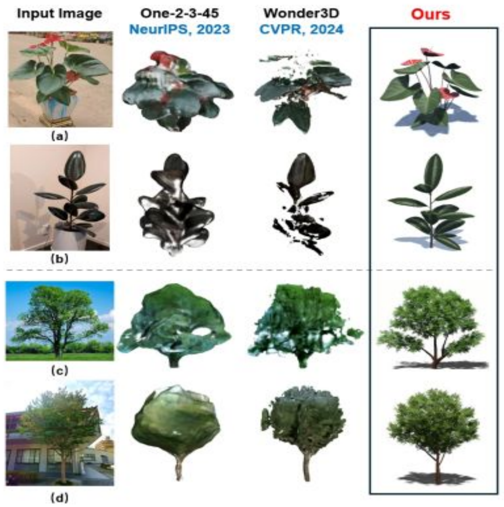


CVPR25: Meta Insights into Trends and Tendencies (60/179)

Neural Hierarchical Decomposition for Single Image Plant Modeling

- Abstract:
  - Constructing high-quality 3D models of biological plants is challenging. They tackled on the 3D decomposition of plants from a single image.
  - segmentation → structure inference → box decomposition → VAE
- Key point:
  - This approach is adaptable for both of the indoor and outdoor trees


Single Image
Mask
(a) Segmentation Net
(b) Structure Inference Module
Feature vector z
Image Feature Network IFN
Hierarchical Graph Decoder Dec
(c) Progressive Box Decomposition
Geometry Gen.
3D Plant Model
(d) VAE-style Training
Hierarchical Graph Encoder Enc
Feature vector z
Hierarchical Graph Decoder Dec

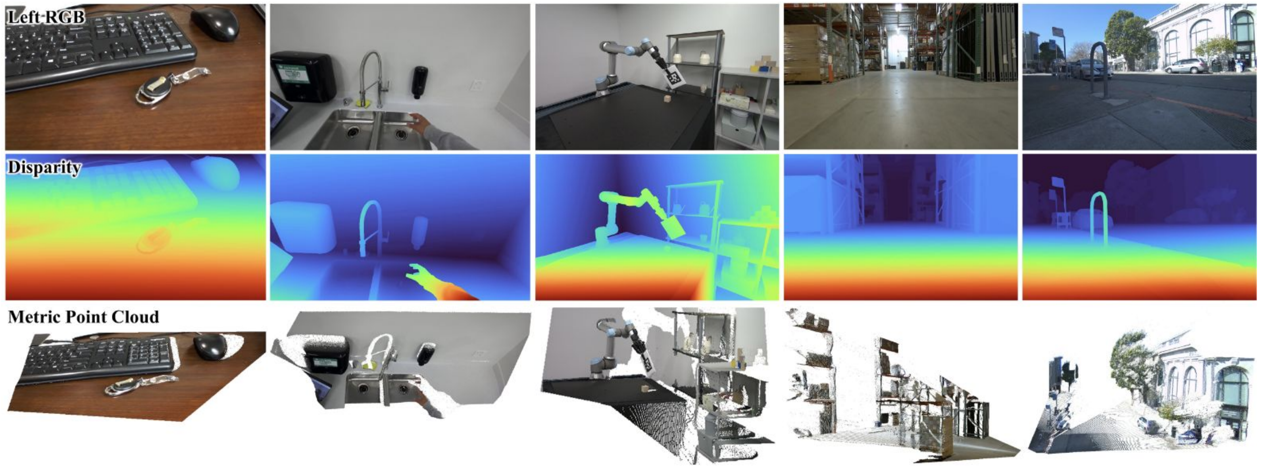


Zhihao et al., “Neural Hierarchical Decomposition for Single Image Plant Modeling”, CVPR 2025, [Link]

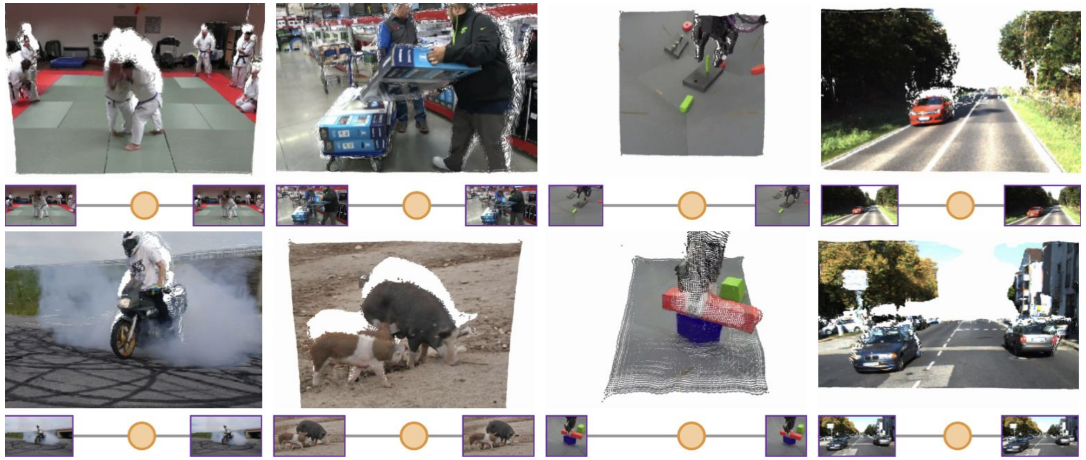
CVPR25: Meta Insights into Trends and Tendencies (61/179)

Synthetic data has high potential

- From the perspective of best paper candidates,
  - 1. Synthetic data – Data generation using world simulator, generative models, and other resources
  - 2. Language-free vision foundation model – Depth, optical flow, matting, segmentation
  - 3. Zero-shot / in-the-wild recognition – Without an additional fine-tuning, the synthetic pre-training is working on real-world data



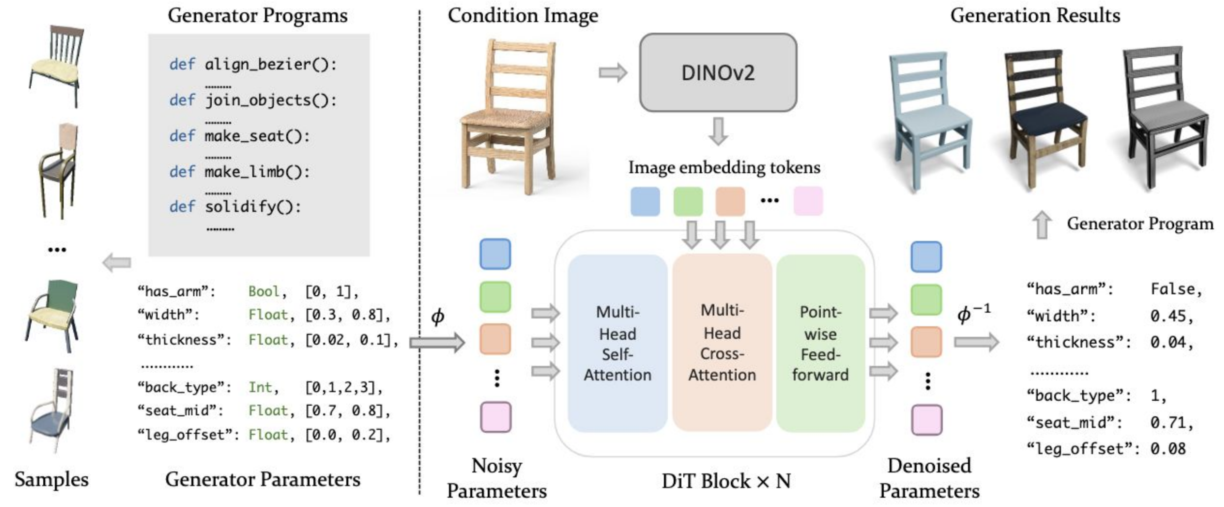
B. Wen et al. "FoundationStereo: Zero-Shot Stereo Matching," CVPR 2025. [Link]



Y. Liang et al. "Zero-Shot Monocular Scene Flow Estimation in the Wild," CVPR 2025. [Link]

Synthetic data has high potential

- ❑ From the perspective of CVPR 2025 papers,
- ❑ Infinigen / Infinigen Indoors have been applied the following papers
  - ❑ DI-PCG: An inverse PCG (Procedural Content Generation) framework that realizes image-to-3D generation via parameter estimation in diffusion models
  - ❑ BlenderGym: A comprehensive benchmark for real-world graphics editing tasks by having VLMs to reconstruct 3D scenes from a start to a goal state.

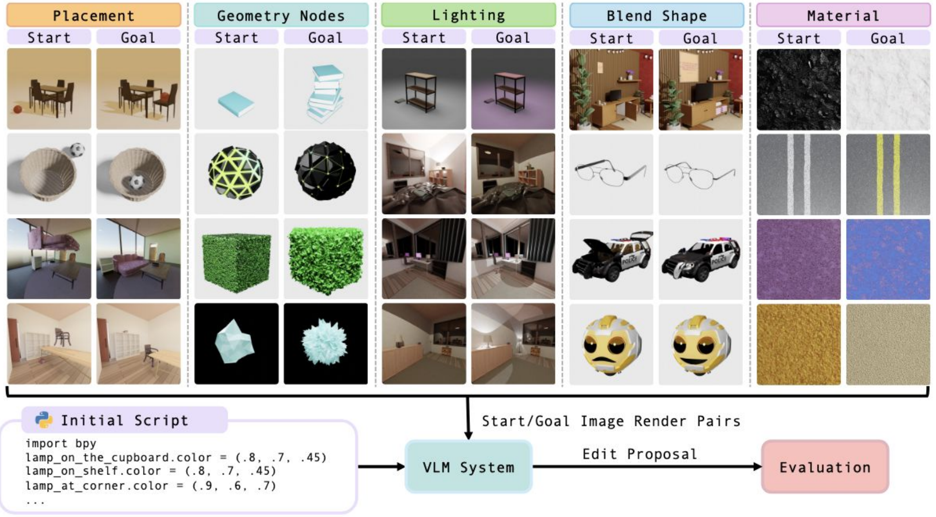


```
def align_bezier():
def join_objects():
def make_seat():
def make_limb():
def solidify():
```

```
"has_arm":    Bool,  [0, 1],
"width":      Float, [0.3, 0.8],
"thickness":  Float, [0.02, 0.1],
............
"back_type":  Int,   [0,1,2,3],
"seat_mid":   Float, [0.7, 0.8],
"leg_offset": Float, [0.0, 0.2],
```

```
"has_arm":    False,
"width":      0.45,
"thickness":  0.04,
............
"back_type":  1,
"seat_mid":   0.71,
"leg_offset": 0.08
```

W. Zhao et al. “DI-PCG: Diffusion-based Efficient Inverse Procedural Content Generation for High-quality 3D Asset Creation,” CVPR 2025. [Link]

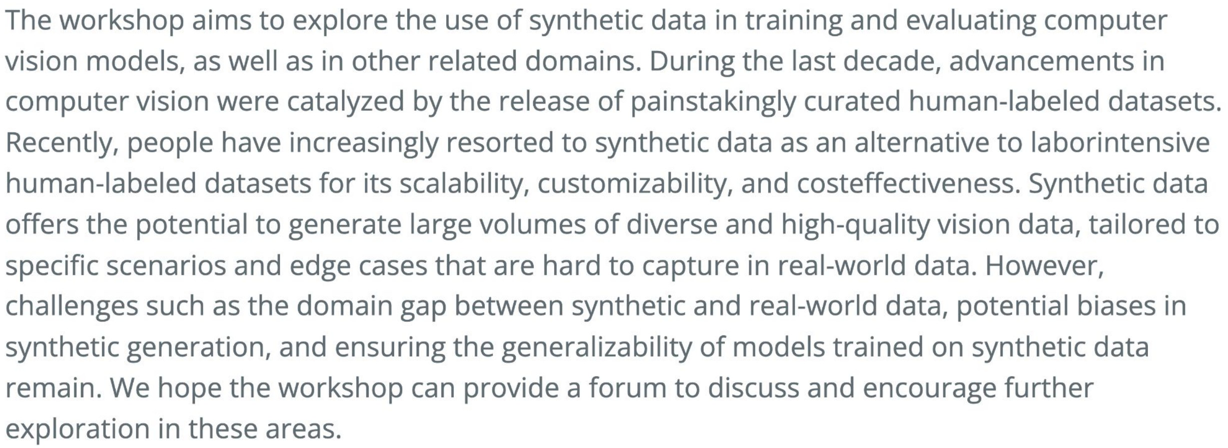


```
import bpy
lamp_on_the_cupboard.color = (.8, .7, .45)
lamp_on_shelf.color = (.8, .7, .45)
lamp_at_corner.color = (.9, .6, .7)
...
```

Y. Gu et al. “BlenderGym: Benchmarking Foundational Model Systems for Graphics Editing,” CVPR 2025. [Link]

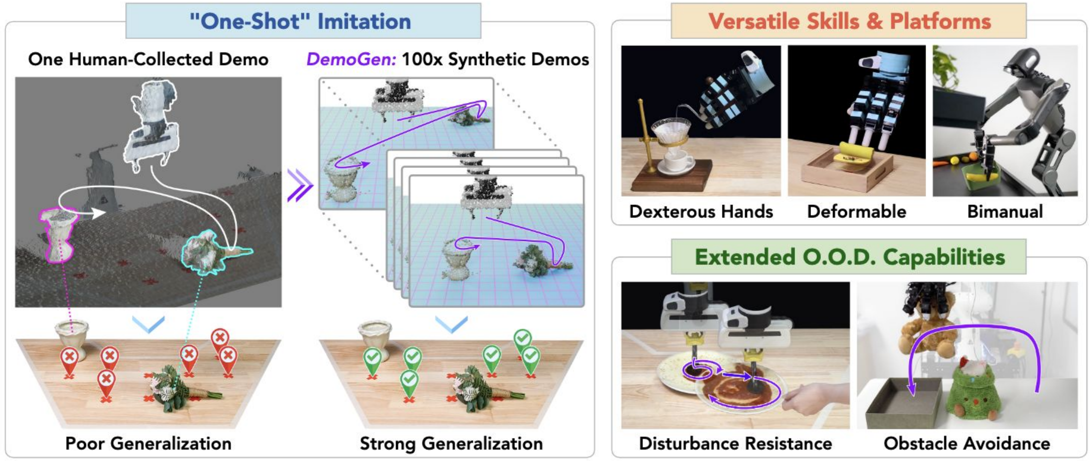
Synthetic data has high potential

- From the perspective of CVPR 2025 SynData4CV Workshop,
- Synthetic / generated data can be used training for vision / VL models
  - Over 60+ posters in the SynData4CV Workshop
  - Many tasks such as classification, detection, segmentation, 3D recognition, anomaly detection, video, robot manipulation



The workshop aims to explore the use of synthetic data in training and evaluating computer vision models, as well as in other related domains. During the last decade, advancements in computer vision were catalyzed by the release of painstakingly curated human-labeled datasets. Recently, people have increasingly resorted to synthetic data as an alternative to laborintensive human-labeled datasets for its scalability, customizability, and costeffectiveness. Synthetic data offers the potential to generate large volumes of diverse and high-quality vision data, tailored to specific scenarios and edge cases that are hard to capture in real-world data. However, challenges such as the domain gap between synthetic and real-world data, potential biases in synthetic generation, and ensuring the generalizability of models trained on synthetic data remain. We hope the workshop can provide a forum to discuss and encourage further exploration in these areas.

Workshop overview

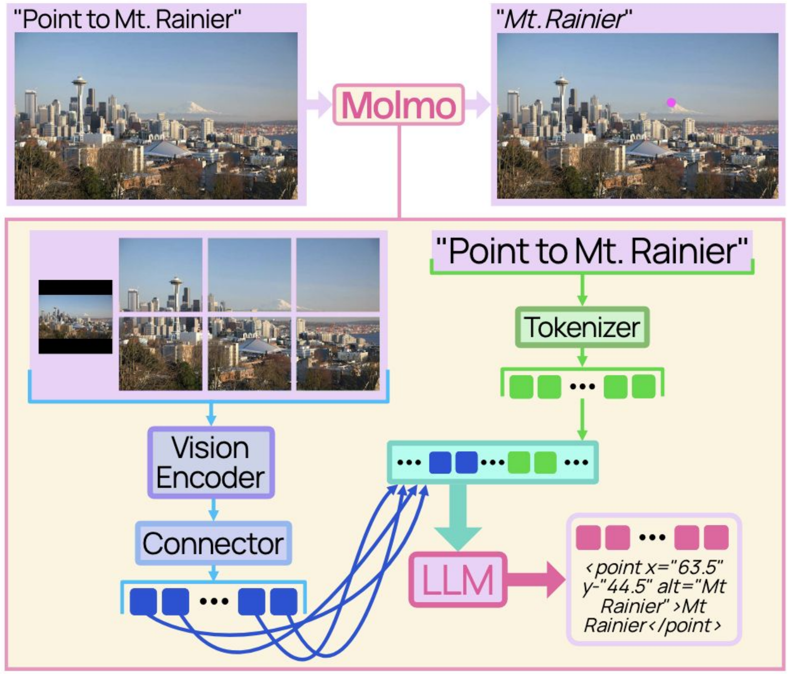


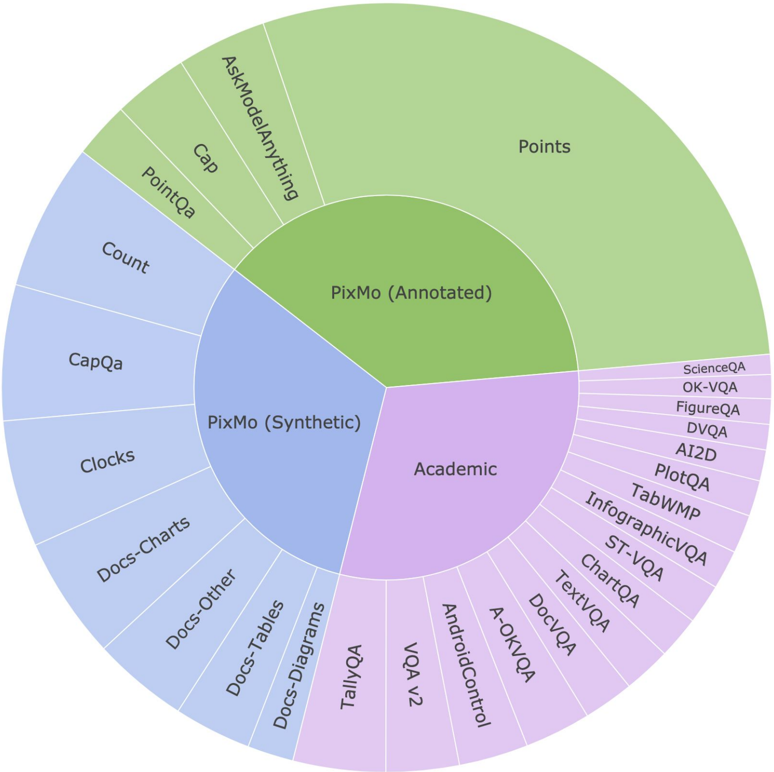
Best Long Paper: Synthetic data for robot policy learning

Z. Xue et al. “DemoGen: Synthetic Demonstration Generation for Data-Efficient Visuomotor Policy Learning,” CVPR 2025 SynData4CV Workshop. [Link]

Molmo & PixMo: State-of-the-art open-sourced VLM

- ❏ Molmo & PixMo have been proven:
  - ❏ Data quality rather than simple model scale (once again!)
  - ❏ Molmo keeps the standard architecture and simple VL connections
  - ❏ For the purpose of open strategy, PixMo consists of public & synthetic data

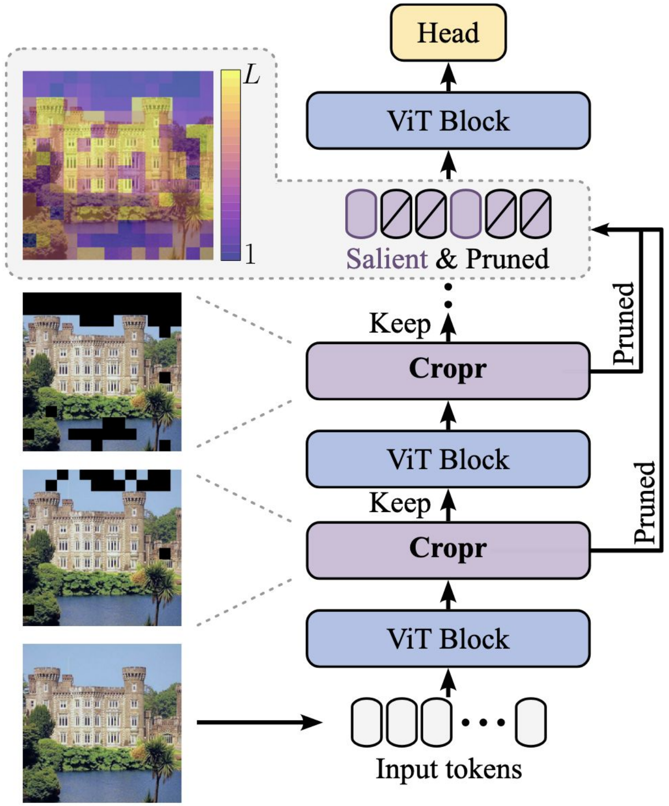


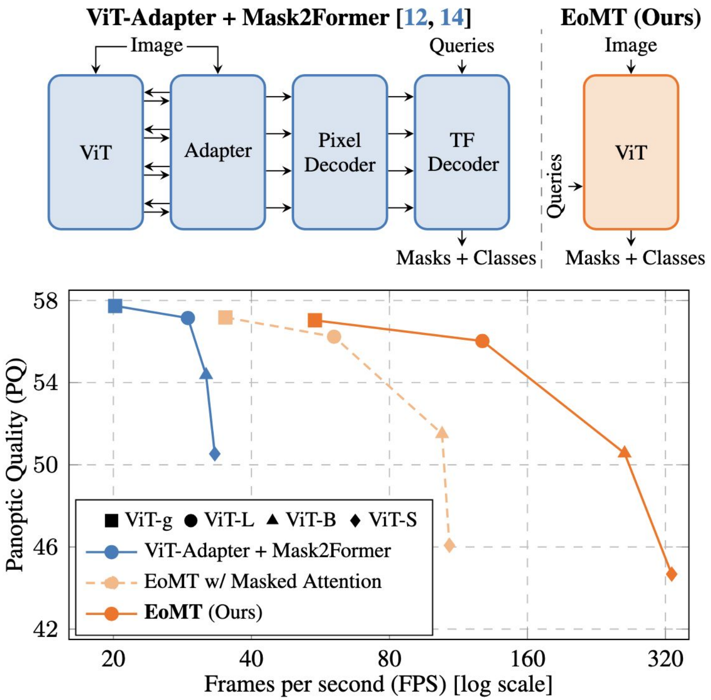


CVPR25: Meta Insights into Trends and Tendencies (65/179)

ViT (Vision Transformer) has been improved several points!

- ❑ Token Cropr ViT
  - ❑ Token pruning that uses auxiliary prediction heads
- ❑ Your ViT is Secretly an Image Segmentation Model
  - ❑ With learnable queries + mask logits, simple ViT can be a segmentor

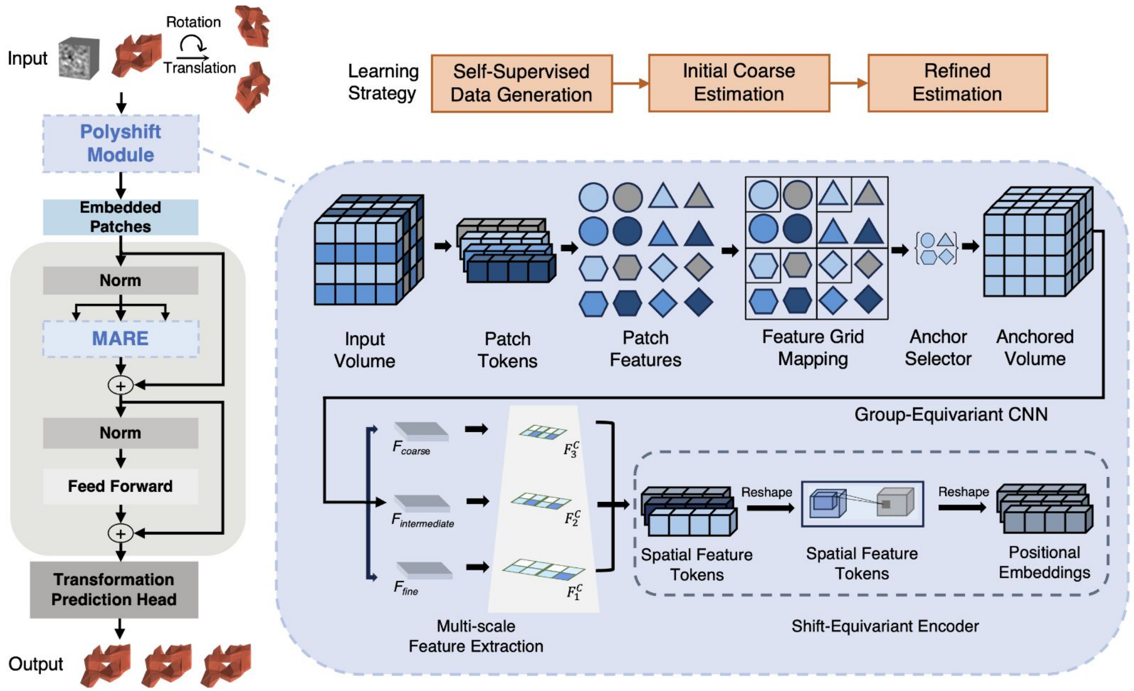


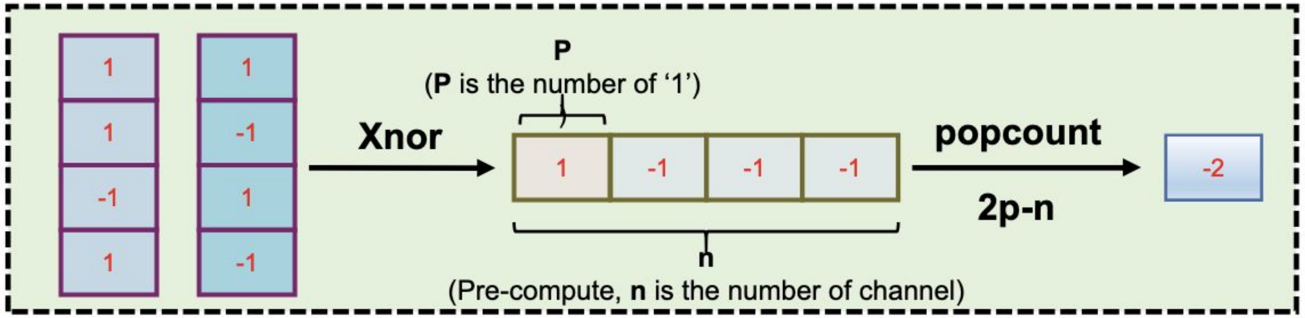


CVPR25: Meta Insights into Trends and Tendencies (66/179)

ViT (Vision Transformer) has been improved several points!

- ❑ BOE-ViT (Boosting Orientation Estimation ViT)
  - ❑ ViT for 3D subtomogram alignment through shift/rotation estimation
- ❑ BHViT (Binarized Hybrid ViT)
  - ❑ Full binarized ViT model / Xnor and popcount operations





…and some more methods regarding ViT architecture

CVPR25: Meta Insights into Trends and Tendencies (67/179)

Synthetic data for computer vision

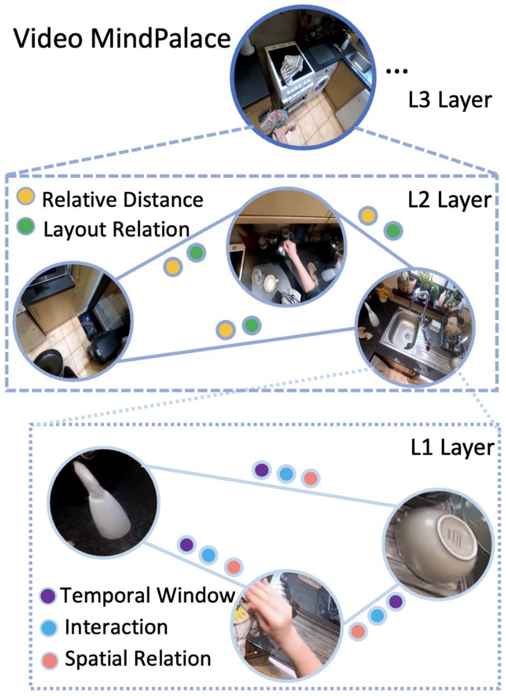
- **CLIPasso: Semantically-Aware Object Sketching**
- Generation Sketch images tasks is important.
  - Many artistic and scientific ideas begin as sketches.
- Challenges in generating sketch images by diffusion models.
  - Key point: collecting a large-scale sketch dataset is challenging because existing sketches are limited in number and often subject to copyright restrictions.
- Approach: Generating sketch images from large real-image datasets with ControlNet.
  - The synthetic sketches can then be used to train new diffusion models, enabling large-scale, copyright-free sketch generation.
- https://clipasso.github.io/clipasso/

Graph structures for video understanding using VLMs

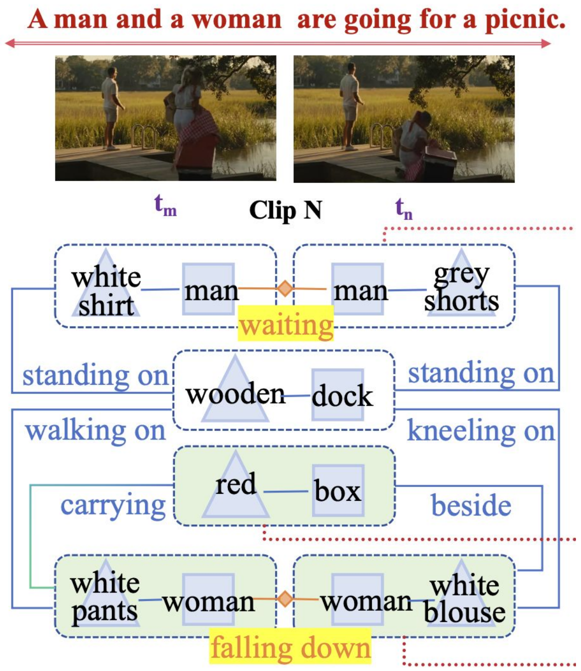
- ❏ [1] Spatio-temporal scene graph, enabling VLM to self-generate detailed multi-step reasoning data and improve their compositional reasoning abilities. (Qui et al.[Link],)

- ❏ [2]Topological semantic graph, empowering VLM to deeply understand the 3D context of complex long-form videos and perform human-like reasoning. (Huang et al.[Link])

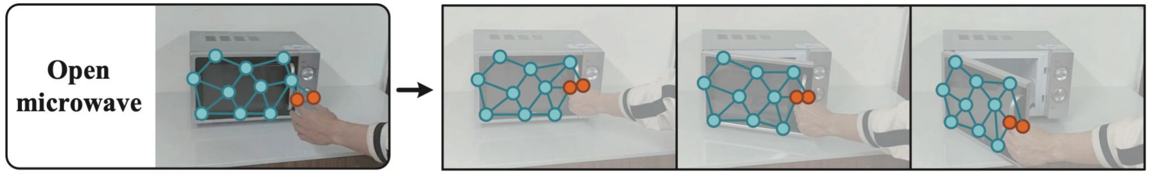
- ❏ [3]Graph structure of objects and visual actions, allowing robots to predict future states and actions (Chen et al.[Link])



[1]



[2]



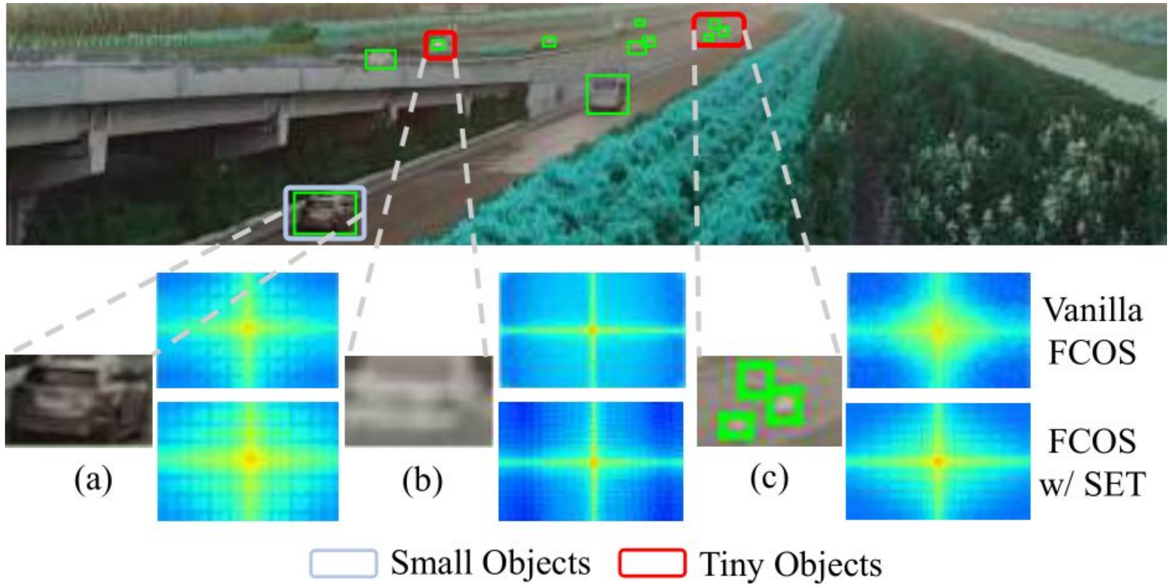
[3]

Featured by a Geoscientist working on microscopic images (1)

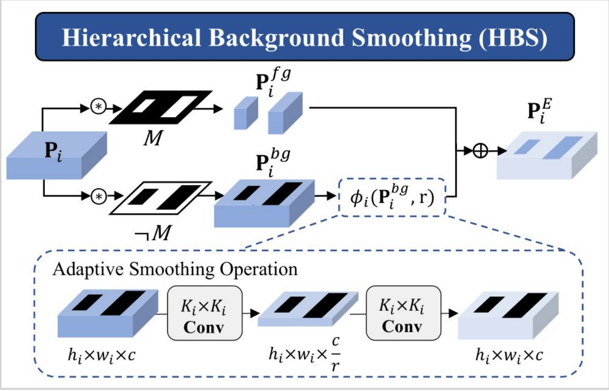
SET: Spectral Enhancement for Tiny Object Detection

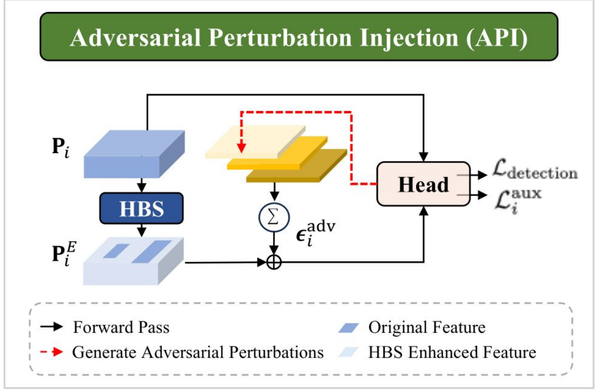
By Huixin Sun et al.
https://cvpr.thecvf.com/virtual/2025/poster/34394

Target: Detecting tiny objects



Proposal: HBS + API


Hierarchical Background Smoothing (HBS)
P_i
M
$\neg M$
P_i^{fg}
P_i^{bg}
$\phi_i(P_i^{bg}, r)$
P_i^E
Adaptive Smoothing Operation
$K_i \times K_i$ Conv
$K_i \times K_i$ Conv
$h_i \times w_i \times c$
$h_i \times w_i \times \frac{c}{r}$
$h_i \times w_i \times c$



Enhancing the distinctiveness of tiny features by suppressing high-frequency noise in the background

Adding adversarial perturbations to the feature map to enhance tiny object features during training.

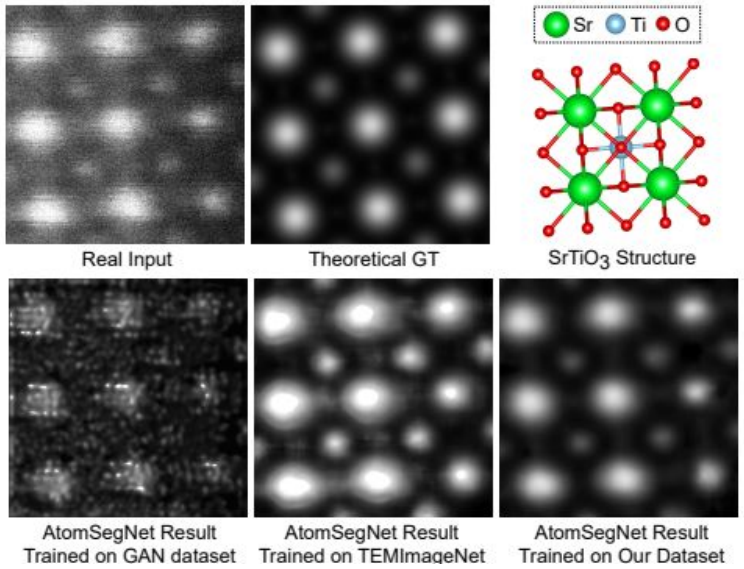
Achievements: +3.2% AP on AI-TOD (Tiny Object Detection) dataset

Featured by a Geoscientist working on microscopic images (2)

Noise Calibration and Spatial-Frequency Interactive Network for STEM Image Enhancement

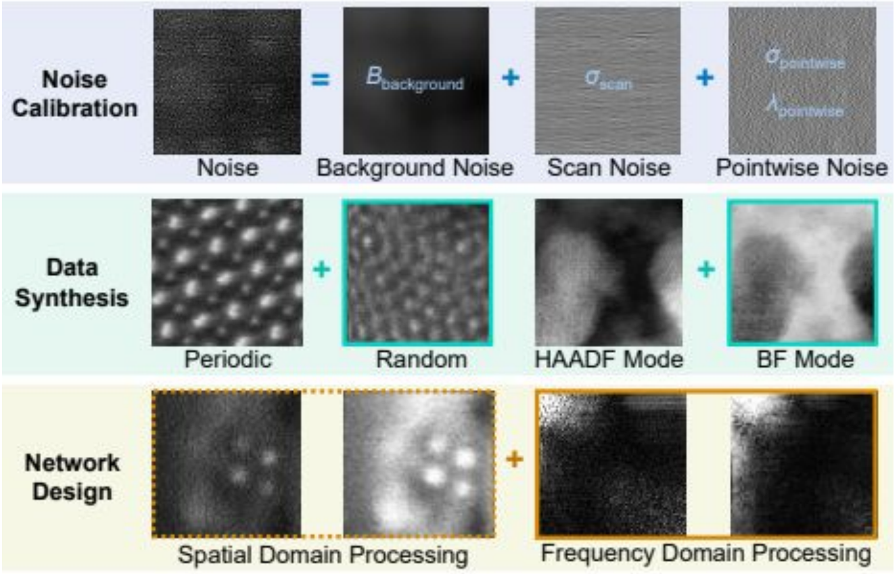
By Hesong Li et al. CVPR 2025 Open Access Repository

Target: Clearer observation by STEM*


Sr
Ti
O
Real Input
Theoretical GT
$SrTiO_3$ Structure
AtomSegNet Result Trained on GAN dataset
AtomSegNet Result Trained on TEMImageNet
AtomSegNet Result Trained on Our Dataset

*STEM: Scanning Transmission Electron **Microscope**

Proposal:



Create Dataset

Achievements: More realistic visualization of minute structures

CVPR25: Meta Insights into Trends and Tendencies (71/179)

New Trends in Image Restoration and Enhancement workshop and associated challenges (NTIRE2025)

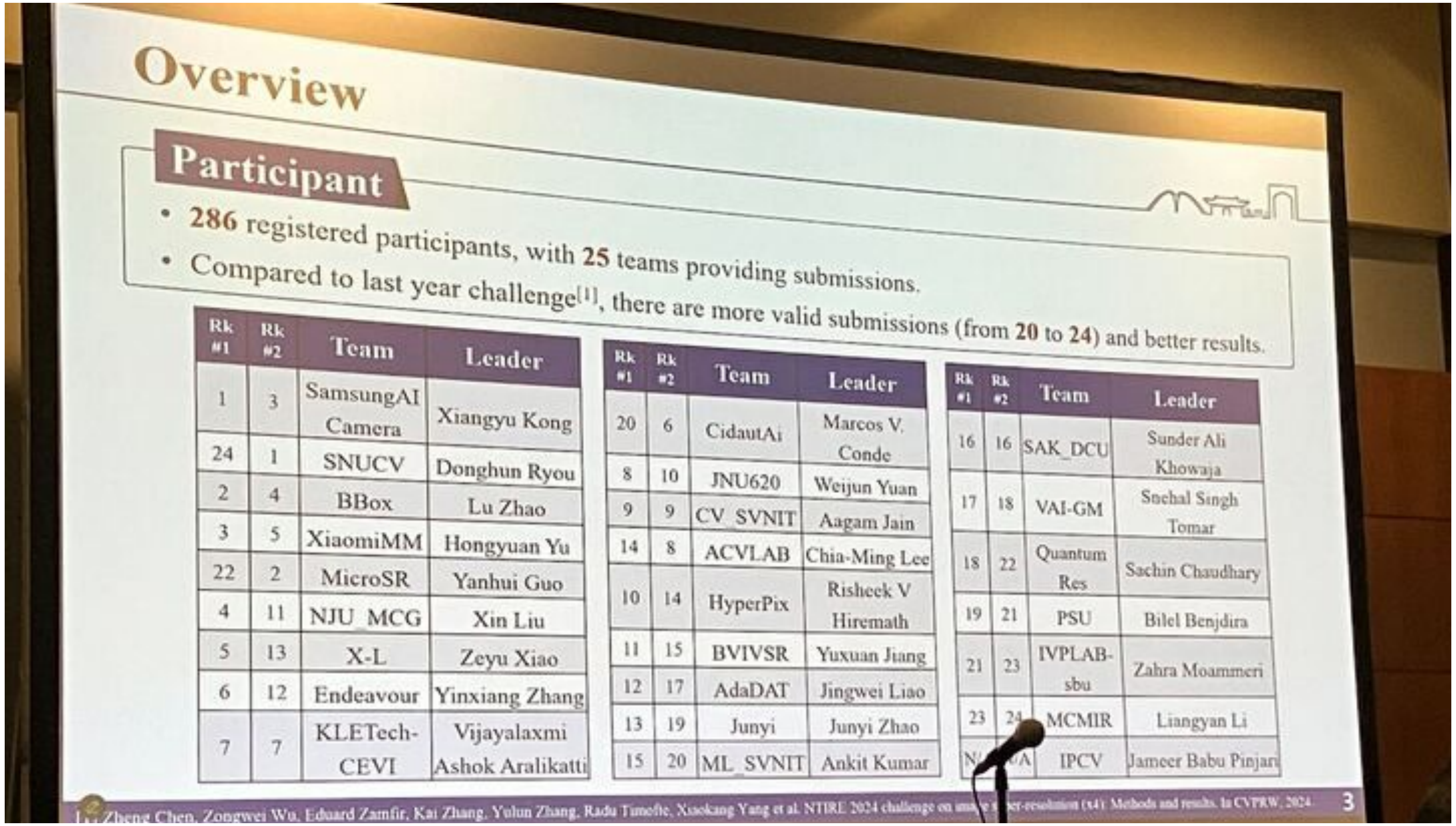
NTIRE 2025 Challenge on Image Super-Resolution (x4)

- ❑ Participants should recover a high-resolution image from a single low-resolution input that is 4× smaller.

- ❑ Dataset
  - ❑ The Challenge provides three official datasets: DIV2K, Flickr2K, and LSDIR.
  - ❑ The LR-HR image pairs are generated using bicubic downsampling.
- ❑ Track
  - ❑ The challenge evaluates performance on the DIV2K test set with two tracks:
  - ❑ Restoration Track: Ranked by PSNR and SSIM on the Y channel.
  - ❑ Perceptual Track: Ranked by a composite score from seven IQA metrics.

New Trends in Image Restoration and Enhancement workshop and associated challenges (NTIRE2025)

NTIRE 2025 Challenge on Image Super-Resolution (x4)



Overview

Participant

- **286** registered participants, with **25** teams providing submissions.
- Compared to last year challenge[1], there are more valid submissions (from **20** to **24**) and better results.

Rk #1	Rk #2	Team	Leader
1	3	SamsungAI Camera	Xiangyu Kong
24	1	SNUCV	Donghun Ryou
2	4	BBox	Lu Zhao
3	5	XiaomiMM	Hongyuan Yu
22	2	MicroSR	Yanhui Guo
4	11	NJU_MCG	Xin Liu
5	13	X-L	Zeyu Xiao
6	12	Endeavour	Yinxiang Zhang
7	7	KLETech-CEVI	Vijayalaxmi Ashok Aralikatti

Rk #1	Rk #2	Team	Leader
20	6	CidautAi	Marcos V. Conde
8	10	JNU620	Weijun Yuan
9	9	CV_SVNIT	Aagam Jain
14	8	ACVLAB	Chia-Ming Lee
10	14	HyperPix	Risheek V Hiremath
11	15	BVIVSR	Yuxuan Jiang
12	17	AdaDAT	Jingwei Liao
13	19	Junyi	Junyi Zhao
15	20	ML_SVNIT	Ankit Kumar

Rk #1	Rk #2	Team	Leader
16	16	SAK_DCU	Sunder Ali Khowaja
17	18	VAI-GM	Snehal Singh Tomar
18	22	Quantum Res	Sachin Chaudhary
19	21	PSU	Bilel Benjdira
21	23	IVPLAB-sbu	Zahra Moammeri
23	24	MCMIR	Liangyan Li
N/A	N/A	IPCV	Jameer Babu Pinjari

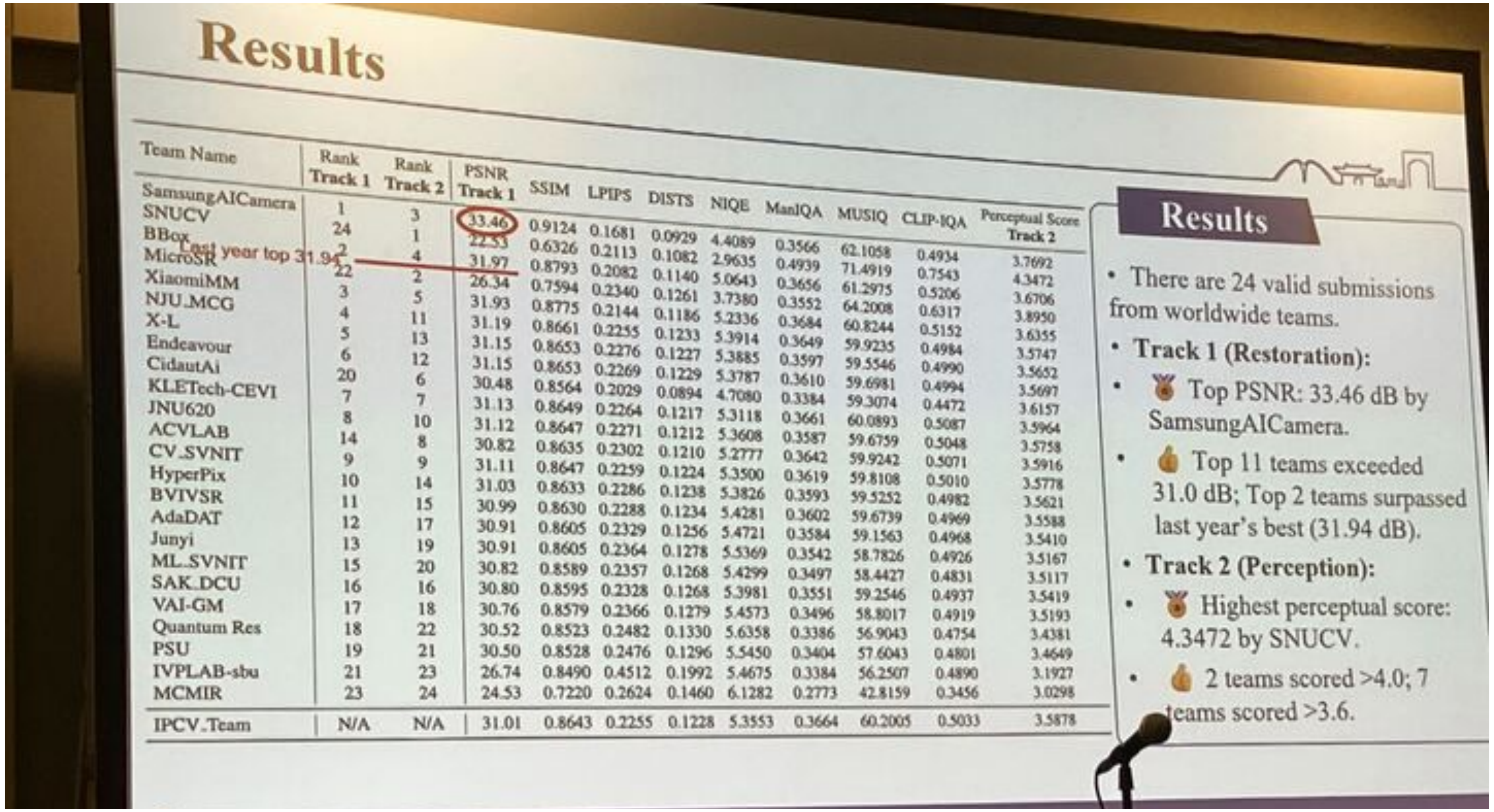
[1] Zheng Chen, Zongwei Wu, Eduard Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang et al. NTIRE 2024 challenge on image super-resolution (x4): Methods and results. In CVPRW, 2024

3

CVPR25: Meta Insights into Trends and Tendencies (73/179)

New Trends in Image Restoration and Enhancement workshop and associated challenges (NTIRE2025)

NTIRE 2025 Challenge on Image Super-Resolution (x4)



Results

Last year top 31.94

Team Name	Rank Track 1	Rank Track 2	PSNR Track 1	SSIM	LPIPS	DISTS	NIQE	ManIQA	MUSIQ	CLIP-IQA	Perceptual Score Track 2
SamsungAICamera	1	3	33.46	0.9124	0.1681	0.0929	4.4089	0.3566	62.1058	0.4934	3.7692
SNUCV	24	1	22.53	0.6326	0.2113	0.1082	2.9635	0.4939	71.4919	0.7543	4.3472
BBox	2	4	31.97	0.8793	0.2082	0.1140	5.0643	0.3656	61.2975	0.5206	3.6706
MicroSR	22	2	26.34	0.7594	0.2340	0.1261	3.7380	0.3552	64.2008	0.6317	3.8950
XiaomiMM	3	5	31.93	0.8775	0.2144	0.1186	5.2336	0.3684	60.8244	0.5152	3.6355
NJU_MCG	4	11	31.19	0.8661	0.2255	0.1233	5.3914	0.3649	59.9235	0.4984	3.5747
X-L	5	13	31.15	0.8653	0.2276	0.1227	5.3885	0.3597	59.5546	0.4990	3.5652
Endeavour	6	12	31.15	0.8653	0.2269	0.1229	5.3787	0.3610	59.6981	0.4994	3.5697
CidautAi	20	6	30.48	0.8564	0.2029	0.0894	4.7080	0.3384	59.3074	0.4472	3.6157
KLETech-CEVI	7	7	31.13	0.8649	0.2264	0.1217	5.3118	0.3661	60.0893	0.5087	3.5964
JNU620	8	10	31.12	0.8647	0.2271	0.1212	5.3608	0.3587	59.6759	0.5048	3.5758
ACVLAB	14	8	30.82	0.8635	0.2302	0.1210	5.2777	0.3642	59.9242	0.5071	3.5916
CV_SVNIT	9	9	31.11	0.8647	0.2259	0.1224	5.3500	0.3619	59.8108	0.5010	3.5778
HyperPix	10	14	31.03	0.8633	0.2286	0.1238	5.3826	0.3593	59.5252	0.4982	3.5621
BVIVSR	11	15	30.99	0.8630	0.2288	0.1234	5.4281	0.3602	59.6739	0.4969	3.5588
AdaDAT	12	17	30.91	0.8605	0.2329	0.1256	5.4721	0.3584	59.1563	0.4968	3.5410
Junyi	13	19	30.91	0.8605	0.2364	0.1278	5.5369	0.3542	58.7826	0.4926	3.5167
ML_SVNIT	15	20	30.82	0.8589	0.2357	0.1268	5.4299	0.3497	58.4427	0.4831	3.5117
SAK_DCU	16	16	30.80	0.8595	0.2328	0.1268	5.3981	0.3551	59.2546	0.4937	3.5419
VAI-GM	17	18	30.76	0.8579	0.2366	0.1279	5.4573	0.3496	58.8017	0.4919	3.5193
Quantum Res	18	22	30.52	0.8523	0.2482	0.1330	5.6358	0.3386	56.9043	0.4754	3.4381
PSU	19	21	30.50	0.8528	0.2476	0.1296	5.5450	0.3404	57.6043	0.4801	3.4649
IVPLAB-sbu	21	23	26.74	0.8490	0.4512	0.1992	5.4675	0.3384	56.2507	0.4890	3.1927
MCMIR	23	24	24.53	0.7220	0.2624	0.1460	6.1282	0.2773	42.8159	0.3456	3.0298
IPCV_Team	N/A	N/A	31.01	0.8643	0.2255	0.1228	5.3553	0.3664	60.2005	0.5033	3.5878

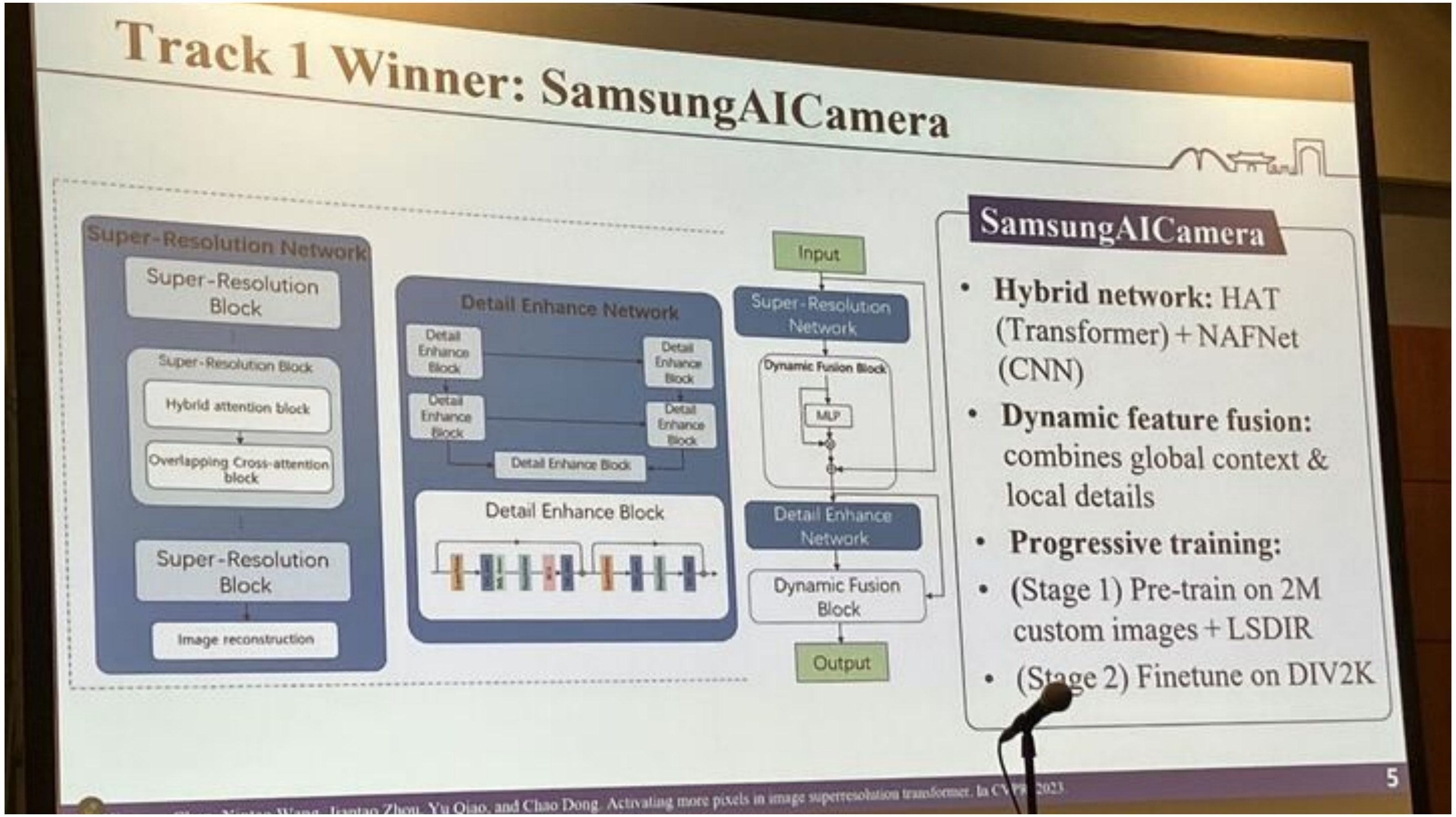
Results

- There are 24 valid submissions from worldwide teams.
- **Track 1 (Restoration):**
- 🏅 Top PSNR: 33.46 dB by SamsungAICamera.
- 👍 Top 11 teams exceeded 31.0 dB; Top 2 teams surpassed last year's best (31.94 dB).
- **Track 2 (Perception):**
- 🏅 Highest perceptual score: 4.3472 by SNUCV.
- 👍 2 teams scored >4.0; 7 teams scored >3.6.

CVPR25: Meta Insights into Trends and Tendencies (74/179)

New Trends in Image Restoration and Enhancement workshop and associated challenges (NTIRE2025)

NTIRE 2025 Challenge on Image Super-Resolution (x4)



Track 1 Winner: SamsungAICamera

SamsungAICamera

- **Hybrid network:** HAT (Transformer) + NAFNet (CNN)
- **Dynamic feature fusion:** combines global context & local details
- **Progressive training:**
- (Stage 1) Pre-train on 2M custom images + LSDIR
- (Stage 2) Finetune on DIV2K

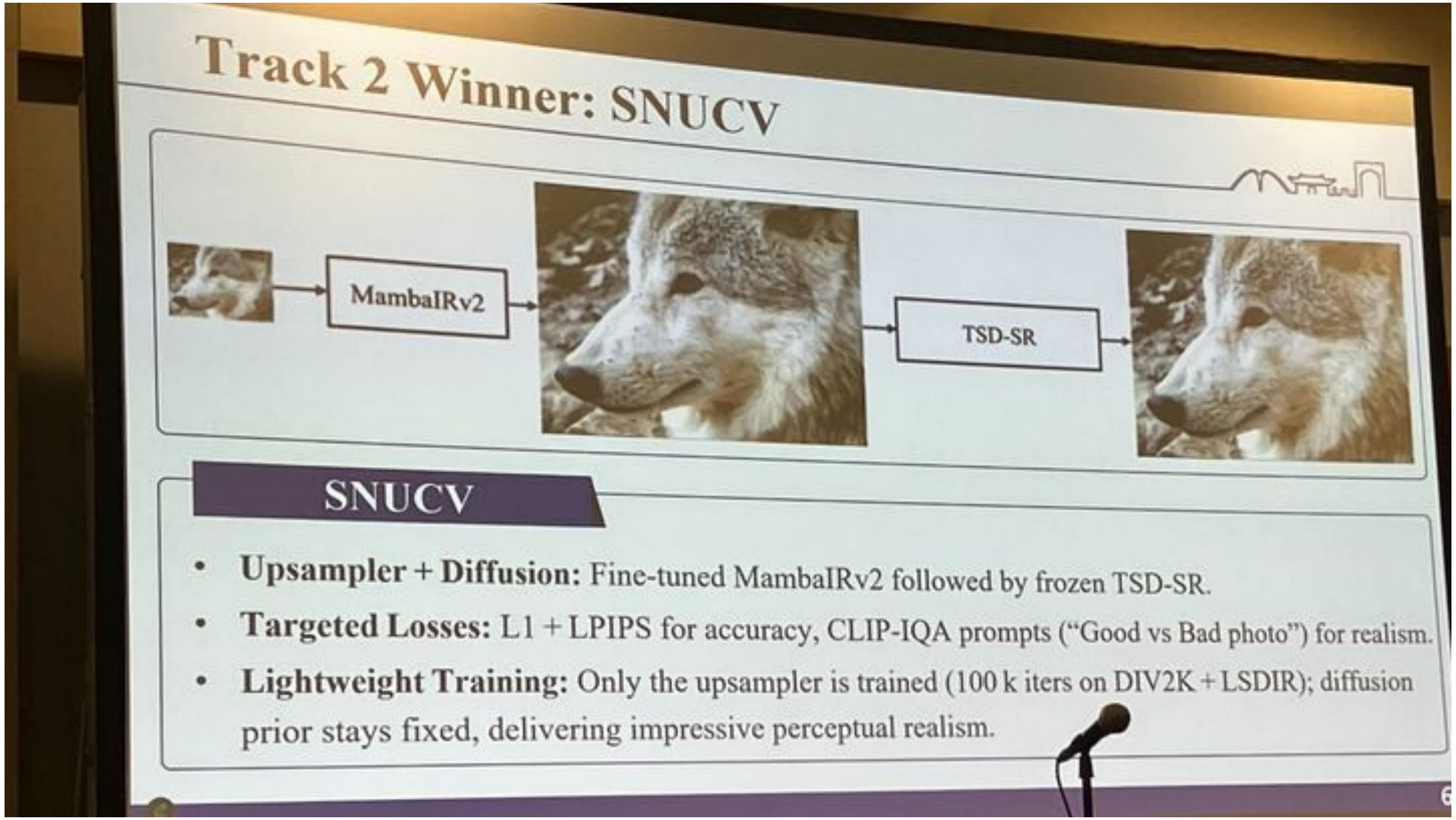
Chen, Xiangyu Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image superresolution transformer. In CVPR 2023

5

CVPR25: Meta Insights into Trends and Tendencies (75/179)

New Trends in Image Restoration and Enhancement workshop and associated challenges (NTIRE2025)

NTIRE 2025 Challenge on Image Super-Resolution (x4)



Track 2 Winner: SNUCV

SNUCV

- **Upsampler + Diffusion:** Fine-tuned MambaIRv2 followed by frozen TSD-SR.
- **Targeted Losses:** L1 + LPIPS for accuracy, CLIP-IQA prompts ("Good vs Bad photo") for realism.
- **Lightweight Training:** Only the upsampler is trained (100 k iters on DIV2K + LSDIR); diffusion prior stays fixed, delivering impressive perceptual realism.

CVPR25: Meta Insights into Trends and Tendencies (76/179)

New Trends in Image Restoration and Enhancement workshop and associated challenges (NTIRE2025)

NTIRE 2025 Challenge on Image Super-Resolution (x4)

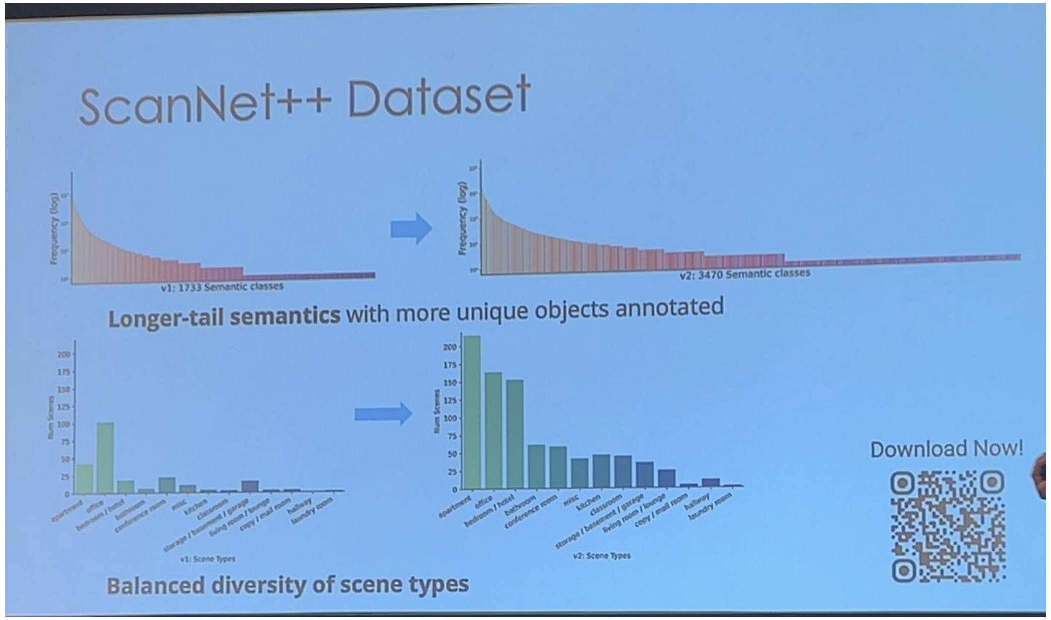
Insights

- ❑ Transformer-based architectures remain a main-stream approach
- ❑ Integrating Mamba architectures for improved global context modeling
- ❑ Advanced training strategies, including multi-stage pipelines, progressive patch training, and CLIP-based semantic filtering, have boosted model generalization and robustness.
- ❑ Generative priors, particularly pre-trained diffusion models combined with CLIP-based perceptual losses, has achieved superior perceptual quality with minimal training.

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

This workshop focuses on the following key areas:

- ❏ High-fidelity, large-vocabulary 3D semantic scene understanding
- ❏ Novel view synthesis in large-scale 3D scenes

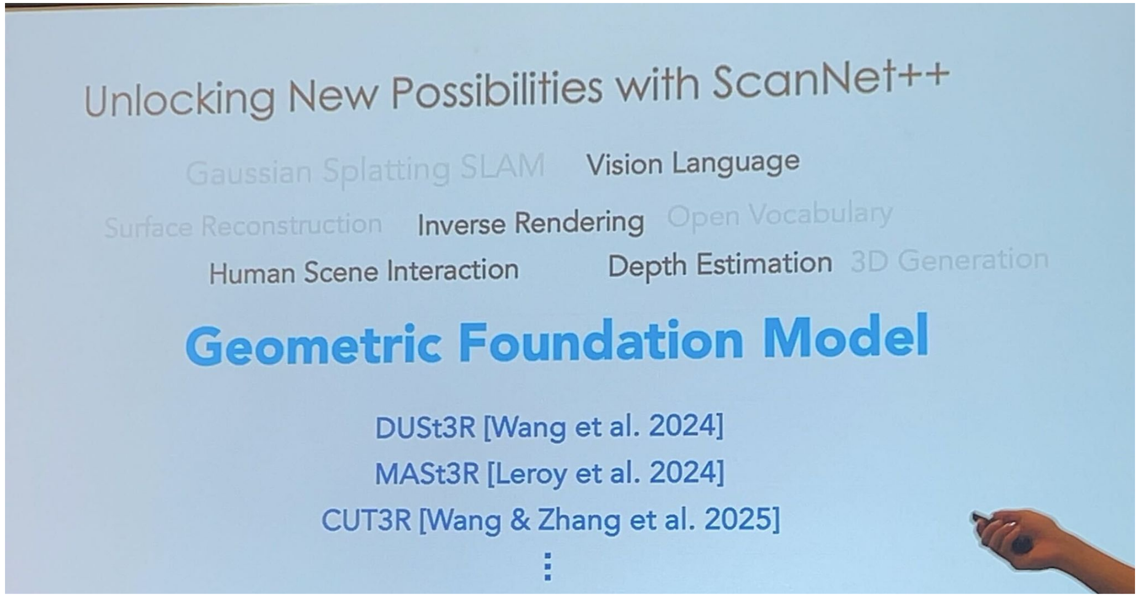


ScanNet++ Dataset

Longer-tail semantics with more unique objects annotated

Balanced diversity of scene types

Download Now!



Unlocking New Possibilities with ScanNet++

Gaussian Splatting SLAM Vision Language

Surface Reconstruction Inverse Rendering Open Vocabulary

Human Scene Interaction Depth Estimation 3D Generation

Geometric Foundation Model

DUSt3R [Wang et al. 2024]

MASt3R [Leroy et al. 2024]

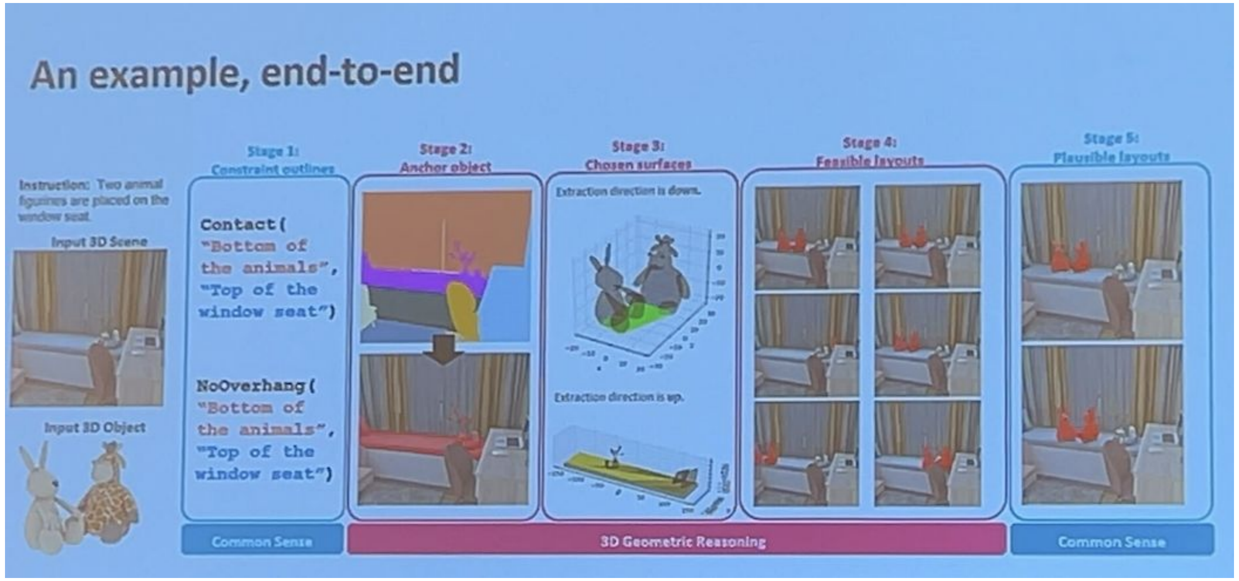
CUT3R [Wang & Zhang et al. 2025]

⋮

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

3D scene layout generation that combines LLMs with spatial reasoning

Enabling 3D scene generation technologies allows for applications in areas such as robotic simulation environments and game assets.



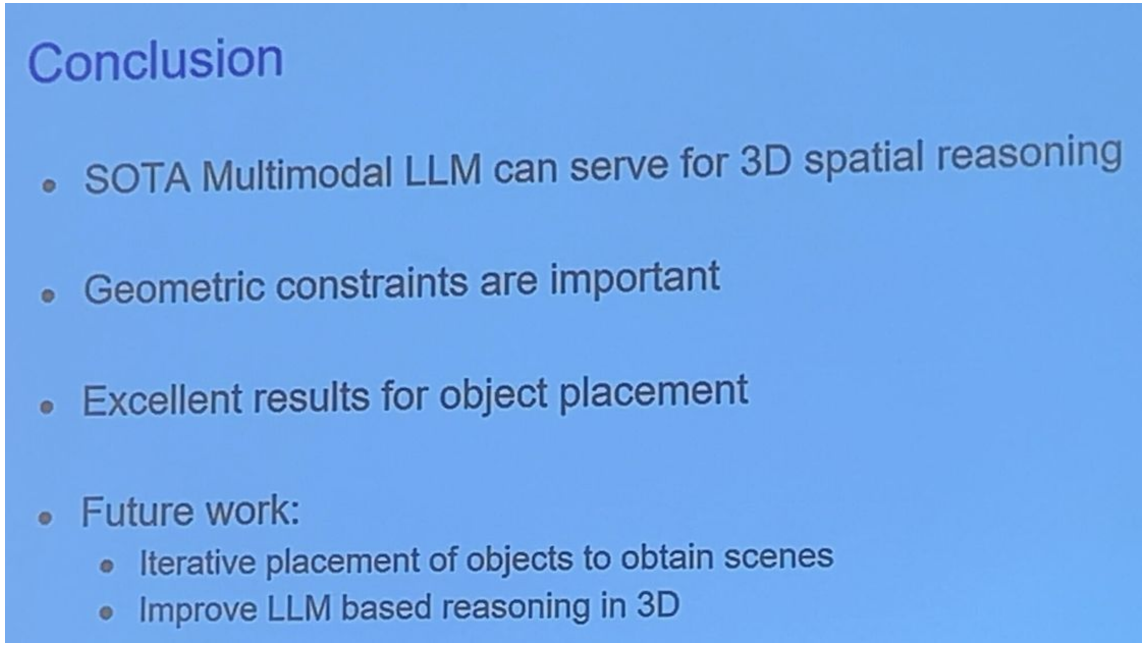
An example, end-to-end

Instruction: Two animal figurines are placed on the window seat.

```
Contact(
"Bottom of
the animals",
"Top of the
window seat")

NoOverhang(
"Bottom of
the animals",
"Top of the
window seat")
```

Input 3D Scene
Input 3D Object
Stage 1: Constraint outlines
Stage 2: Anchor object
Stage 3: Chosen surfaces
Extraction direction is down.
Extraction direction is up.
Stage 4: Feasible layouts
Stage 5: Plausible layouts
Common Sense
3D Geometric Reasoning
Common Sense



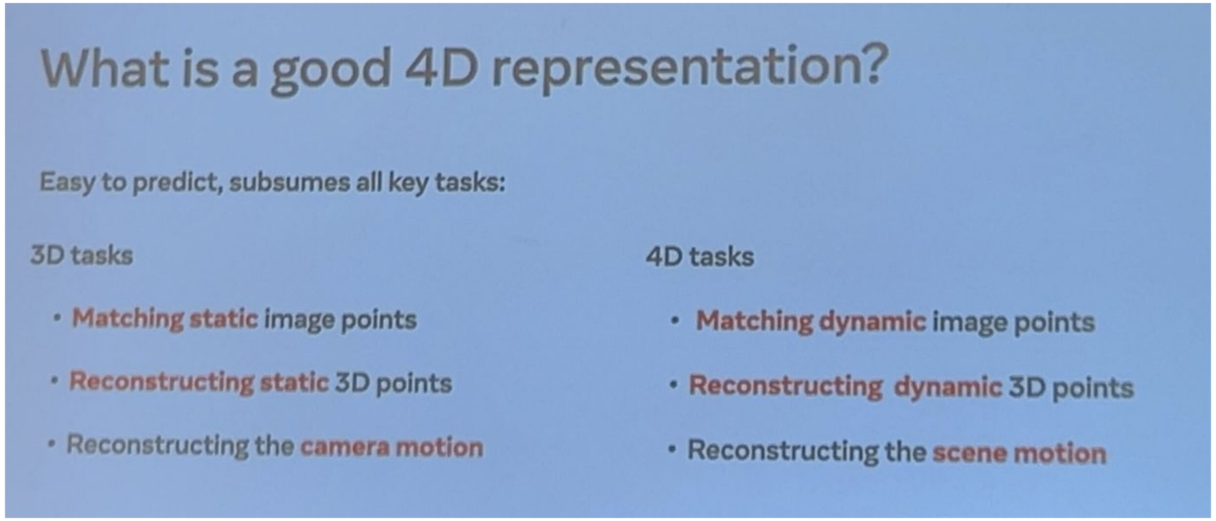
Conclusion

- SOTA Multimodal LLM can serve for 3D spatial reasoning
- Geometric constraints are important
- Excellent results for object placement
- Future work:
  - Iterative placement of objects to obtain scenes
  - Improve LLM based reasoning in 3D

CVPR25: Meta Insights into Trends and Tendencies (79/179)

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

The trend is expected to shift toward 4D, which combines 3D and video.



What is a good 4D representation?

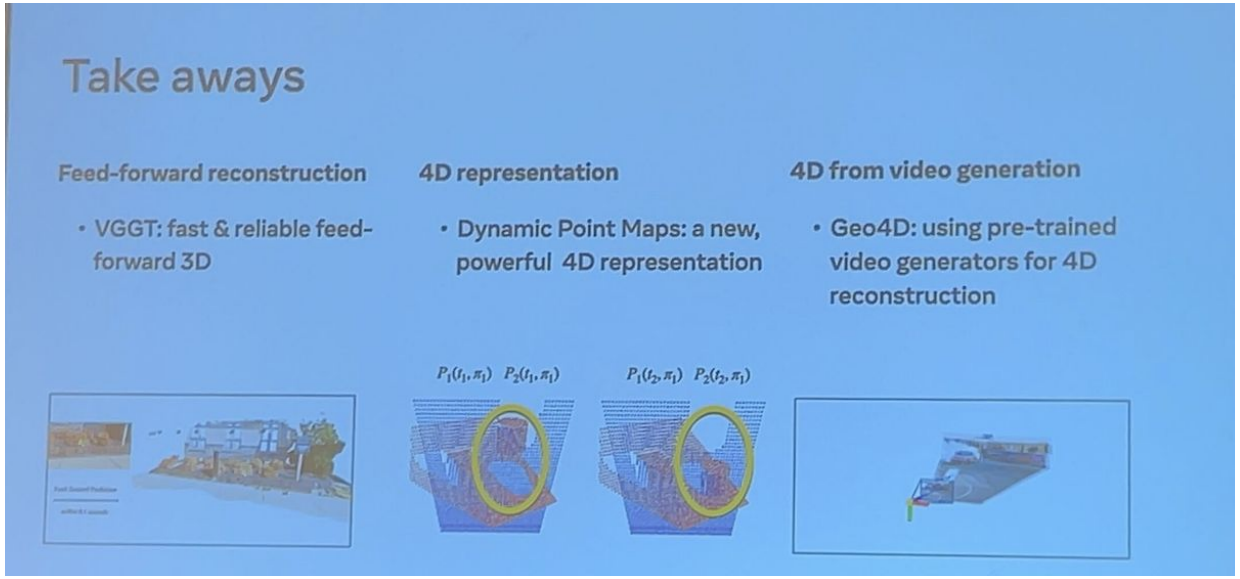
Easy to predict, subsumes all key tasks:

3D tasks

- Matching static image points
- Reconstructing static 3D points
- Reconstructing the camera motion

4D tasks

- Matching dynamic image points
- Reconstructing dynamic 3D points
- Reconstructing the scene motion



Take aways

Feed-forward reconstruction

- VGGT: fast & reliable feed-forward 3D

4D representation

- Dynamic Point Maps: a new, powerful 4D representation

4D from video generation

- Geo4D: using pre-trained video generators for 4D reconstruction

$P_1(t_1, \pi_1)$ $P_2(t_1, \pi_1)$
$P_1(t_2, \pi_1)$ $P_2(t_2, \pi_1)$

CVPR25: Meta Insights into Trends and Tendencies (80/179)

Workshop: Sight and Sound (paper session 1/2)

- CAV-MAE Sync: Improving Contrastive Audio-Visual Mask Autoencoders via Fine-Grained Alignment main conference paper
  - Audio-visual masked autoencoder for audio-visual representation; temporal correspondences matter; improve CAV-MAE by a finer-grained temporal resolution; better audio-visual correspondence compared with ImageBind.
- Diagnosing and Treating Audio-Video Fake Detection
  - Deepfakes in audio-video; propose a benchmark dataset, a simple baseline, and evaluation protocols.
- UWAV: Uncertainty-weighted Weakly-supervised Audio-Visual Video Parsing main conference paper
  - Task: audio-visual video parsing; Proposed model: training the pseudo label models with larger training set; training with uncertainty-weighted feature mixup; Evaluation dataset: LLP dataset which contains both audio events, visual events, and audio-visual events.
- STM2DVG: Synthetically Trained Music to Dance Video Generation leveraging Latent Diffusion Framework
  - Pose-conditioned LDM; Synthetic dataset generation pipeline; music-to-pose encoder.
- Seeing Speech and Sound: Distinguishing and Locating Audio Sources in Visual Scenes main conference paper
  - Task: simultaneous audio-visual grounding, including overlapping spoken language and non-speech sounds; Method:

Workshop: Sight and Sound (paper session 2/2)

- ❑ AVS-Net: Audio-Visual Scale Net for Self-supervised Monocular Metric Depth Estimation
  - ❑ Concept: using echoes for enhancing depth estimation; Especially using echos to obtain higher scale accuracy.
- ❑ SAVGBench: Benchmarking Spatially Aligned Audio-Video Generation
  - ❑ Concept: spatially aligned audio-video generation; Propose dataset (with ambisonics audio, 360-degree videos), with human speeches and instrument sounds; Method: MM Diffusion based.
- ❑ BGM2Pose: Active 3D Human Pose Estimation with Non-Stationary Sounds
  - ❑ Concept: using non-stationary sound to aid active 3D human pose estimation;
- ❑ Visual Sound Source Localization: Assessing Performance with Both Positive and Negative Audio
  - ❑ Visual sound source localization; Motivation: enhance robustness to offscreen and white noise; Extended datasets by adding negative audio; Introduced several evaluation metrics;
- ❑ VGGSounder: Audio-Visual Evaluations for Foundation Models
  - ❑ Introduce a comprehensive av benchmark, which is human labeled and each video has annotated classes in each modality.

CVPR25: Meta Insights into Trends and Tendencies (82/179)

Workshop: Sight and Sound (invited speaker)

- Learning to Infer Audio-Visual Attention in Social Communication (James Rehg)
  - Egocentric perception is important in social communication
    - Paper: Listen to Look into the Future: Audio-Visual Egocentric Gaze Anticipation (ECCV2024)
      - Task: gaze anticipation
      - Dataset (Ego4D Social, Aria)
      - Method: using video and audio, spatial and temporal fusion
  - Vision foundation model for gaze estimation
    - Paper: Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders (CVPR 2025)
      - Task: recognition human gazes in all people in videos
      - Method: utilizing foundation model + finetuning decoder part
  - Group conversation recognition
    - Paper: The Audio-Visual Conversational Graph: From an Egocentric-Exocentric Perspective (CVPR2024)
      - Motivation: ego-centric, exo-centric both important
      - Task: ego-exocentric conversational graph prediction
      - Model: multiple encoders (image, audio), plus cross, and self-attention
  - Social gestures, and other social behaviors, combined with computer vision tasks
    - Paper: SocialGesture: Delving into Multi-person Gesture Understanding (CVPR2025)
      - A benchmark to recognize and localize gestures
      - Current multimodal LLMs, foundation models still fall short in recognizing multiple person social interactions
    - Paper: Werewolf Among Us: A Multimodal Dataset for Modeling Persuasion Behaviors in Social Deduction Games
      - Persuasion strategy prediction task
    - Other datasets: speaking target identification,

CVPR25: Meta Insights into Trends and Tendencies (83/179)

Workshop: Sight and Sound (invited speaker)

- Sight and Sound with Large Language Models: Applications to Video Dubbing and Spatial Sound Understanding (David Harwath)
- Expanding the ability of multimodal LLMs in two applications
- Video dubbing
  - Paper: VOICECRAFT: Zero-Shot Speech Editing and Text-to-Speech in the Wild
    - Task: voice-cloning TTS
    - Issues: not enough talking-head datasets
    - Approach: firstly using language-based data to train text-to-speech backbone; the second, using video data to finetune
    - Other ideas: tokenizing speech with neural codecs
  - Paper: VoiceCraft-Dub: Automated Video Dubbing with Neural Codec Language Models
    - Task: dubbing, which should generate audio aligned to person, text, and video
    - Method: initialize from VoiceCraft model; add lip/face encoders and finetune on video-based datasets
- Spatial sound understanding
  - Paper: BAT:Learning to Reason about Spatial Sounds with Large Language Models
    - Task: spatial sound recognition; sound event detection; direction; distances, …
    - Dataset: SoundSpaces 2.0 simulator
    - Method: adding spatial sound encoding to LLM + pretraining (Spatial-AST model)

CVPR25: Meta Insights into Trends and Tendencies (84/179)

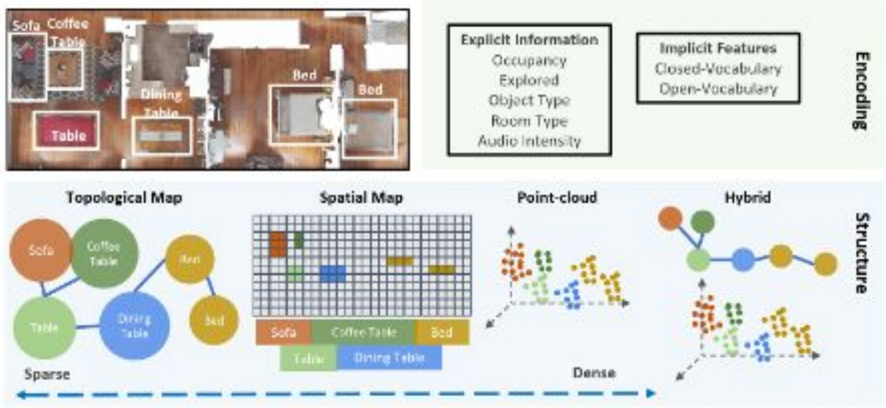
Workshop: Sight and Sound (invited speaker)

- ❑ Learning Sight and Sound through Generative Models (Ziyang Chen)
  - ❑ Paper: Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models
    - ❑ LDF for sound generation from video;
  - ❑ Paper: Video-Guided Foley Sound Generation with Multimodal Controls (CVPR 2025)
    - ❑ multimodal conditional; joint training using paired video+text+audio / text+audio data.
    - ❑ Very impressive generation result
  - ❑ Paper: Composing Images and Sounds on a Single Canvas (NeurIPS 2025)
    - ❑ Intersection between image representation and spectrogram
  - ❑ Idea: images that sound (image looking like the image caption, and sound like the sound caption)
  - ❑ Method: using the diffusion model

CVPR25: Meta Insights into Trends and Tendencies (85/179)

Workshop: 3D Vision Language Models (VLMs) for Robotic Manipulation: Opportunities and Challenges (invited speaker)

- ❏ Building vision-language maps for embodied AI (Angel Chang)
  - ❏ Semantic Mapping in Indoor Embodied AI (survey paper)



- ❏ Building maps for object-centric navigation
  - ❏ Multi-Object Navigation (MultiON) Task, LangNavBench, GOAT-Bench
  - ❏ Multi-Layered Feature Map
    - ❏ Two phase: explore-and-explore to get object information (build map) -> explore only (find goal in map)
  - ❏ Paper: Zero-shot Object-Centric Instruction Following: Integrating Foundation Models with Traditional Navigation
    - ❏ Using foundation models; Proposed Language-Inferred Factor Graph for Instruction Following, a graph structure, leading to more structured navigation and good accuracy.

CVPR25: Meta Insights into Trends and Tendencies (86/179)

Workshop: 3D Vision Language Models (VLMs) for Robotic Manipulation: Opportunities and Challenges (invited speaker)

- ❑ Hierarchical Action Models for Open-World 3D Policies (Dieter Fox)
    - ❑ Perceiver Actor
        - ❑ Anticipating the 3D poses / positions before executing actions
        - ❑ Paper: RVT: Robotic View Transformer for 3D Object Manipulation
            - ❑ Building explicit 3D representation by generating robotic observation by combining multiple viewpoint information
        - ❑ Paper: RVT2: RVT-2: Learning Precise Manipulation from Few Demonstrations
            - ❑ Predicting the next key-frame pose
    - ❑ Guiding 3D policies by utilizing vision language models
        - ❑ Input -> Vision-Language-Action Model -> Context for low-level policy -> control
        - ❑ Paper: HAMSTER: Hierarchical Action Models For Open-World Robot Manipulation (ICLR 2025)
            - ❑ VLM generates 2D trajectory sketch for policy learning
            - ❑ 3D policy models for 3D policy training
            - ❑ Shows promising generation ability in real-world evaluation of generalization

CVPR25: Meta Insights into Trends and Tendencies (87/179)

Workshop: 3D Vision Language Models (VLMs) for Robotic Manipulation: Opportunities and Challenges (invited speaker)

- Genesis: An Unified and Generative Physics Simulation for Robotics (Chuang Gan)
    - How far are we from an embodied generalist agent?
        - Generalist agents: multi-modal; multi-task; multi-environment.
        - At least, we need a world model
    - How to model physical world
        - Generative AI + Differentiable Physical engines
        - Forward Simulations Using Differentiable Physics
        - Joint geometry and physics estimation from multi-view videos
        - Forward simulation with actions
        - Result is better than RL-based learning
    - Applications
        - Soft-object animation; Thin-shell manipulation
    - Can we scale up the above approach?
        - Genesis: general-purpose physics simulator; Fully differentiable;
            - To scale up data generation, the authors proposed RoboGen, capable of task proposal (using language, using language as prompt), scene generation, training supervision generation.
            - How to solve the issue that human tools could not be suitable for robotics? -> Generative tools for robotics

CVPR25: Meta Insights into Trends and Tendencies (88/179)

Workshop: 3D Vision Language Models (VLMs) for Robotic Manipulation: Opportunities and Challenges (spotlight papers)

- ❑ The One RING: A Robotic Indoor Navigation Generalist
  - ❑ Universal navigation policy for all different robotics? – different robots observe and behave differently. The authors proposed One Ring, both data and model; Method: large-scale training + RL fine-tuning on random embodiments; Trained on sim but can be adapted to real-env.
- ❑ Manual2Skill: Learning to Read Manuals and Acquire Robotic Skills for Furniture Assembly Using Vision-Language Models
  - ❑ Manipulation skills learning by utilizing vision-language models; Use 2D images and language as input, generate hierarchical graph for assembly.
- ❑ Agentic Language-Grounded Adaptive Robotic Assembly
  - ❑ Issue: the ability to adapt to different model shape and structure change accordingly is essential in assembly applications; Resolve: using LLM to generate hierarchical guide for adapting assembly skills for different parts and models change.
- ❑ ZeroMimic: Distilling Robotic Manipulation Skills from Web Videos
  - ❑ Issue: datasets for robotic training; Resolve: learn general skill policies from online videos; Model: step1: knowing what human hands are doing; step2: training BC policies to operate human hands; step3: deploying human arm policies on robot arms. Foundation models facilitate distilling human videos into robot policies.

CVPR25: Meta Insights into Trends and Tendencies (89/179)

Workshop: Visual Generative Modeling: What’s After Diffusion?

- After Diffusion Models (Bill Freeman)
    - Problems of diffusion models
        - Slow, not easy control, not easy editing, …
    - Possible methods that might work
        - Conventional Graphics, hiring human artists, Simple explainable generative algorithms, Engineering, Better reasoning models, …
    - Conventional Graphics
        - Why not: slow, complicated, lots of manual work
    - Simple explainable generative algorithms
        - Paper: Infinite Images: Creating and Exploring a Large Photorealistic Virtual Space
            - Unwrap the motion into an infinite image
        - Paper: WonderWorld: Interactive 3D Scene Generation from a Single Image (CVPR 2025)
    - Computer vision by asking nicely
        - Foundation model that has enough knowledge about the physical world
        - Paper: Alchemist: Parametric Control of Material Properties with Diffusion Models (CVPR 2024)
        - Coaxing LLM to do what you ask, infuriating but sometimes powerful
    - Engineering better diffusion models
        - Paper: Improved Distribution Matching Distillation for Fast Image Synthesis (NeurIPS 2024)
        - Distillation of diffusion model

CVPR25: Meta Insights into Trends and Tendencies (90/179)

Workshop: Visual Generative Modeling: What’s After Diffusion?

- Controllable, Intuitive Generation of 4D Objects and Scenes (1/2) (Jiajun Wu)
  - Perception and Generation beyond 2D pixels
    - De-render to physical object intrinsics – allow better controllableness
  - Idea1: intrinsics as inductive bias for SSL
    - De-Render to intrinsics and then re-render to images
    - Physical intrinsics (lighting, 3D shape, camera, ..)
    - Paper: Seeing a Rose in Five Thousand Ways (CVPR 2023)
      - Single image, multiple object instances (like bouquet) –all roses have similar intrinsics
      - Make it possible to controllability
    - Paper: Learning the 3D Fauna of the Web (CVPR 2024)
      - From a single category to multiple categories, extending the above
  - Idea2: intrinsics as distilling targets for FMs (only vision)
    - Paper: Birth and Death of a Rose (CVPR 2025)
      - Generating temporal object intrinsics—temporally evolving sequences of object geometry, reflectance, and texture, such as a blooming rose—from pre-trained 2D foundation models.
    - Paper: PhycsDream: From Object Motion to Action (ECCV 2024)
      - Action-conditioned dynamics prediction of 3D objects

CVPR25: Meta Insights into Trends and Tendencies (91/179)

Workshop: Visual Generative Modeling: What’s After Diffusion?

- ❏ Controllable, Intuitive Generation of 4D Objects and Scenes (2/2) ([Jiajun Wu](https://jiajunwu.com/))
  - ❏ Idea2A: intrinsics as distilling targets for FMs (vision and language)
    - ❏ From vision-language models observations to reconstruction
    - ❏ Paper: WonderJourney: Going from Anywhere to Everywhere (CVPR 2024)
      - ❏ Model illustration: using LLMs to generate long sequence of scene descriptions, a text-driven point cloud generation pipeline to synthesize 3D scenes, and a VLM to verify generated examples
    - ❏ Paper: WonderWorld: Real-Time, Interactive 3D Scene Generation (CVPR2025)
      - ❏ User can specify where and what to generate in an interactive manner
    - ❏ Paper: WonderPlay: Dynamic 3D Scene Generation from a Single Image and Actions
      - ❏ Allow editing and manipulation in 4D manner
    - ❏ Paper: The Scene Language: Representing Scenes with Programs, Words, and Embeddings (CVPR 2025)
      - ❏ Chessboard generation as an example, allowing generating scenes expressively, allowing high-level interactive editing
      - ❏ Structure: program function dependency
      - ❏ Semantics: natural language words
      - ❏ Visual identity: Neural embeddings
      - ❏ Allowing Text-to-3D/4D Generation and Editing
- ❏ What’s the next:
  - ❏ Questions: what needs to be modeled? How do they help visual generation? What’s the roles of FMs
  - ❏ Rendering/conditioning based on causal, physical, and universal object intrinsics
  - ❏ Effective use of the in-the-wild data, enhanced interpretability

CVPR25: Meta Insights into Trends and Tendencies (92/179)

Workshop: Visual Generative Modeling: What’s After Diffusion?

- Language as a Visual Format (Phillip Isola)
  - How similar is the way a vision model sees the world to the way a language model sees the world
  - Paper: The Platonic Representation Hypothesis (ICML, 2024)
    - Characterizing representations using kernels
    - Comparing sim (DINO, Llama)
    - The bigger the DINO model is, the similarity is higher
    - More detailed captions align better with vision representations
  - Paper: Cycle Consistency as Reward: Learning Image-Text Alignment without Human Preferences
    - Image and language cycle consistency to improve image-text alignment
    - Light-weight, fast, differentiable
    - Dataset: Cycle consistency preference collection (CyclePrefDB)
      - Useful for evaluating and improving image-text alignment
      - Direct Preference Optimization using CyclePrefDB-T2I
      - Usage: pip install cyclereward
- What’s after diffusion
  - Language as an alternative to pixels but requires visually descriptive language

CVPR25: Meta Insights into Trends and Tendencies (93/179)

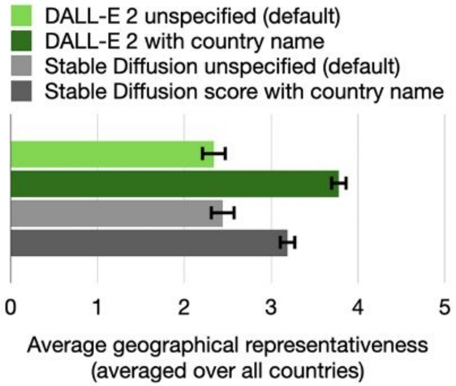
Workshop: VLMs-4-All 2025

- ❑ Re-scaling cultural knowledge in a UK heritage and museum context (Maya Indira Ganesh)
  - ❑ AI for cultural heritage hub (ArCH)
    - ❑ Important question: how to use current AI to let people be more engaged in cultural heritage recognition
  - ❑ GLAM sector and cultural heritage collection challenges
    - ❑ Challenges: Accessibility, reconstructing fragmentary and dispersed objects, requiring expert knowledge, legacy technical systems across multiple providers, legacies of colonialism and contested provenance of artifacts and cultural knowledge
    - ❑ Some other issues: negative perception of AI, uncertainty, errors, digitization, undigitized backlog, legacy infrastructure, provisional semantics
  - ❑ Street observatories of everyday AI
    - ❑ Street recognition and embodied tasks in streets still have lots of unresolved issues

CVPR25: Meta Insights into Trends and Tendencies (94/179)

Workshop: VLMs-4-All 2025

- Richer Outputs for Richer Countries: Geographical Disparities in Language and Image Generation (1/2) (Danish Pruthi)
  - Mononyms (single word names)
    - Common in some modern societies (like India, Myanmar)
    - Digital forms rarely account for single names
    - Issues: might not be recognized by recent LLMs
    - Other AI system issues: biases related to gender, race, geographical
  - Generating Representative
    - Generating an image / images representative of visual concepts, which is important in various applications and real-world scenarios
  - Issue: Are current models generating geographical representative?
    - Paper: Inspecting the Geographical Representativeness of Images from Text-to-Image Models (ICCV 2023)
      - Current models show low-level of geographical representativeness

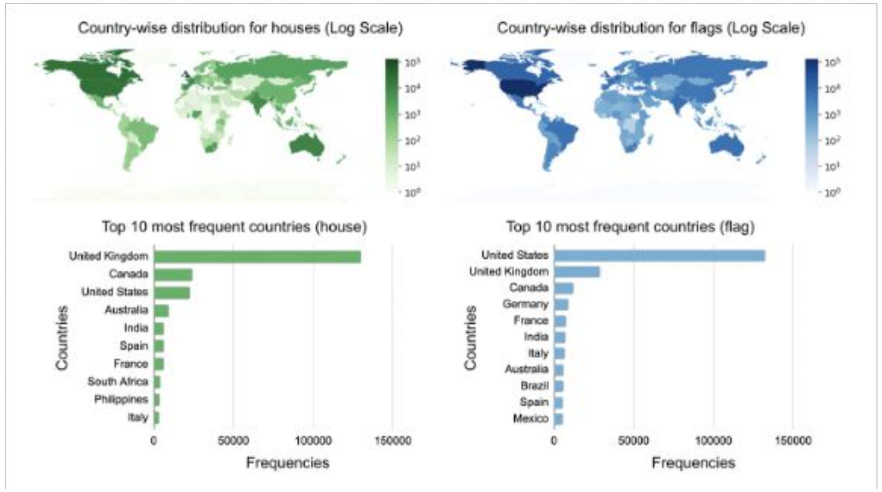


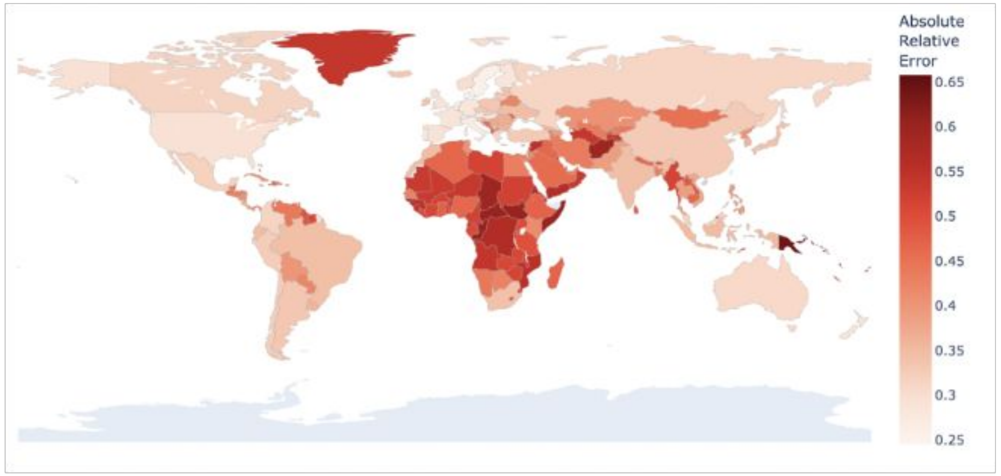
    - Paper: Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale (FAccT 2023)
      - A broad range of ordinary prompts produce stereotypes, including prompts simply mentioning traits, descriptors, occupations, or objects
      - Stereotypes are present regardless of whether prompts explicitly mention identity and demographic language or avoid such language.

CVPR25: Meta Insights into Trends and Tendencies (95/179)

Workshop: VLMs-4-All 2025

- Richer Outputs for Richer Countries: Geographical Disparities in Language and Image Generation (2/2) (Danish Pruthi)
  - Issue: Are current models generating geographical representative? - No
    - Paper: Where Do Images Come From? Analyzing Captions to Geographically Profile Datasets (left image)
      - Task: Geographically profiling datasets, given an image-caption pair, map them to a location
      - Geo-profiling LAION2B-EN proves heavy geographical biases, data distribution greatly biased where Wealthy countries have more data



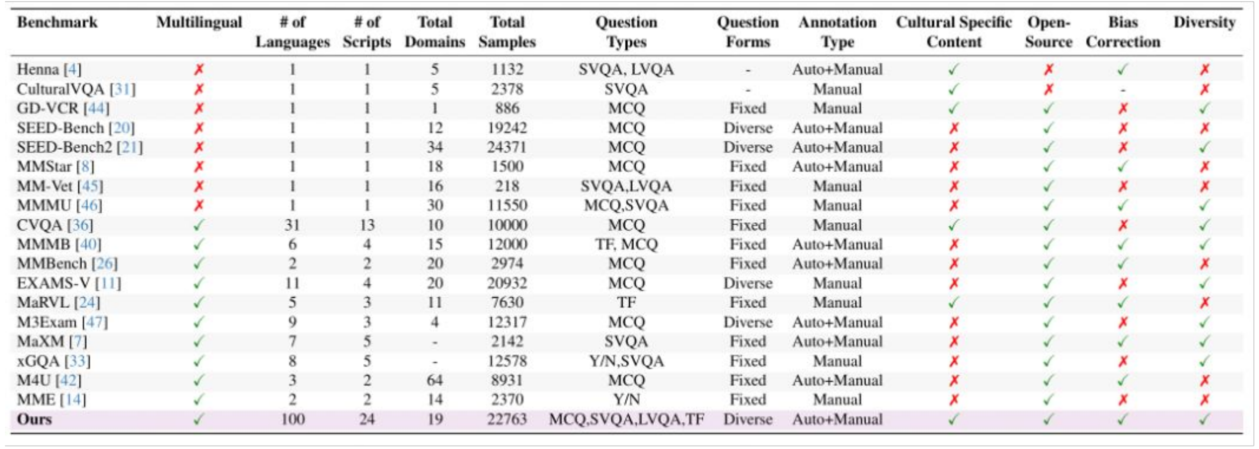


-
  -
    - Language oppression: erasure of minority languages
      - Paper: Geographical Erasure in Language generation (EMNLP 2023)
        - Many countries suffer from geographical erasure
      - Paper: WorldBench: Quantifying Geographic Disparities in LLM Factual Recall (right image)
        - Assess the ability of large language models (LLMs) to recall factual information about specific countries
    - Paper: Richer Output for Richer Countries: Uncovering Geographical Disparities in Generated Stories and Travel Recommendations
      - Reveal the issues of geographical erasure, regarding to applications such as travel recommendations and story generation

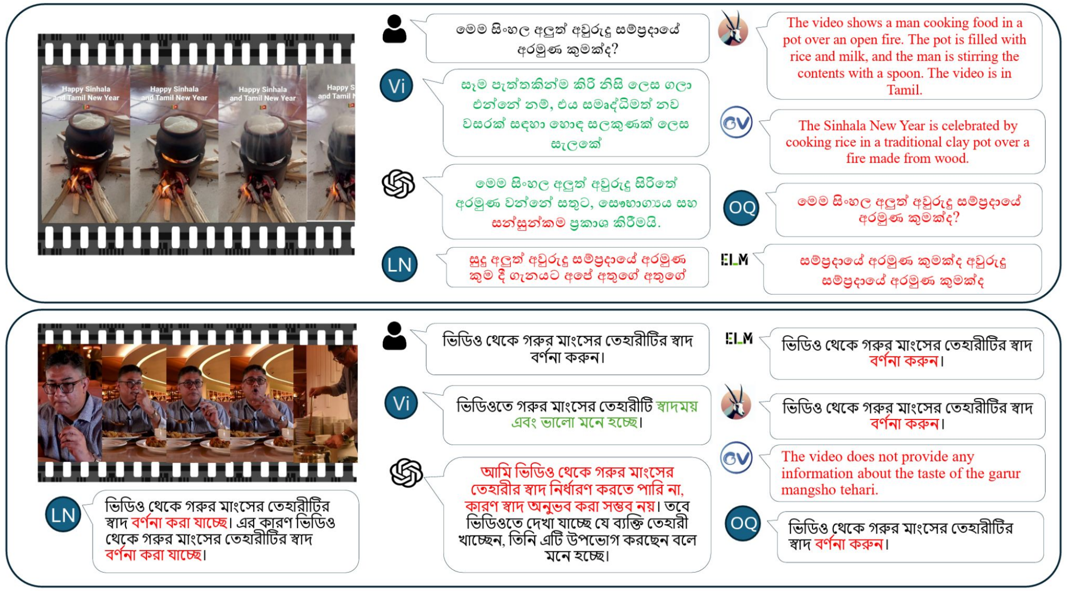
CVPR25: Meta Insights into Trends and Tendencies (96/179)

Workshop: VLMs-4-All 2025

- ❏ Building Culturally Aware Multilingual LMM Benchmarks (1/2) (Fahad Shahbaz Khan)
  - ❏ Efforts 1: multilingual multimodal understanding benchmarks
    - ❏ Paper: All Languages Matter: Evaluating LMMs on Culturally Diverse 100 Languages (CVPR2025) (left image)
      - ❏ Culturally Aware Multilingual Image LMM Benchmark
      - ❏ 100 languages, including many low-resource ones, native-language experts (800 hours annotations)
      - ❏ 13 cultural aspects covered in the benchmark, such as heritage, customs, architecture, literature, music, and sports.
      - ❏ Generation pipeline with AI generated QAs with human validation
    - ❏ Paper: A culturally-diverse multilingual multimodal video benchmark and model (right image)
      - ❏ Issue: Most video datasets are concentrating in English, models struggle with low-resource languages
      - ❏ Propose: Culturally Aware Multilingual Video LMM Benchmark



Benchmark	Multilingual	# of Languages	# of Scripts	Total Domains	Total Samples	Question Types	Question Forms	Annotation Type	Cultural Specific Content	Open-Source	Bias Correction	Diversity
Henna [4]	✗	1	1	5	1132	SVQA, LVQA	-	Auto+Manual	✓	✗	✓	✗
CulturalVQA [31]	✗	1	1	5	2378	SVQA	-	Manual	✓	✗	-	✗
GD-VCR [44]	✗	1	1	1	886	MCQ	Fixed	Manual	✓	✓	✗	✓
SEED-Bench [20]	✗	1	1	12	19242	MCQ	Diverse	Auto+Manual	✗	✓	✗	✗
SEED-Bench2 [21]	✗	1	1	34	24371	MCQ	Diverse	Auto+Manual	✗	✓	✗	✓
MMStar [8]	✗	1	1	18	1500	MCQ	Fixed	Auto+Manual	✗	✓	✓	✗
MM-Vet [45]	✗	1	1	16	218	SVQA,LVQA	Fixed	Manual	✗	✓	✗	✗
MMMU [46]	✗	1	1	30	11550	MCQ,SVQA	Fixed	Manual	✗	✓	✓	✓
CVQA [36]	✓	31	13	10	10000	MCQ	Fixed	Manual	✓	✓	✗	✓
MMMB [40]	✓	6	4	15	12000	TF, MCQ	Fixed	Auto+Manual	✗	✓	✓	✓
MMBench [26]	✓	2	2	20	2974	MCQ	Fixed	Auto+Manual	✗	✓	✓	✗
EXAMS-V [11]	✓	11	4	20	20932	MCQ	Diverse	Manual	✗	✓	✗	✓
MaRVL [24]	✓	5	3	11	7630	TF	Fixed	Manual	✓	✓	✓	✗
M3Exam [47]	✓	9	3	4	12317	MCQ	Diverse	Auto+Manual	✗	✓	✗	✓
MaXM [7]	✓	7	5	-	2142	SVQA	Fixed	Auto+Manual	✗	✓	✓	✓
xGQA [33]	✓	8	5	-	12578	Y/N,SVQA	Fixed	Manual	✗	✓	✗	✓
M4U [42]	✓	3	2	64	8931	MCQ	Fixed	Auto+Manual	✗	✓	✓	✗
MME [14]	✓	2	2	14	2370	Y/N	Fixed	Manual	✗	✓	✗	✗
Ours	✓	100	24	19	22763	MCQ,SVQA,LVQA,TF	Diverse	Auto+Manual	✓	✓	✓	✓



මෙම සිංහල අලුත් අවුරුදු සම්ප්‍රදායේ අරමුණ කුමක්ද?

සෑම පැත්තකින්ම කිරි නිසි ලෙස ගලා එන්නේ නම්, එය සමෘද්ධිමත් නව වසරක් සඳහා හොඳ සලකුණක් ලෙස සැලකේ

මෙම සිංහල අලුත් අවුරුදු සිරිතේ අරමුණ වන්නේ සතුට, සෞභාග්‍යය සහ සන්සුන්කම ප්‍රකාශ කිරීමයි.

සුදු අලුත් අවුරුදු සම්ප්‍රදායේ අරමුණ කුම දී ගැනයට අපේ අතුගේ අතුගේ

The video shows a man cooking food in a pot over an open fire. The pot is filled with rice and milk, and the man is stirring the contents with a spoon. The video is in Tamil.

The Sinhala New Year is celebrated by cooking rice in a traditional clay pot over a fire made from wood.

මෙම සිංහල අලුත් අවුරුදු සම්ප්‍රදායේ අරමුණ කුමක්ද?

සම්ප්‍රදායේ අරමුණ කුමක්ද අවුරුදු සම්ප්‍රදායේ අරමුණ කුමක්ද

ভিডিও থেকে গরুর মাংসের তেহারীটির স্বাদ বর্ণনা করুন।

ভিডিওতে গরুর মাংসের তেহারীটি স্বাদময় এবং ভালো মনে হচ্ছে।

আমি ভিডিও থেকে গরুর মাংসের তেহারীর স্বাদ নির্ধারণ করতে পারি না, কারণ স্বাদ অনুভব করা সম্ভব নয়। তবে ভিডিওতে দেখা যাচ্ছে যে ব্যক্তি তেহারী খাচ্ছেন, তিনি এটি উপভোগ করছেন বলে মনে হচ্ছে।

ভিডিও থেকে গরুর মাংসের তেহারীটির স্বাদ বর্ণনা করা যাচ্ছে। এর কারণ ভিডিও থেকে গরুর মাংসের তেহারীটির স্বাদ বর্ণনা করা যাচ্ছে।

ভিডিও থেকে গরুর মাংসের তেহারীটির স্বাদ বর্ণনা করুন।

ভিডিও থেকে গরুর মাংসের তেহারীটির স্বাদ বর্ণনা করুন।

The video does not provide any information about the taste of the garur mangsho tehari.

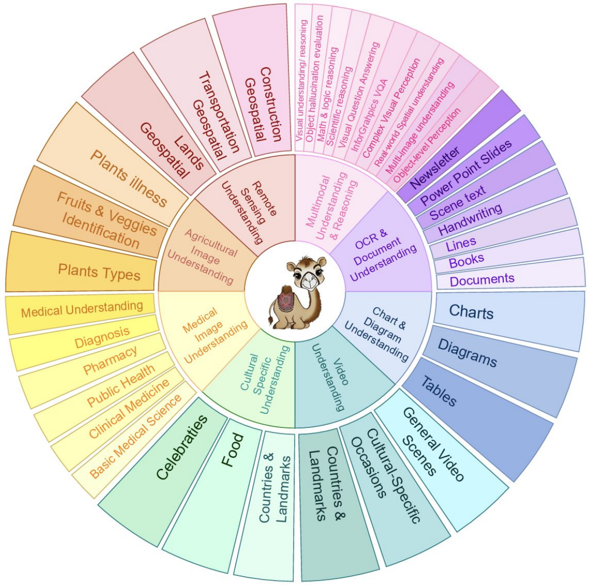
ভিডিও থেকে গরুর মাংসের তেহারীটির স্বাদ বর্ণনা করুন।

LIMIT.LAB
https://limitlab.xyz/

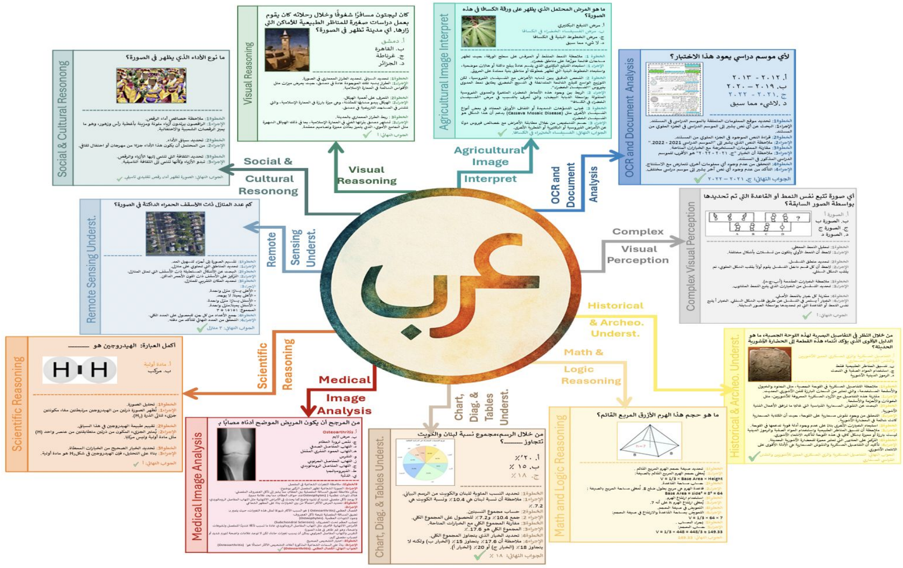
CVPR25: Meta Insights into Trends and Tendencies (97/179)

Workshop: VLMs-4-All 2025

- Building Culturally Aware Multilingual LMM Benchmarks (2/2) (Fahad Shahbaz Khan)
  - Efforts 2: Arabic LMM benchmarks
    - Paper: CAMEL-Bench: A Comprehensive Arabic LMM Benchmark (left image)
      - CAMEL-Bench, comprises eight diverse domains and 38 sub-domains including, multi-image understanding, complex visual perception, handwritten document understanding, video understanding, medical imaging, plant diseases, and remote sensing-based land use understanding
      - Arabic language has over 400M users, however, even the closed-source GPT-4o achieving an overall score of 62%
    - Paper: ARB: A Comprehensive Arabic Multimodal Reasoning Benchmark (right image)
      - The first benchmark designed to evaluate step-by-step reasoning in Arabic across both textual and visual modalities.
      - ARB spans 11 diverse domains, including visual reasoning, document understanding, OCR, scientific analysis, and cultural interpretation.



Multimodal Understanding & Reasoning
Visual understanding/ reasoning
Object hallucination evaluation
Math & logic reasoning
Scientific reasoning
Visual Question Answering
InforGrahpics VQA
Complex Visual Perception
Real-world Spatial understanding
Multi-image understanding
Object-level Perception
OCR & Document Understanding
Newsletter
Power Point Slides
Scene text
Handwriting
Lines
Books
Documents
Chart & Diagram Understanding
Charts
Diagrams
Tables
Video Understanding
General Video Scenes
Cultural-Specific Occasions
Countries & Landmarks
Cultural Specific Understanding
Countries & Landmarks
Food
Celebrities
Medical Image Understanding
Basic Medical Science
Clinical Medicine
Public Health
Pharmacy
Diagnosis
Medical Understanding
Agricultural Image Understanding
Plants Types
Fruits & Veggies Identification
Plants illness
Remote Sensing Understanding
Geospatial
Lands
Transportation Geospatial
Construction Geospatial



عرب
Social & Cultural Resonong
Visual Reasoning
Agricultural Image Interpret
OCR and Document Analysis
Complex Visual Perception
Historical & Archeo. Underst.
Math & Logic Reasoning
Chart, Diag. & Tables Underst.
Medical Image Analysis
Scientific Reasoning
Remote Sensing Underst.
H I H

CVPR25: Meta Insights into Trends and Tendencies (98/179)

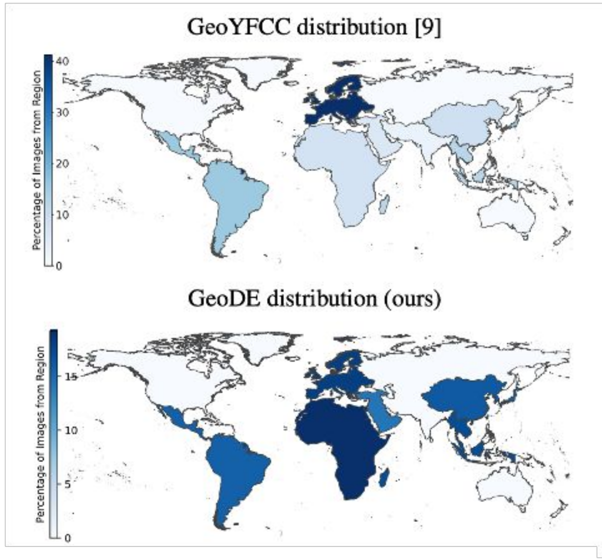
Workshop: VLMs-4-All 2025

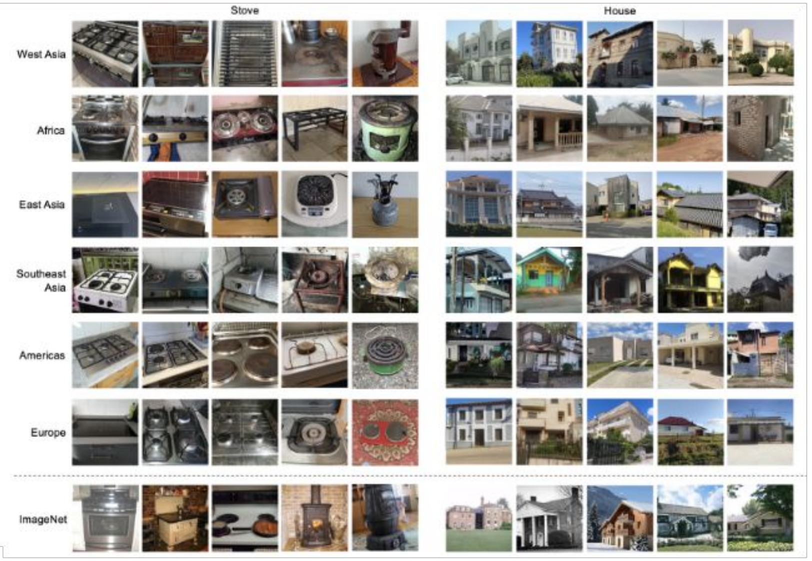
- ❑ Challenges with Geo-Cultural Understanding and Generation (Roopal Garg)
  - ❑ GeoCultural Multimodal Content Understanding and Generation
  - ❑ Issue: current models fall short on a truly global user base
    - ❑ Impact: affects accessibility, fairness, utility, and trust of billions of users worldwide
  - ❑ Data quality gap – comprehensiveness vs. cultural specificity
    - ❑ Not having enough high-quality scaled datasets
    - ❑ Data generation / collection: collect data that highlights cultural distinctiveness, e.g., food, festivals.
    - ❑ Generate Locale Aware human annotations, instead of direct translations, using locale reference
    - ❑ Search from the internet to get correct labels, and use local person annotations or validations
  - ❑ Generation model for enhanced geo-cultural understanding
    - ❑ Issue: image generation could be culture-related, text-to-image generation from english translations contain multiple issues and is less accurate for concepts with different cultural backgrounds
    - ❑ Model structure: prompt expansion (get local language and modification based on the context of the target language) before generation

CVPR25: Meta Insights into Trends and Tendencies (99/179)

Workshop: VLMs-4-All 2025

- Building geo-diverse model (Olga Russakovsky)
  - Current issue: Lack of cultural diversity in datasets
  - GeoDE: Geographically Diverse Evaluation Dataset (NeurIPS2023) (left image)
    - Partnered with Appen to solicit photographs from people around the world
    - 61940 images from 6 different regions
    - Generated using crowdsourcing where each image has consent and no recognizable people
    - Findings:
      - Crowd-sourced images look different in different regions (right image)
      - Features representations between crowd-sourced and web-scraped images are different
      - GeoDE can find gaps in models, for example, CLIP shows stereotypes and errors in recognizing non-western objects, building
      - Training on GeoDE improves performance

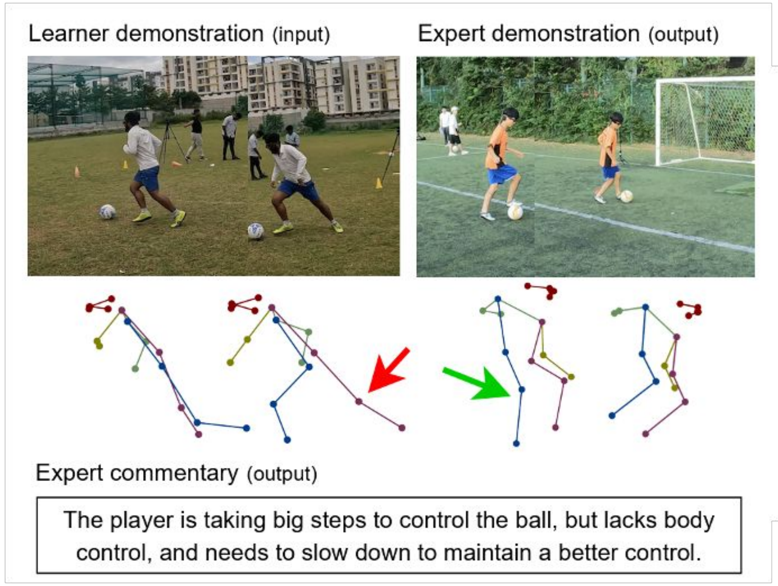

GeoYFCC distribution [9]
Percentage of Images from Region
40
30
20
10
0
GeoDE distribution (ours)
Percentage of Images from Region
15
10
5
0



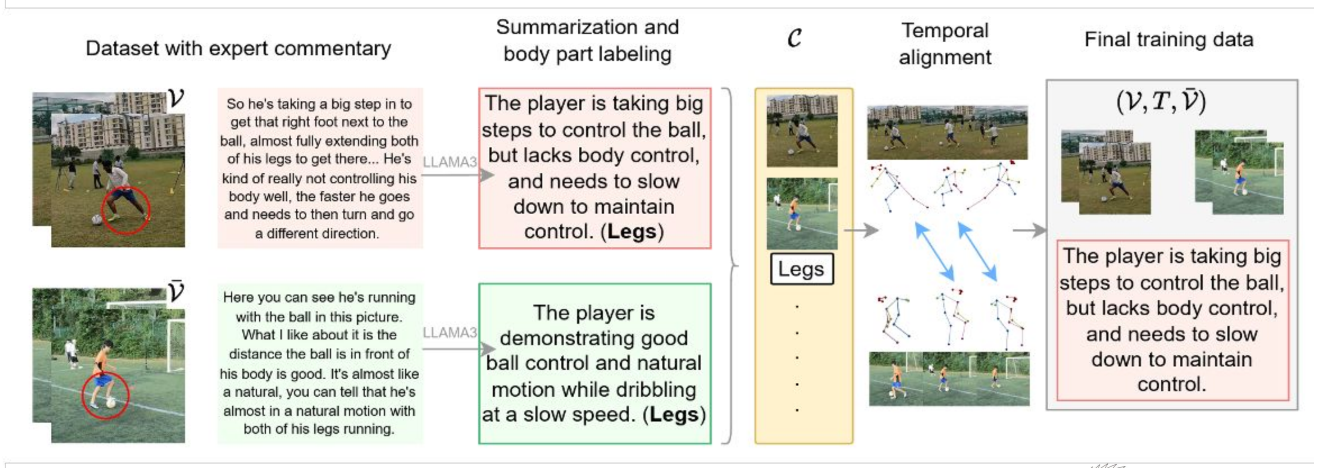
CVPR25: Meta Insights into Trends and Tendencies (100/179)

ExpertAF: Expert Actionable Feedback from Video

- **Summary:** This work proposes a task that generates video commentary, movement guidance (text), and correct posture (skeleton) from amateur videos (e.g., playing soccer) and corresponding pose (skeleton) data. From the Ego-Exo4D dataset for skill learning, they paired similar samples (amateur video + pose, expert video + pose) and used an LLM to generate movement guidance, thereby semi-automatically constructing a dataset. The method is simple, consisting of separate modules: an encoder for each input, an LLM-based motion guidance generator, and a Retrieval-Augmented Generator-based pose generator.
- **Novelty:** Task definition and dataset.
- **Thoughts:** The part that also outputs poses is interesting. There are still many challenges in tasks that aim to generate high-quality and visually meaningful feedback. Both the method and the dataset seem to have room for improvement. It's also worth exploring whether training with a large amount of data can produce precise feedback. Generating demonstrations for group activities may be particularly challenging.


Learner demonstration (input)
Expert demonstration (output)
Expert commentary (output)

The player is taking big steps to control the ball, but lacks body control, and needs to slow down to maintain a better control.


Dataset with expert commentary
Summarization and body part labeling
$\mathcal{C}$
Temporal alignment
Final training data
$\mathcal{V}$
$\bar{\mathcal{V}}$
LLAMA3
LLAMA3
Legs
$(\mathcal{V}, \mathcal{T}, \bar{\mathcal{V}})$

So he's taking a big step in to get that right foot next to the ball, almost fully extending both of his legs to get there... He's kind of really not controlling his body well, the faster he goes and needs to then turn and go a different direction.

The player is taking big steps to control the ball, but lacks body control, and needs to slow down to maintain control. (**Legs**)

Here you can see he's running with the ball in this picture. What I like about it is the distance the ball is in front of his body is good. It's almost like a natural, you can tell that he's almost in a natural motion with both of his legs running.

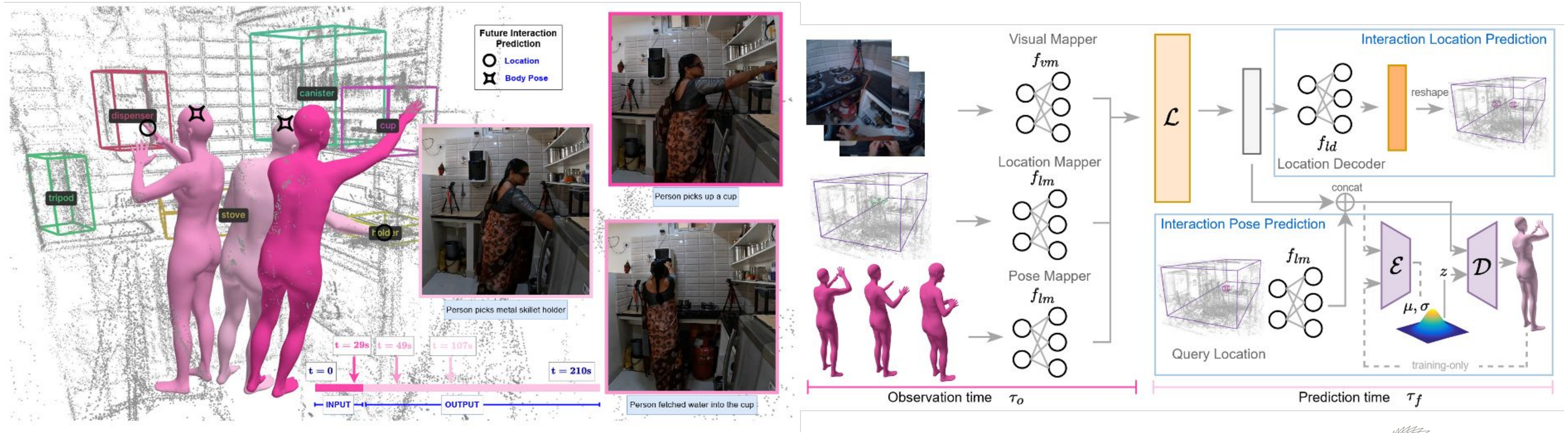
The player is demonstrating good ball control and natural motion while dribbling at a slow speed. (**Legs**)

The player is taking big steps to control the ball, but lacks body control, and needs to slow down to maintain control.

CVPR25: Meta Insights into Trends and Tendencies (101/179)

FICTION: 4D Future Interaction Prediction from Video

- **Summary:** This paper proposes *FICTION*, a method for 4D estimation from video—i.e., determining *when*, *where*, *which* object is being interacted with, and *how* (its pose). FICTION uses two VAE structures to estimate location and pose respectively from the embedding of past observations (see figure on the right).
- **Novelty:** While previous studies have explored estimating Video Scene Graphs from video, this work extends that task into 4D estimation.
- **Thoughts:** Research combining video and 3D is increasing. Papers from Kristen Grauman's group often use simple and easy-to-understand methods, especially for task-setting works—where they tend to first establish baselines using straightforward approaches. Also, since the same group is working on other 4D generation tasks, it seems they are shifting from video-only to video + 3D (i.e., 4D) representations.



CVPR25: Meta Insights into Trends and Tendencies (102/179)

Video-3D LLM: Learning Position-Aware Video Representation for 3D Scene Understanding

- **Summary:** This work proposes *Video-3D LLM*, a novel MLLM (Multimodal Large Language Model) method capable of 3D recognition. It treats 3D as dynamic video and aligns video representations with 3D representations. The model achieves high accuracy on several 3D recognition benchmarks.
- **Novelty:** The key novelty lies in recognizing 3D information from video, and enabling MLLMs to understand 3D using this approach.
- **Thoughts:** The ability to perceive 3D from video alone feels highly practical. In real-world applications, it's not always necessary to have explicit 3D input if the system can understand 3D from videos. Being able to recognize 3D using only video makes the method much more user-friendly and applicable.

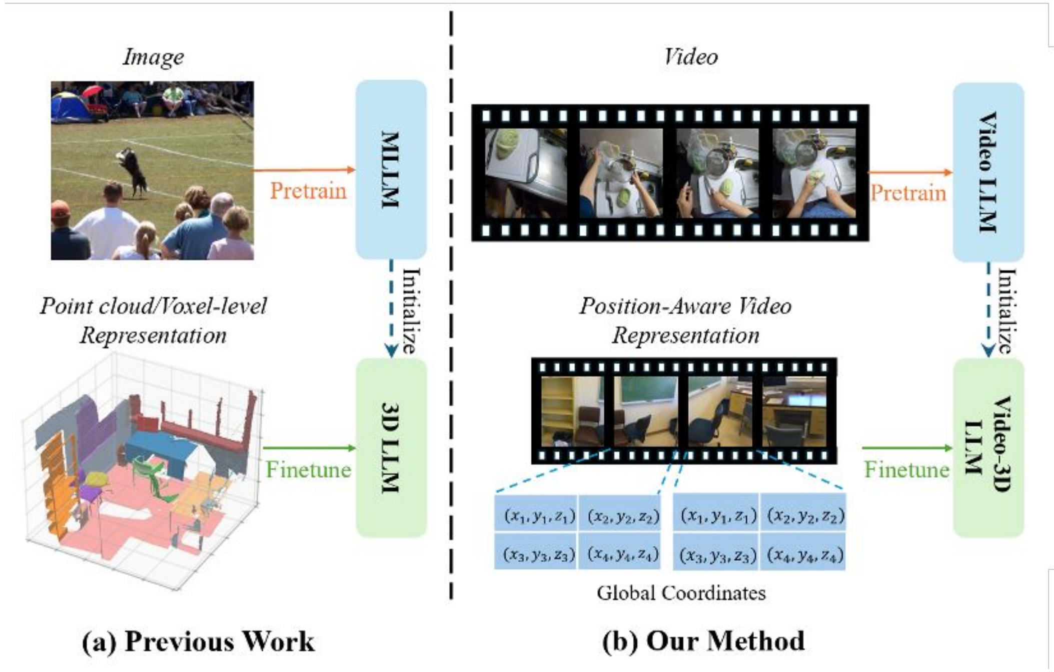
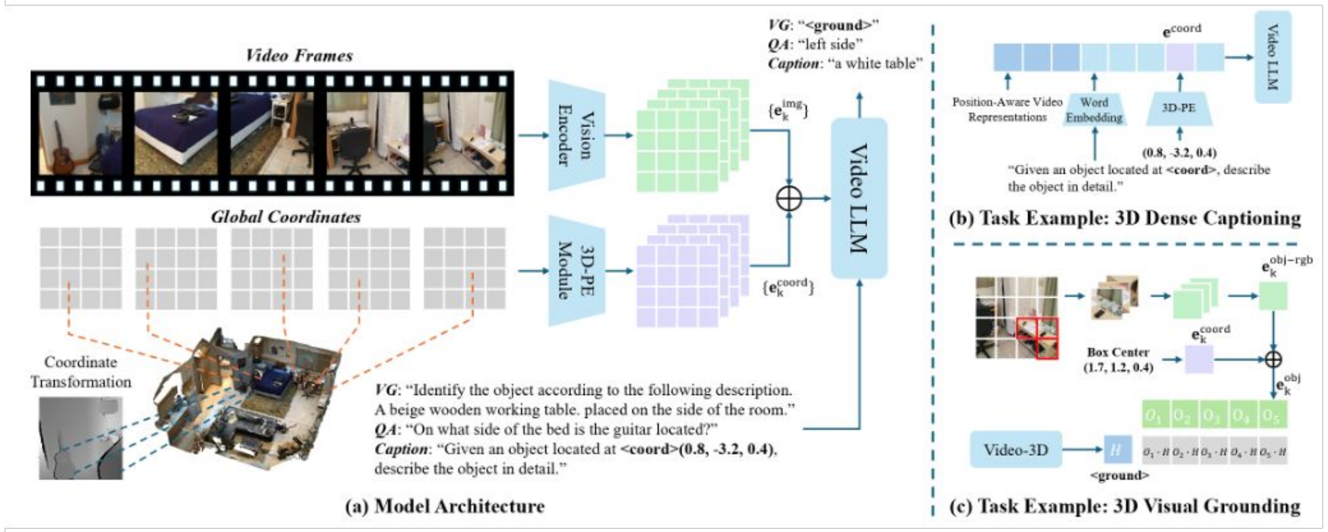


Image
MLLM
Pretrain
Initialize
Point cloud/Voxel-level Representation
Finetune
3D LLM
Video
Video LLM
Pretrain
Initialize
Position-Aware Video Representation
Finetune
Video-3D LLM
(x_1, y_1, z_1) (x_2, y_2, z_2) (x_1, y_1, z_1) (x_2, y_2, z_2)
(x_3, y_3, z_3) (x_4, y_4, z_4) (x_3, y_3, z_3) (x_4, y_4, z_4)
Global Coordinates

(a) Previous Work

(b) Our Method



VG: "**<ground>**"
QA: "left side"
Caption: "a white table"

VG: "Identify the object according to the following description. A beige wooden working table. placed on the side of the room."
QA: "On what side of the bed is the guitar located?"
Caption: "Given an object located at **<coord>(0.8, -3.2, 0.4)**, describe the object in detail."

"Given an object located at **<coord>**, describe the object in detail."

(a) Model Architecture

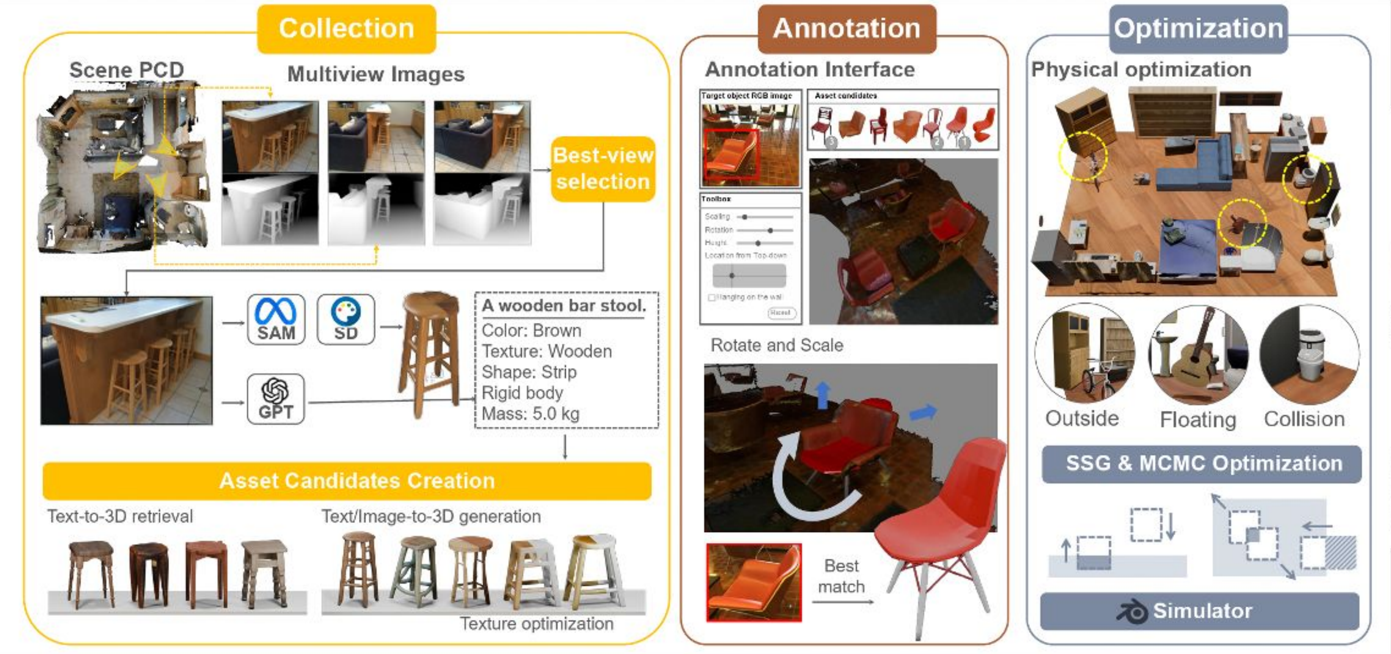
(b) Task Example: 3D Dense Captioning

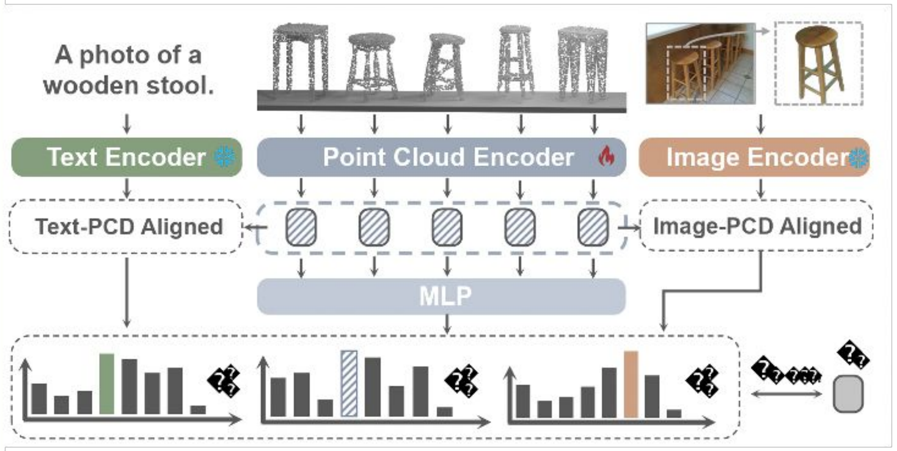
(c) Task Example: 3D Visual Grounding

LIMIT.LAB
https://limitlab.xyz/

METASCENES: Towards Automated Replica Creation for Real-world 3D Scans

- **Summary:** This work proposes a new dataset, *METASCENES*, and a retrieval-based method called *Scan2Sim*. METASCENES is built upon ScanNet, augmented with human annotations and enhanced through physical optimization by replacing ScanNet assets with CAD models. Experiments demonstrate the effectiveness of METASCENES in improving accuracy.
- **Novelty:** The work introduces a pipeline that easily converts real scanned data into simulation data. This significantly reduces the cost of generating 3D indoor environment datasets.
- **Thoughts:** It would be interesting to see comparisons with other methods for automatic creation of indoor 3D simulation data. Text-based generation methods might be more user-friendly in some cases.

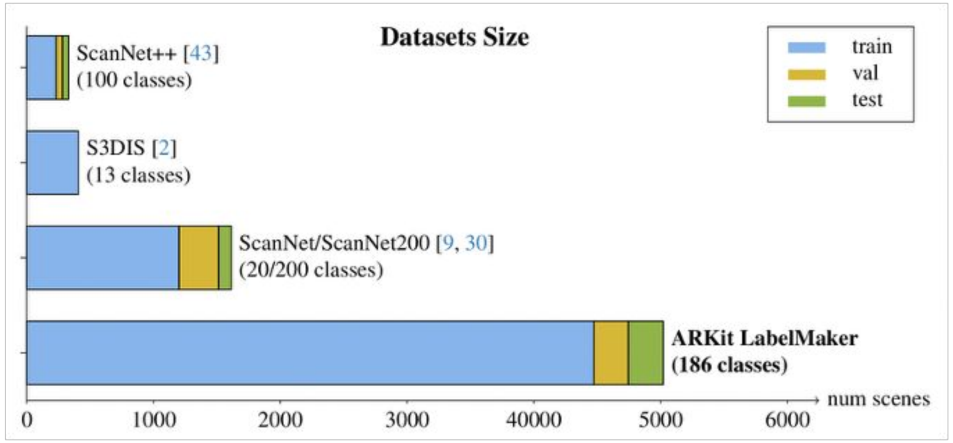


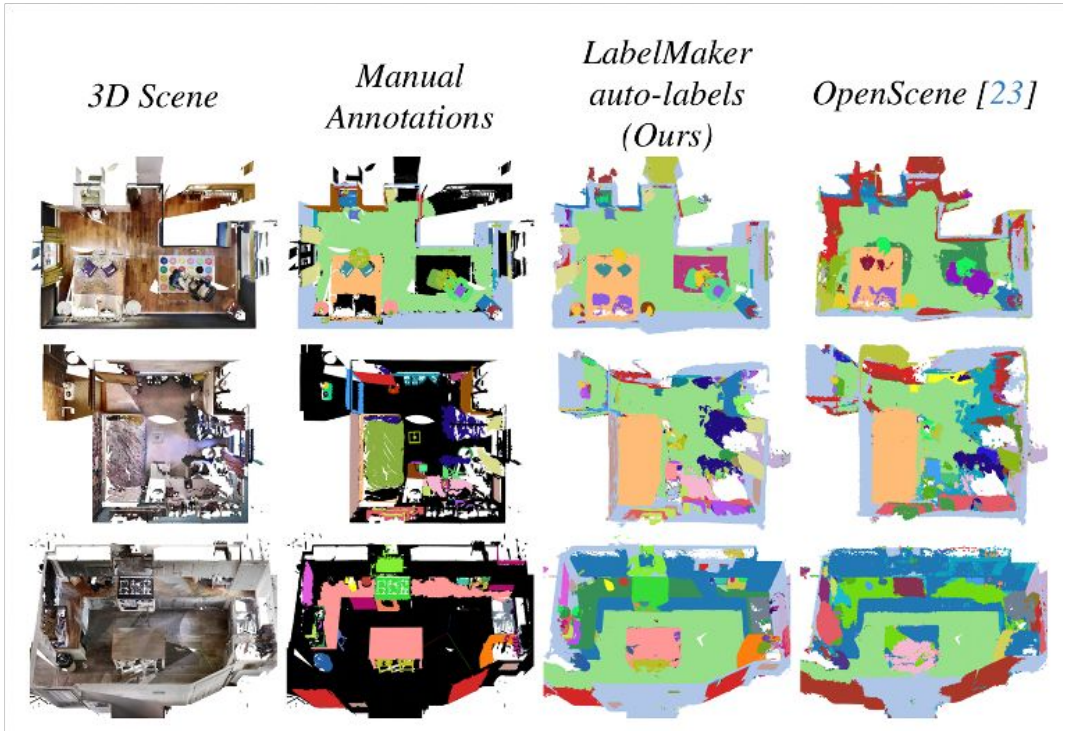


CVPR25: Meta Insights into Trends and Tendencies (104/179)

ARKit LabelMaker: A New Scale for Indoor 3D Scene Understanding

- **Summary:** This work proposes *ARKit LabelMaker*, a large-scale 3D semantic segmentation dataset—over three times larger than existing datasets. It also introduces an automated pipeline to generate ARKit LabelMaker, leveraging several high-performing models (OVSeg, Grounded-SAM, InternImage, Mask3D). Methods pre-trained on ARKit LabelMaker outperform those trained on existing datasets.
- **Novelty:** The dataset's scale is a key contribution, along with the automated pipeline that enables further scaling of data generation.
- **Thoughts:** If this approach could also scale language + 3D data together, performance might improve even further.

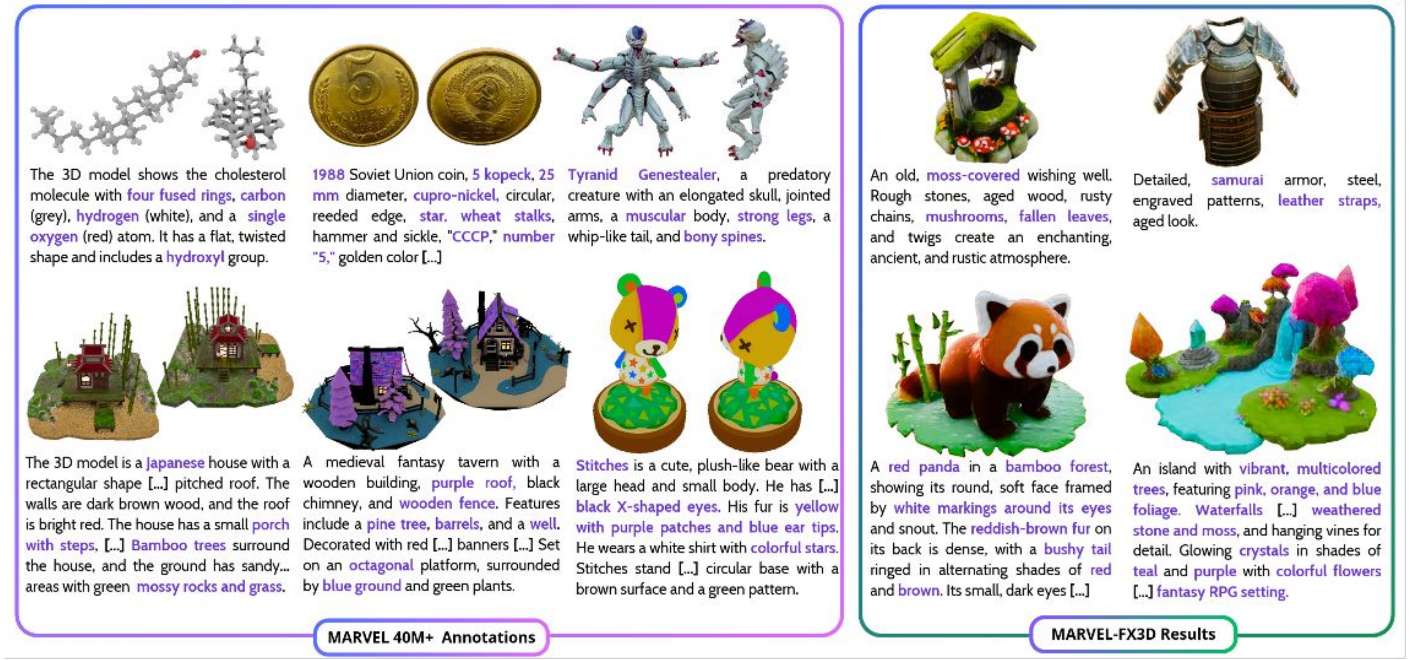




CVPR25: Meta Insights into Trends and Tendencies (105/179)

MARVEL-40M+: Multi-Level Visual Elaboration for High-Fidelity Text-to-3D Content Creation Creation

- ❏ **Summary:** This work proposes *MARVEL-40M+*, a large-scale dataset of text and 3D models (40 million text entries and 8.9 million 3D shapes). It introduces an automated dataset generation pipeline using a multimodal LLM to generate tags and detailed descriptions for each 3D model. The authors also propose a two-stage *Text-to-3D* generation model: Stage 1: Text-to-Image, and Stage 2: Image-to-3D.
- ❏ **Novelty:** The key novelty lies in performing both detailed 3D annotation and 3D model generation. The strong experimental results further validate the approach.
- ❏ **Thoughts:** It would be interesting to see tasks that combine 3D prompts (bounding boxes, segmentation, text) with 3D recognition.



The 3D model shows the cholesterol molecule with four fused rings, carbon (grey), hydrogen (white), and a single oxygen (red) atom. It has a flat, twisted shape and includes a hydroxyl group.

1988 Soviet Union coin, 5 kopeck, 25 mm diameter, cupro-nickel, circular, reeded edge, star, wheat stalks, hammer and sickle, "CCCP," number "5," golden color [...]

Tyranid Genestealer, a predatory creature with an elongated skull, jointed arms, a muscular body, strong legs, a whip-like tail, and bony spines.

The 3D model is a Japanese house with a rectangular shape [...] pitched roof. The walls are dark brown wood, and the roof is bright red. The house has a small porch with steps, [...] Bamboo trees surround the house, and the ground has sandy... areas with green mossy rocks and grass.

A medieval fantasy tavern with a wooden building, purple roof, black chimney, and wooden fence. Features include a pine tree, barrels, and a well. Decorated with red [...] banners [...] Set on an octagonal platform, surrounded by blue ground and green plants.

Stitches is a cute, plush-like bear with a large head and small body. He has [...] black X-shaped eyes. His fur is yellow with purple patches and blue ear tips. He wears a white shirt with colorful stars. Stitches stand [...] circular base with a brown surface and a green pattern.

MARVEL 40M+ Annotations

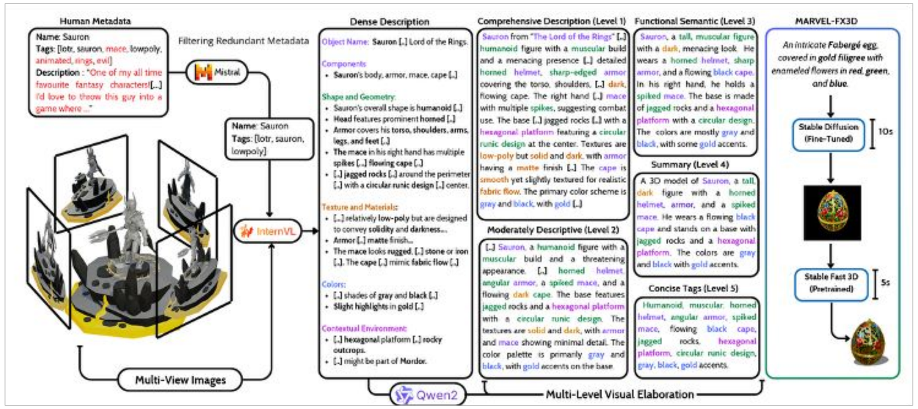
An old, moss-covered wishing well. Rough stones, aged wood, rusty chains, mushrooms, fallen leaves, and twigs create an enchanting, ancient, and rustic atmosphere.

Detailed, samurai armor, steel, engraved patterns, leather straps, aged look.

A red panda in a bamboo forest, showing its round, soft face framed by white markings around its eyes and snout. The reddish-brown fur on its back is dense, with a bushy tail ringed in alternating shades of red and brown. Its small, dark eyes [...]

An island with vibrant, multicolored trees, featuring pink, orange, and blue foliage. Waterfalls [...] weathered stone and moss, and hanging vines for detail. Glowing crystals in shades of teal and purple with colorful flowers [...] fantasy RPG setting.

MARVEL-FX3D Results



Human Metadata
Filtering Redundant Metadata
Mistral
InternVL
Multi-View Images
Dense Description
Qwen2
Multi-Level Visual Elaboration
MARVEL-FX3D
Stable Diffusion (Fine-Tuned)
10s
Stable Fast 3D (Pretrained)
5s

Human Metadata

Name: Sauron
Tags: [lotr, sauron, mace, lowpoly, animated, rings, evil]
Description : "One of my all time favourite fantasy characters![...] I'd love to throw this guy into a game where ..."

Name: Sauron
Tags: [lotr, sauron, lowpoly]

Dense Description

Object Name: Sauron [..] Lord of the Rings.

Components
- Sauron's body, armor, mace, cape [..]

Shape and Geometry:
- Sauron's overall shape is humanoid [..]
- Head features prominent horned [..]
- Armor covers his torso, shoulders, arms, legs, and feet [..]
- The mace in his right hand has multiple spikes [..] flowing cape [..]
- [..] jagged rocks [..] around the perimeter [..] with a circular runic design [..] center.

Texture and Materials:
- [..] relatively low-poly but are designed to convey solidity and darkness...
- Armor [..] matte finish...
- The mace looks rugged. [..] stone or iron [..]. The cape [..] mimic fabric flow [..]

Colors:
- [..] shades of gray and black [..]
- Slight highlights in gold [..]

Contextual Environment:
- [..] hexagonal platform [..] rocky outcrops.
- [..] might be part of Mordor.

Comprehensive Description (Level 1)

Sauron from "The Lord of the Rings" [..] humanoid figure with a muscular build and a menacing presence [..] detailed horned helmet, sharp-edged armor covering the torso, shoulders, [..] dark, flowing cape. The right hand [..] mace with multiple spikes, suggesting combat use. The base [..] jagged rocks [..] with a hexagonal platform featuring a circular runic design at the center. Textures are low-poly but solid and dark, with armor having a matte finish [..] The cape is smooth yet slightly textured for realistic fabric flow. The primary color scheme is gray and black, with gold [..]

Moderately Descriptive (Level 2)

[..] Sauron, a humanoid figure with a muscular build and a threatening appearance. [..] horned helmet, angular armor, a spiked mace, and a flowing dark cape. The base features jagged rocks and a hexagonal platform with a circular runic design. The textures are solid and dark, with armor and mace showing minimal detail. The color palette is primarly gray and black, with gold accents on the base.

Functional Semantic (Level 3)

Sauron, a tall, muscular figure with a dark, menacing look. He wears a horned helmet, sharp armor, and a flowing black cape. In his right hand, he holds a spiked mace. The base is made of jagged rocks and a hexagonal platform with a circular design. The colors are mostly gray and black, with some gold accents.

Summary (Level 4)

A 3D model of Sauron, a tall, dark figure with a horned helmet, armor, and a spiked mace. He wears a flowing black cape and stands on a base with jagged rocks and a hexagonal platform. The colors are gray and black with gold accents.

Concise Tags (Level 5)

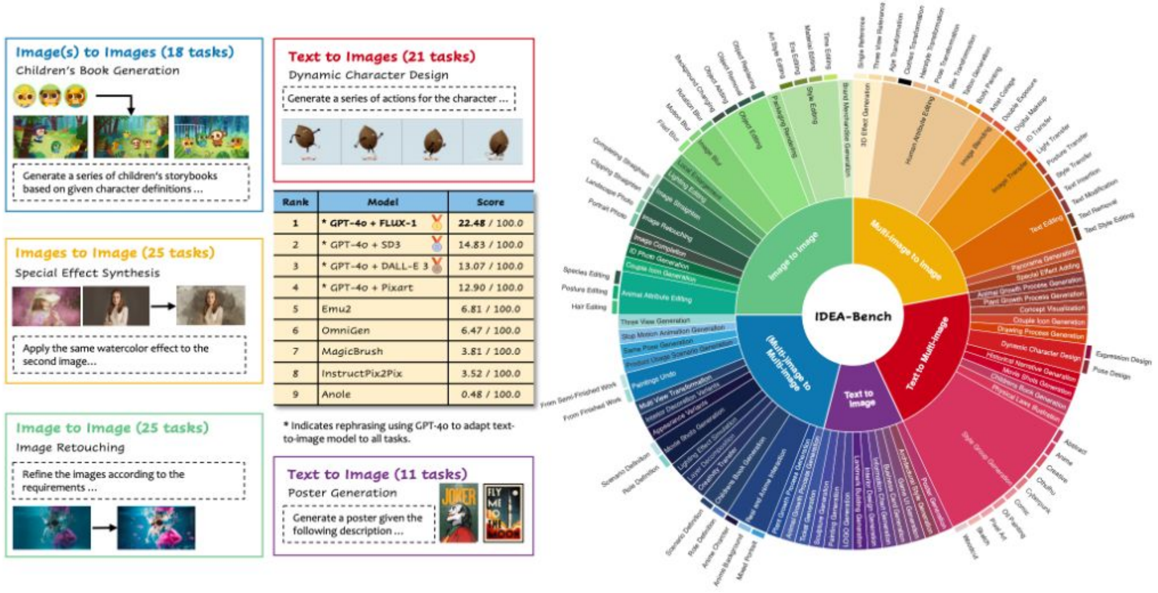
Humanoid, muscular, horned helmet, angular armor, spiked mace, flowing black cape, jagged rocks, hexagonal platform, circular runic design, gray, black, gold accents.

*An intricate **Fabergé egg** covered in **gold filigree** with enameled flowers in **red**, **green**, and **blue**.*

CVPR25: Meta Insights into Trends and Tendencies (106/179)

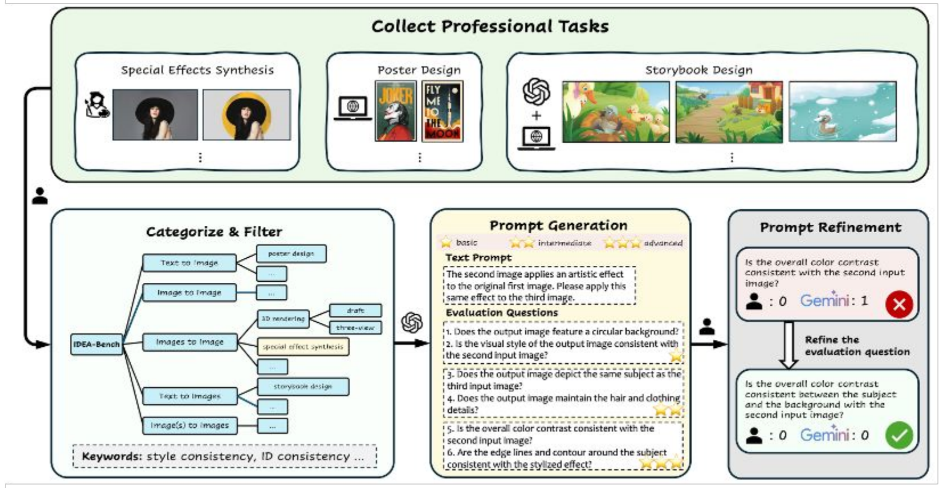
IDEA-BENCH: HOW FAR ARE GENERATIVE MODELS FROM PROFESSIONAL DESIGNING?

- **Summary:** A new benchmark called IDEA-BENCH was created to examine how closely current generative models can perform design work in a “professional designer” setting. The benchmark contains 100 tasks spanning text-to-image, image-to-image, and image-editing scenarios, and it was used to evaluate several prominent models. The study offers a detailed analysis of the shortcomings those models exhibit when asked to solve genuine design problems.
- **Novelty:** IDEA-BENCH assesses real-world design tasks along multiple axes. Even powerful systems such as DALL-E 3 and InstructPix2Pix achieve only about 20 points out of 100 on this benchmark, revealing systematic weaknesses that had not been fully documented.
- **Thoughts:** Evaluation tasks that gauge a model’s proximity to human expertise are becoming more common. If a dataset can reliably expose crucial failure modes, then a moderate dataset size could be enough. Developing robust evaluation schemes for generative models remains an intriguing research direction.


Image(s) to Images (18 tasks)
Children's Book Generation
Generate a series of children's storybooks based on given character definitions ...
Images to Image (25 tasks)
Special Effect Synthesis
Apply the same watercolor effect to the second image...
Image to Image (25 tasks)
Image Retouching
Refine the images according to the requirements ...
Text to Images (21 tasks)
Dynamic Character Design
Generate a series of actions for the character ...
Text to Image (11 tasks)
Poster Generation
Generate a poster given the following description ...
IDEA-Bench
Image to Image
Multi-image to Image
Text to Multi-image
Text to Image
(Multi-)image to Multi-image

Rank	Model	Score
1	*** GPT-4o + FLUX-1**	**22.48** / 100.0
2	* GPT-4o + SD3	14.83 / 100.0
3	* GPT-4o + DALL-E 3	13.07 / 100.0
4	* GPT-4o + Pixart	12.90 / 100.0
5	Emu2	6.81 / 100.0
6	OmniGen	6.47 / 100.0
7	MagicBrush	3.81 / 100.0
8	InstructPix2Pix	3.52 / 100.0
9	Anole	0.48 / 100.0

* Indicates rephrasing using GPT-4o to adapt text-to-image model to all tasks.



Text Prompt

The second image applies an artistic effect to the original first image. Please apply this same effect to the third image.

Evaluation Questions

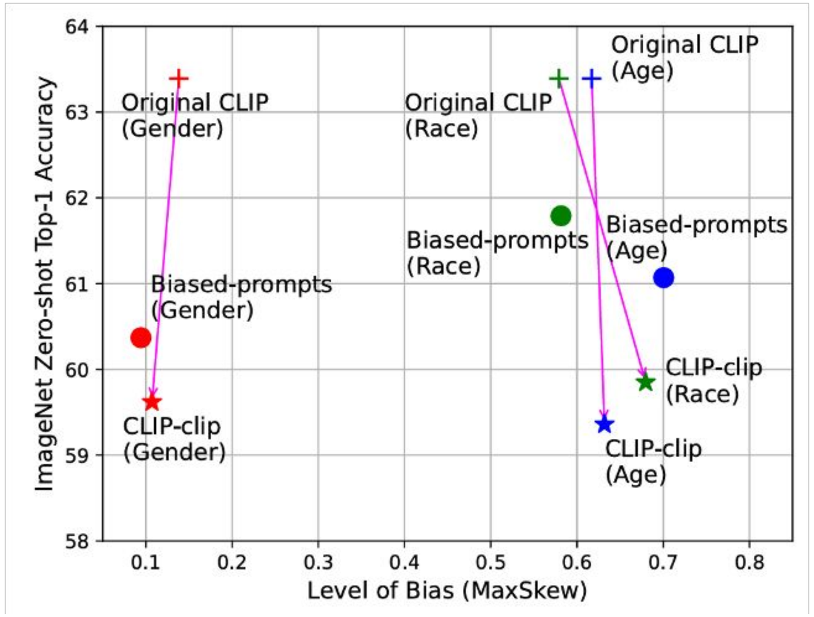
1. Does the output image feature a circular background?
2. Is the visual style of the output image consistent with the second input image?
3. Does the output image depict the same subject as the third input image?
4. Does the output image maintain the hair and clothing details?
5. Is the overall color contrast consistent with the second input image?
6. Are the edge lines and contour around the subject consistent with the stylized effect?

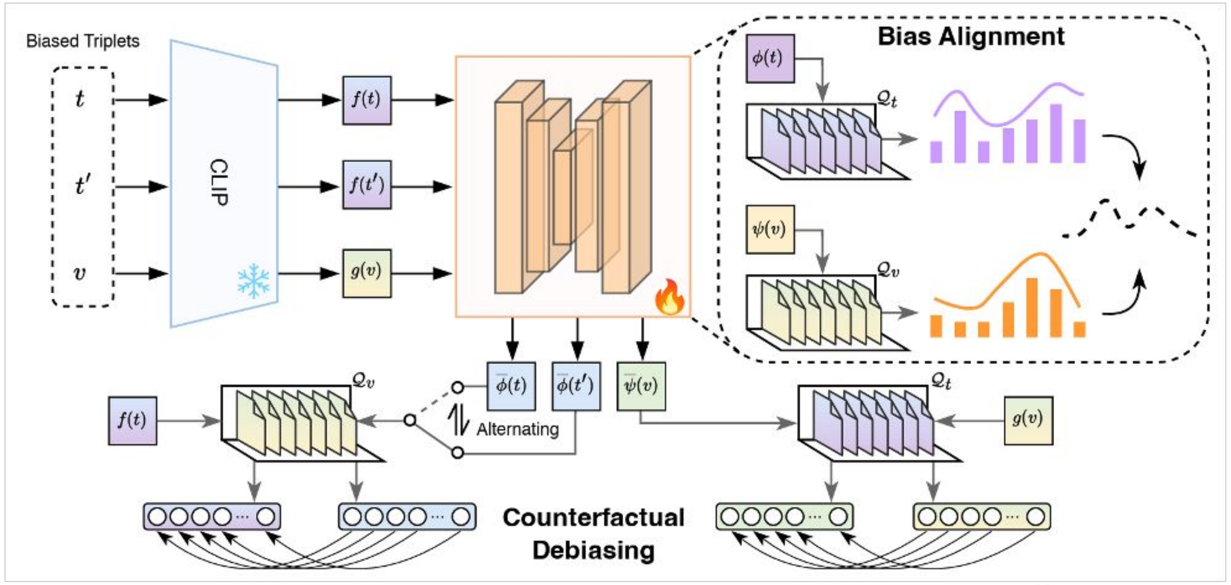
Is the overall color contrast consistent with the second input image?

Is the overall color contrast consistent between the subject and the background with the second input image?

Joint Vision-Language Social Bias Removal for CLIP

- **Summary:** A method is introduced for removing gender, race, and age bias from CLIP embeddings. Previous approaches that strip biased attributes from embeddings often degrade downstream task performance. Through a comprehensive analysis, the study presents a technique that debiases both the image and text branches of CLIP, achieving strong bias mitigation while largely preserving accuracy on downstream tasks.
- **Novelty:** Debiasing CLIP remains relatively under-explored despite its importance. The work offers a thorough critique of existing methods and addresses their limitations, delivering an approach that balances fairness and task performance.
- **Thoughts:** Investigating bias in large (multi)-modal language models is a pressing challenge, and this contribution highlights both the difficulties and the potential of targeted debiasing strategies.

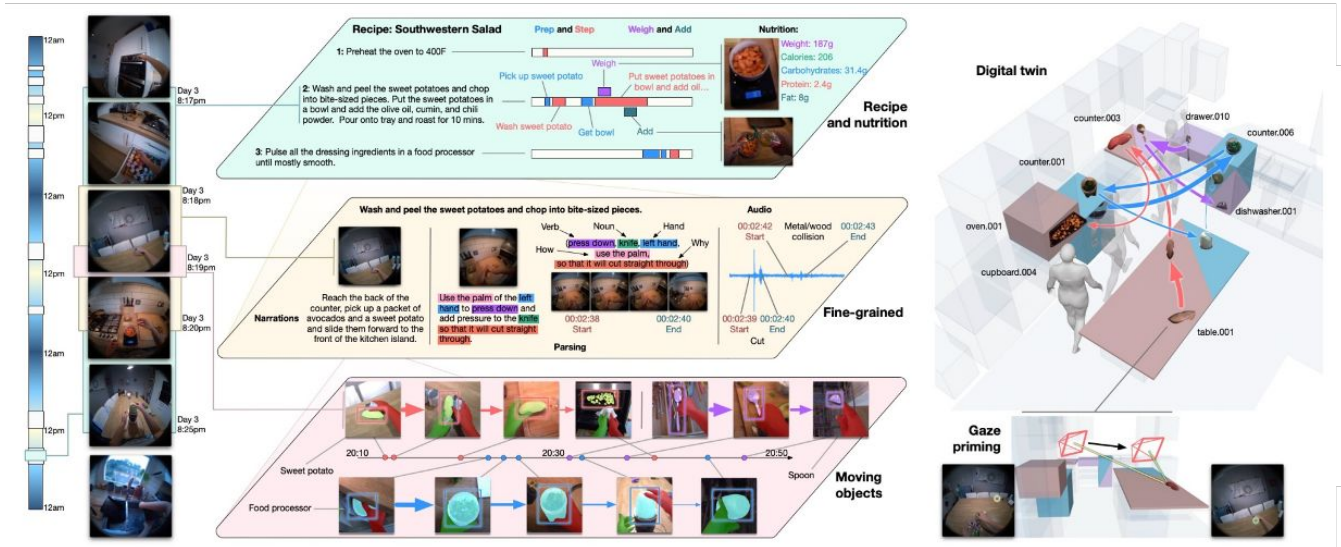




CVPR25: Meta Insights into Trends and Tendencies (108/179)

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

- **Summary:** The study presents HD-EPIC, a fine-grained video dataset enriched with recipe steps, atomic actions, ingredient lists with nutritional values, object-tracking labels, audio annotations, and fully aligned 3-D digital-twin data. Even the strong Gemini Pro model achieves only 37.6 % accuracy on the accompanying benchmark, underscoring the dataset's difficulty.
- **Novelty:** HD-EPIC extends existing resources by combining fine-grained video annotations with explicit 3-D twins, creating a unified test bed for vision-language and embodied-AI research. The tight linkage between 2-D video and 3-D geometry is a particularly novel contribution.
- **Thoughts:** The results highlight how current multi-modal language models still fall short on detailed video understanding. Producing fine-grained, 3-D-aligned labels is labor-intensive, yet the resulting dataset should prove valuable for a wide range of video, robotics, and embodied-reasoning tasks.



1: Preheat the oven to 400F

2: Wash and peel the sweet potatoes and chop into bite-sized pieces. Put the sweet potatoes in a bowl and add the olive oil, cumin, and chili powder. Pour onto tray and roast for 10 mins.

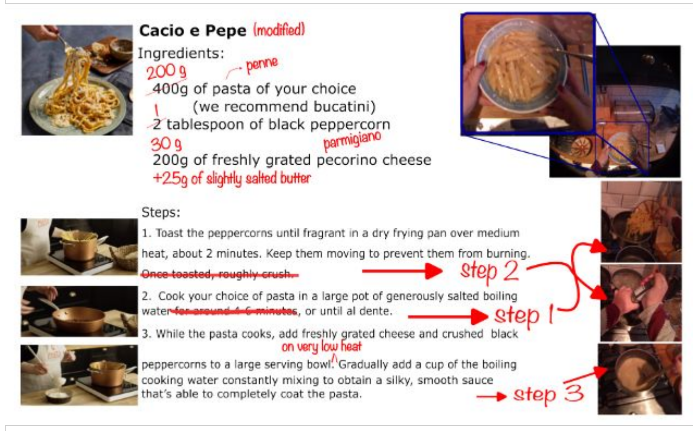
3: Pulse all the dressing ingredients in a food processor until mostly smooth.

Wash and peel the sweet potatoes and chop into bite-sized pieces.

Reach the back of the counter, pick up a packet of avocados and a sweet potato and slide them forward to the front of the kitchen island.

Use the palm of the left hand to press down and add pressure to the knife so that it will cut straight through.

(press down, knife, left hand, use the palm, so that it will cut straight through)



Cacio e Pepe (modified)

Ingredients:

200 g ~~400g~~ of pasta of your choice → penne
(we recommend bucatini)
1 ~~2~~ tablespoon of black peppercorn
30 g 200g of freshly grated pecorino cheese parmigiano
+25g of slightly salted butter

Steps:

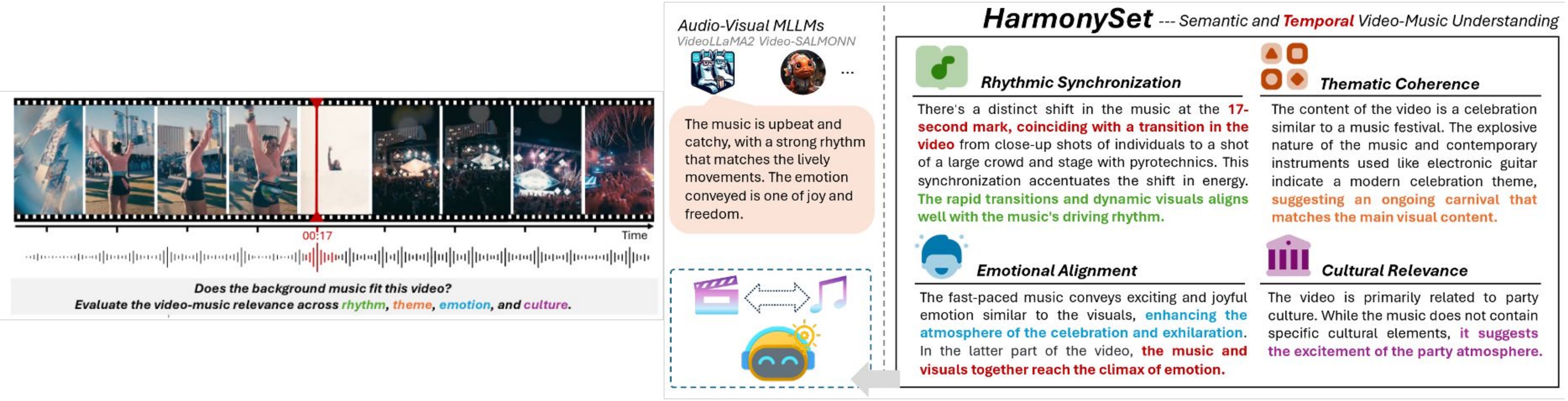
1. Toast the peppercorns until fragrant in a dry frying pan over medium heat, about 2 minutes. Keep them moving to prevent them from burning. ~~Once toasted, roughly crush.~~ → step 2
2. Cook your choice of pasta in a large pot of generously salted boiling water ~~for around 4-6 minutes~~, or until al dente. → step 1
3. While the pasta cooks, add freshly grated cheese and crushed black peppercorns to a large serving bowl. on very low heat Gradually add a cup of the boiling cooking water constantly mixing to obtain a silky, smooth sauce that's able to completely coat the pasta. → step 3

LIMIT.LAB
https://limitlab.xyz/

CVPR25: Meta Insights into Trends and Tendencies (109/179)

HarmonySet: A Comprehensive Dataset for Understanding Video–Music Semantic Alignment and Temporal Synchronization

- **Summary:** HarmonySet is introduced as an instruction-tuning dataset for training and evaluating multi-modal language models on joint music–video understanding. The collection supports assessment of rhythmic synchronization, emotional alignment, thematic coherence, and cultural relevance. All annotations are generated automatically with the aid of existing MLLMs, enabling large-scale coverage.
- **Novelty:** Very few benchmarks explicitly target the fusion of musical and visual semantics for MLLMs. HarmonySet fills this gap and offers new evaluation axes that go beyond conventional audiovisual alignment.
- **Thoughts:** There is a visible trend toward building datasets and metrics for challenging domains such as music and design, where subjective judgments make evaluation difficult. HarmonySet represents a timely step in that direction.



Does the background music fit this video?
Evaluate the video-music relevance across rhythm, theme, emotion, and culture.

The music is upbeat and catchy, with a strong rhythm that matches the lively movements. The emotion conveyed is one of joy and freedom.

Rhythmic Synchronization: There's a distinct shift in the music at the 17-second mark, coinciding with a transition in the video from close-up shots of individuals to a shot of a large crowd and stage with pyrotechnics. This synchronization accentuates the shift in energy. The rapid transitions and dynamic visuals aligns well with the music's driving rhythm.

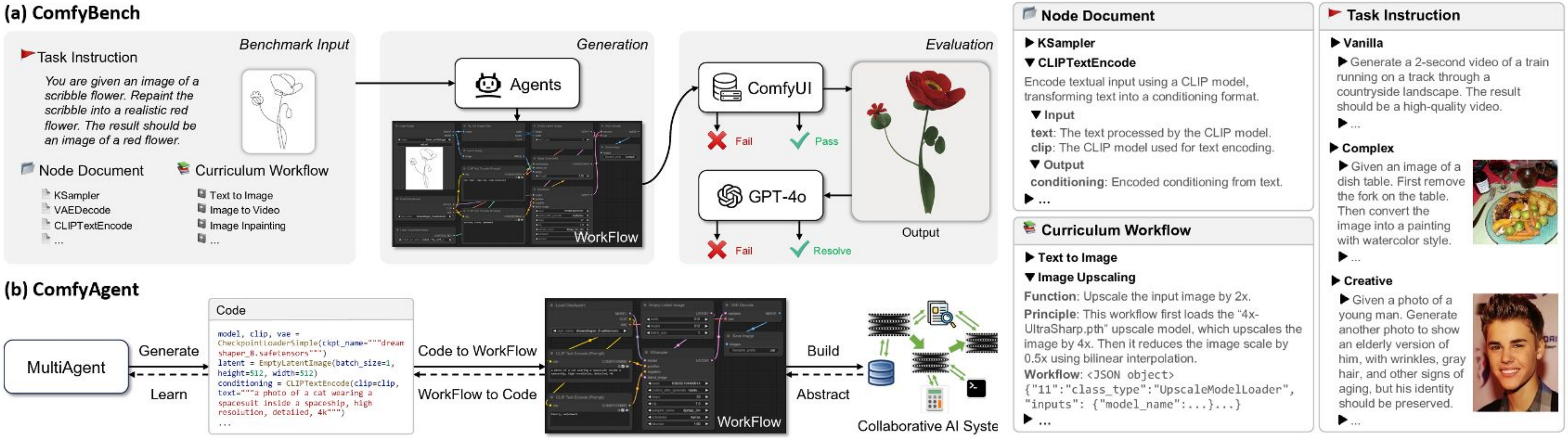
Thematic Coherence: The content of the video is a celebration similar to a music festival. The explosive nature of the music and contemporary instruments used like electronic guitar indicate a modern celebration theme, suggesting an ongoing carnival that matches the main visual content.

Emotional Alignment: The fast-paced music conveys exciting and joyful emotion similar to the visuals, enhancing the atmosphere of the celebration and exhilaration. In the latter part of the video, the music and visuals together reach the climax of emotion.

Cultural Relevance: The video is primarily related to party culture. While the music does not contain specific cultural elements, it suggests the excitement of the party atmosphere.

ComfyBench: Benchmarking LLM-based Agents in ComfyUI for Autonomously Designing Collaborative AI Systems

- ❏ **Summary**: A benchmark called ComfyBench and an accompanying method named ComfyAgent are introduced for automatically generating—given a high-level task—the required tools, step-by-step workflow, and intermediate prompts. The resulting “flow” is expressed in an easily editable code format, and ComfyAgent reaches accuracy comparable to the O1-Preview system.
- ❏ **Novelty:** The work extends visual-programming ideas to a newer tool ecosystem; although the conceptual leap is modest, matching O1-Preview’s performance is noteworthy.
- ❏ **Thoughts:** Building agents by combining predefined toolsets with LLM-generated prompts reflects an interesting trade-off and resembles modern visual-programming paradigms.



You are given an image of a scribble flower. Repaint the scribble into a realistic red flower. The result should be an image of a red flower.

```
model, clip, vae =
CheckpointLoaderSimple(ckpt_name="""dream
shaper_8.safetensors""")
latent = EmptyLatentImage(batch_size=1,
height=512, width=512)
conditioning = CLIPTextEncode(clip=clip,
text="""a photo of a cat wearing a
spacesuit inside a spaceship, high
resolution, detailed, 4k""")
...
```

Node Document

▶ **KSampler**

▼ **CLIPTextEncode**

Encode textual input using a CLIP model, transforming text into a conditioning format.

▼ **Input**

text: The text processed by the CLIP model.
clip: The CLIP model used for text encoding.

▼ **Output**

conditioning: Encoded conditioning from text.

▶ ...

Curriculum Workflow

▶ **Text to Image**

▼ **Image Upscaling**

Function: Upscale the input image by 2x.

Principle: This workflow first loads the “4x-UltraSharp.pth” upscale model, which upscales the image by 4x. Then it reduces the image scale by 0.5x using bilinear interpolation.

Workflow: `<JSON object>`
`{"11":"class_type":"UpscaleModelLoader", "inputs": {"model_name":...}...}`

▶ ...

Task Instruction

▶ **Vanilla**

▶ Generate a 2-second video of a train running on a track through a countryside landscape. The result should be a high-quality video.

▶ ...

▶ **Complex**

▶ Given an image of a dish table. First remove the fork on the table. Then convert the image into a painting with watercolor style.

▶ ...

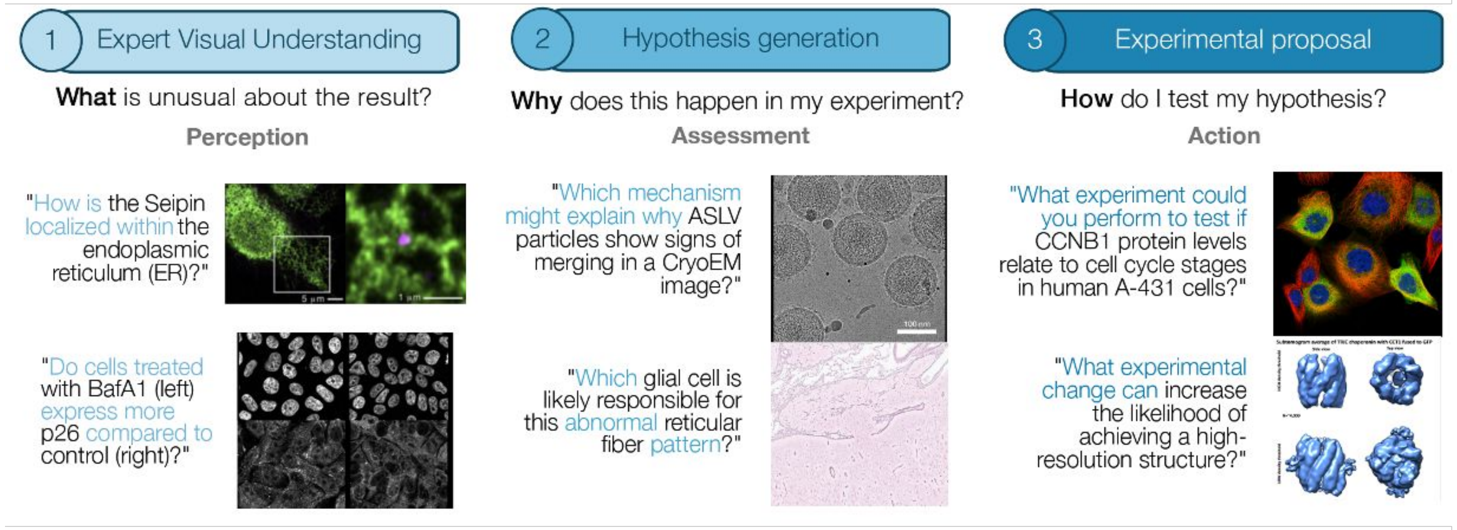
▶ **Creative**

▶ Given a photo of a young man. Generate another photo to show an elderly version of him, with wrinkles, gray hair, and other signs of aging, but his identity should be preserved.

▶ ...

MicroVQA: A Multimodal Reasoning Benchmark for Microscopy-Based Scientific Research

- **Summary:** The paper presents MicroVQA, a microscopy-based visual-question-answering dataset. Unlike conventional VQA corpora, MicroVQA is organized around the full scientific workflow—observation, hypothesis generation, and experimental verification—and was curated by professional biologists. Several multi-modal language models are evaluated on the benchmark, revealing a clear advantage for larger models.
- **Novelty:** MicroVQA goes beyond simple question–answer pairs by embedding scientific reasoning steps into the data structure. Evaluation shows that most errors stem from incorrect visual recognition, indicating that interpreting microscopy images remains a bottleneck.
- **Thoughts:** “AI for Science” resources are on the rise, and datasets that mirror the entire research process, such as MicroVQA, appear especially valuable. Results suggest that domain-specific imagery (microscopy, remote sensing, etc.) still suffers from insufficient pre-training in current models.



1 Expert Visual Understanding

What is unusual about the result?

Perception

"How is the Seipin localized within the endoplasmic reticulum (ER)?"

"Do cells treated with BafA1 (left) express more p26 compared to control (right)?"

2 Hypothesis generation

Why does this happen in my experiment?

Assessment

"Which mechanism might explain why ASLV particles show signs of merging in a CryoEM image?"

"Which glial cell is likely responsible for this abnormal reticular fiber pattern?"

3 Experimental proposal

How do I test my hypothesis?

Action

"What experiment could you perform to test if CCNB1 protein levels relate to cell cycle stages in human A-431 cells?"

"What experimental change can increase the likelihood of achieving a high-resolution structure?"

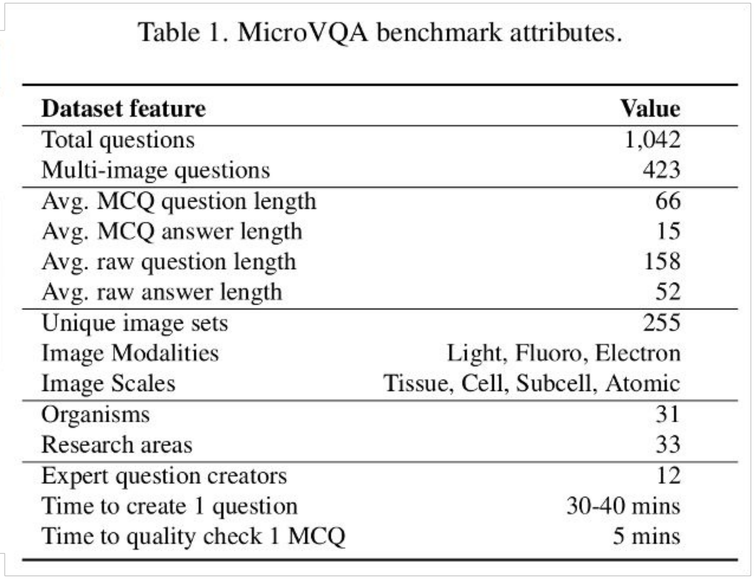


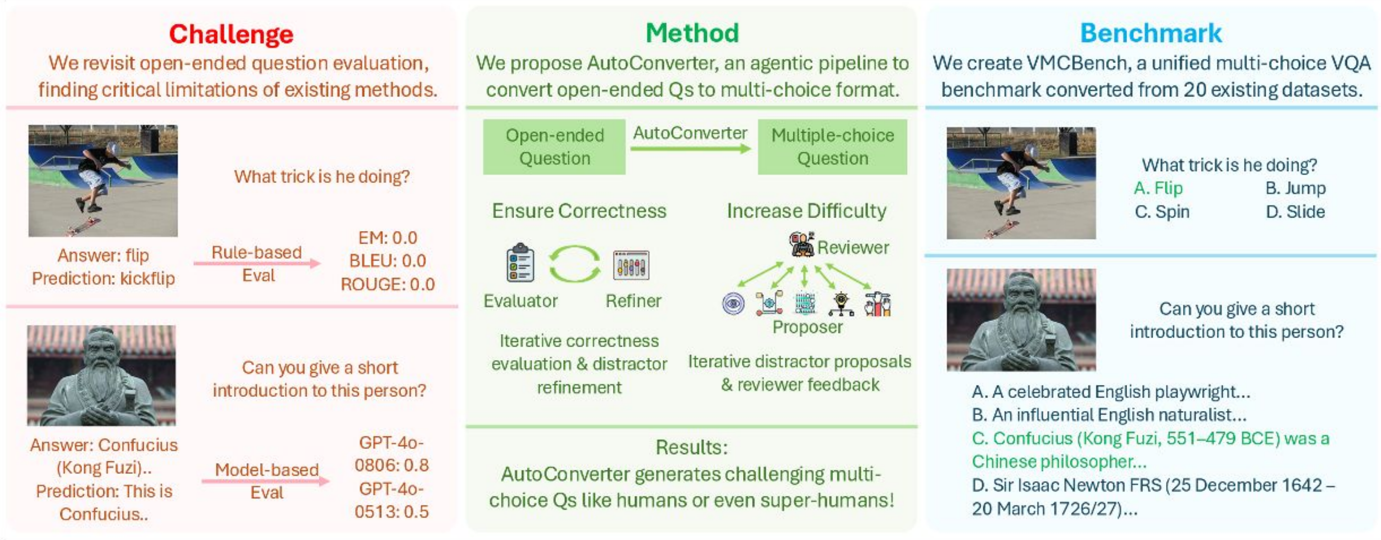
Table 1. MicroVQA benchmark attributes.

Dataset feature	Value
Total questions	1,042
Multi-image questions	423
Avg. MCQ question length	66
Avg. MCQ answer length	15
Avg. raw question length	158
Avg. raw answer length	52
Unique image sets	255
Image Modalities	Light, Fluoro, Electron
Image Scales	Tissue, Cell, Subcell, Atomic
Organisms	31
Research areas	33
Expert question creators	12
Time to create 1 question	30-40 mins
Time to quality check 1 MCQ	5 mins

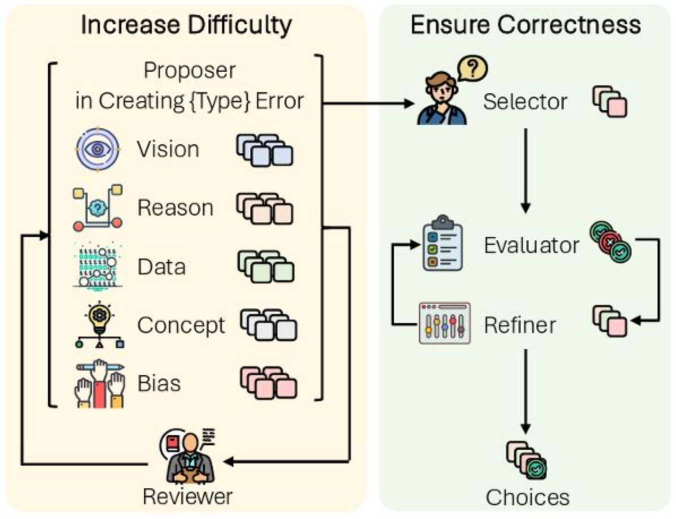
CVPR25: Meta Insights into Trends and Tendencies (112/179)

Automated Generation of Challenging Multiple-Choice Questions for Vision Language Model Evaluation

- ❏ **Summary:** The paper introduces AutoConverter, an MLLM-based method that automatically transforms existing open-ended VQA datasets into multiple-choice format. The converted data enable the creation of much larger multiple-choice VQA benchmarks, making it easier to evaluate multi-modal language models at scale.
- ❏ **Novelty:** AutoConverter's main contribution lies in repurposing open-ended VQA data for multiple-choice evaluation. While useful, the conceptual advance is relatively modest.
- ❏ **Thoughts:** The same idea could plausibly be extended to video VQA. Although image and video datasets are abundant, comprehensive benchmarks for 3-D recognition remain scarce. In the long run, it would be valuable to develop AI systems that can identify their own knowledge gaps, acquire relevant data, and continue learning autonomously.



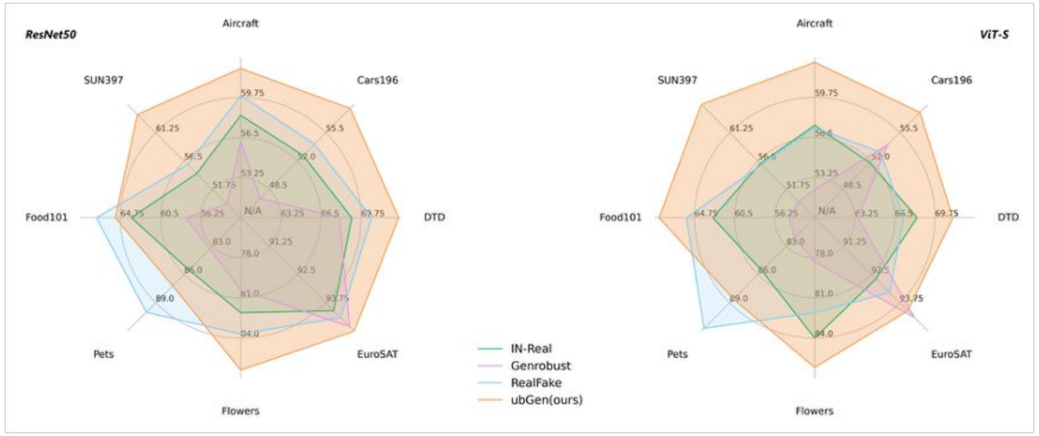
Challenge
We revisit open-ended question evaluation, finding critical limitations of existing methods.
What trick is he doing?
Answer: flip
Prediction: kickflip
Rule-based Eval
EM: 0.0
BLEU: 0.0
ROUGE: 0.0
Can you give a short introduction to this person?
Answer: Confucius (Kong Fuzi)..
Prediction: This is Confucius..
Model-based Eval
GPT-4o-0806: 0.8
GPT-4o-0513: 0.5
Method
We propose AutoConverter, an agentic pipeline to convert open-ended Qs to multi-choice format.
Open-ended Question
AutoConverter
Multiple-choice Question
Ensure Correctness
Increase Difficulty
Evaluator
Refiner
Reviewer
Proposer
Iterative correctness evaluation & distractor refinement
Iterative distractor proposals & reviewer feedback
Results:
AutoConverter generates challenging multi-choice Qs like humans or even super-humans!
Benchmark
We create VMCBench, a unified multi-choice VQA benchmark converted from 20 existing datasets.
What trick is he doing?
A. Flip
B. Jump
C. Spin
D. Slide
Can you give a short introduction to this person?
A. A celebrated English playwright...
B. An influential English naturalist...
C. Confucius (Kong Fuzi, 551–479 BCE) was a Chinese philosopher...
D. Sir Isaac Newton FRS (25 December 1642 – 20 March 1726/27)...

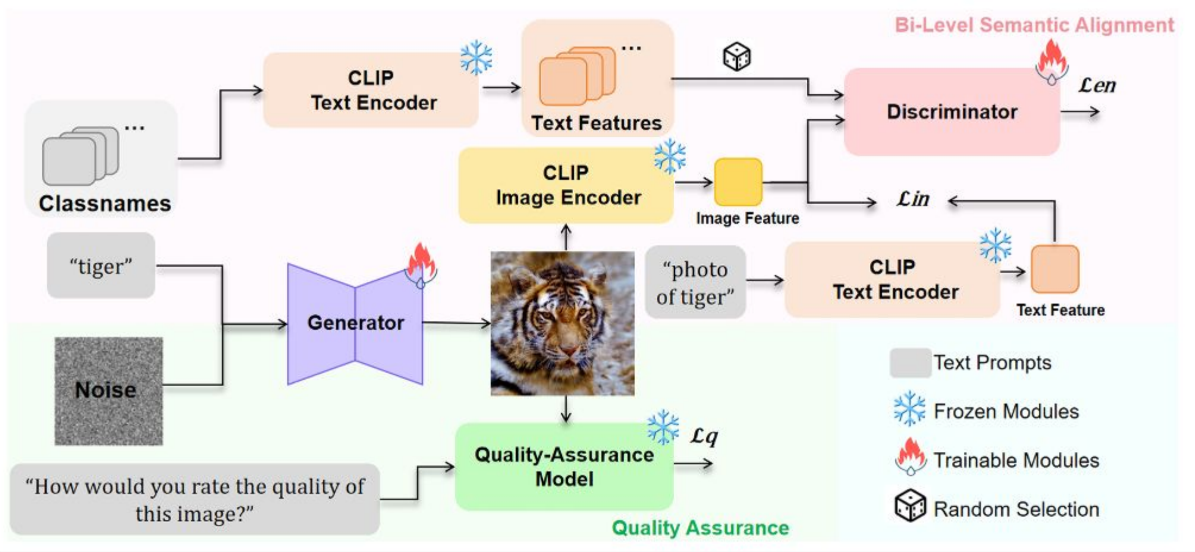


(a) *AutoConverter* framework.

Unbiased General Annotated Dataset Generation

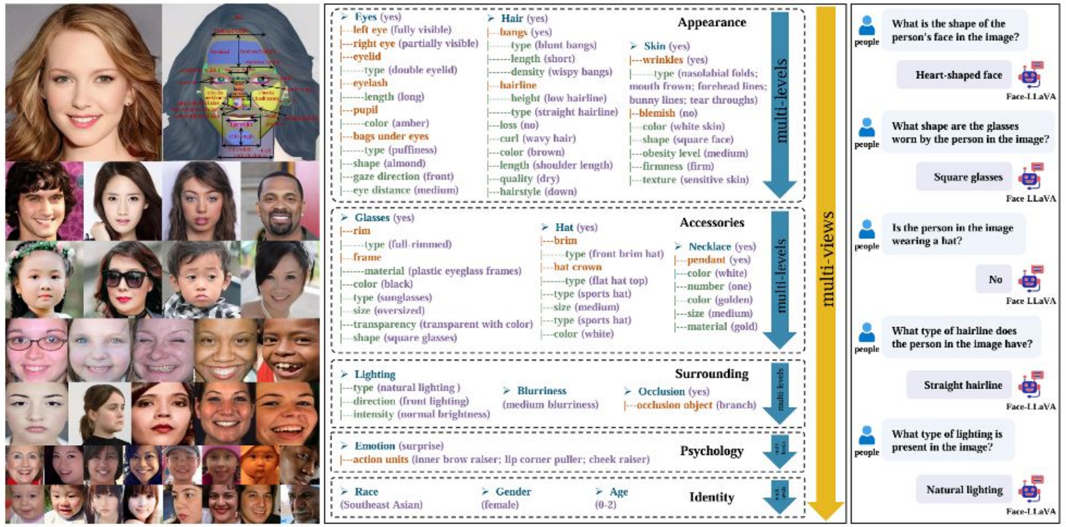
- **Summary:** A data-generation method called ubGen is introduced to improve the generalization of image-recognition models. The approach generates images along directions in CLIP's semantic space and employs a quality–text-alignment function to ensure both fidelity and diversity. Models trained on ubGen-augmented data achieve markedly better transfer performance across several benchmark datasets.
- **Novelty:** ubGen can produce highly transferable training images using nothing more than the category names, offering a lightweight alternative to large-scale manual collection.
- **Thoughts:** Work on boosting diffusion-model expressiveness, often by leveraging MLLMs, appears to be gaining momentum, and ubGen fits neatly into this emerging trend.

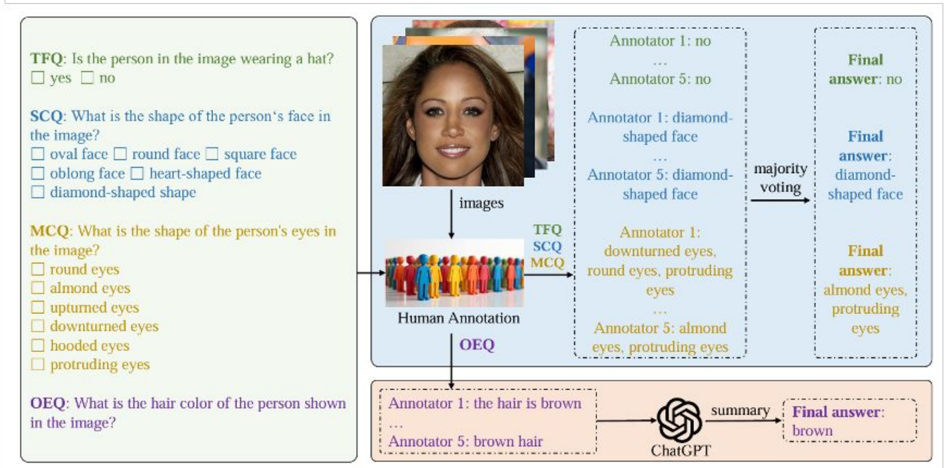




FaceBench: A Multi-View Multi-Level Facial Attribute VQA Dataset for Benchmarking Face Perception MLLMs

- **Summary:** A dataset named FaceBench is introduced for multi-view, fine-grained face recognition. It contains 210 annotated facial attributes and more than 70 000 question, answer pairs. A model pre-trained on FaceBench attains accuracy comparable to GPT-4o and Gemini on the benchmark.
- **Novelty:** FaceBench provides substantially finer facial-attribute coverage than previous resources, enabling more detailed recognition and analysis.
- **Thoughts:** While FaceBench targets static face images, a complementary video corpus for nuanced expression recognition would also be valuable. Growing demand is evident for fine-grained datasets across many domains, though it is somewhat surprising that the FaceBench-trained model does not surpass GPT-4o.

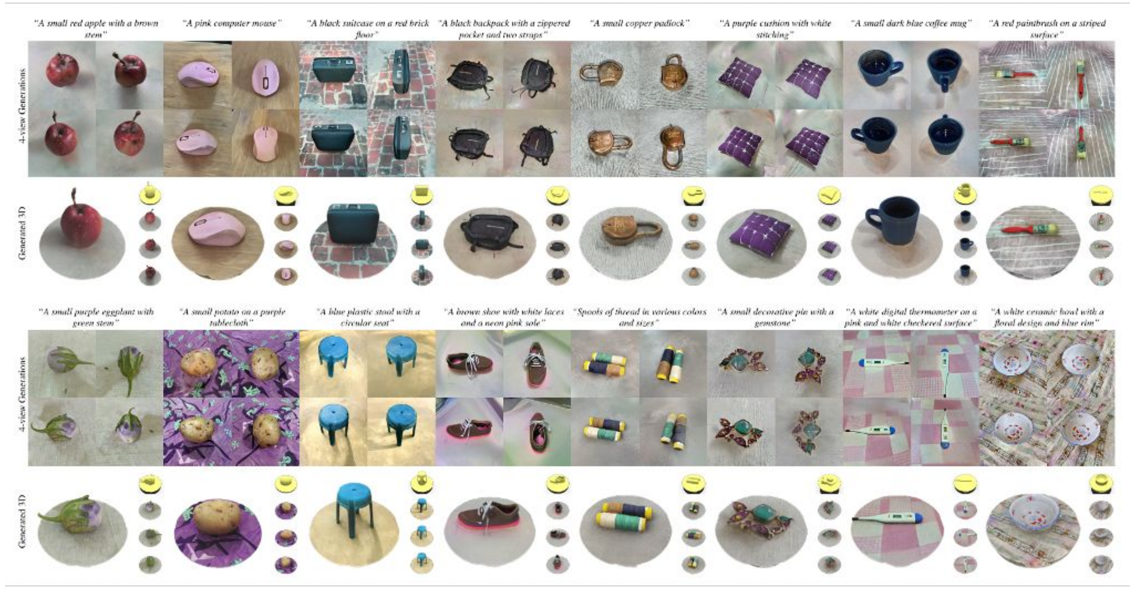



TFQ: Is the person in the image wearing a hat?
☐ yes ☐ no
SCQ: What is the shape of the person's face in the image?
☐ oval face ☐ round face ☐ square face
☐ oblong face ☐ heart-shaped face
☐ diamond-shaped shape
MCQ: What is the shape of the person's eyes in the image?
☐ round eyes
☐ almond eyes
☐ upturned eyes
☐ downturned eyes
☐ hooded eyes
☐ protruding eyes
OEQ: What is the hair color of the person shown in the image?
images
Human Annotation
TFQ
SCQ
MCQ
OEQ
Annotator 1: no
…
Annotator 5: no
Annotator 1: diamond-shaped face
…
Annotator 5: diamond-shaped face
Annotator 1: downturned eyes, round eyes, protruding eyes
…
Annotator 5: almond eyes, protruding eyes
majority voting
Final answer: no
Final answer: diamond-shaped face
Final answer: almond eyes, protruding eyes
Annotator 1: the hair is brown
…
Annotator 5: brown hair
ChatGPT
summary
Final answer: brown

CVPR25: Meta Insights into Trends and Tendencies (115/179)

UnCommon Objects in 3D

- **Summary:** A large-scale 3-D object dataset named uCO3D is introduced. Spanning more than 1 000 object categories, each sample includes accurate structure-from-motion camera poses, a point cloud, and a 3-D Gaussian-splat reconstruction. Experiments show that models trained on uCO3D achieve markedly better performance in novel-view synthesis and 3-D reconstruction than when trained on MVImageNet or CO3Dv2.
- **Novelty:** uCO3D surpasses existing datasets in both scale and annotation quality, offering denser geometry and more object diversity than previous resources.
- **Thoughts:** Pairing such high-fidelity geometry with strong language supervision could be an interesting next step, although complex shapes and textures remain difficult to describe textually.

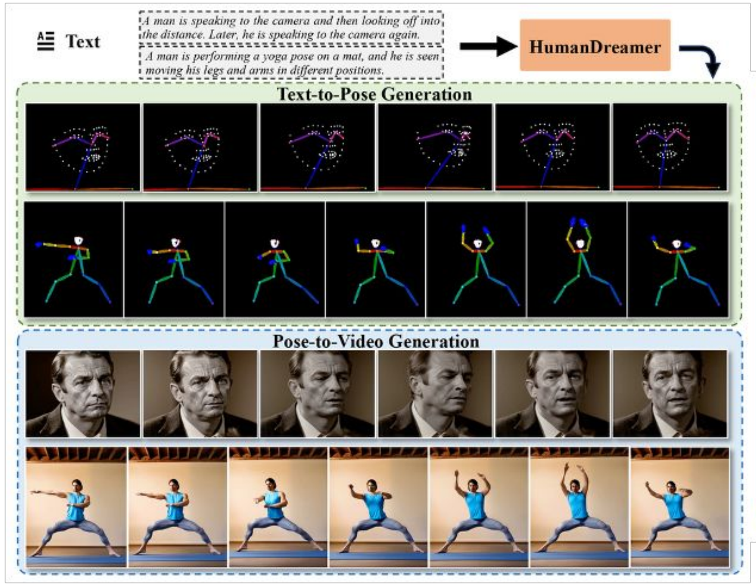


"A small red apple with a brown stem"
"A pink computer mouse"
"A black suitcase on a red brick floor"
"A black backpack with a zippered pocket and two straps"
"A small copper padlock"
"A purple cushion with white stitching"
"A small dark blue coffee mug"
"A red paintbrush on a striped surface"
4-view Generations
Generated 3D
"A small purple eggplant with green stem"
"A small potato on a purple tablecloth"
"A blue plastic stool with a circular seat"
"A brown shoe with white laces and a neon pink sole"
"Spools of thread in various colors and sizes"
"A small decorative pin with a gemstone"
"A white digital thermometer on a pink and white checkered surface"
"A white ceramic bowl with a floral design and blue rim"
4-view Generations
Generated 3D

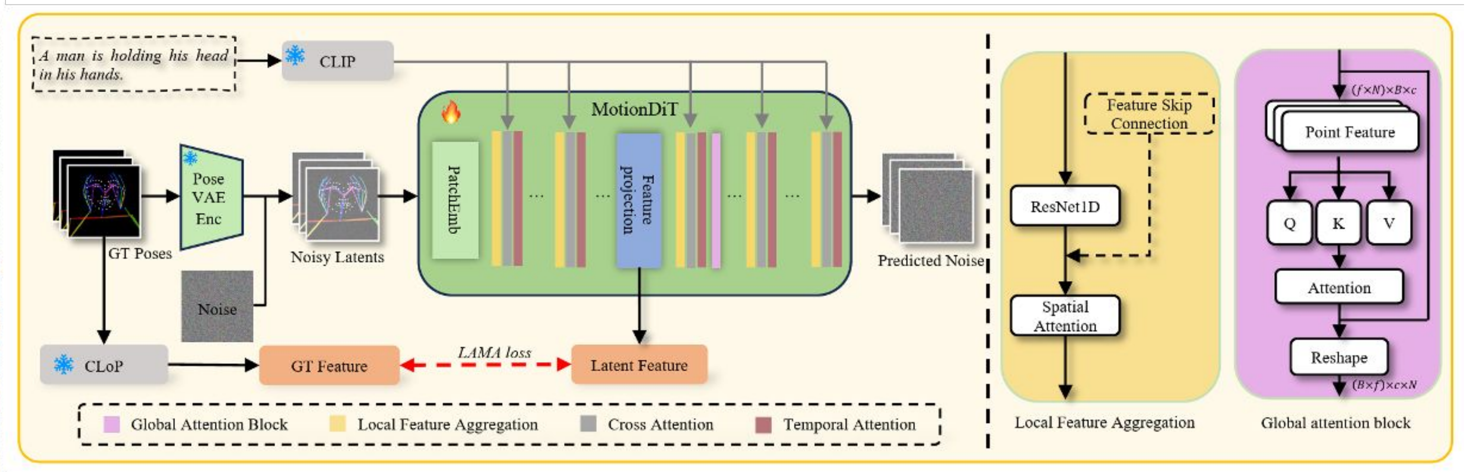
CVPR25: Meta Insights into Trends and Tendencies (116/179)

HumanDreamer: Generating Controllable Human-Motion Videos via Decoupled Generation

- **Summary:** This paper introduces a diffusion-based video generation method that produces both poses and corresponding videos from text inputs in a two-stage process: first generating poses from text, then generating video from the predicted poses. To enable the text-to-pose stage, a large-scale, semi-automatically constructed dataset is presented. Additionally, a new loss function, LAMA loss, is proposed for the text-to-pose model to better align predicted poses with pose features.
- **Novelty:** Key contributions include the construction process and release of a large-scale Text-to-Pose dataset, as well as the introduction of the LAMA loss tailored for pose prediction from text.
- **Thoughts:** Text-to-pose generation has broad potential applications. This work also follows a growing trend of using pre-existing models and large language models (LLMs) to automate dataset construction.



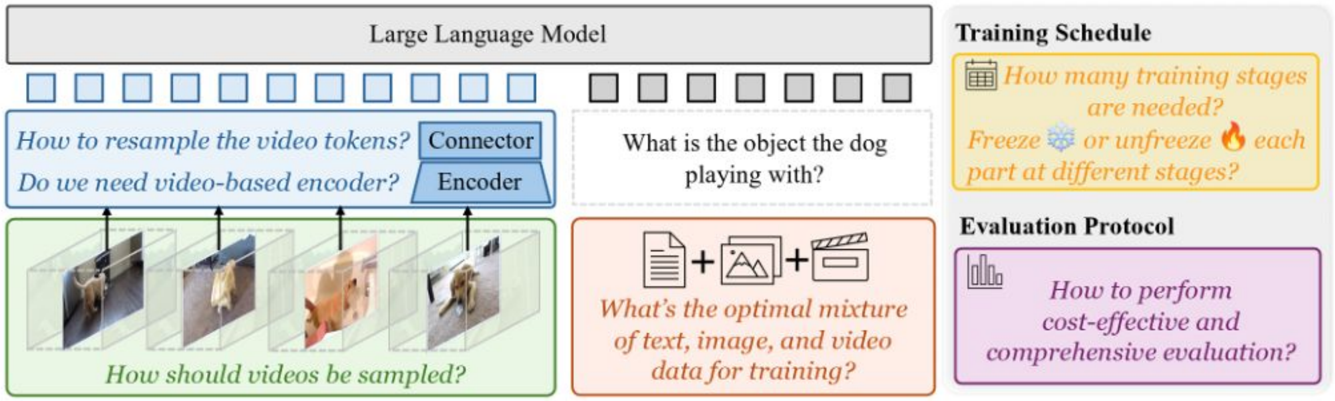
Text
A man is speaking to the camera and then looking off into the distance. Later, he is speaking to the camera again.
A man is performing a yoga pose on a mat, and he is seen moving his legs and arms in different positions.
HumanDreamer
Text-to-Pose Generation
Pose-to-Video Generation



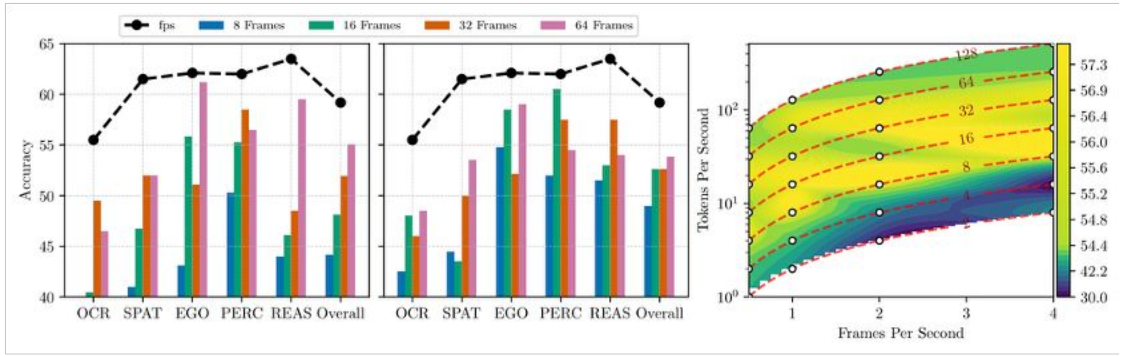
CVPR25: Meta Insights into Trends and Tendencies (117/179)

Apollo: An Exploration of Video Understanding in Large Multi-Modal Models

- **Summary:** This Meta research conducts a comprehensive experimental study and comparison of Video LLM design choices, culminating in the introduction of the Apollo model series. Apollo achieves higher accuracy compared to models with similar computational cost.
- **Novelty:** Key findings include the observation of scaling consistency—performance trends seen on relatively large datasets continue to hold on even larger ones. Additionally, FPS-based frame sampling is shown to outperform uniform sampling, yielding performance gains proportional to the number of frames, which is not the case for uniform sampling. At the time of publication, models using InternVideo2 + SigLIP SO400M features achieved the best results.
- **Thoughts:** The study involves an extensive amount of experimentation and provides valuable insights into the design of video recognition models.



Large Language Model
How to resample the video tokens?
Connector
Do we need video-based encoder?
Encoder
What is the object the dog playing with?
How should videos be sampled?
What's the optimal mixture of text, image, and video data for training?
Training Schedule
How many training stages are needed?
Freeze ❄ or unfreeze 🔥 each part at different stages?
Evaluation Protocol
How to perform cost-effective and comprehensive evaluation?

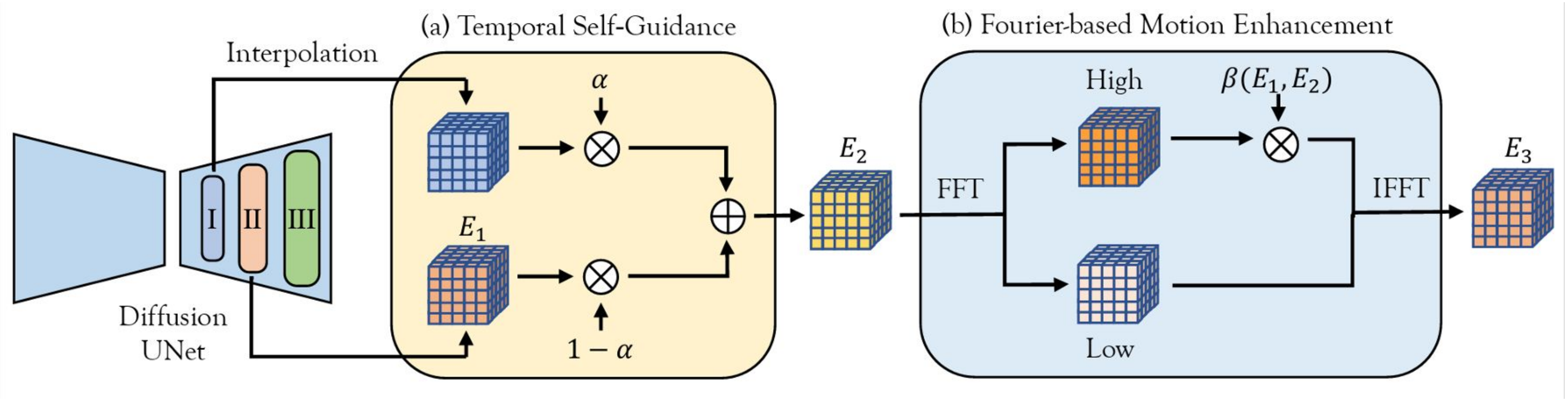


LIMIT.LAB
https://limitlab.xyz/

CVPR25: Meta Insights into Trends and Tendencies (118/179)

ByTheWay: Boost Your Text-to-Video Generation Model to Higher Quality in a Training-free Way

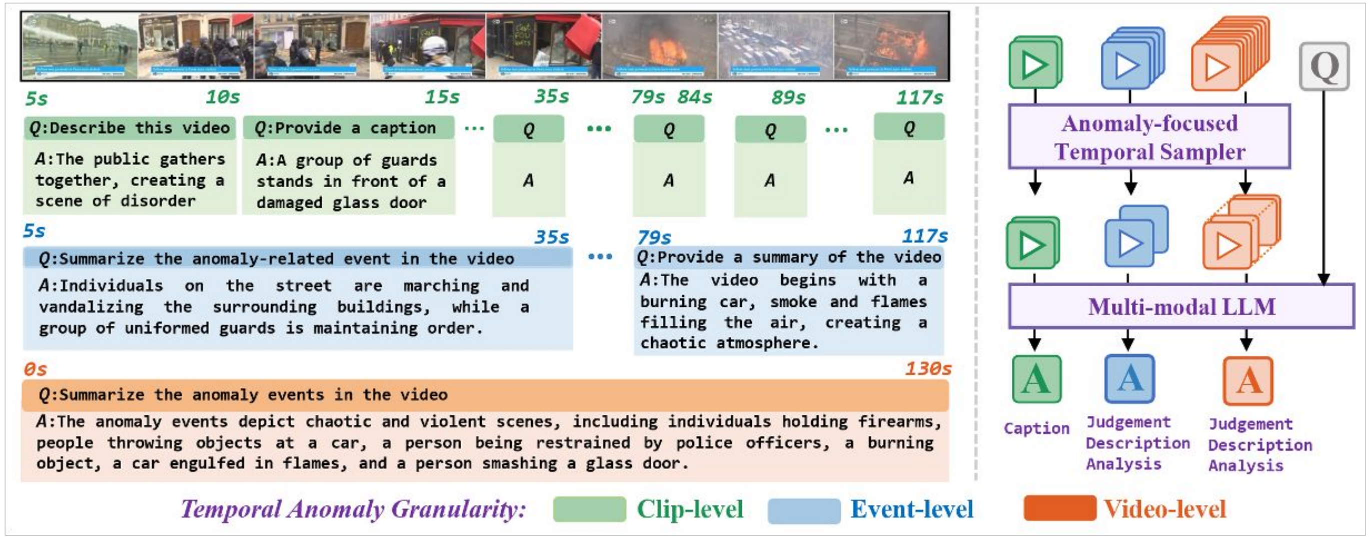
- **Summary:** This work introduces a training-free method to improve generation consistency and motion magnitude in diffusion-based video generation. Through an analysis of temporal attention, the authors identify that differences between temporal attention maps strongly affect temporal consistency. By adjusting these maps to enhance consistency, overall generation stability improves. Additionally, the energy of temporal attention maps is found to influence motion magnitude, leading to a method that amplifies motion strength. The proposed technique significantly improves the output quality of existing models such as AnimationDiff and VideoCrafter2.
- **Novelty:** The study offers a comprehensive analysis of temporal attention maps and introduces a novel, training-free method for enhancing consistency and motion dynamics in generated videos.
- **Thoughts:** It would be valuable to verify whether the proposed method generalizes to other diffusion-based video generation models.



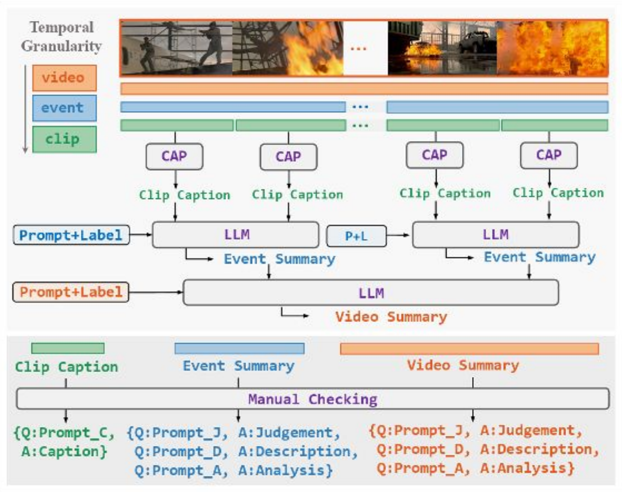
CVPR25: Meta Insights into Trends and Tendencies (119/179)

Holmes-VAU: Towards Long-term Video Anomaly Understanding at Any Granularity

- **Summary:** This paper introduces **Holmes-VAU**, a large-scale dataset designed to recognize anomalies in videos across **multiple temporal granularities**. The dataset is semi-automatically constructed using existing multi-modal language models (MLLMs). In addition, a method combining a **temporal sampler** with an MLLM-based model is proposed for anomaly detection at various time scales.
- **Novelty:** The key contribution lies in formulating anomaly recognition across **different temporal granularities**, which offers a new perspective compared to standard fixed-scale video analysis.
- **Thoughts:** The idea of analyzing video at varying temporal resolutions is clearly structured and meaningful. However, since related datasets like temporal localization benchmarks already exist, the degree of novelty may be somewhat limited.



5s
10s
15s
35s
79s
84s
89s
117s
Q:Describe this video
A:The public gathers together, creating a scene of disorder
Q:Provide a caption
A:A group of guards stands in front of a damaged glass door
Q
A
5s
35s
79s
117s
Q:Summarize the anomaly-related event in the video
A:Individuals on the street are marching and vandalizing the surrounding buildings, while a group of uniformed guards is maintaining order.
Q:Provide a summary of the video
A:The video begins with a burning car, smoke and flames filling the air, creating a chaotic atmosphere.
0s
130s
Q:Summarize the anomaly events in the video
A:The anomaly events depict chaotic and violent scenes, including individuals holding firearms, people throwing objects at a car, a person being restrained by police officers, a burning object, a car engulfed in flames, and a person smashing a glass door.
Q
Anomaly-focused Temporal Sampler
Multi-modal LLM
A
Caption
Judgement Description Analysis
Judgement Description Analysis
Temporal Anomaly Granularity:
Clip-level
Event-level
Video-level

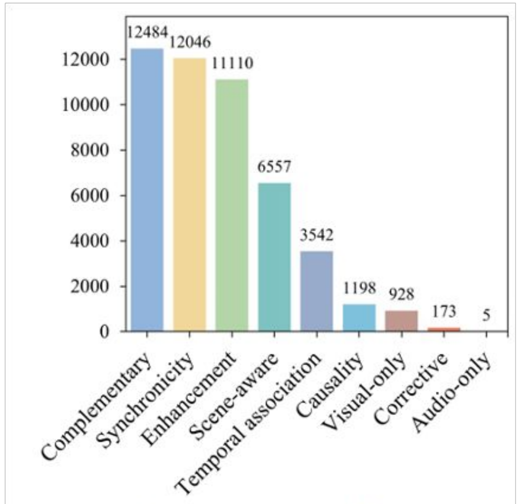


Temporal Granularity
video
event
clip
CAP
Clip Caption
Prompt+Label
LLM
P+L
Event Summary
Video Summary
Manual Checking
{Q:Prompt_C, A:Caption}
{Q:Prompt_J, A:Judgement, Q:Prompt_D, A:Description, Q:Prompt_A, A:Analysis}
{Q:Prompt_J, A:Judgement, Q:Prompt_D, A:Description, Q:Prompt_A, A:Analysis}

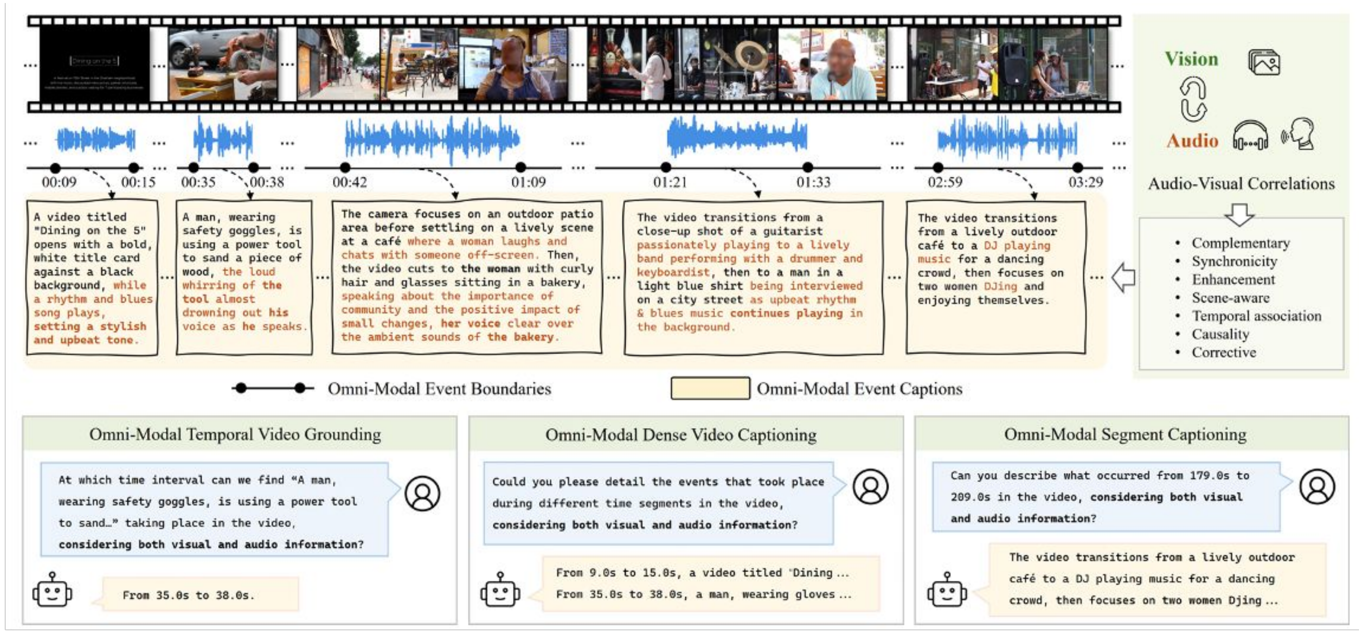
LIMIT.LAB
https://limitlab.xyz/

LongVALE: Vision-Audio-Language-Event Benchmark Towards Time-Aware Omni-Modal Perception of Long Videos

- **Summary:** This paper introduces LongVALE, a benchmark designed to evaluate comprehensive understanding in omni-modal models that integrate vision, audio, and language. Models pre-trained on LongVALE demonstrate strong performance in both omni-modal reasoning and fine-grained reasoning tasks.
- **Novelty:** LongVALE enables systematic evaluation of omni-modal perception and supports assessment across a wide range of reasoning abilities, including fine-grained and cross-modal understanding.
- **Thoughts:** LongVALE was constructed through the combination of existing models and automated processes. Despite recent progress, fine-grained understanding of visual data—such as images and videos—remains a significant challenge for current models.



(d) Audio-visual correlation types


00:09
00:15
00:35
00:38
00:42
01:09
01:21
01:33
02:59
03:29
Vision
Audio
Audio-Visual Correlations
Omni-Modal Event Boundaries
Omni-Modal Event Captions

A video titled "Dining on the 5" opens with a bold, white title card against a black background, while a rhythm and blues song plays, setting a stylish and upbeat tone.

A man, wearing safety goggles, is using a power tool to sand a piece of wood, the loud whirring of the tool almost drowning out his voice as he speaks.

The camera focuses on an outdoor patio area before settling on a lively scene at a café where a woman laughs and chats with someone off-screen. Then, the video cuts to the woman with curly hair and glasses sitting in a bakery, speaking about the importance of community and the positive impact of small changes, her voice clear over the ambient sounds of the bakery.

The video transitions from a close-up shot of a guitarist passionately playing to a lively band performing with a drummer and keyboardist, then to a man in a light blue shirt being interviewed on a city street as upbeat rhythm & blues music continues playing in the background.

The video transitions from a lively outdoor café to a DJ playing music for a dancing crowd, then focuses on two women DJing and enjoying themselves.

- Complementary
- Synchronicity
- Enhancement
- Scene-aware
- Temporal association
- Causality
- Corrective

Omni-Modal Temporal Video Grounding

At which time interval can we find "A man, wearing safety goggles, is using a power tool to sand…" taking place in the video, **considering both visual and audio information?**

From 35.0s to 38.0s.

Omni-Modal Dense Video Captioning

Could you please detail the events that took place during different time segments in the video, **considering both visual and audio information?**

From 9.0s to 15.0s, a video titled 'Dining …
From 35.0s to 38.0s, a man, wearing gloves …

Omni-Modal Segment Captioning

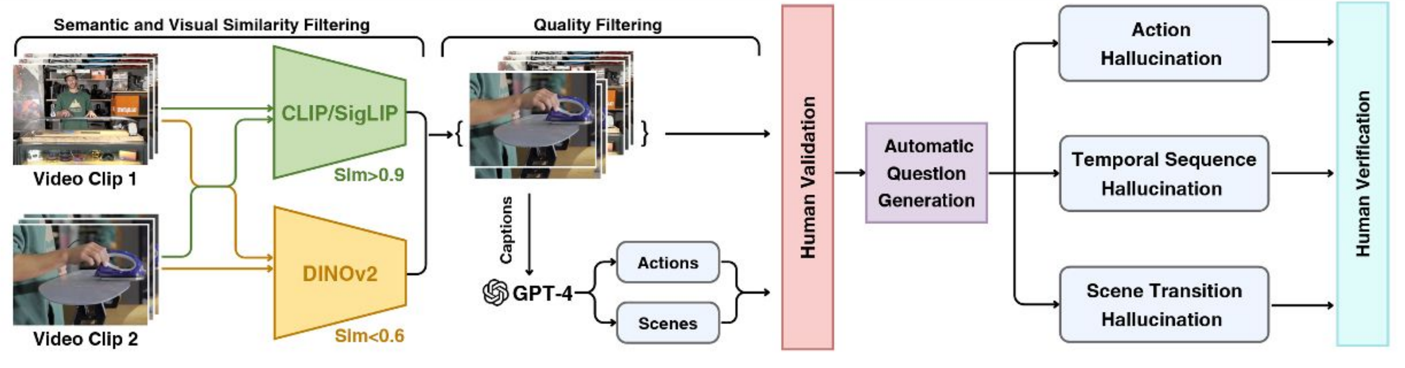
Can you describe what occurred from 179.0s to 209.0s in the video, **considering both visual and audio information?**

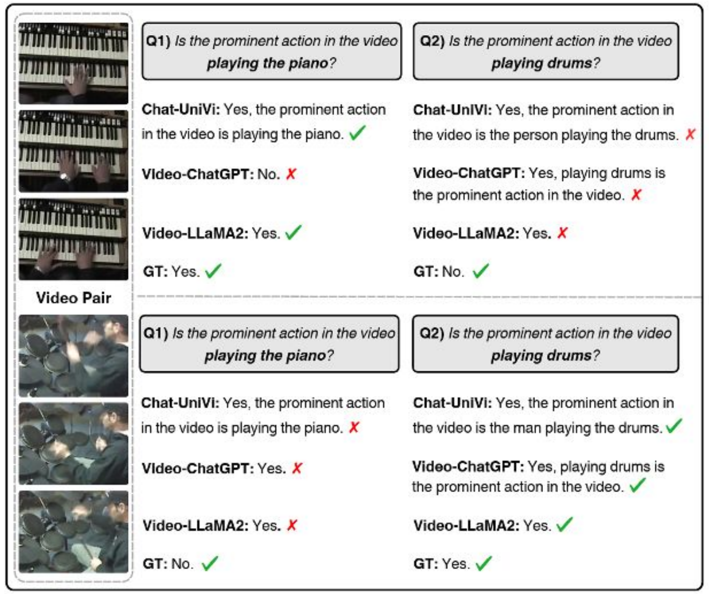
The video transitions from a lively outdoor café to a DJ playing music for a dancing crowd, then focuses on two women Djing …

CVPR25: Meta Insights into Trends and Tendencies (121/179)

VIDHALLUC: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding

- **Summary:** This paper proposes VIDHALLUC, a dataset for evaluating temporal reasoning and temporal hallucinations in video large language models (Video LLMs). The dataset is automatically constructed through the process illustrated in the figure (left), with final verification performed by human annotators.
- **Novelty:** VIDHALLUC enables a comprehensive assessment of how current Video LLMs handle temporal hallucinations, revealing notable performance gaps compared to human reasoning.
- **Thoughts:** Temporal reasoning and fine-grained video understanding remain challenging areas, and VIDHALLUC highlights the limitations of existing models in capturing accurate temporal dynamics.





Q1) *Is the prominent action in the video* ***playing the piano****?*

Chat-UniVi: Yes, the prominent action in the video is playing the piano. ✓

Video-ChatGPT: No. ✗

Video-LLaMA2: Yes. ✓

GT: Yes. ✓

Q2) *Is the prominent action in the video* ***playing drums****?*

Chat-UniVi: Yes, the prominent action in the video is the person playing the drums. ✗

Video-ChatGPT: Yes, playing drums is the prominent action in the video. ✗

Video-LLaMA2: Yes. ✗

GT: No. ✓

Q1) *Is the prominent action in the video* ***playing the piano****?*

Chat-UniVi: Yes, the prominent action in the video is playing the piano. ✗

Video-ChatGPT: Yes. ✗

Video-LLaMA2: Yes. ✗

GT: No. ✓

Q2) *Is the prominent action in the video* ***playing drums****?*

Chat-UniVi: Yes, the prominent action in the video is the man playing the drums. ✓

Video-ChatGPT: Yes, playing drums is the prominent action in the video. ✓

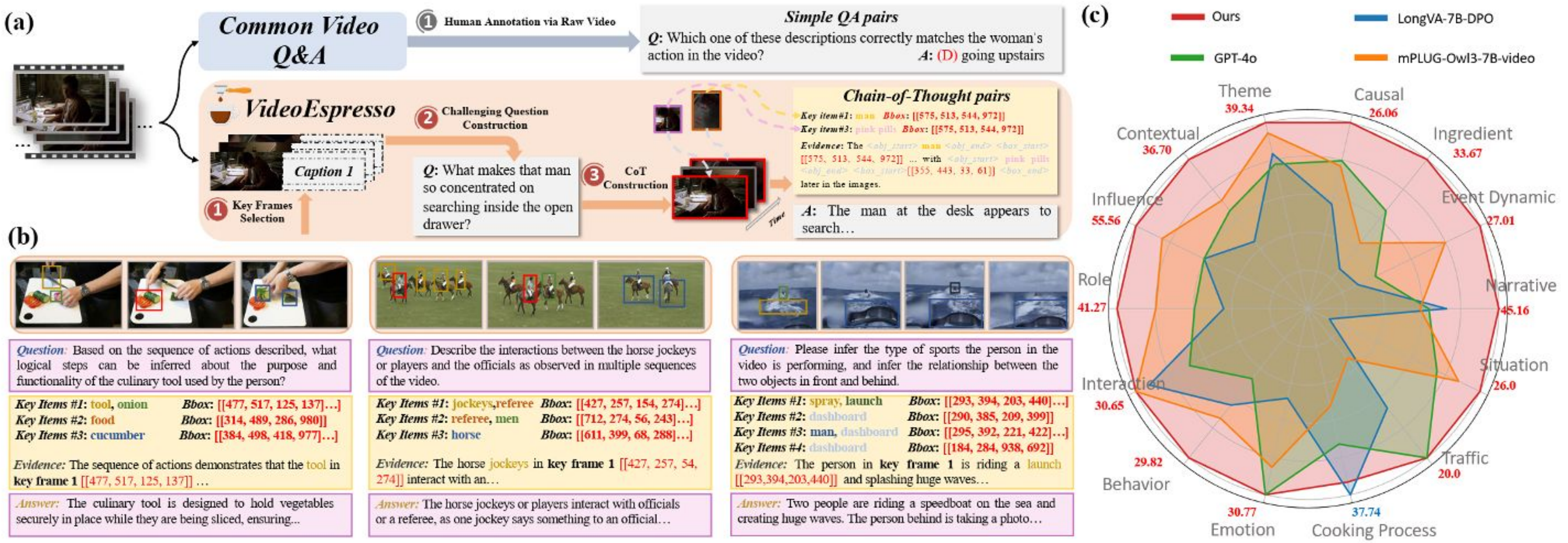
Video-LLaMA2: Yes. ✓

GT: Yes. ✓

CVPR25: Meta Insights into Trends and Tendencies (122/179)

VideoEspresso: A Large-Scale Chain-of-Thought Dataset for Fine-Grained Video Reasoning via Core Frame Selection

- **Summary:** This paper introduces VideoExpresso, a benchmark for evaluating fine-grained video understanding, automatically constructed using Chain-of-Thought reasoning combined with existing multi-modal language models. A model trained on VideoExpresso achieves state-of-the-art performance on several existing datasets.
- **Novelty:** Key contributions include a focus on fine-grained video recognition and the use of Chain-of-Thought reasoning for large-scale data generation—an approach that enhances both interpretability and training quality.
- **Thoughts:** Several benchmarks for fine-grained video recognition were presented at CVPR, reflecting growing interest in this area. Pretraining on fine-grained tasks appears essential to achieving high accuracy, and both fine-grained recognition and Chain-of-Thought reasoning remain open challenges.



Q: Which one of these descriptions correctly matches the woman's action in the video? *A*: (D) going upstairs

Q: What makes that man so concentrated on searching inside the open drawer?

Key item#1: man *Bbox*: [[575, 513, 544, 972]]
Key item#3: pink pills *Bbox*: [[575, 513, 544, 972]]
Evidence: The <obj_start> man <obj_end> <box_start> [[575, 513, 544, 972]] ... with <obj_start> pink pills <obj_end> <box_start>[[355, 443, 33, 61]] <box_end> later in the images.

A: The man at the desk appears to search…

Question: Based on the sequence of actions described, what logical steps can be inferred about the purpose and functionality of the culinary tool used by the person?
Key Items #1: tool, onion *Bbox*: [[477, 517, 125, 137]…]
Key Items #2: food *Bbox*: [[314, 489, 286, 980]]
Key Items #3: cucumber *Bbox*: [[384, 498, 418, 977]…]
Evidence: The sequence of actions demonstrates that the tool in **key frame 1** [[477, 517, 125, 137]] …
Answer: The culinary tool is designed to hold vegetables securely in place while they are being sliced, ensuring...

Question: Describe the interactions between the horse jockeys or players and the officials as observed in multiple sequences of the video.
Key Items #1: jockeys,referee *Bbox*: [[427, 257, 154, 274]…]
Key Items #2: referee, men *Bbox*: [[712, 274, 56, 243]…]
Key Items #3: horse *Bbox*: [[611, 399, 68, 288]…]
Evidence: The horse jockeys in **key frame 1** [[427, 257, 54, 274]] interact with an…
Answer: The horse jockeys or players interact with officials or a referee, as one jockey says something to an official…

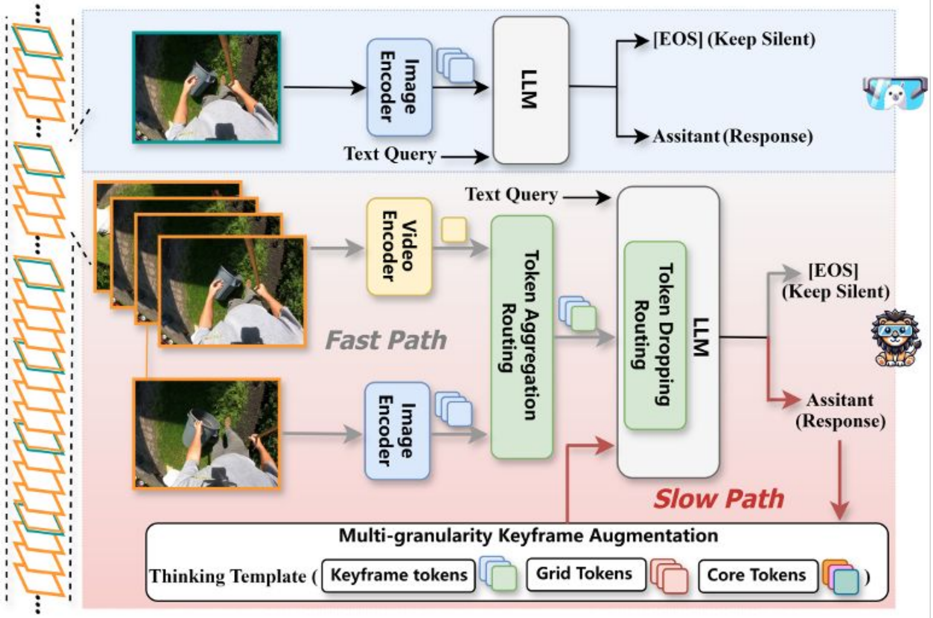
Question: Please infer the type of sports the person in the video is performing, and infer the relationship between the two objects in front and behind.
Key Items #1: spray, launch *Bbox*: [[293, 394, 203, 440]…]
Key Items #2: dashboard *Bbox*: [[290, 385, 209, 399]]
Key Items #3: man, dashboard *Bbox*: [[295, 392, 221, 422]…]
Key Items #4: dashboard *Bbox*: [[184, 284, 938, 692]]
Evidence: The person in **key frame 1** is riding a launch [[293,394,203,440]] and splashing huge waves…
Answer: Two people are riding a speedboat on the sea and creating huge waves. The person behind is taking a photo…

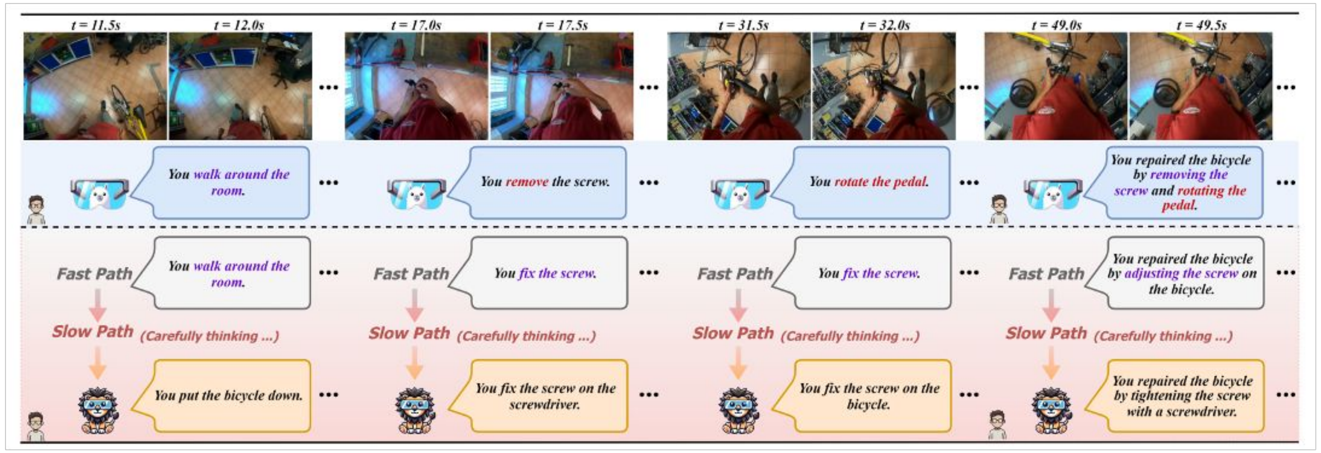
LIMIT.LAB
https://limitlab.xyz/

CVPR25: Meta Insights into Trends and Tendencies (123/179)

LION-FS: Fast & Slow Video-Language Thinker as Online Video Assistant

- **Summary:** This paper proposes LION-FS (Fast-Slow), a new method for online video assistants, particularly in the context of first-person view (FPV) videos. The Fast track provides rapid responses using high frame-rate processing, while the Slow track performs deeper, fine-grained reasoning by hierarchically processing more detailed features.
- **Novelty:** The core novelty lies in adapting the SlowFast architecture—originally used for action recognition—to the domain of online video assistants, enabling both speed and detailed understanding.
- **Thoughts:** While the SlowFast concept itself is not new, its application to online video assistance is practical and promising. The approach could be extended to other video recognition tasks, and explicitly incorporating detection into the Slow track could further enhance performance.



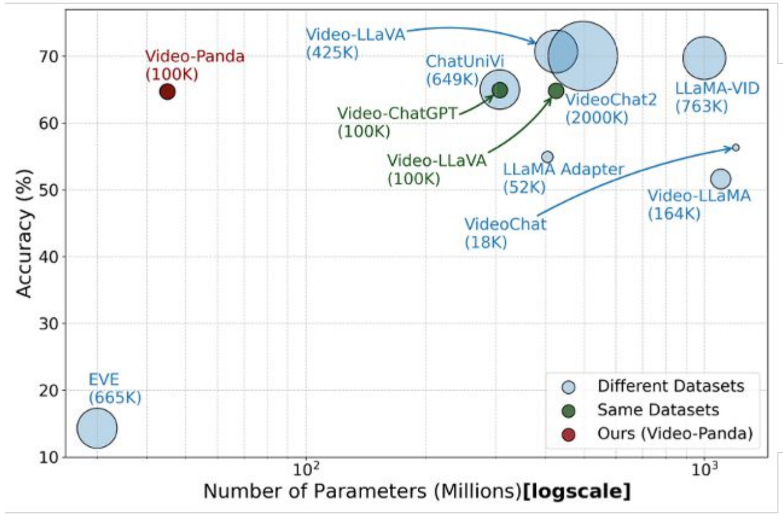


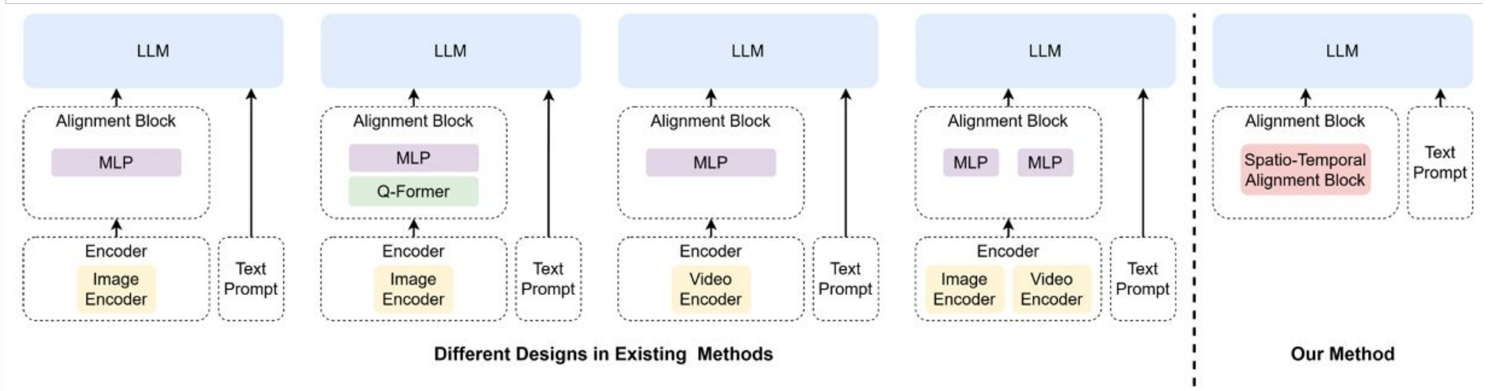
t = 11.5s
t = 12.0s
t = 17.0s
t = 17.5s
t = 31.5s
t = 32.0s
t = 49.0s
t = 49.5s
You walk around the room.
You remove the screw.
You rotate the pedal.
You repaired the bicycle by removing the screw and rotating the pedal.
Fast Path
You walk around the room.
You fix the screw.
You fix the screw.
You repaired the bicycle by adjusting the screw on the bicycle.
Slow Path (Carefully thinking ...)
You put the bicycle down.
You fix the screw on the screwdriver.
You fix the screw on the bicycle.
You repaired the bicycle by tightening the screw with a screwdriver.

LIMIT.LAB
https://limitlab.xyz/

Video-Panda: Parameter-efficient Alignment for Encoder-free Video-Language Models

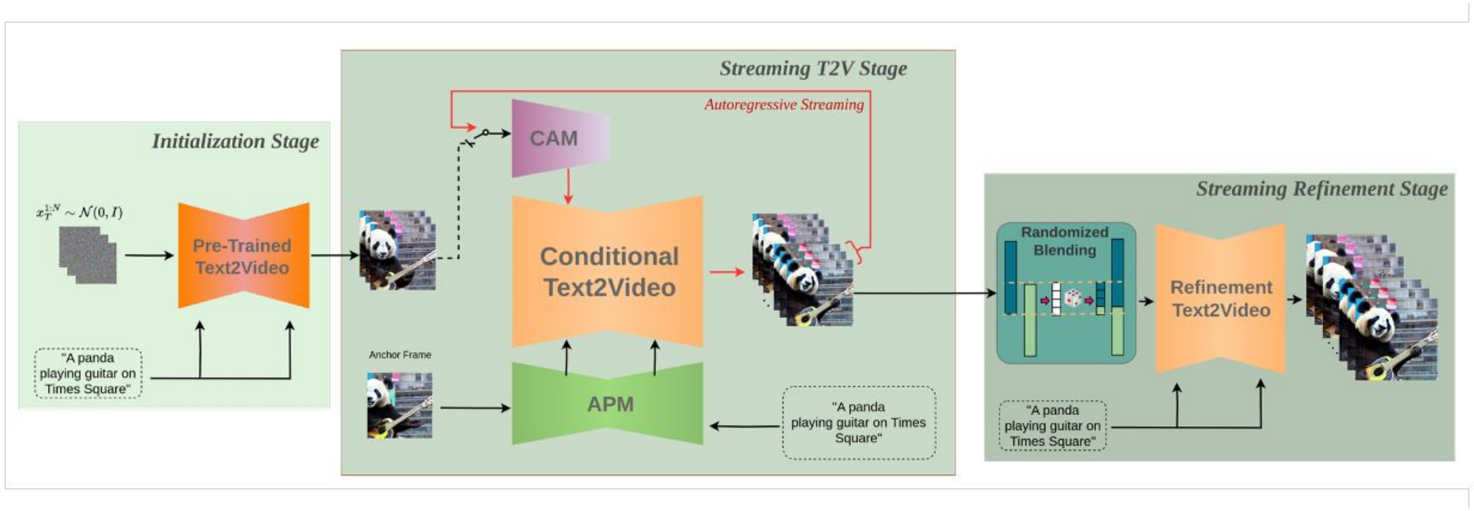
- **Summary:** This paper introduces Video-Panda, an encoder-free video LLM architecture. Instead of using a conventional video encoder, the model employs a Spatio-Temporal Alignment Block (STAB) to extract fine-grained spatial and temporal features directly from video inputs. Video-Panda outperforms models like Video-LLaMA and Video-ChatGPT, while using only 45 million parameters for visual encoding.
- **Novelty:** The key innovation lies in removing the typically large video encoder module, resulting in a lightweight yet high-performing model that maintains strong accuracy with significantly fewer parameters.
- **Thoughts:** Video-Panda is notably compact but still performs competitively across multiple benchmarks. A more comprehensive evaluation showing how it achieves this performance would be insightful, especially compared to similarly sized models. It would also be valuable to explore how the model scales and how its performance evolves with larger configurations. The proposed STAB module appears somewhat complex and warrants further examination.





StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text

- ❏ **Summary:** This paper presents StreamingT2V, a text-to-video generation method structured in three stages, capable of generating videos around 2 minutes long. The model incorporates a Condition Attention Module to align current frames with recent ones and an Appearance Preserve Module to maintain global consistency over long sequences. Finally, long videos are generated in an autoregressive manner using randomized blending. The approach significantly outperforms prior methods such as OpenSora and OpenSoraPlan.
- ❏ **Novelty:** The improvements stem from engineering refinements and training strategies that substantially boost generation quality over existing baselines.
- ❏ **Thoughts:** The method has already gained notable attention. It raises interesting questions about how different camera settings (e.g., movement, FPV) may require distinct design considerations. It also remains to be seen whether a multi-stage approach is truly optimal. Training such a system likely involves significant complexity.

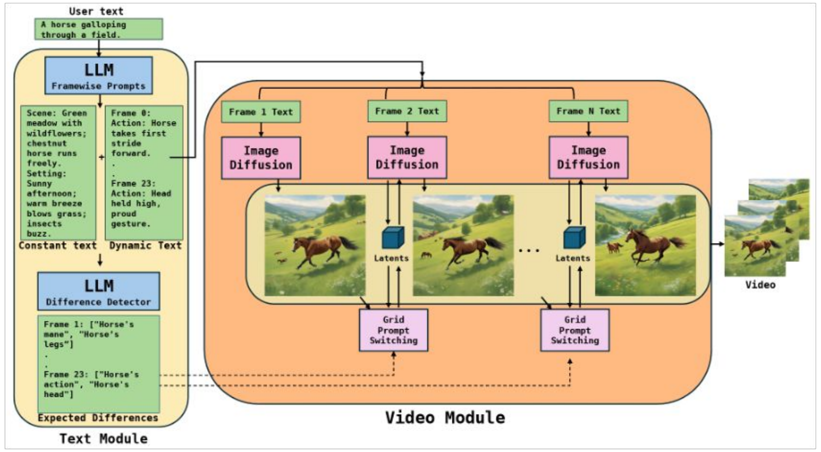

Initialization Stage
$x_T^{1:N} \sim \mathcal{N}(0, I)$
Pre-Trained Text2Video
"A panda playing guitar on Times Square"
Streaming T2V Stage
Autoregressive Streaming
CAM
Conditional Text2Video
Anchor Frame
APM
"A panda playing guitar on Times Square"
Streaming Refinement Stage
Randomized Blending
Refinement Text2Video
"A panda playing guitar on Times Square"

"Fishes swimming in ocean while the camera moving."
"A knight riding on a horse through the countryside."
"A beagle walking down the street."
"Beside the ancient amphitheater of Taormina, a group of friends enjoyed a leisurely picnic."
frame 1-3 @1s
frame 300-302 @30s
frame 600-602 @60s

CVPR25: Meta Insights into Trends and Tendencies (126/179)

EIDT-V: Exploiting Intersections in Diffusion Trajectories for Model-Agnostic, Zero-Shot, Training-Free Text-to-Video Generation

- **Summary:** This paper introduces EIDT-V, a training-free optimization method for improving existing image-based text-to-video generation models. By incorporating frame-wise descriptions and inter-frame motion prompts, EIDT-V adjusts generation in the latent space, enhancing both per-frame quality and temporal coherence. The method achieves strong performance without requiring model retraining.
- **Novelty:** Training-free approaches for text-to-video generation remain relatively underexplored. EIDT-V enables image-level control through textual prompts, offering a creative and flexible mechanism with potential across various applications.
- **Thoughts:** Using distinct prompts for each frame is a compelling idea and may also support temporal alignment. It would be interesting to explore whether this approach could be extended to video-native generation models beyond image-based frameworks.

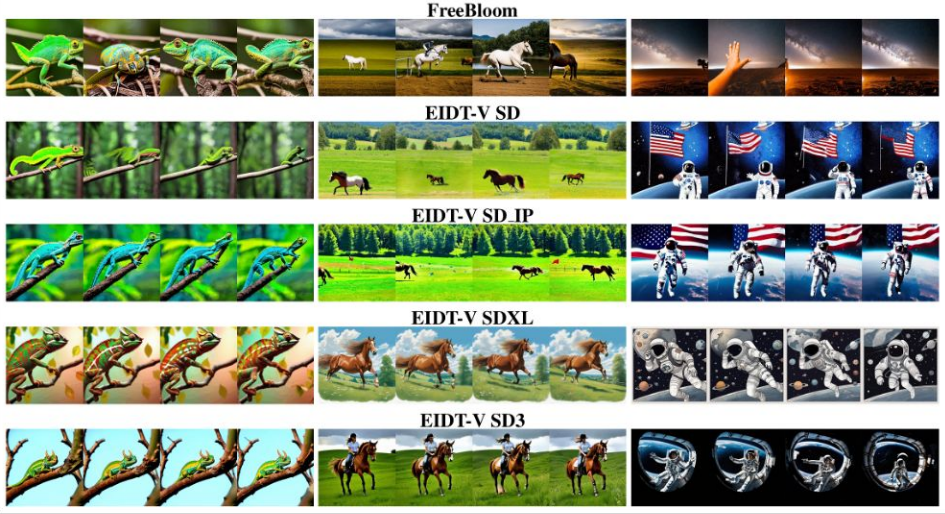


A horse galloping through a field.

Scene: Green meadow with wildflowers; chestnut horse runs freely. Setting: Sunny afternoon; warm breeze blows grass; insects buzz.

Frame 0: Action: Horse takes first stride forward.
.
.
Frame 23: Action: Head held high, proud gesture.

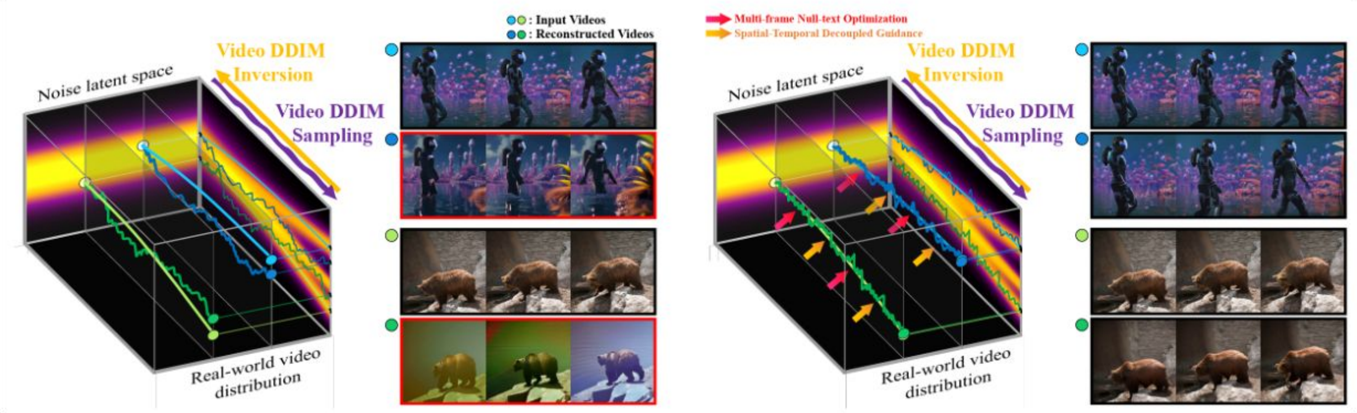
Frame 1: ["Horse's mane", "Horse's legs"]
.
.
Frame 23: ["Horse's action", "Horse's head"]

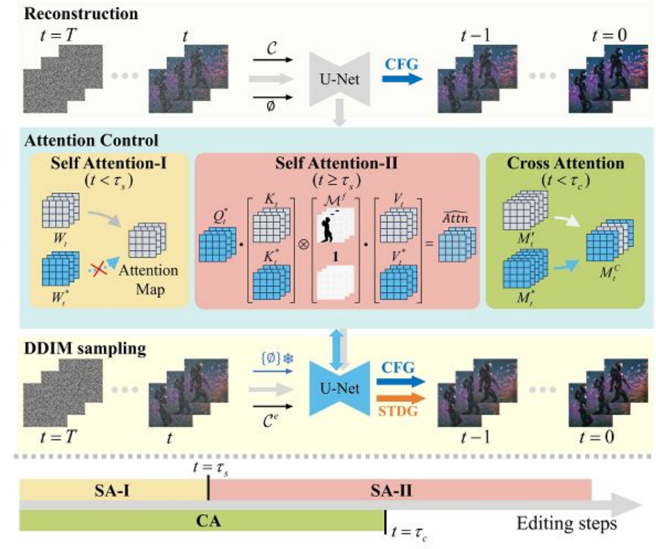


CVPR25: Meta Insights into Trends and Tendencies (127/179)

VideoDirector: Precise Video Editing via Text-to-Video Models

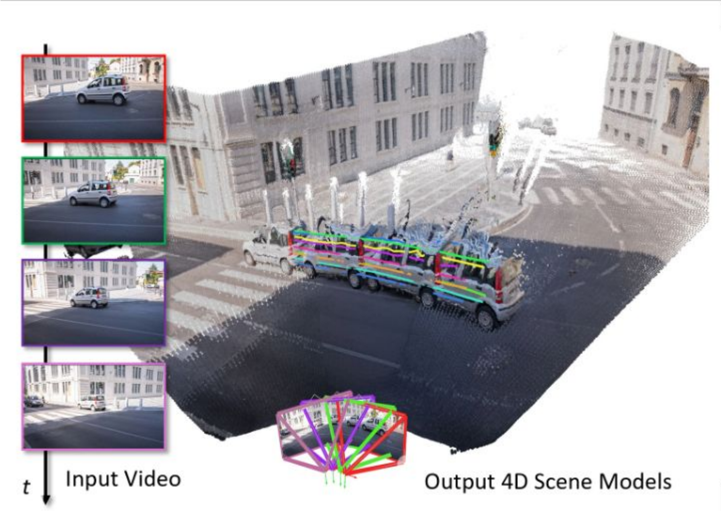
- **Summary:** This paper presents VideoDirector, a method for direct editing of text-to-video (T2V) generation models. The approach introduces separate guidance for spatial and temporal features, making it easier to edit each aspect independently. Additionally, it proposes multi-frame null-text optimization, enabling fine-grained temporal editing. The method achieves state-of-the-art performance.
- **Novelty:** The paper offers a thorough analysis of limitations in existing inversion-then-editing approaches to video editing, such as tightly coupled spatial-temporal features and complex layout entanglement. VideoDirector addresses these issues by explicitly decoupling spatial and temporal editing mechanisms.
- **Thoughts:** The challenge of spatial-temporal coupling in traditional T2V generation is well known. This work provides a clear and structured approach to mitigating it, offering a promising direction for more controllable and precise video editing.

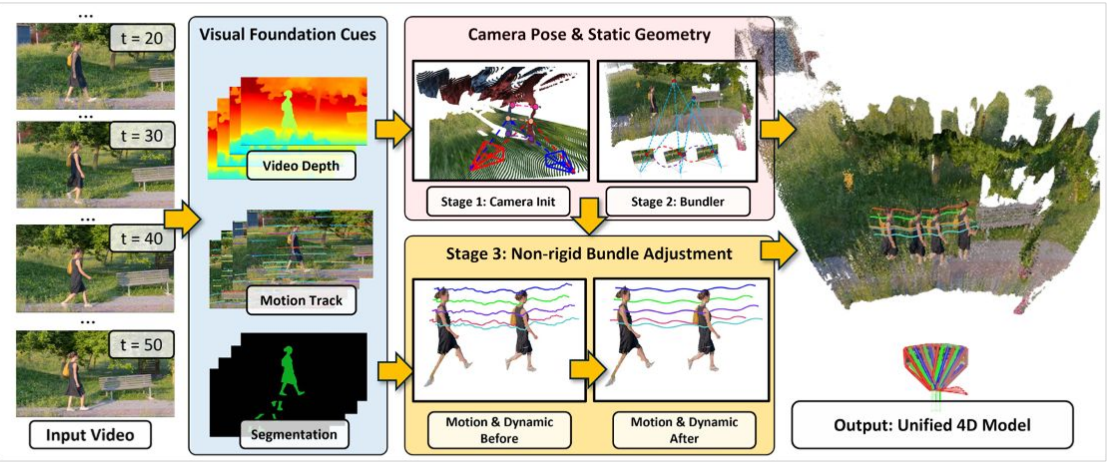




Uni4D: Unifying Visual Foundation Models for 4D Modeling from a Single Video

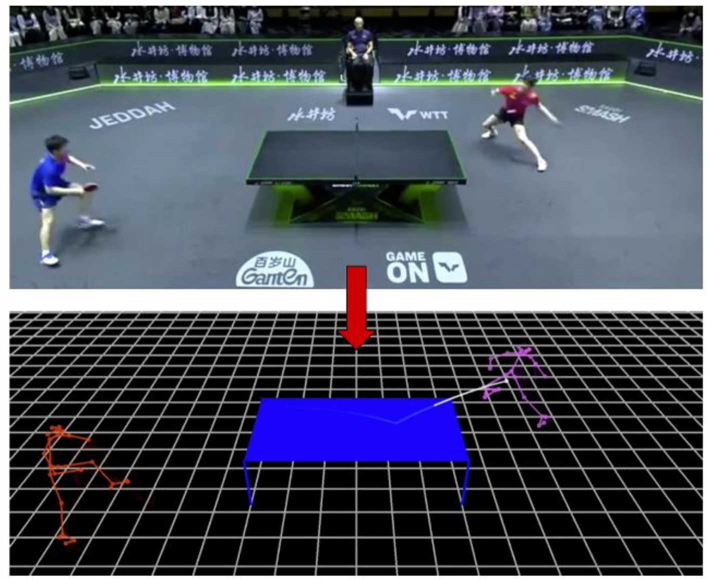
- **Summary:** This paper introduces Uni4D, a method for generating 4D scenes (3D geometry over time) directly from video. Uni4D leverages a visual foundation model to extract depth, motion tracking, and segmentation from video frames. Using these outputs, the system estimates camera poses and scene geometry, followed by bundle adjustment to refine results. Notably, the visual foundation model is used without retraining, yet Uni4D achieves top performance across multiple benchmarks.
- **Novelty:** The key innovation lies in applying a pre-trained visual foundation model to the video-to-4D task, enabling high-quality results without additional training or supervision.
- **Thoughts:** Though built from existing components, the method effectively combines them to deliver state-of-the-art performance. The integration of foundation models into dynamic scene reconstruction is promising and practical.

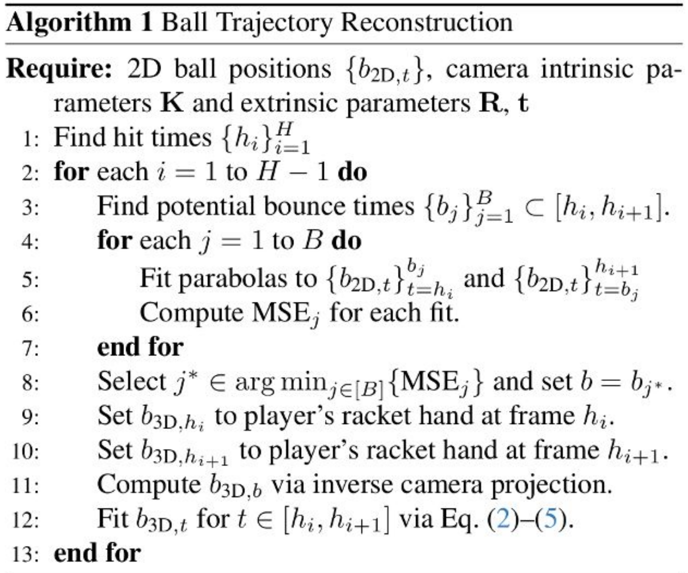




LATTE-MV: Learning to Anticipate Table Tennis Hits from Monocular Videos

- **Summary:** This work makes two key contributions toward developing a table tennis agent. First, it proposes a pipeline to reconstruct the 3D scene from video, including mask extraction for the table and ball, estimation of the table's four corners, camera calibration, pose estimation, and final ball trajectory generation. Second, it introduces an uncertainty-aware controller based on generative modeling to predict the opponent's future actions. A dataset of approximately 50 hours of professional table tennis matches was also collected from online sources.
- **Novelty:** The inclusion of anticipation modeling—predicting opponent behavior under uncertainty—distinguishes this work from prior approaches.
- **Thoughts:** Incorporating physical dynamics would further enhance realism. Achieving full 4D scene reconstruction (geometry + motion over time) would be a promising next step.





Algorithm 1 Ball Trajectory Reconstruction

```
Require: 2D ball positions $\{b_{2D,t}\}$, camera intrinsic parameters $\mathbf{K}$ and extrinsic parameters $\mathbf{R}$, $\mathbf{t}$
Find hit times $\{h_i\}_{i=1}^{H}$
for each $i = 1$ to $H - 1$ do
    Find potential bounce times $\{b_j\}_{j=1}^{B} \subset [h_i, h_{i+1}]$.
    for each $j = 1$ to $B$ do
        Fit parabolas to $\{b_{2D,t}\}_{t=h_i}^{b_j}$ and $\{b_{2D,t}\}_{t=b_j}^{h_{i+1}}$
        Compute MSE$_j$ for each fit.
    end for
    Select $j^* \in \arg\min_{j\in[B]}\{\text{MSE}_j\}$ and set $b = b_{j^*}$.
    Set $b_{3D,h_i}$ to player's racket hand at frame $h_i$.
    Set $b_{3D,h_{i+1}}$ to player's racket hand at frame $h_{i+1}$.
    Compute $b_{3D,b}$ via inverse camera projection.
    Fit $b_{3D,t}$ for $t \in [h_i, h_{i+1}]$ via Eq. (2)–(5).
end for
```

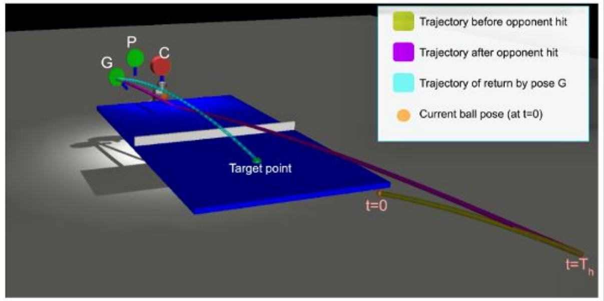
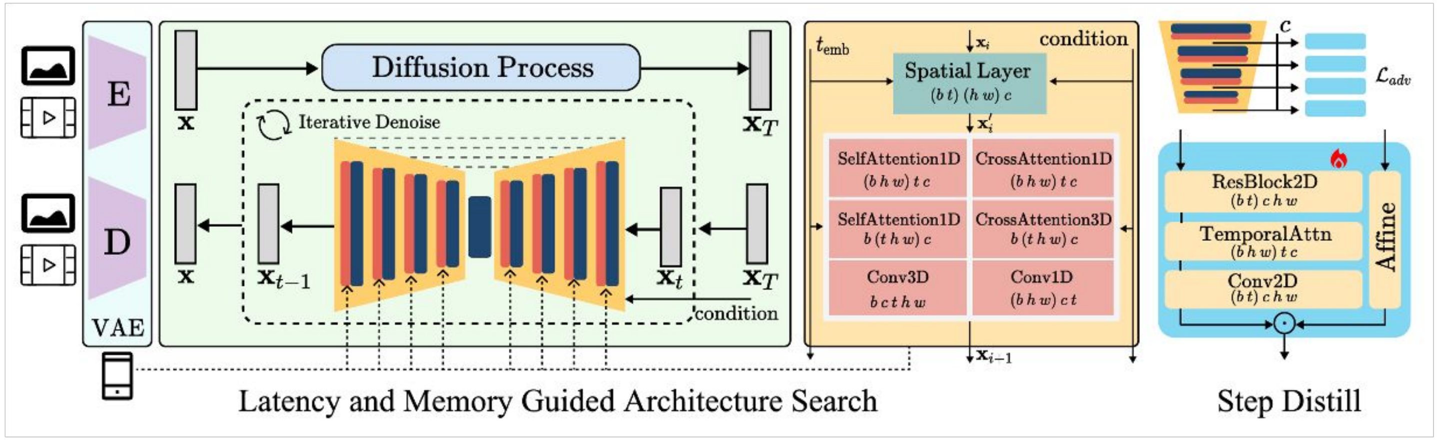

G
P
C
Target point
t=0
t=T_h
Trajectory before opponent hit
Trajectory after opponent hit
Trajectory of return by pose G
Current ball pose (at t=0)

Figure 7. Simulated target poses and ball trajectory.

CVPR25: Meta Insights into Trends and Tendencies (130/179)

SnapGen-V: Generating a Five-Second Video within Five Seconds on a Mobile

- ❑ **Summary:** This paper proposes SnapGen-V, a lightweight yet high-performing text-to-video generation model capable of generating 5-second videos within 5 seconds on a mobile device (iPhone 16 Pro Max). To achieve this efficiency, the authors conducted extensive experiments to identify a compact yet effective temporal backbone, and designed efficient temporal layers. Additionally, a new adversarial fine-tuning strategy for diffusion models is introduced, along with distillation techniques that reduce the required denoising steps to just four.
- ❑ **Novelty:** The work stands out for its systematic exploration of lightweight architectures and presents the first text-to-video generation model that runs effectively on mobile devices.
- ❑ **Thoughts:** Model efficiency is an increasingly popular research direction. This work opens the door for real-time, on-device video generation, and may pave the way for a wave of mobile applications in the near future.

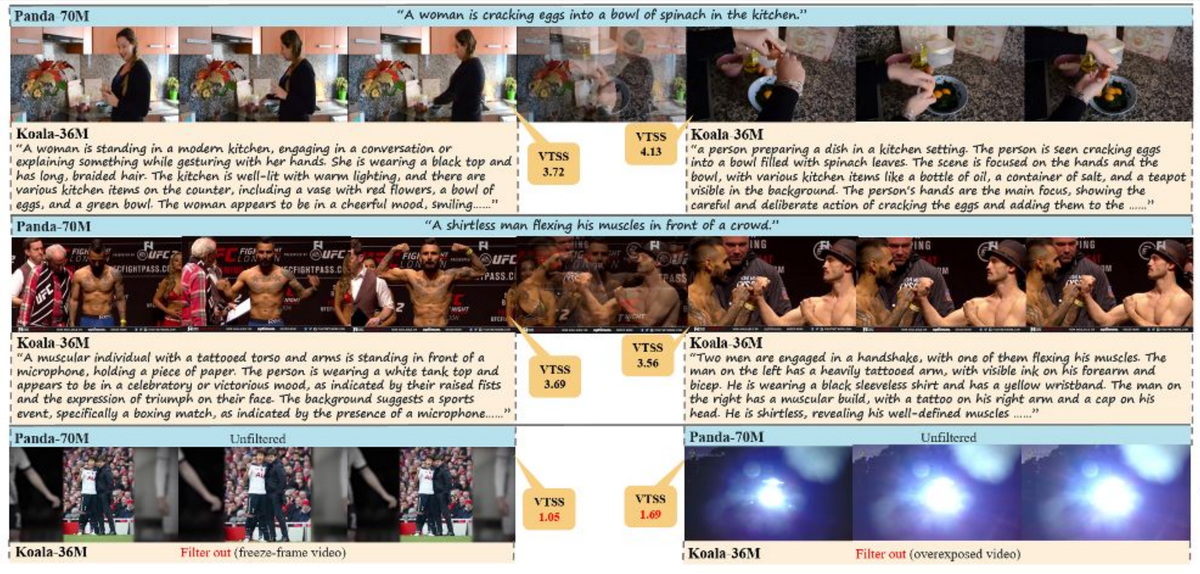


"A Labrador Retriever astronaut captain in a spacesuit"
"An erupting volcano"
"Pirate ship trapped in a cosmic maelstrom nebula"
"Beautiful lady, big smile, blue eyes, short ginger hair"
"A space rocket with trails of smoke behind it launching"
"Sunset from Mystic Beach"

CVPR25: Meta Insights into Trends and Tendencies (131/179)

Koala-36M: A Large-scale Video Dataset Improving Consistency between Fine-grained Conditions and Video Content

- **Summary:** This paper introduces Koala-36M, a large-scale dataset designed for both video recognition and generation. Each video clip (around 10 sec) is annotated with detailed textual descriptions averaging around 200 words, generated automatically by models. The authors also propose a Video Training Suitability Score (VTSS) to assess and enhance video quality. Furthermore, diffusion models are employed to improve temporal consistency between the video and its corresponding text.
- **Novelty:** Koala-36M provides more detailed and higher-quality captions compared to similar datasets such as Panda70M. The use of VTSS and diffusion models for quality and consistency further enhances its utility.
- **Thoughts:** High-quality, fine-grained datasets for video and image understanding are becoming increasingly essential. A key question is how to ensure the quality of automatically generated text. Still, generating large volumes of reasonably good captions with current models seems to be an effective strategy for training.



Panda 70M "A woman is cracking eggs into a bowl of spinach in the kitchen."

Koala-36M "A woman is standing in a modern kitchen, engaging in a conversation or explaining something while gesturing with her hands. She is wearing a black top and has long, braided hair. The kitchen is well-lit with warm lighting, and there are various kitchen items on the counter, including a vase with red flowers, a bowl of eggs, and a green bowl. The woman appears to be in a cheerful mood, smiling......"

Koala-36M "a person preparing a dish in a kitchen setting. The person is seen cracking eggs into a bowl filled with spinach leaves. The scene is focused on the hands and the bowl, with various kitchen items like a bottle of oil, a container of salt, and a teapot visible in the background. The person's hands are the main focus, showing the careful and deliberate action of cracking the eggs and adding them to the ......"

Panda-70M "A shirtless man flexing his muscles in front of a crowd."

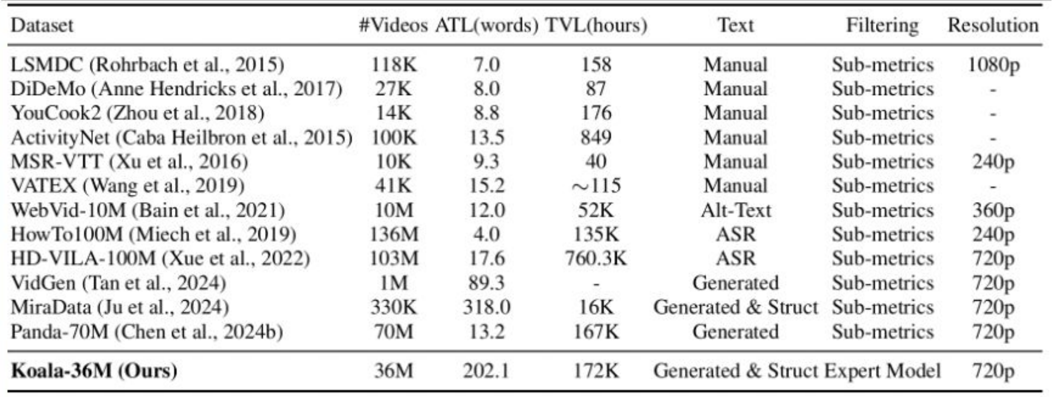
Koala-36M "A muscular individual with a tattooed torso and arms is standing in front of a microphone, holding a piece of paper. The person is wearing a white tank top and appears to be in a celebratory or victorious mood, as indicated by their raised fists and the expression of triumph on their face. The background suggests a sports event, specifically a boxing match, as indicated by the presence of a microphone......"

Koala-36M "Two men are engaged in a handshake, with one of them flexing his muscles. The man on the left has a heavily tattooed arm, with visible ink on his forearm and bicep. He is wearing a black sleeveless shirt and has a yellow wristband. The man on the right has a muscular build, with a tattoo on his right arm and a cap on his head. He is shirtless, revealing his well-defined muscles ......"

Panda-70M Unfiltered — Koala-36M Filter out (freeze-frame video)

Panda-70M Unfiltered — Koala-36M Filter out (overexposed video)

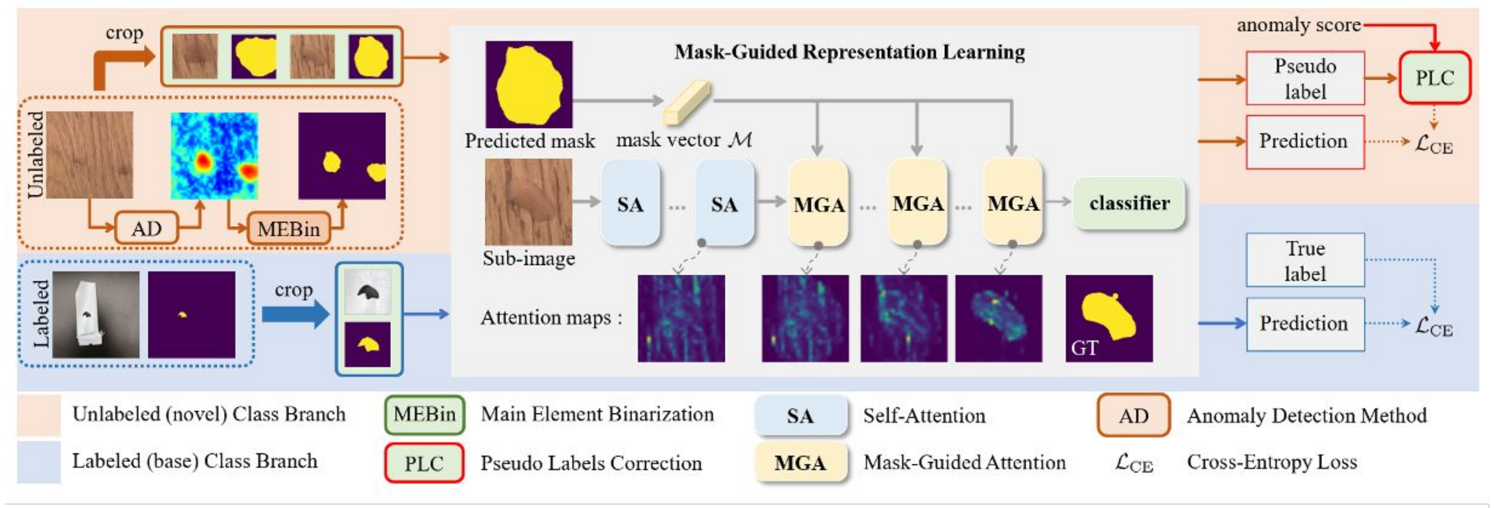
VTSS 3.72
VTSS 4.13
VTSS 3.69
VTSS 3.56
VTSS 1.05
VTSS 1.69

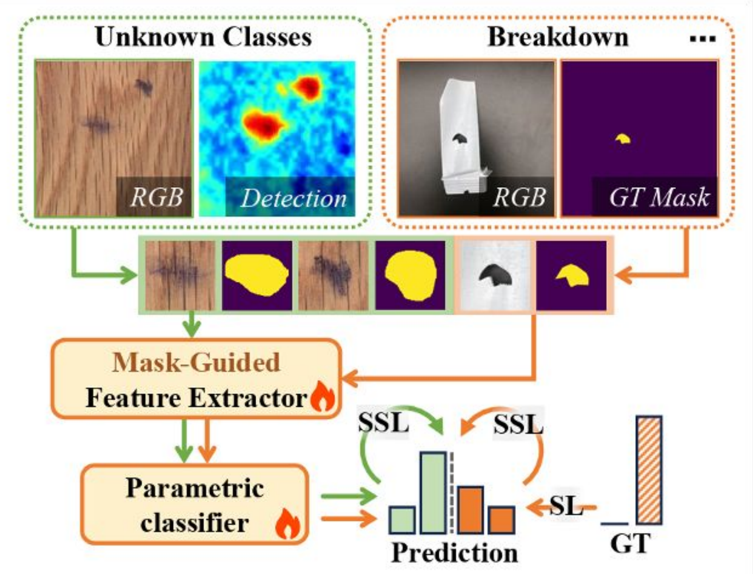


Dataset	#Videos	ATL(words)	TVL(hours)	Text	Filtering	Resolution
LSMDC (Rohrbach et al., 2015)	118K	7.0	158	Manual	Sub-metrics	1080p
DiDeMo (Anne Hendricks et al., 2017)	27K	8.0	87	Manual	Sub-metrics	-
YouCook2 (Zhou et al., 2018)	14K	8.8	176	Manual	Sub-metrics	-
ActivityNet (Caba Heilbron et al., 2015)	100K	13.5	849	Manual	Sub-metrics	-
MSR-VTT (Xu et al., 2016)	10K	9.3	40	Manual	Sub-metrics	240p
VATEX (Wang et al., 2019)	41K	15.2	∼115	Manual	Sub-metrics	-
WebVid-10M (Bain et al., 2021)	10M	12.0	52K	Alt-Text	Sub-metrics	360p
HowTo100M (Miech et al., 2019)	136M	4.0	135K	ASR	Sub-metrics	240p
HD-VILA-100M (Xue et al., 2022)	103M	17.6	760.3K	ASR	Sub-metrics	720p
VidGen (Tan et al., 2024)	1M	89.3	-	Generated	Sub-metrics	720p
MiraData (Ju et al., 2024)	330K	318.0	16K	Generated & Struct	Sub-metrics	720p
Panda-70M (Chen et al., 2024b)	70M	13.2	167K	Generated	Sub-metrics	720p
Koala-36M (Ours)	36M	202.1	172K	Generated & Struct	Expert Model	720p

AnomalyNCD: Towards Novel Anomaly Class Discovery in Industrial Scenarios

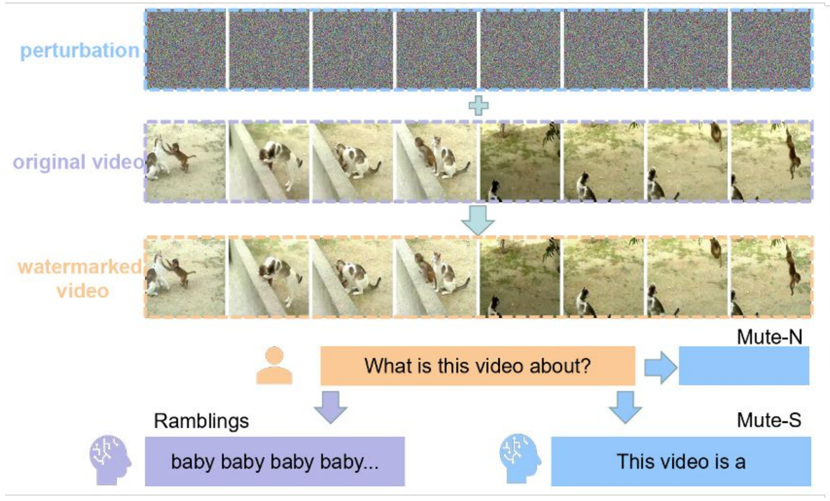
- **Summary:** This paper proposes AnomalyNCD, a new method for multiclass anomaly detection. Since anomaly regions typically occupy only a small portion of the image, the authors introduce a main element binarization (MEBin) module that emphasizes detecting central regions of anomalies. The method also incorporates mask-guided feature extraction to enrich semantic information through region segmentation. The approach achieves strong performance across several benchmarks.
- **Novelty:** The innovation mainly lies in the methodological design—while the individual components are not entirely novel, their integration leads to notably high accuracy.
- **Thoughts:** Applying multi-modal language models (MLLMs) to anomaly detection could be a promising direction, potentially enabling not just detection but also interpretation of anomaly types. There's also potential to unify anomaly detection and object counting within a single framework.

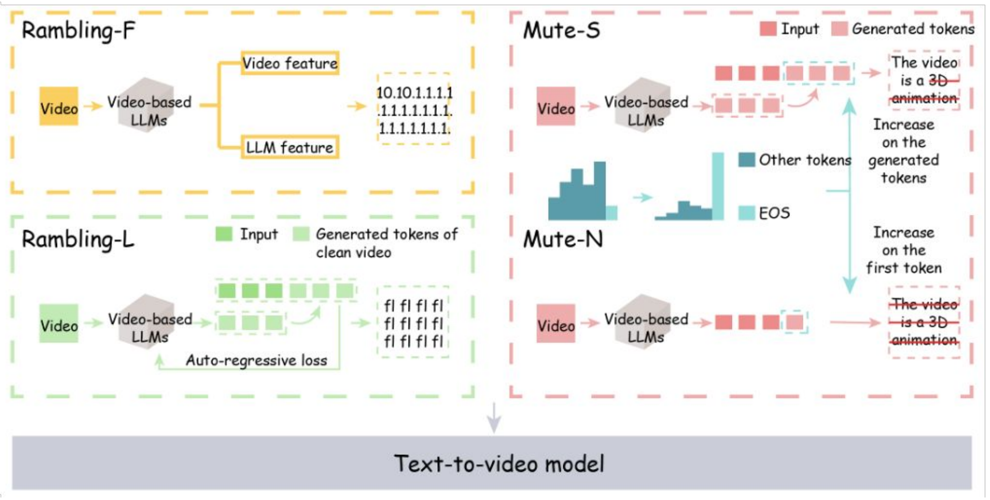

crop
Unlabeled
AD
MEBin
Labeled
crop
Mask-Guided Representation Learning
Predicted mask
mask vector $\mathcal{M}$
Sub-image
SA
SA
MGA
MGA
MGA
classifier
Attention maps :
GT
anomaly score
Pseudo label
PLC
Prediction
$\mathcal{L}_{CE}$
True label
Prediction
$\mathcal{L}_{CE}$
Unlabeled (novel) Class Branch
Labeled (base) Class Branch
MEBin Main Element Binarization
PLC Pseudo Labels Correction
SA Self-Attention
MGA Mask-Guided Attention
AD Anomaly Detection Method
$\mathcal{L}_{CE}$ Cross-Entropy Loss



Protecting Your Video Content: Disrupting Automated Video-based LLM Annotations

- ❏ **Summary:** This paper introduces two watermarking-style methods—Ramblings and Mute—to address the issue of video LLMs misinterpreting or over-accessing video content, thus aiming to protect data privacy. The Ramblings method steers Video LLMs toward incorrect responses by embedding misleading signals, while Mute focuses on reducing output quality by targeting the model's end-of-sequence (EOS) behavior. Both methods demonstrate strong performance against models such as Video-ChatGPT, Video-LLaMA, and Video-Vicuna.
- ❏ **Novelty:** The work addresses a new and important research problem: protecting data privacy in video content from video LLMs. The proposed methods introduce adversarial-style defenses specific to multimodal large language models.
- ❏ **Thoughts:** As video recognition capabilities of LLMs rapidly improve, the demand for privacy-preserving and adversarial techniques targeting MLLMs is likely to grow. This work highlights a critical emerging direction in responsible AI development.

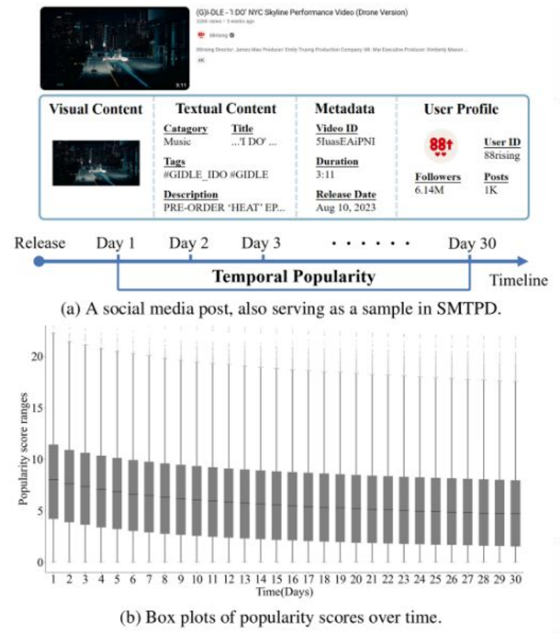




CVPR25: Meta Insights into Trends and Tendencies (134/179)

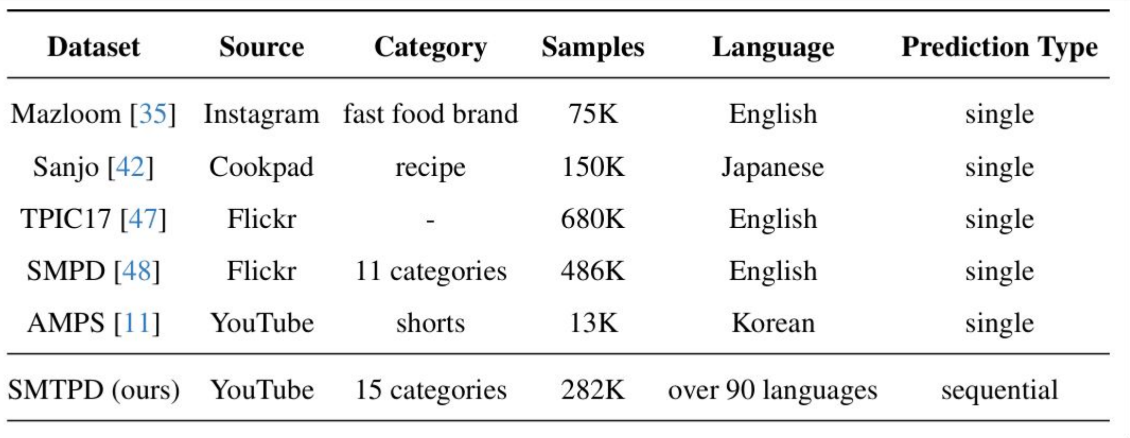
SMTPD:ANewBenchmarkfor Temporal Prediction of Social Media Popularity

- ❑ **Summary:** This paper proposes SMTPD, a new benchmark for evaluating social media video popularity, with a particular focus on the temporal dynamics of popularity. Compared to existing datasets, SMTPD captures how video popularity evolves over time.
- ❑ **Novelty:** The task itself is novel and socially relevant—social media popularity has significant implications for public influence and economic value, making it an important area for machine learning research.
- ❑ **Thoughts:** The topic of social popularity is compelling, with potential for exploring factors behind popularity, transferring popularity to other content, and more. However, since popularity is not static and can change due to real-world user interactions, it's unclear whether the task can be fully isolated as a learning problem. This work also reflects a broader trend of applying MLLMs to complex, real-world systems like social media.



(a) A social media post, also serving as a sample in SMTPD.

(b) Box plots of popularity scores over time.

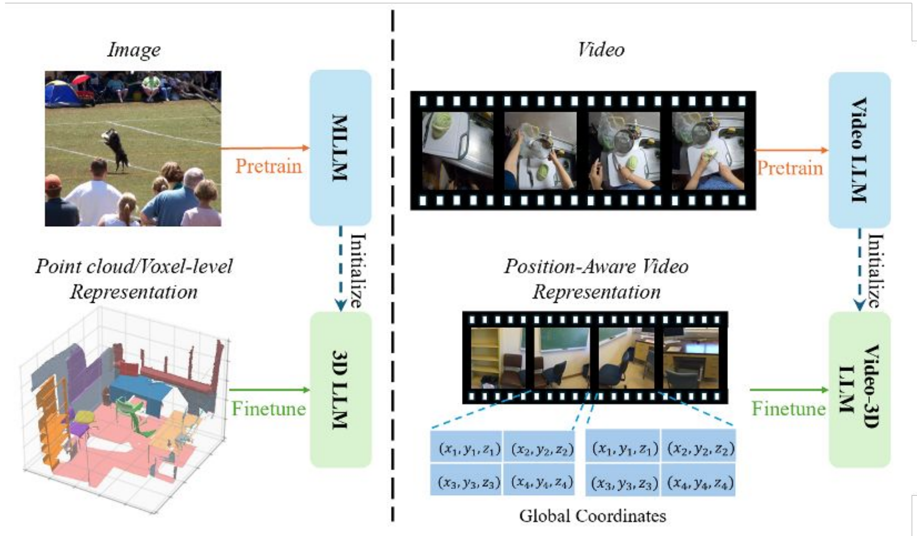


Dataset	Source	Category	Samples	Language	Prediction Type
Mazloom [35]	Instagram	fast food brand	75K	English	single
Sanjo [42]	Cookpad	recipe	150K	Japanese	single
TPIC17 [47]	Flickr	-	680K	English	single
SMPD [48]	Flickr	11 categories	486K	English	single
AMPS [11]	YouTube	shorts	13K	Korean	single
SMTPD (ours)	YouTube	15 categories	282K	over 90 languages	sequential

LIMIT.LAB
https://limitlab.xyz/

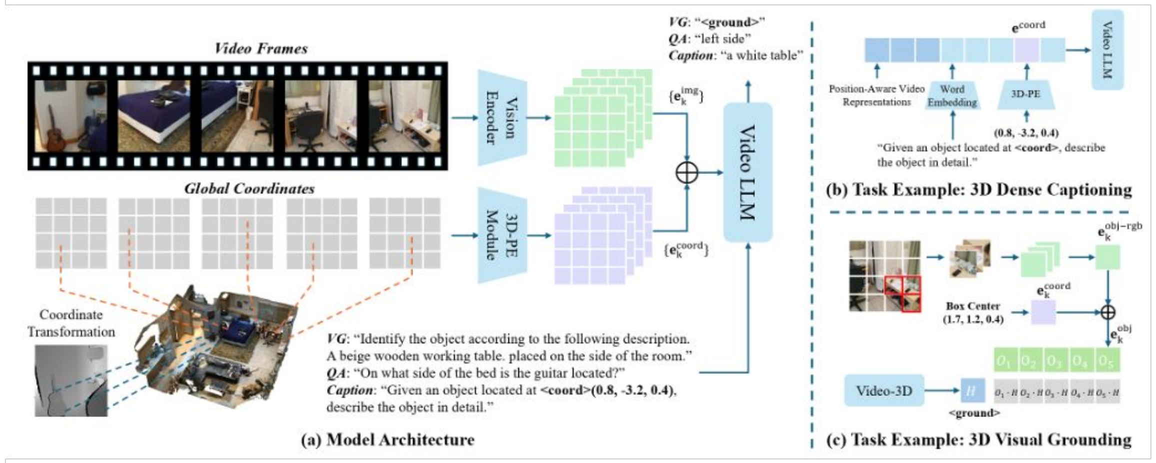
Video-3D LLM: Learning Position-Aware Video Representation for 3D Scene Understanding

- **Summary:** This paper proposes Video-3D LLM, a new multimodal large language model capable of 3D recognition using video. The method treats 3D understanding as a form of dynamic video analysis, aligning video representations with 3D representations. The model achieves strong performance across several 3D recognition benchmarks.
- **Novelty:** The key innovation lies in enabling 3D understanding from video, allowing an MLLM to perceive 3D structure without requiring explicit 3D input, by bridging video and 3D modalities.
- **Thoughts:** Recognizing 3D information from video alone is highly practical. This approach makes 3D perception more accessible in real-world scenarios, where dedicated 3D input is often unavailable. It enhances usability and opens up new application possibilities for MLLMs in spatial understanding.



(a) Previous Work

(b) Our Method


Video Frames
Global Coordinates
Coordinate Transformation
Vision Encoder
3D-PE Module
$\{e_k^{img}\}$
$\{e_k^{coord}\}$
Video LLM
VG: "<ground>"
QA: "left side"
Caption: "a white table"
VG: "Identify the object according to the following description. A beige wooden working table. placed on the side of the room."
QA: "On what side of the bed is the guitar located?"
Caption: "Given an object located at <coord>(0.8, -3.2, 0.4), describe the object in detail."
e^{coord}
Video LLM
Position-Aware Video Representations
Word Embedding
3D-PE
(0.8, -3.2, 0.4)
"Given an object located at <coord>, describe the object in detail."
$e_k^{obj-rgb}$
e_k^{coord}
Box Center (1.7, 1.2, 0.4)
e_k^{obj}
O_1 O_2 O_3 O_4 O_5
Video-3D
H
$O_1 \cdot H$ $O_2 \cdot H$ $O_3 \cdot H$ $O_4 \cdot H$ $O_5 \cdot H$
<ground>

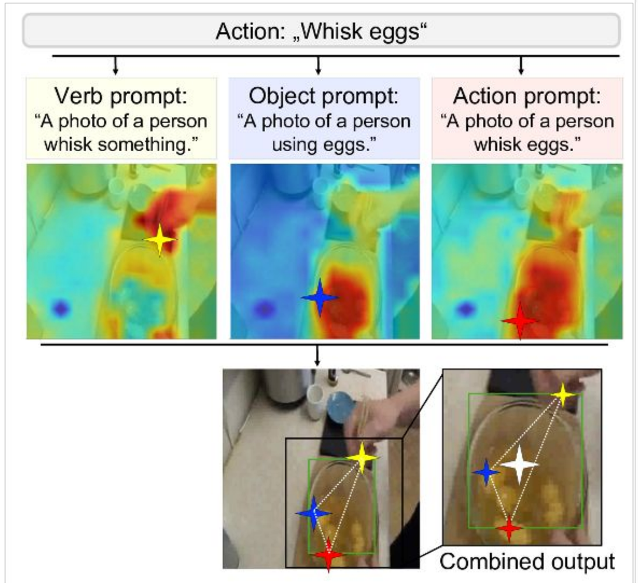
(a) Model Architecture

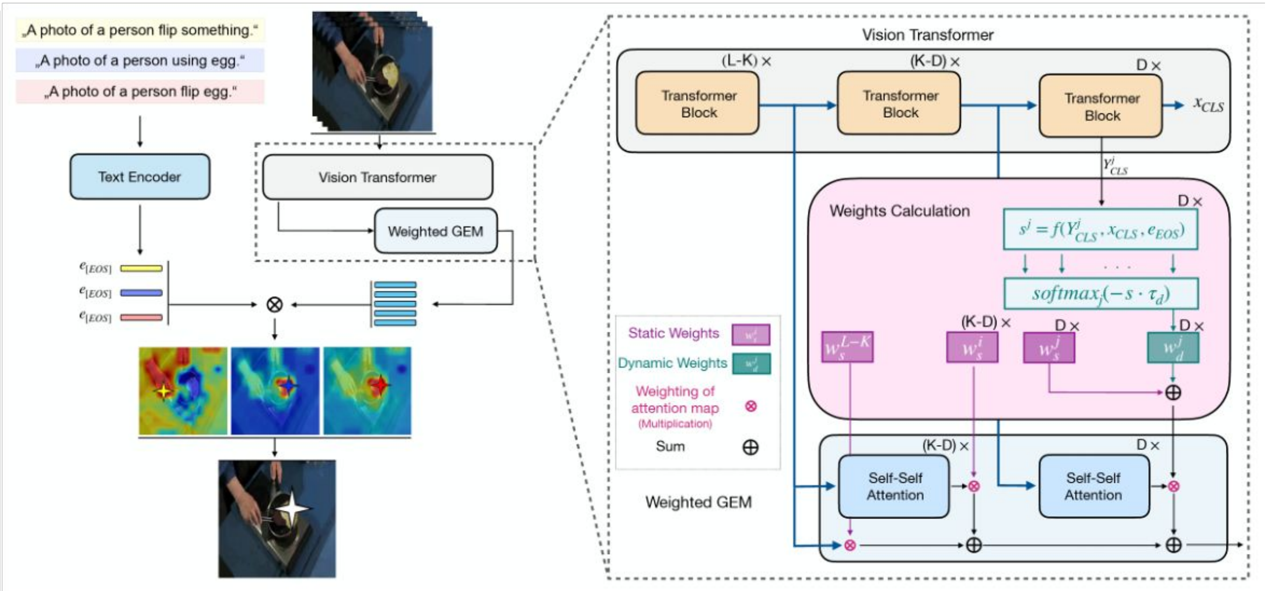
(b) Task Example: 3D Dense Captioning

(c) Task Example: 3D Visual Grounding

VideoGEM: Training-free Action Grounding in Videos

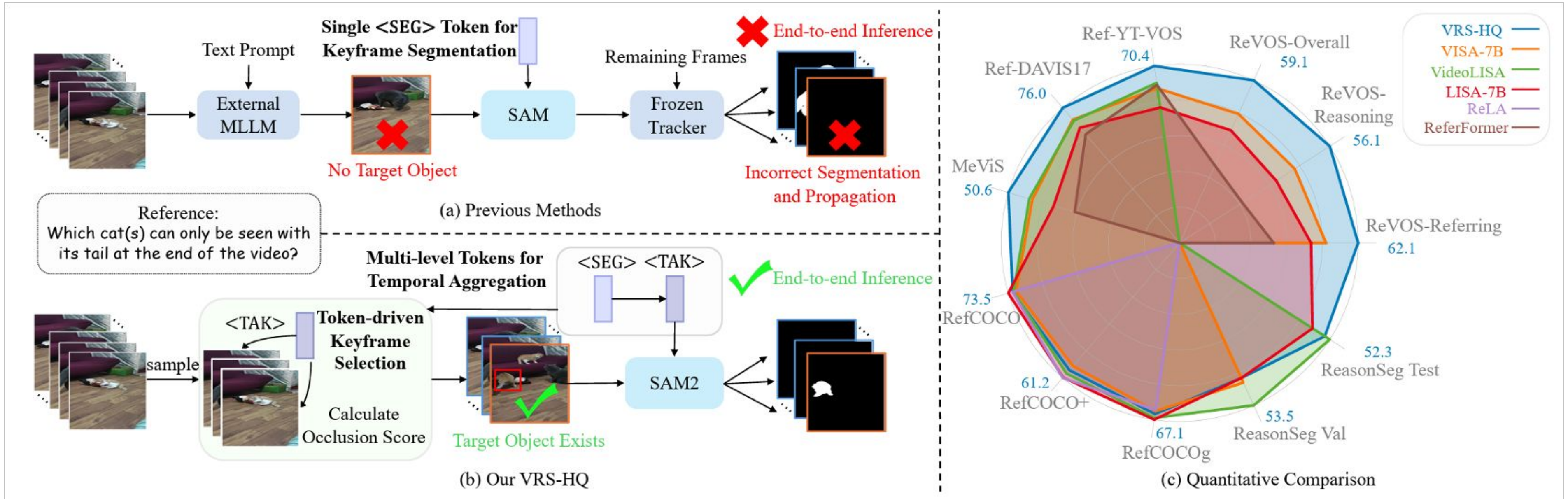
- **Summary:** This paper introduces VideoGEM, the first training-free video grounding method that localizes actions and objects in video without additional model training. It extends the image-based grounding method GEM to video. Based on the observation that different attention layers correspond to actions and objects, the method dynamically adjusts layer weights. It also generates prompts for actions, verbs, and objects from action labels, using these prompts to guide attention and extract corresponding attention maps.
- **Novelty:** This is the first approach to achieve video grounding without training, leveraging attention layer dynamics and prompt-based guidance to localize relevant content.
- **Thoughts:** The training-free nature of VideoGEM is highly appealing, especially for scalable or low-resource settings. Its flexibility with prompts could enable fine-grained, customizable grounding and support the automatic construction of video-based datasets, opening up broader applications.


Action: „Whisk eggs“
Verb prompt: “A photo of a person whisk something.”
Object prompt: “A photo of a person using eggs.”
Action prompt: “A photo of a person whisk eggs.”
Combined output



The Devil is in Temporal Token: High Quality Video Reasoning Segmentation

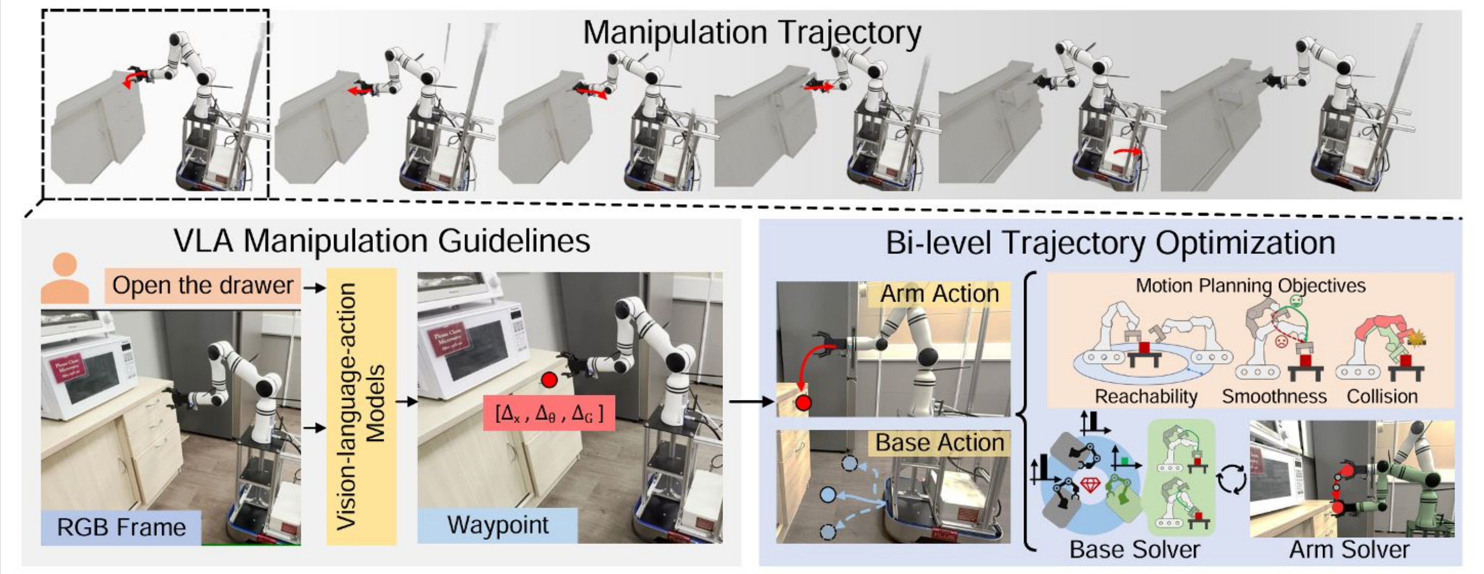
- **Summary:** This paper proposes VRS-HQ, a new method for video reasoning segmentation. Unlike previous methods that use a single [SEG] token to detect target objects across video frames, VRS-HQ emphasizes the importance of temporal localization. It introduces a temporal attention token ([TAK]) to identify which frames contain the target object. The method also integrates similarity-based fusion and frame selection mechanisms to enhance temporal reasoning.
- **Novelty:** The key innovation is the introduction of the [TAK] temporal token, enabling more precise temporal grounding of objects. Despite its simplicity, the method achieves strong performance across multiple datasets and tasks.
- **Thoughts:** The [TAK] token shows potential for broader use in other video reasoning tasks, and it raises interesting questions about whether similar temporal tokens could be applied to audio or other sequential modalities. The approach is practical and generalizable.



CVPR25: Meta Insights into Trends and Tendencies (138/179)

MoManipVLA: Transferring Vision-language-action Models for General Mobile Manipulation

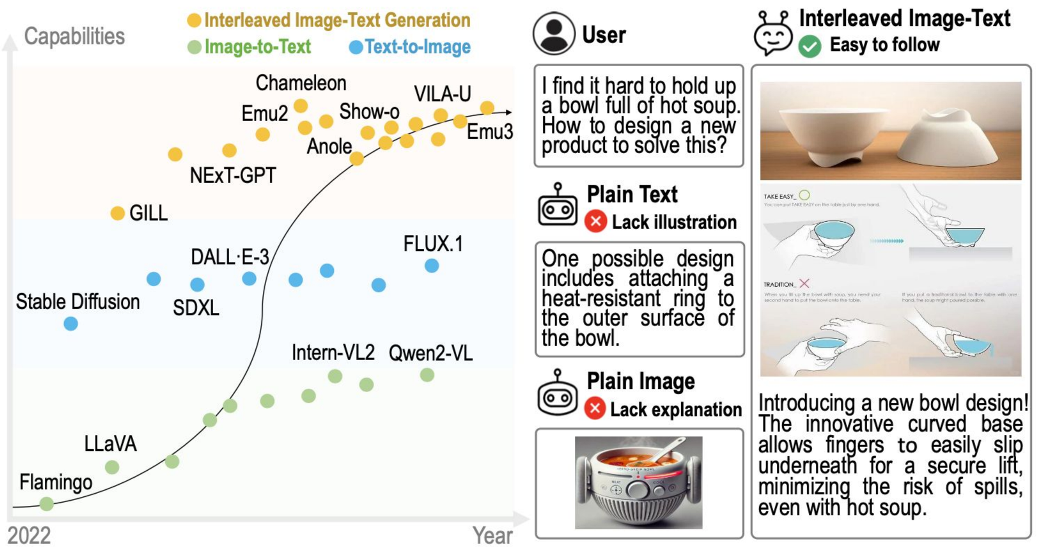
- **Summary:** This paper presents a framework for transferring large-scale Vision-Language-Action (VLA) models to mobile manipulation tasks in robotics. The proposed method predicts waypoint representations from VLA outputs, which generalize well across various manipulation tasks. It also incorporates motion planning for both arm and base actions, improving the accuracy of movement and manipulation in real-world settings. The system achieves mobile manipulation with minimal additional training.
- **Novelty:** The key innovation is the transfer of VLA models to real-world mobile manipulation, using waypoints as an effective and generalizable intermediate representation.
- **Thoughts:** There's a growing interest in applying MLLMs to robotics. This work raises important questions about what scene representations are most effective for real-world tasks—waypoints, scene graphs, or others. Integrating structured representations may further enhance precision and generalization in robotic planning.



Manipulation Trajectory
VLA Manipulation Guidelines
Open the drawer
RGB Frame
Vision-language-action Models
[Δ_x , Δ_θ , Δ_G]
Waypoint
Bi-level Trajectory Optimization
Arm Action
Base Action
Motion Planning Objectives
Reachability
Smoothness
Collision
Base Solver
Arm Solver

OpenING: A Comprehensive Benchmark for Judging Open-ended Interleaved Image-Text Generation (Oral)

- ❑ **Summary:** This paper introduces OpenING, a comprehensive benchmark designed to evaluate interleaved image-text generation in multimodal large language models (MLLMs). OpenING includes 5,400 human-annotated instances spanning 56 real-world tasks such as travel guidance, design, and brainstorming. To support reliable evaluation, the authors also propose IntJudge, a novel automatic judge model trained via a custom data pipeline, which achieves 82.42% agreement with human ratings, surpassing GPT-based evaluators. Experiments demonstrate that current interleaved generation models still fall short, and the paper provides insights to drive future advancements in the field.
- ❑ **Novelty:** Interleaved generation still remains less discussed.
- ❑ **Thoughts:** Dataset for detailed generation is still insufficient therefore important. Pretraining models with interleaved multimodal data could possibly improve model performance. In real-world user inputs, the data could be an incorporation of various data types, therefore it is important to carefully think about what kind of input data is important, less addressed, and fundamentally hard.



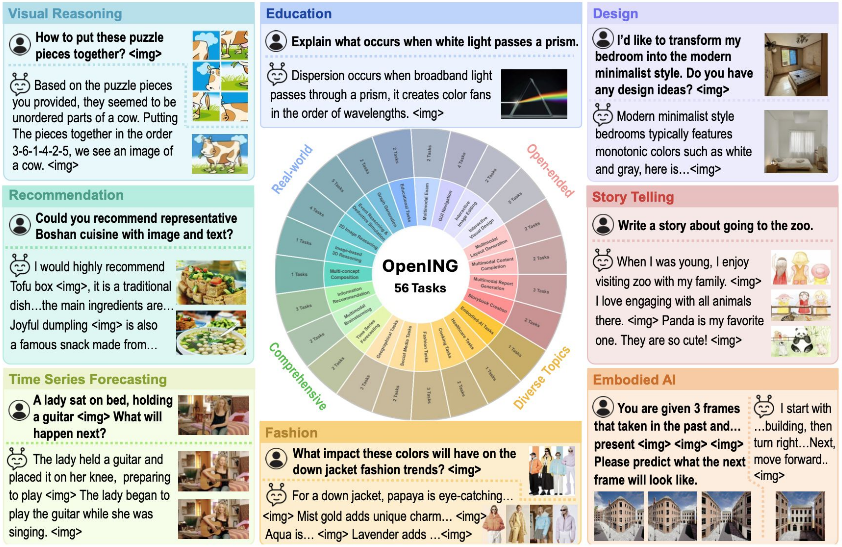
I find it hard to hold up a bowl full of hot soup. How to design a new product to solve this?

One possible design includes attaching a heat-resistant ring to the outer surface of the bowl.

Introducing a new bowl design! The innovative curved base allows fingers to easily slip underneath for a secure lift, minimizing the risk of spills, even with hot soup.

(a) Trends in Generative Model Development

(b) Benefits of Interleaved Image-Text Generation



Visual Reasoning

How to put these puzzle pieces together? <img>

Based on the puzzle pieces you provided, they seemed to be unordered parts of a cow. Putting The pieces together in the order 3-6-1-4-2-5, we see an image of a cow. <img>

Recommendation

Could you recommend representative Boshan cuisine with image and text?

I would highly recommend Tofu box <img>, it is a traditional dish…the main ingredients are… Joyful dumpling <img> is also a famous snack made from…

Time Series Forecasting

A lady sat on bed, holding a guitar <img> What will happen next?

The lady held a guitar and placed it on her knee, preparing to play <img> The lady began to play the guitar while she was singing. <img>

Education

Explain what occurs when white light passes a prism.

Dispersion occurs when broadband light passes through a prism, it creates color fans in the order of wavelengths. <img>

Fashion

What impact these colors will have on the down jacket fashion trends? <img>

For a down jacket, papaya is eye-catching… <img> Mist gold adds unique charm… <img> Aqua is… <img> Lavender adds …<img>

Design

I'd like to transform my bedroom into the modern minimalist style. Do you have any design ideas? <img>

Modern minimalist style bedrooms typically features monotonic colors such as white and gray, here is…<img>

Story Telling

Write a story about going to the zoo.

When I was young, I enjoy visiting zoo with my family. <img> I love engaging with all animals there. <img> Panda is my favorite one. They are so cute! <img>

Embodied AI

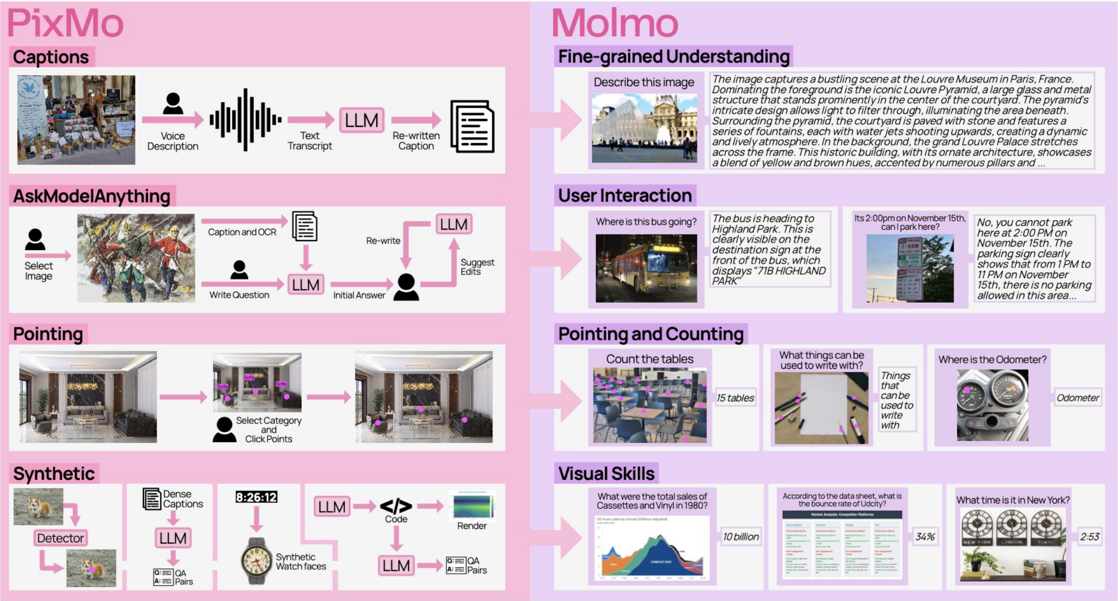
You are given 3 frames that taken in the past and… present <img> <img> <img> Please predict what the next frame will look like.

I start with …building, then turn right…Next, move forward.. <img>

CVPR25: Meta Insights into Trends and Tendencies (140/179)

Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models (Oral)

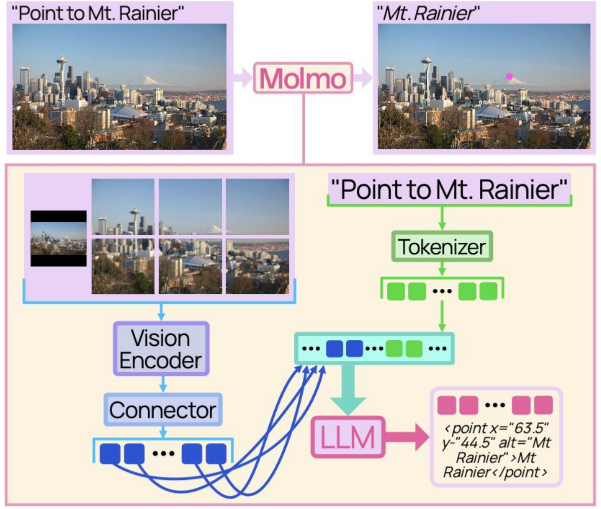
- **Summary:** This paper introduces Molmo, a family of open-weight vision-language models trained entirely from scratch using a newly collected dataset suite called PixMo. Without relying on proprietary VLMs, Molmo achieves state-of-the-art performance among open models and even outperforms several leading proprietary systems such as Claude 3.5 Sonnet and Gemini 1.5 Pro. Molmo involves detailed caption, impressive grounding ability, detailed knowledge-based reasoning.
- **Novelty:** Openness of data and model. While Molmo is open-sourced, it obtained with state-of-the-art result even compared with closed source commercial models.
- **Thoughts:** Detailed recognition and grounding ability is very import, with them several recent works including molmo obtained improved performance.



The image captures a bustling scene at the Louvre Museum in Paris, France. Dominating the foreground is the iconic Louvre Pyramid, a large glass and metal structure that stands prominently in the center of the courtyard. The pyramid's intricate design allows light to filter through, illuminating the area beneath. Surrounding the pyramid, the courtyard is paved with stone and features a series of fountains, each with water jets shooting upwards, creating a dynamic and lively atmosphere. In the background, the grand Louvre Palace stretches across the frame. This historic building, with its ornate architecture, showcases a blend of yellow and brown hues, accented by numerous pillars and ...

The bus is heading to Highland Park. This is clearly visible on the destination sign at the front of the bus, which displays "71B HIGHLAND PARK"

No, you cannot park here at 2:00 PM on November 15th. The parking sign clearly shows that from 1 PM to 11 PM on November 15th, there is no parking allowed in this area...



Rethinking Vision-Language Model in Face Forensics: Multi-Modal Interpretable Forged Face Detector

- **Summary:** This paper proposes M2F2-Det, a multi-modal deepfake detection model that not only classifies images as real or fake but also provides detailed natural language explanations for its decisions, enhancing interpretability. The model introduces a bridge adapter to align the image encoder with the LLM-based explanation generator, and uses forgery prompt learning and layer-wise LF-Tokens to better capture subtle facial manipulations. Through multi-stage training, M2F2-Det achieves strong generalization to unseen forgeries and sets new benchmarks in both detection accuracy and explanatory quality.
- **Novelty:** The task of simultaneously generating score and interpretations (left image). Also, the model is complicated but new (right image).
- **Thoughts:** Faces, gestures, gazes are important in social interaction recognition, but they are not really well addressed in data side. Instead of score, the ability to ground regions to explain the reason for forged faces could be more interesting. The model is a bit complicated.

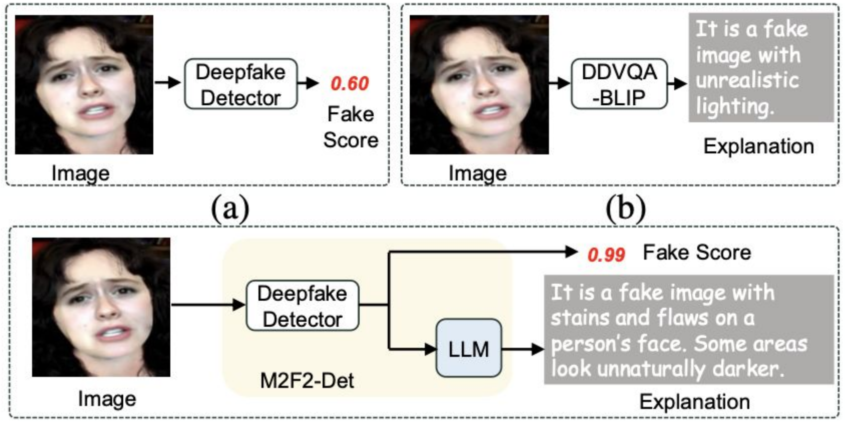
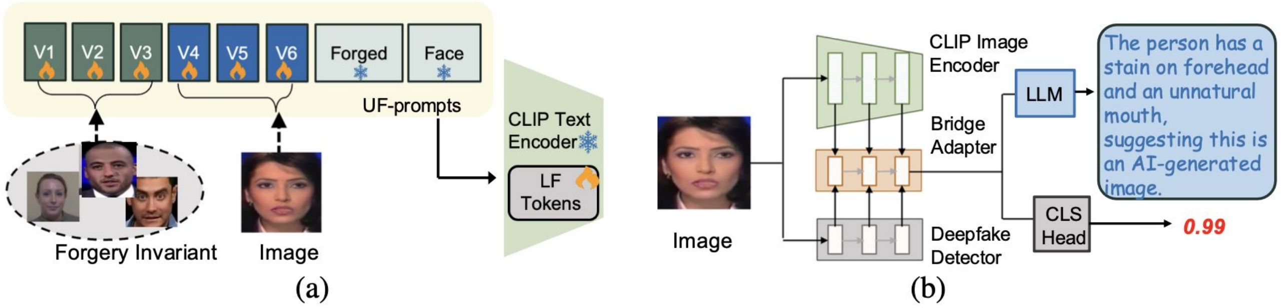


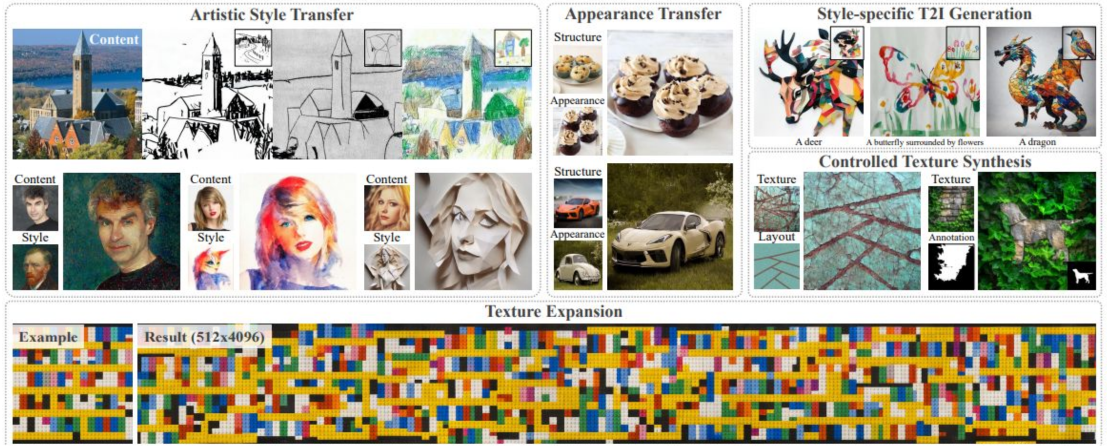
Image
Deepfake Detector
0.60
Fake Score
(a)
Image
DDVQA -BLIP
It is a fake image with unrealistic lighting.
Explanation
(b)
Image
Deepfake Detector
LLM
M2F2-Det
0.99
Fake Score
It is a fake image with stains and flaws on a person's face. Some areas look unnaturally darker.
Explanation

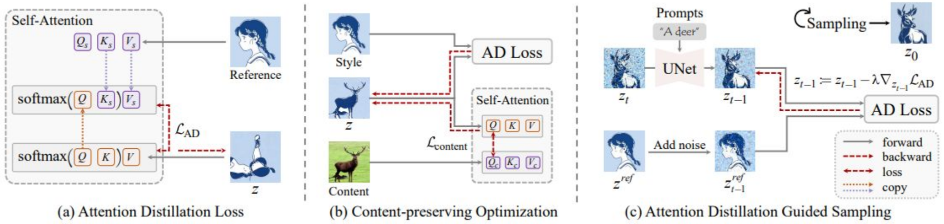


CVPR25: Meta Insights into Trends and Tendencies (142/179)

Attention Distillation: A Unified Approach to Visual Characteristics Transfer

- **Summary:** The self-attention mechanism in diffusion models inherently captures image style and semantic structure. This paper proposes a unified framework that utilizes this attention information to transfer visual characteristics—such as style and texture—from a reference image to a newly generated one.
- **Novelty:** The technical novelty lies in the introduction of an *Attention Distillation Loss*, which is defined as the discrepancy between the attention map of an ideally stylized image and that of the generated image. This enables effective transfer of visual properties through image optimization in the latent space via backpropagation. Additionally, by integrating the attention distillation loss into the diffusion sampling process and enhancing the traditional classifier guidance, the method achieves faster and more efficient image generation.
- **Thoughts:** While the model excels in transferring style and texture, it still struggles with conveying higher-level semantic attributes such as composition and content. Nonetheless, due to its general applicability, the method holds strong potential for future development and practical utility—**especially in use cases that require consistent visual styling across images**.



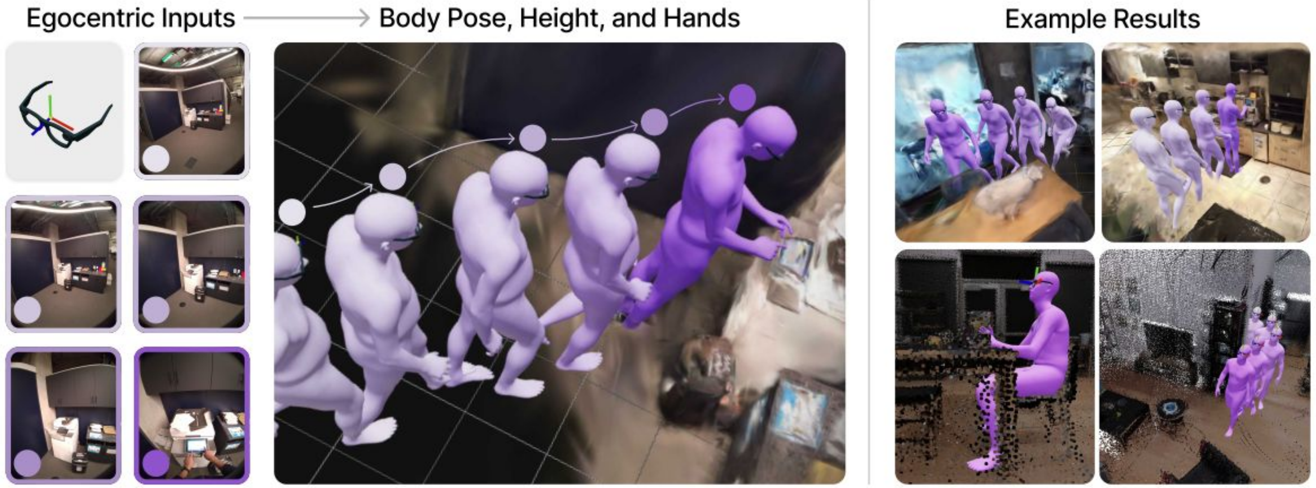


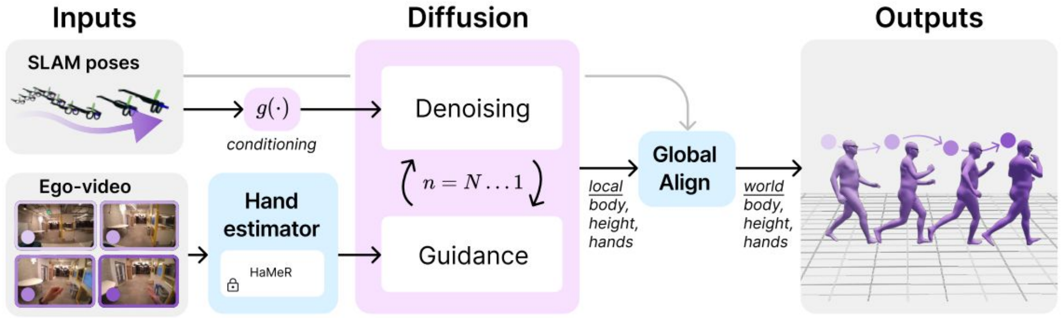
https://github.com/xugao97/AttentionDistillation

LIMIT.LAB
https://limitlab.xyz/

Estimating Body and Hand Motion in an Ego-sensed World

- **Summary:**This work proposes a system that estimates full-body 3D pose, body height, and hand motion in a global (world) coordinate system using only head-mounted device inputs—namely, 6DoF head poses from egocentric SLAM and onboard camera images. Crucially, the system **constructs the world map online**, without requiring any pre-scanned environments or external sensors.
- **Novelty:**The method introduces a **diffusion-based motion prior** conditioned on head movement. By performing **pose estimation in a locally normalized space relative to the head**, rather than in absolute coordinates, the model achieves significantly improved accuracy through spatial and temporal invariance.
- **Thoughts:**It is impressive that the method achieves full-body and hand motion estimation using only egocentric head pose and images. Height estimation is enabled under the assumption of a **flat ground plane**, which simplifies the problem. However, this also imposes a limitation: the method may not generalize well to **uneven terrains or stairs** due to this assumption.



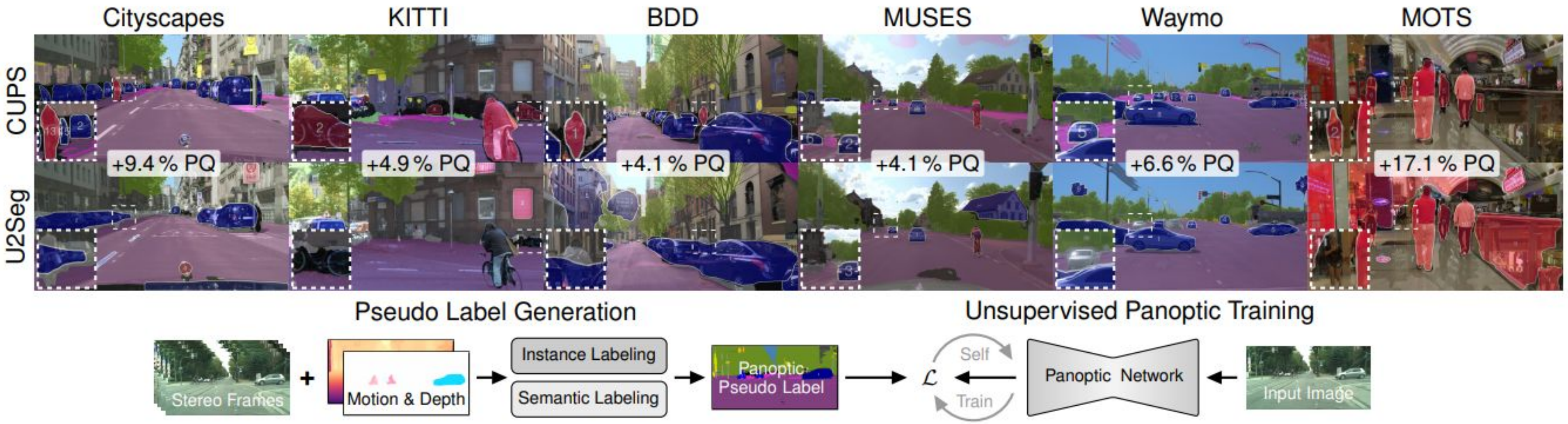


There's a great video!

https://egoallo.github.io/

Scene-Centric Unsupervised Panoptic Segmentation

- **Summary:** This paper proposes a novel framework for *scene-centric* ***unsupervised panoptic segmentation***, which enables holistic understanding of complex scenes *without any manual annotations*.
- **Novelty:** The method leverages stereo videos and self-supervised techniques (e.g., SMURF) to extract motion cues, followed by clustering via SF2SE3 to segment moving objects in 3D space. It further employs self-supervised features from DINO and Depth-G, incorporating stereo depth to generate high-resolution semantic masks. By combining motion-based instance masks ("things") with semantics-based region masks ("stuff"), the framework constructs high-quality pseudo panoptic labels.
- **Thoughts:** This method is highly promising for generating reliable pseudo labels even in complex real-world scenes. Its practicality is notable, as it can serve as an **effective pre-labeling step before manual annotation**, reducing annotation costs significantly.

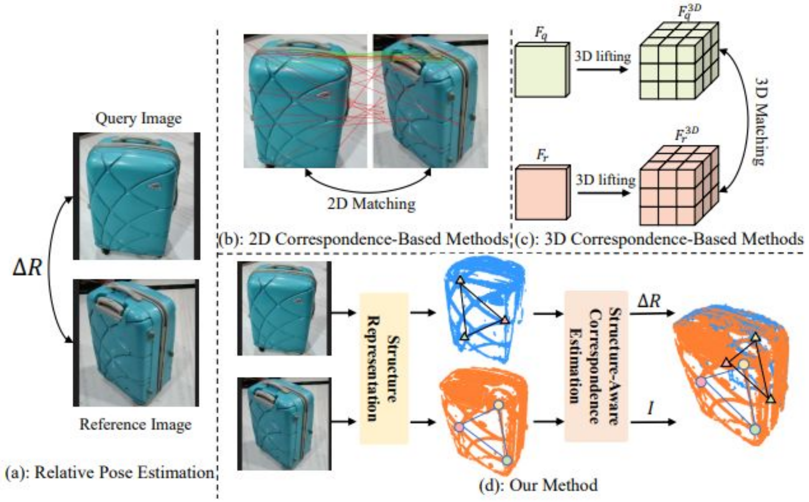


https://visinf.github.io/cups/

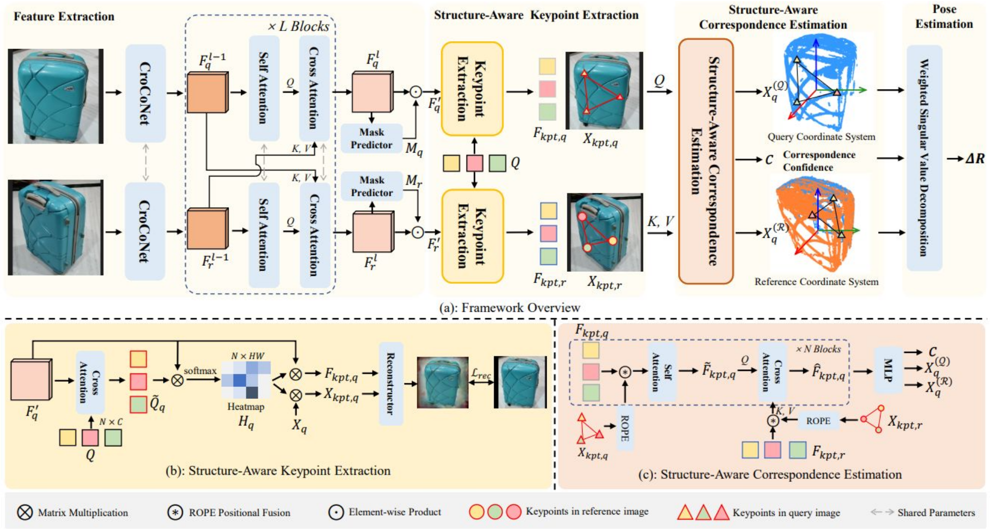
LIMIT.LAB
https://limitlab.xyz/

Structure-Aware Correspondence Learning for Relative Pose Estimation

- **Summary:**
- This paper addresses the task of estimating the 3D relative pose—rotation and translation—between a reference image and a query image, even when the object is previously unseen.
- **Novelty:**
- The method automatically extracts structure-representative keypoints from each image, allowing it to capture object shape structures even across significant appearance differences. By modeling the relationship between images as a graph, the approach directly predicts 3D-to-3D correspondences without requiring explicit feature matching.
- **Thoughts:**
- A key strength of this method lies in its ability to represent object structures as graphs and directly infer correspondences. However, since the structural estimation for a reference image changes depending on the query image, the method lacks a notion of absolute structure. On the other hand, this dependency allows it to reinterpret an image dynamically based on the query, making it practical for real-world applications such as object tracking or viewpoint adaptation.



(a): Relative Pose Estimation (b): 2D Correspondence-Based Methods (c): 3D Correspondence-Based Methods (d): Our Method



(a): Framework Overview

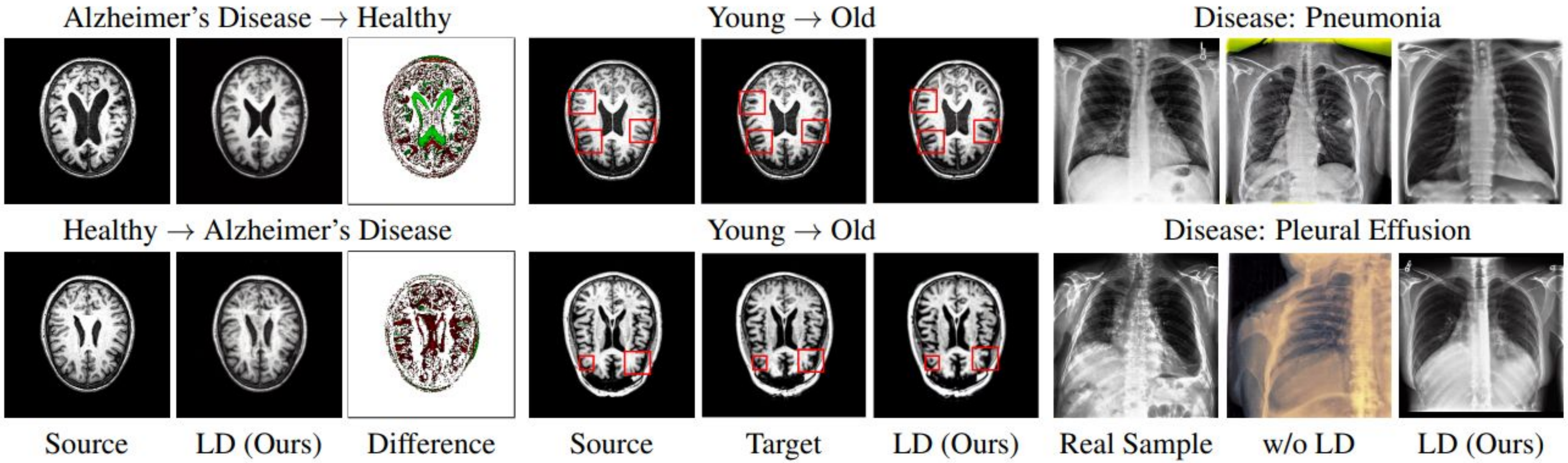
(b): Structure-Aware Keypoint Extraction

(c): Structure-Aware Correspondence Estimation

https://cyhhzo02.github.io/SAC-Pose/

Latent Drifting in Diffusion Models for Counterfactual Medical Image Synthesis

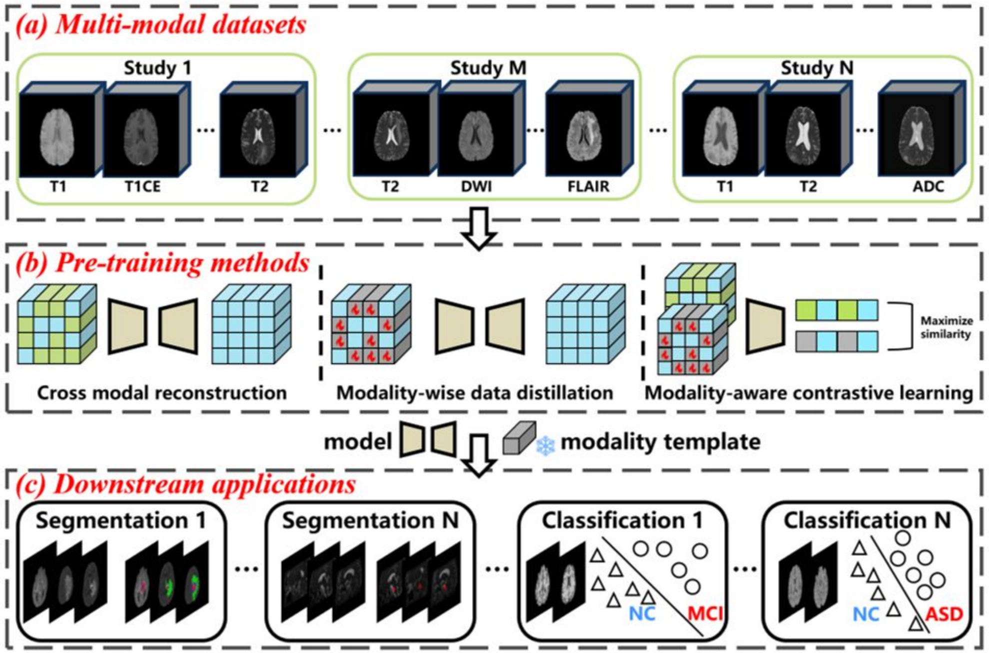
- **Summary:** Generating high-quality medical images is challenging due to the scarcity of labeled data. This paper introduces a novel method called **Latent Drifting (LD)** that enables counterfactual medical image synthesis by gradually shifting the latent distribution between source (e.g., natural) and target (e.g., medical) domains.
- **Novelty:** The key innovation lies in the introduction of a drift parameter δ into the denoising and reconstruction process of diffusion models. This allows controlled deviation in the latent space, facilitating domain adaptation. Furthermore, the method formalizes counterfactual generation as a min-max optimization problem—modeling “what-if” scenarios by modifying semantic attributes in a principled way.
- **Thoughts:** The approach is notable not only for bridging the distribution gap between natural and medical domains through latent control, but also for its structured treatment of counterfactual generation. This enables reliable medical image synthesis even in low-data regimes, with promising applications in clinical decision support and disease progression modeling.

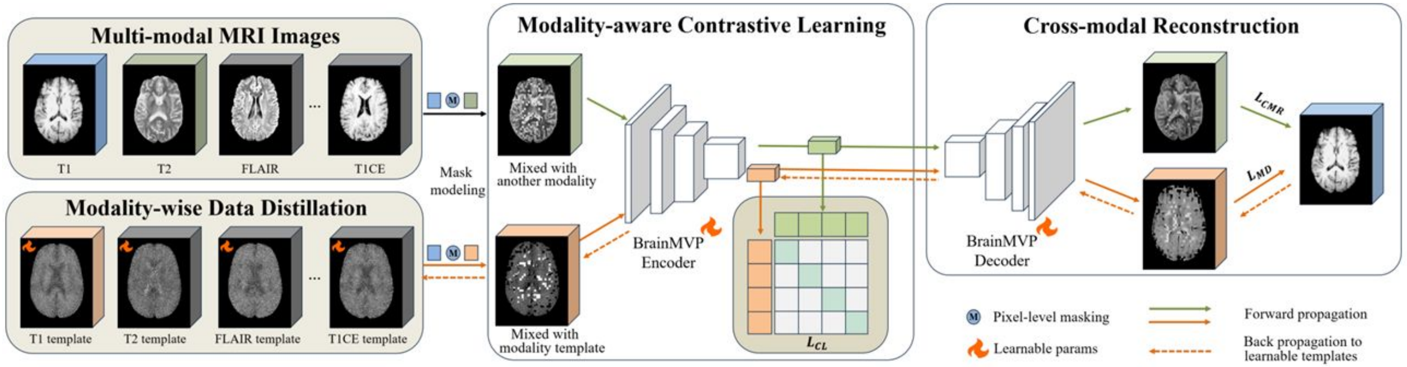


https://latentdrifting.github.io/

Multi-modal Vision Pre-training for Medical Image Analysis

- **Summary:** MRI's key strength is its ability to capture multiple modalities from the same anatomical cross-section, which inherently creates correlations between them. For effective self-supervised learning on MRI images, it's crucial to utilize datasets that include these diverse modalities.
- **Novelty:** To leverage these inter-modal correlations, BrainMVP constructed an impressive dataset: **16,022 scans (over 2.4 million images)** across **8 distinct MRI modalities**. They then conducted multi-modal image pre-training using three novel proxy tasks—**cross-modal image reconstruction, modality-aware contrastive learning, and modality template distillation**—to boost the learning of cross-modality representations and correlations, resulting in excellent performance on downstream tasks.
- **Significance:** The sheer scale of the dataset built for this project is truly remarkable. This dataset alone represents a significant contribution to medical imaging research.



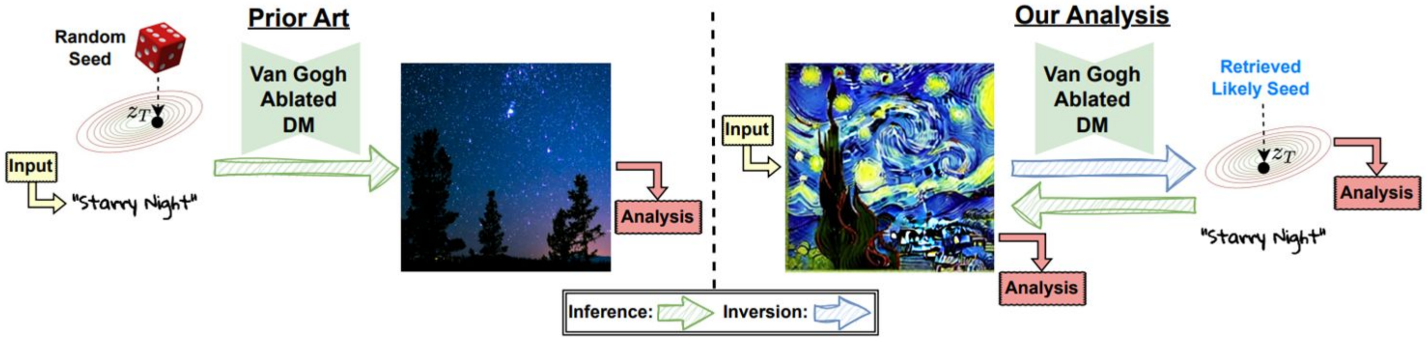


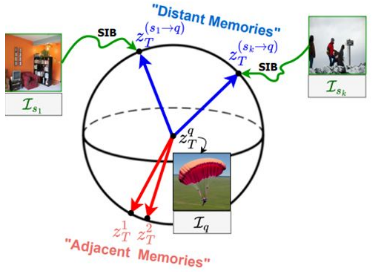
https://github.com/shaohao011/BrainMVP

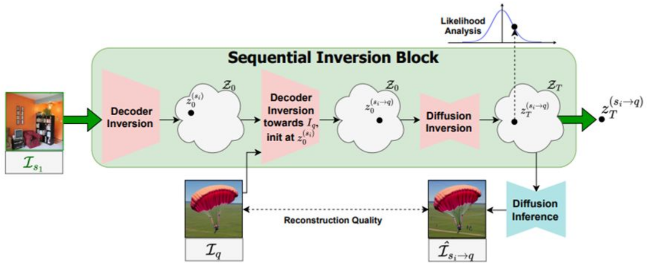
LIMIT.LAB
https://limitlab.xyz/

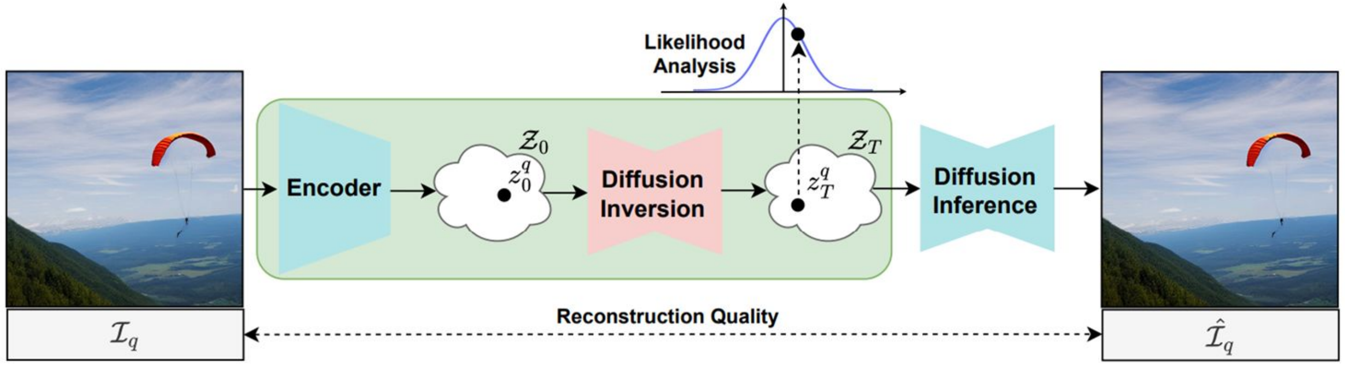
Memories of Forgotten Concepts

- **Summary:** The paper reveals that erased concept information persists within AI models and can be revived. It shows that specific "latent seeds" and "inversion methods" can generate high-quality images of forgotten concepts.
- **Novelty:** This research uniquely demonstrates the intractability of completely erasing concept information from AI. It highlights potential vulnerabilities in current concept ablation techniques by showing how erased data can be reconstructed.
- **Thoughts:** I was surprised to learn that concepts exist in the latent space, not as discrete units, but as overlapping and dispersed information. This makes true "erasure" an incredibly challenging task, like trying to remove a specific color from a painting by only adjusting a few pixels.





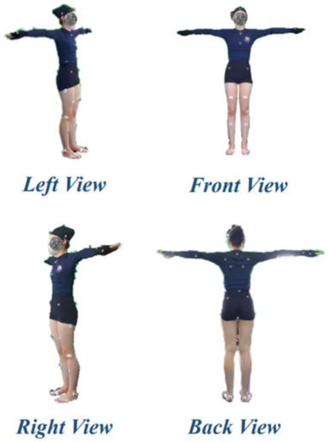


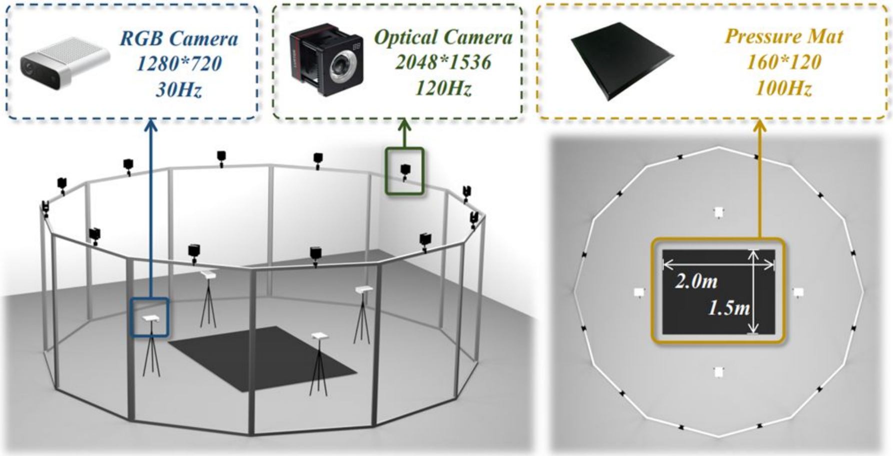


https://matanr.github.io/Memories_of_Forgotten_Concepts/

MotionPRO: Exploring the Role of Pressure in Human MoCap and Beyond

- **Summary:** This paper introduces MotionPRO, a large-scale human motion capture dataset integrating **pressure, RGB, and optical sensors**. It demonstrates the **necessity and effectiveness of pressure signals** for improving pose and trajectory estimation, and for driving realistic virtual humans.
- **Novelty:** The core novelty lies in creating a **multi-modal dataset** that rigorously explores the **critical role of pressure information** in human motion capture, thereby enhancing the physical realism and accuracy of AI models for motion understanding and generation.
- **Thoughts:** The inclusion of simultaneous pressure sensor data offers significant practical utility, especially for **time-series analysis of dynamic movements like jumps**. It provides crucial physical insights beyond visual cues, enabling more accurate and robust understanding of ground interaction and force distribution during such complex actions.

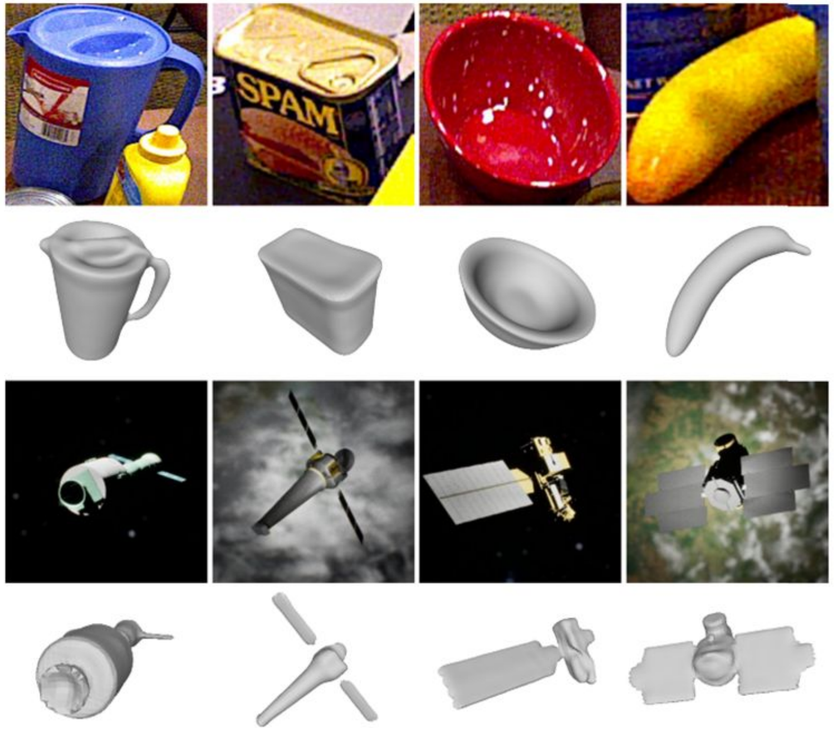


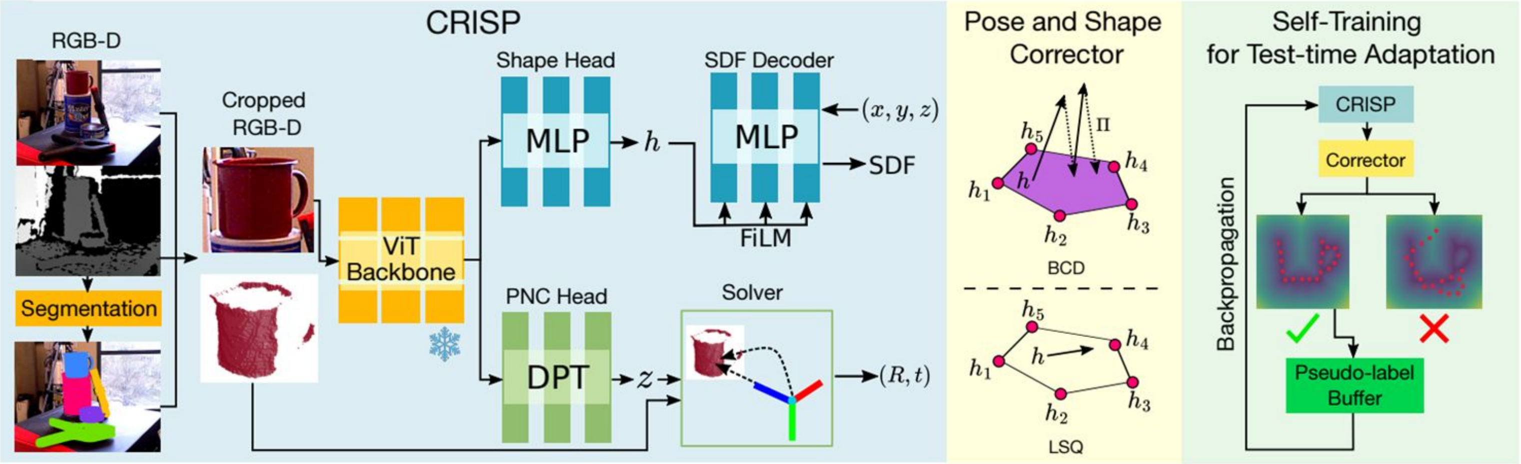


https://nju-cite-mocaphumanoid.github.io/MotionPRO/

CRISP: Object Pose and Shape Estimation with Test-Time Adaptation

- **Summary:** CRISP is a new method that estimates 6D object pose and shape from a single RGB-D image. It excels at adapting to unseen objects and bridging large domain gaps through test-time self-training.
- **Novelty:** CRISP's primary novelty lies in its category-agnostic approach to pose and shape estimation, eliminating the need for prior object knowledge. It also introduces an effective test-time adaptation mechanism, allowing for strong generalization to novel objects and environments.
- **Thoughts:** It's genuinely impressive how CRISP manages to perform object pose and shape estimation without needing to know the object's category beforehand. This category independence is a significant leap forward, making it incredibly versatile for real-world scenarios with unknown items.



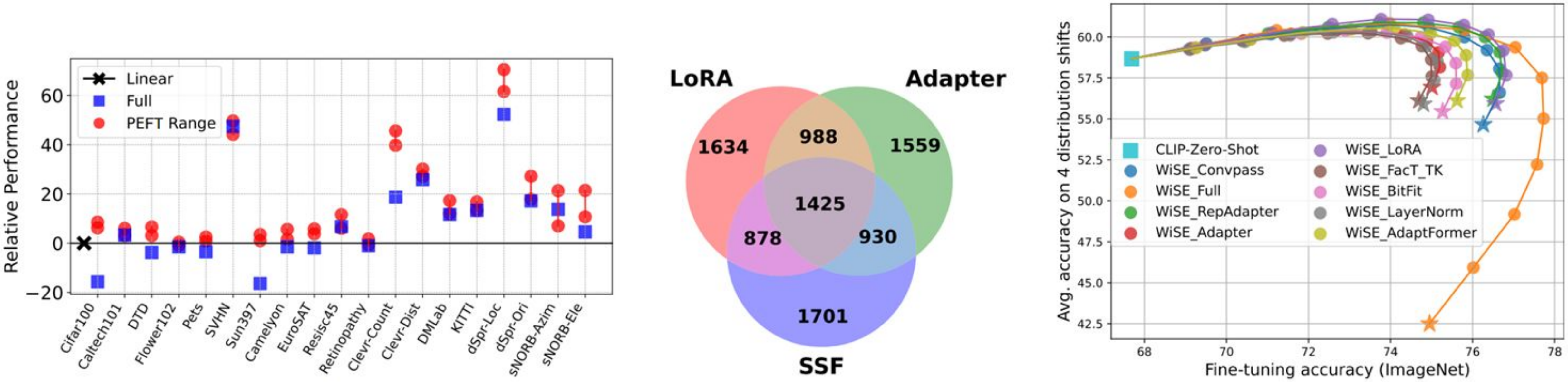


https://github.com/MIT-SPARK/CRISP?tab=readme-ov-file

CVPR25: Meta Insights into Trends and Tendencies (151/179)

Lessons and Insights from a Unifying Study of Parameter-Efficient Fine-Tuning (PEFT) in Visual Recognition

- ❏ **Summary:** This paper unifies the study of Parameter-Efficient Fine-Tuning (PEFT) methods in visual recognition, showing that various PEFT approaches achieve similar accuracy after careful tuning.
- ❏ **Novelty:** It highlights that despite similar performance, different PEFT methods make distinct predictions and errors, suggesting varying inductive biases.
- ❏ **Thoughts:** It's surprising to learn that each **PEFT method has a unique functional bias**, implying no single "one-size-fits-all" solution and necessitating method-specific applications.:



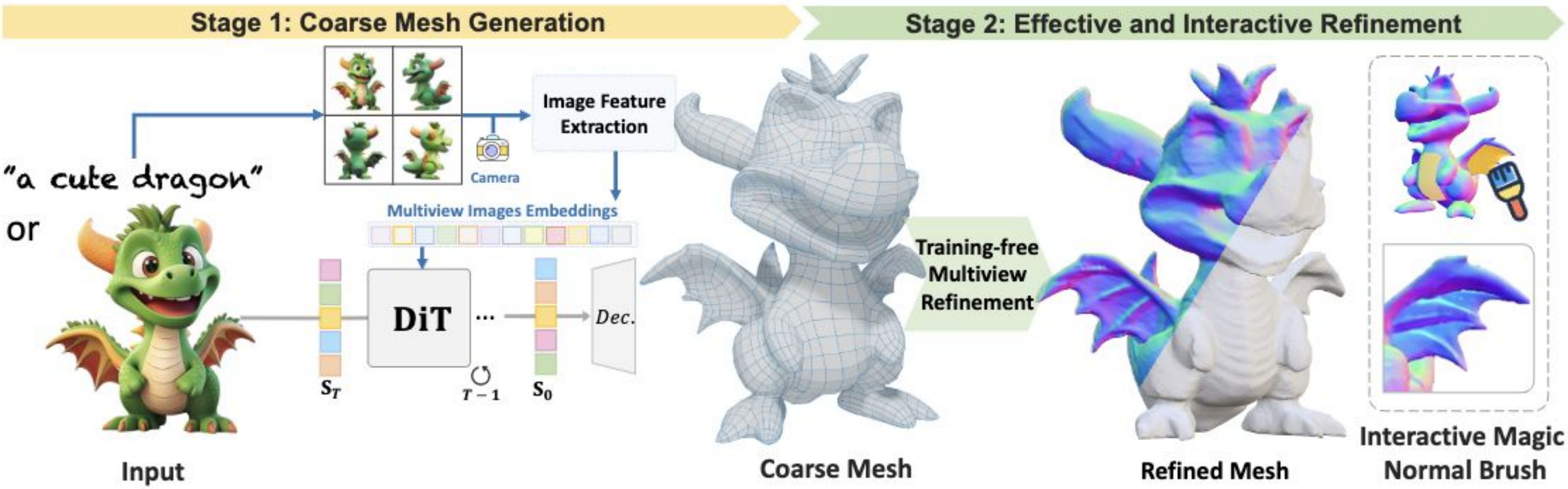
(a) Accuracy gain vs. linear probing on VTAB-1K (19 tasks) (b) Prediction overlaps (5K most confident) (c) Target distribution vs. distribution shifts

https://zheda-mai.github.io/PEFT_Vision_CVPR25/

CVPR25: Meta Insights into Trends and Tendencies (152/179)

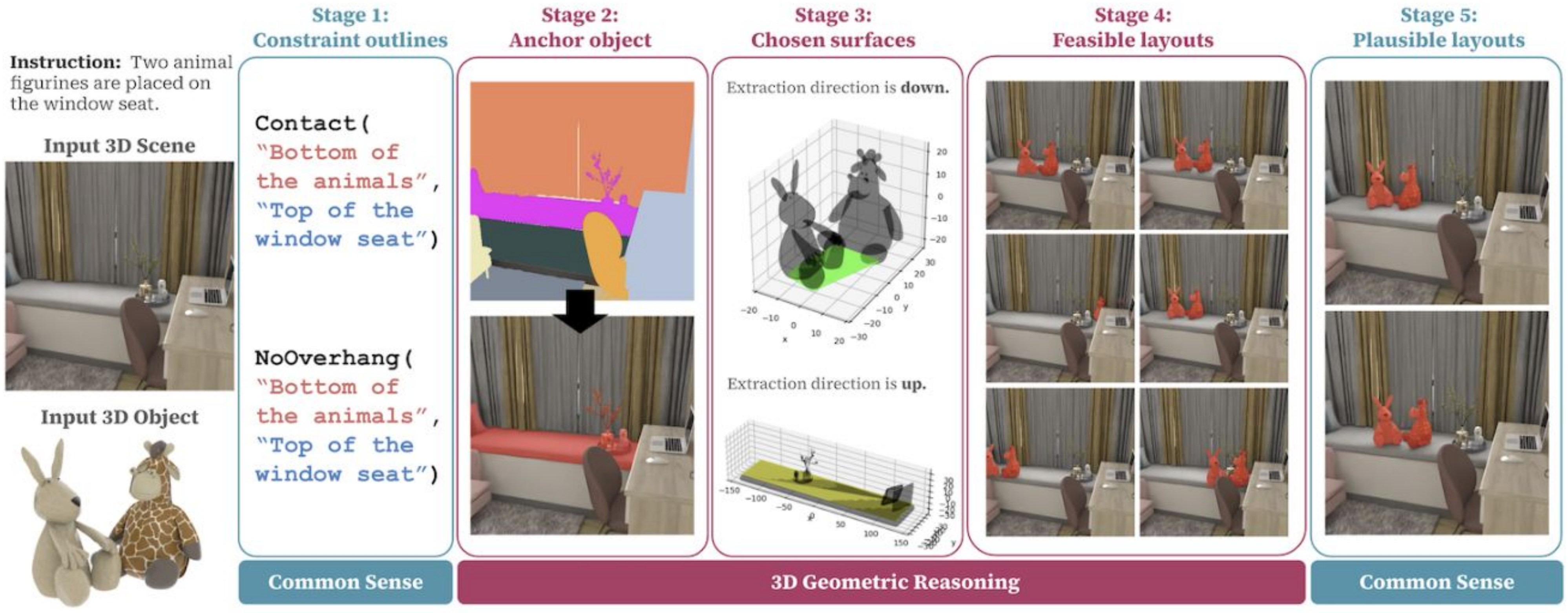
CraftsMan3D: High-fidelity Mesh Generation with 3D Native Diffusion and Interactive Geometry Refiner

- ❑ Summary: This work developed a new 3D generation system that can produce high-quality 3D shapes in a short time and allows interactive shape editing.
- ❑ Novelty:This work propose a generation model for 3D data that is fast, high-quality, and editable.
- ❑ Thoughts:I believe this is a state-of-the-art method capable of automatically generating 3D assets. If combined with LLM-based 3D scene generation, it could potentially solve the problem of the shortage of 3D data all at once.


Stage 1: Coarse Mesh Generation
Stage 2: Effective and Interactive Refinement
"a cute dragon"
or
Input
Image Feature Extraction
Camera
Multiview Images Embeddings
DiT
S_T
$T-1$
S_0
Dec.
Coarse Mesh
Training-free Multiview Refinement
Refined Mesh
Interactive Magic Normal Brush

FirePlace: Geometric Refinements of LLM Common Sense Reasoning for 3D Object Placement

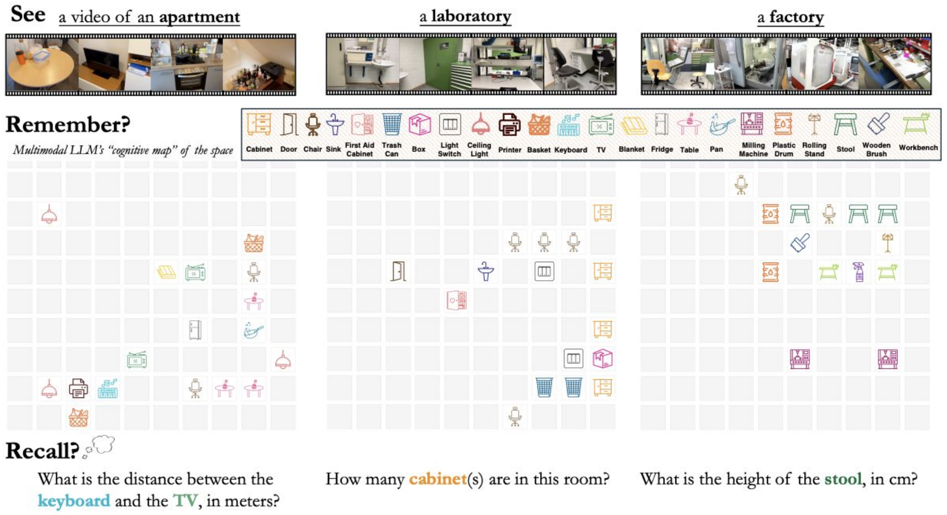
- Summary: This work investigate how to optimally leverage MLLMs for object placement tasks. This work proposed FirPlace.
- Novelty:This work show that presenting the options in multiple stages improves the inference accuracy of the MLLM.
- Thoughts:

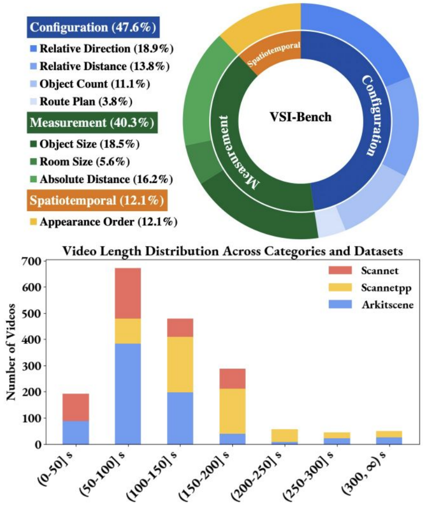

Instruction: Two animal figurines are placed on the window seat.
Input 3D Scene
Input 3D Object
Stage 1: Constraint outlines
Contact("Bottom of the animals", "Top of the window seat")
NoOverhang("Bottom of the animals", "Top of the window seat")
Stage 2: Anchor object
Stage 3: Chosen surfaces
Extraction direction is down.
Extraction direction is up.
Stage 4: Feasible layouts
Stage 5: Plausible layouts
Common Sense
3D Geometric Reasoning
Common Sense

CVPR25: Meta Insights into Trends and Tendencies (154/179)

Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces

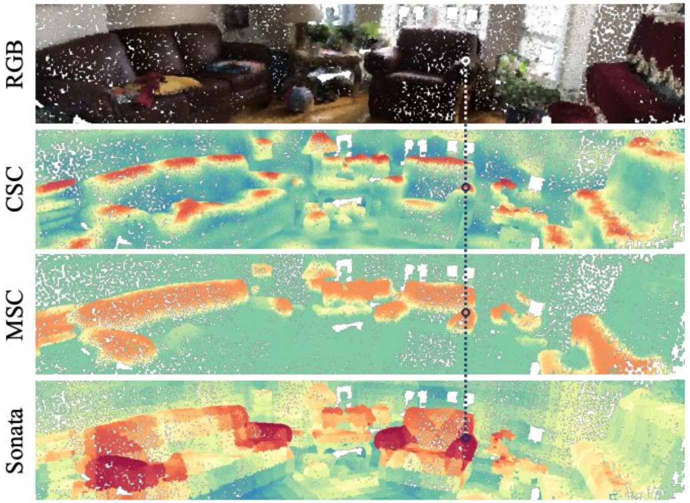
- Summary: MLLMs show emerging spatial intelligence from video but struggle with reasoning, and explicit cognitive maps help
- Motivation & Objective: To investigate whether MLLMs can develop human-like spatial understanding from videos and identify key limitations in their visual-spatial reasoning
- Approach:
  - The study introduces VSI-Bench, a benchmark with diverse spatial tasks from real-world videos, and evaluates MLLMs by analyzing both their linguistic self-explanations and visual cognitive maps.
  - The dataset is built by unifying video and 3D annotation data from ScanNet, ScanNet++, and ARKitScenes, and generating over 5,000 QA pairs across eight spatial tasks using auto-annotation and human-in-the-loop refinement.

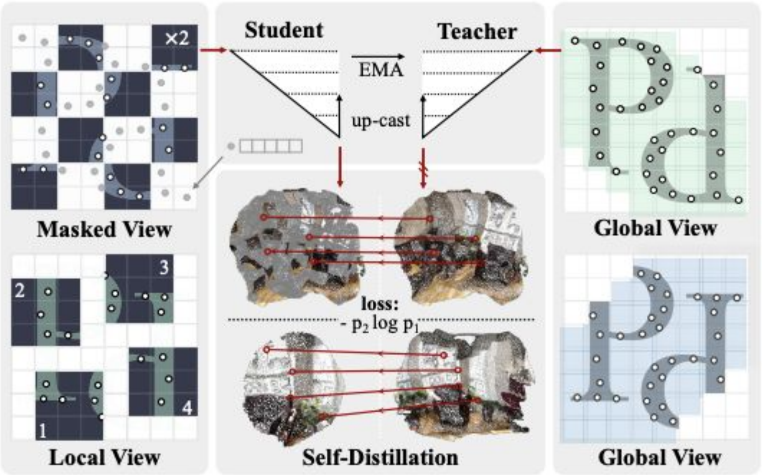




Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces

- ❑ Summary: Sonata mitigates the "geometric shortcut" problem in point cloud self-supervised learning by obscuring spatial cues and strengthening reliance on input features, leading to reliable and generalizable 3D representations that excel in linear probing and transfer tasks.
- ❑ Approach:
  - ❑ Macro Framework: Point Self-Distillation
  - ❑ Two-view strategy: Generate augmented point cloud views (crop, jitter, mask)
  - ❑ Align features: Use EMA teacher–student to match features between:
    - ❑ Local/masked views (student)
    - ❑ Global views (teacher)
  - ❑ Self-distillation replaces contrastive/generative learning
  - ❑ Loss: Sinkhorn clustering + KoLeo regularization for robust alignment



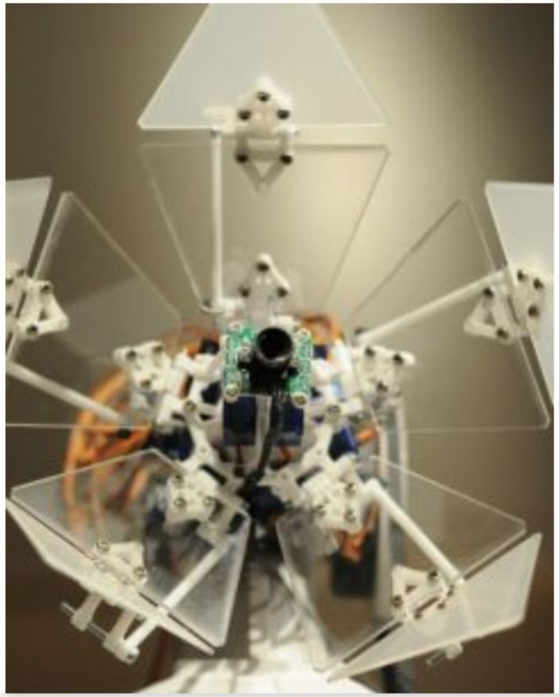


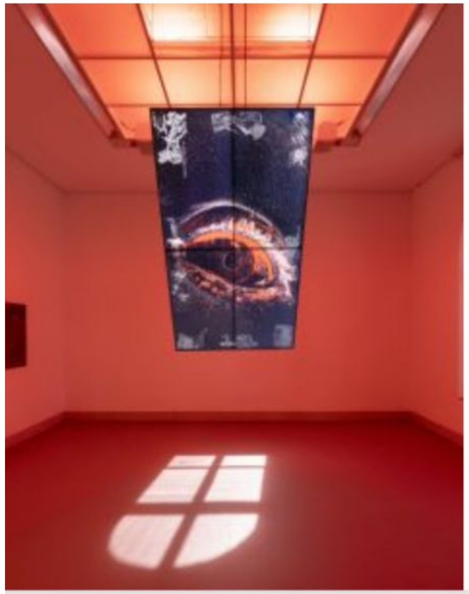
CVPR25: Meta Insights into Trends and Tendencies (156/179)

CVPR AI arts

❑ AI art installations were exhibited in Hall A2. The most impressive piece for me was “The Flower.”

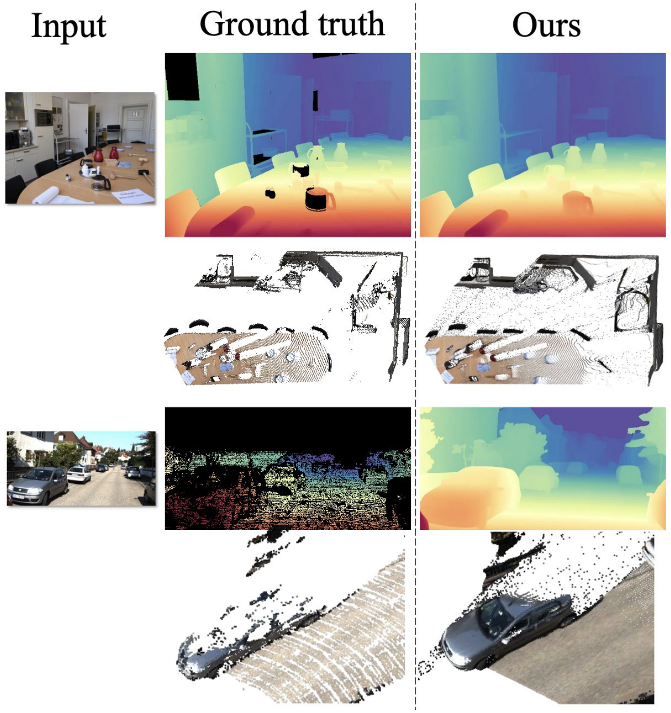
❑ “The Flower” responds to human presence by tracking the viewer’s face, and its petals move in a way that seems to convey emotion or intention — as if the flower were silently reacting to the observer.

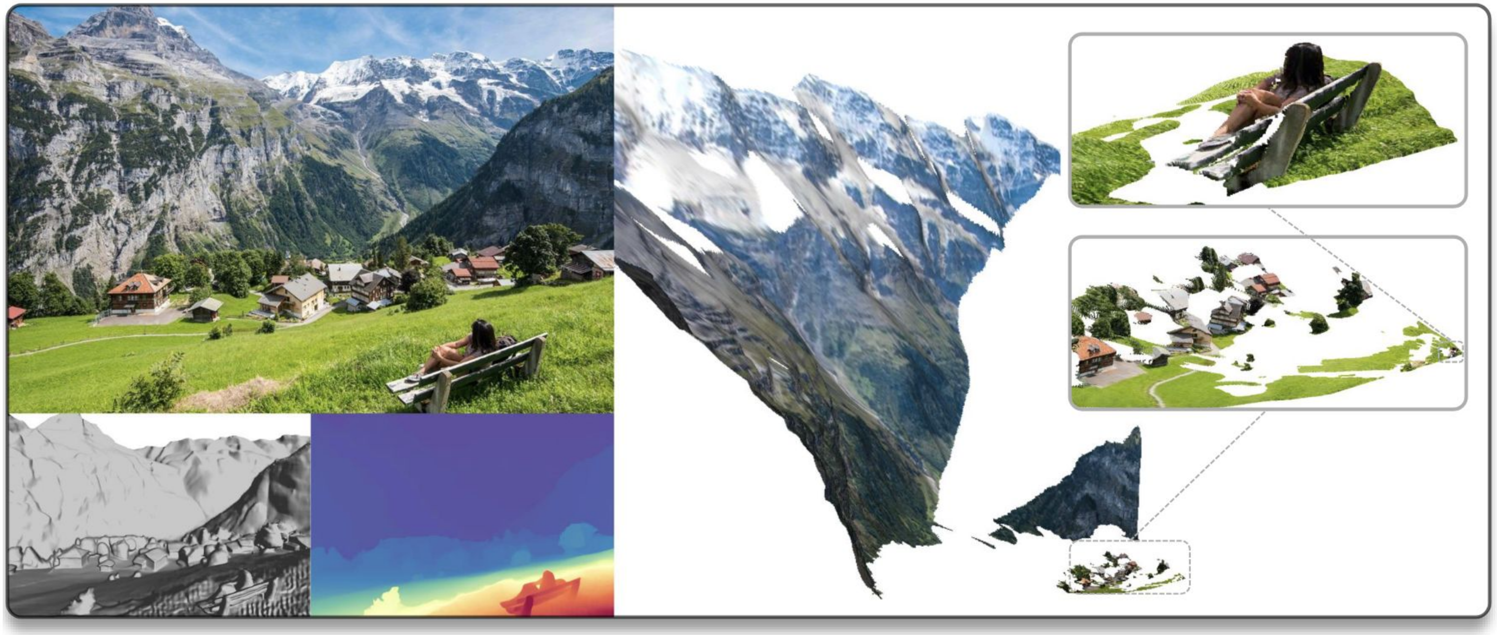




MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision:

- ❏ Summary:This work propose an architecture that generates a 3D point cloud map from a single input image.
- ❏ Novelty: This work introduce an “affine-invariant” point map that is invariant to scale and translation.This removes focal-distance ambiguity and enables more stable learning,allowing direct 3D reconstruction from a single image without the need for camera parameters.
- ❏ Thoughts:I was particularly interested in how this could be applied to recognition tasks. Specifically, I wondered whether there might be a method to automatically generate annotations when reconstructing 3D scenes from a single image. For instance, if the input image already has semantic supervision or labels, it might be possible to propagate those annotations during the scene reconstruction process.





Multi-view Reconstruction via SfM-guided Monocular Depth Estimation

- ❑ Summary: This work propose a novel approach that incorporates prior information from SfM into diffusion-based depth estimation, enabling high-precision and multi-view consistent depth prediction for each viewpoint.
- ❑ Novelty: Using a diffusion model conditioned on SfM point clouds, we predict scale-consistent depth from multi-view images without the need for matching, and use this to achieve high-precision 3D reconstruction.
- ❑ Thoughts:

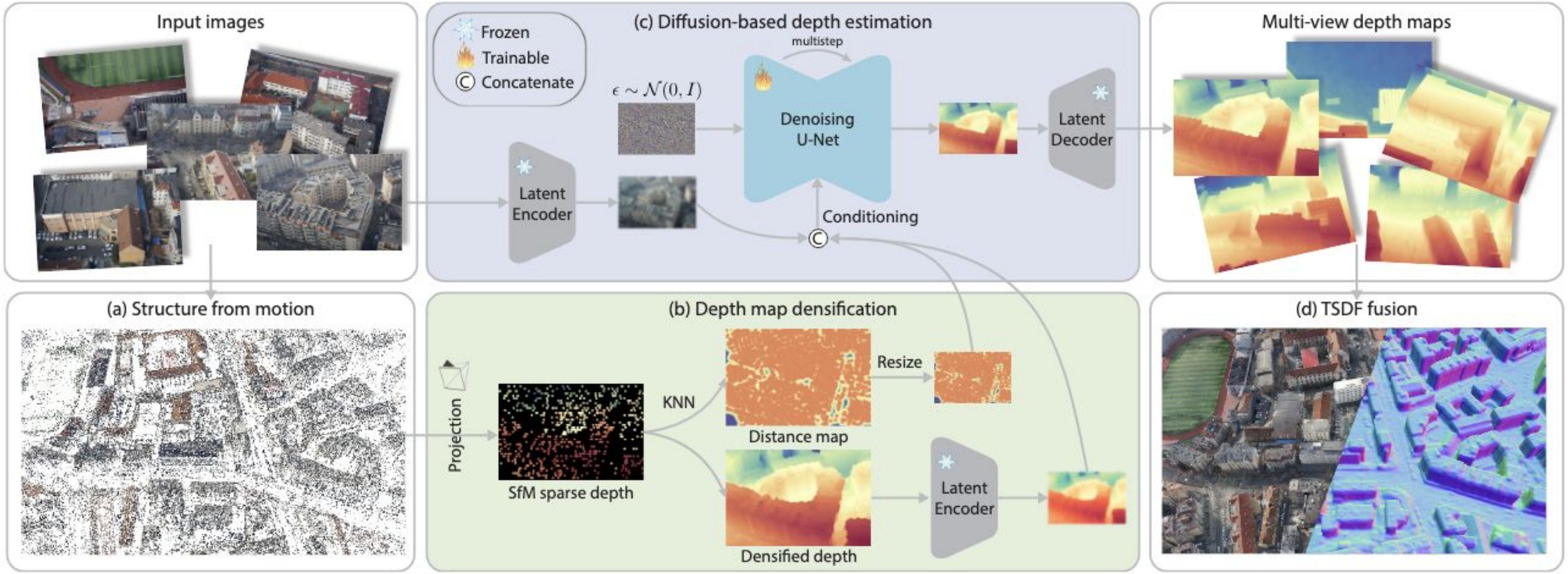
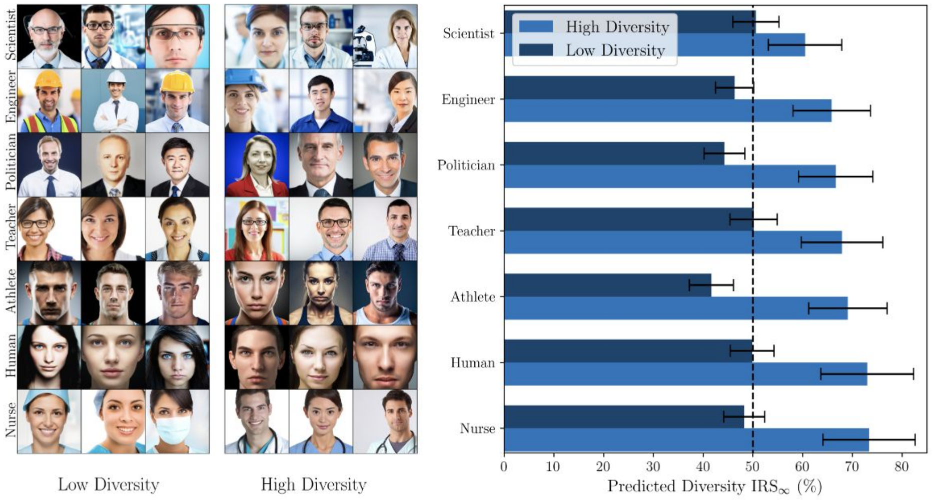
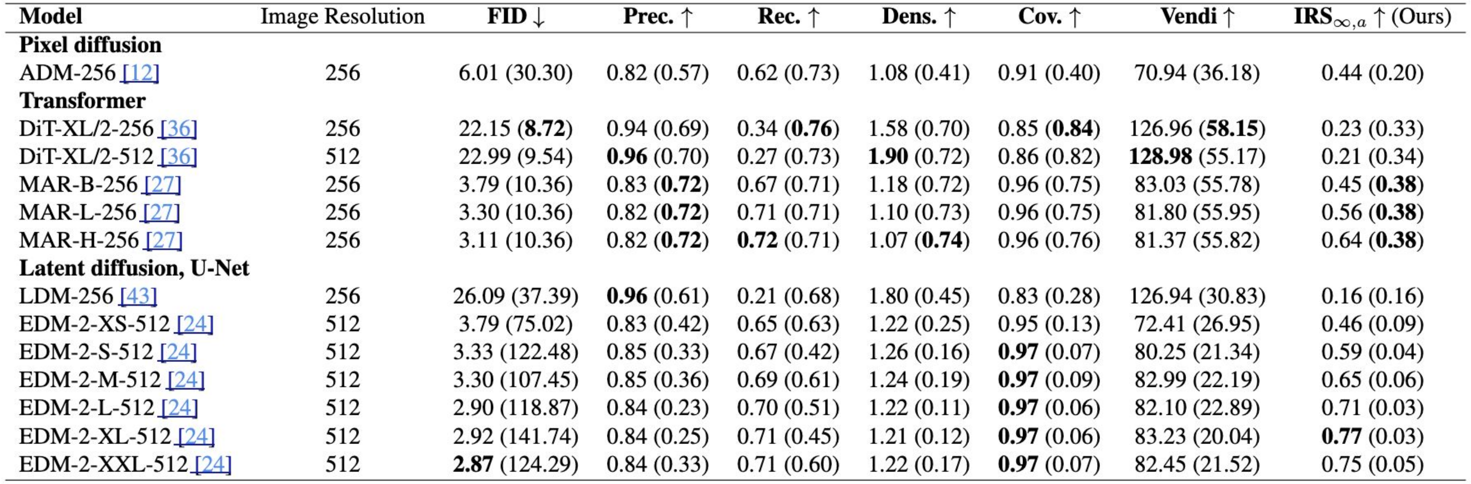


Image Generation Diversity Issues and How to Tame Them

- Summary: This work proposes a new measurement method IRS for assessing diversity in image generation tasks. Using this metric, it is observed that conventional diffusion models—even when conditioned—are only able to cover up to 77% of the diversity present in the training data. Unconditional models perform even worse. Motivated by this insight, the authors propose a pseudo-labeling approach for generation, aiming to enhance diversity without sacrificing image quality.
- Novelty: This work defines new metrics of diversity and indicates conventional models remain low-diversity, even using conditional-generation.

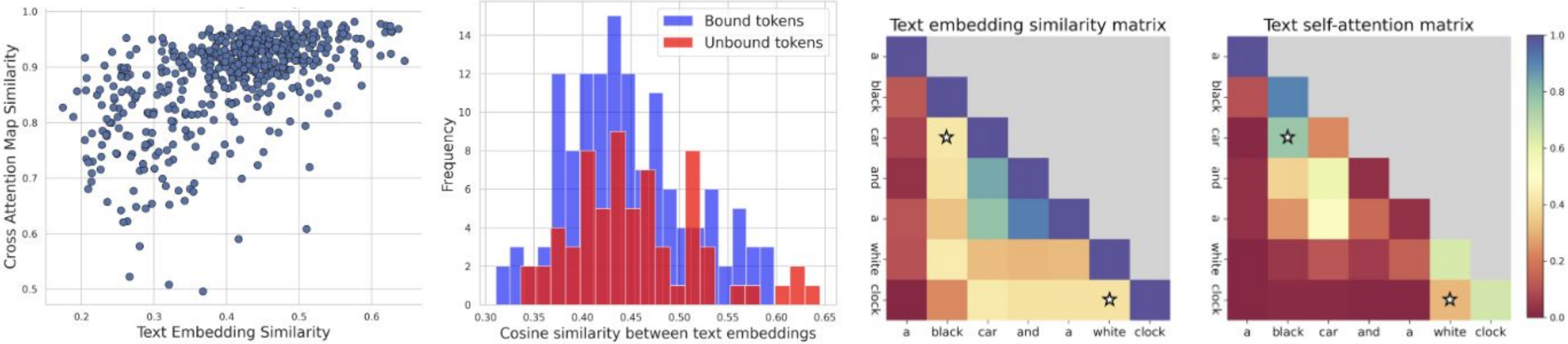




Model	Image Resolution	FID ↓	Prec. ↑	Rec. ↑	Dens. ↑	Cov. ↑	Vendi ↑	$\mathrm{IRS}_{\infty,a}$ ↑ (Ours)
Pixel diffusion								
ADM-256 [12]	256	6.01 (30.30)	0.82 (0.57)	0.62 (0.73)	1.08 (0.41)	0.91 (0.40)	70.94 (36.18)	0.44 (0.20)
Transformer								
DiT-XL/2-256 [36]	256	22.15 (**8.72**)	0.94 (0.69)	0.34 (**0.76**)	1.58 (0.70)	0.85 (**0.84**)	126.96 (**58.15**)	0.23 (0.33)
DiT-XL/2-512 [36]	512	22.99 (9.54)	**0.96** (0.70)	0.27 (0.73)	**1.90** (0.72)	0.86 (0.82)	**128.98** (55.17)	0.21 (0.34)
MAR-B-256 [27]	256	3.79 (10.36)	0.83 (**0.72**)	0.67 (0.71)	1.18 (0.72)	0.96 (0.75)	83.03 (55.78)	0.45 (**0.38**)
MAR-L-256 [27]	256	3.30 (10.36)	0.82 (**0.72**)	0.71 (0.71)	1.10 (0.73)	0.96 (0.75)	81.80 (55.95)	0.56 (**0.38**)
MAR-H-256 [27]	256	3.11 (10.36)	0.82 (**0.72**)	**0.72** (0.71)	1.07 (**0.74**)	0.96 (0.76)	81.37 (55.82)	0.64 (**0.38**)
Latent diffusion, U-Net								
LDM-256 [43]	256	26.09 (37.39)	**0.96** (0.61)	0.21 (0.68)	1.80 (0.45)	0.83 (0.28)	126.94 (30.83)	0.16 (0.16)
EDM-2-XS-512 [24]	512	3.79 (75.02)	0.83 (0.42)	0.65 (0.63)	1.22 (0.25)	0.95 (0.13)	72.41 (26.95)	0.46 (0.09)
EDM-2-S-512 [24]	512	3.33 (122.48)	0.85 (0.33)	0.67 (0.42)	1.26 (0.16)	**0.97** (0.07)	80.25 (21.34)	0.59 (0.04)
EDM-2-M-512 [24]	512	3.30 (107.45)	0.85 (0.36)	0.69 (0.61)	1.24 (0.19)	**0.97** (0.09)	82.99 (22.19)	0.65 (0.06)
EDM-2-L-512 [24]	512	2.90 (118.87)	0.84 (0.23)	0.70 (0.51)	1.22 (0.11)	**0.97** (0.06)	82.10 (22.89)	0.71 (0.03)
EDM-2-XL-512 [24]	512	2.92 (141.74)	0.84 (0.25)	0.71 (0.45)	1.21 (0.12)	**0.97** (0.06)	83.23 (20.04)	**0.77** (0.03)
EDM-2-XXL-512 [24]	512	**2.87** (124.29)	0.84 (0.33)	0.71 (0.60)	1.22 (0.17)	**0.97** (0.07)	82.45 (21.52)	0.75 (0.05)

Text Embedding is Not All You Need: Attention Control for Text-to-Image Semantic Alignment with Text Self-Attention Maps

- Summary: This study investigated the issue where existing text-to-image models fail to faithfully follow the input text and proposed a solution. It identified that the problem arises because incorrect regions tend to be emphasized in the cross-attention map. The study revealed that, in text embeddings, tokens with high similarity are more likely to have high values in the cross-attention map.
- Novelty: Since the self-attention in the text encoder captures the structure of the text, the study improves faithfulness by encouraging the cross-attention to mimic the self-attention map through the exploration of latent variables.



(a) (b) (c)

CVPR25: Meta Insights into Trends and Tendencies (161/179)

DepthCrafter: Generating Consistent Long Depth Sequences for Open-world Videos

- Summary: Depth estimation for open-world videos has been challenging due to the diversity in camera motion and content layout. Existing methods like Depth Anything-v2 lack temporal consistency when applied to videos. To address this, this paper proposes a method that leverages a pre-trained image-to-video diffusion model for temporally consistent depth estimation.
- Novelty: To handle diverse videos, the model is trained using real videos for content diversity and synthetic videos for accurate depth supervision. Additionally, for precise conditioning, it leverages not only cross-attention but also conditioning in the VAE latent space. Moreover, the model adopts a training strategy that selectively fine-tunes either temporal or spatial layers, enabling it to handle long-frame videos efficiently.

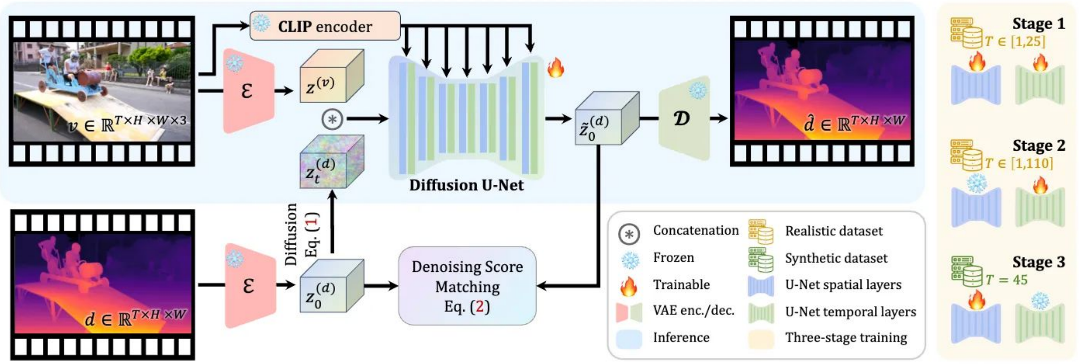


Figure 2. Overview of our *DepthCrafter*. It is a conditional diffusion model that models the distribution $p(\mathbf{d} \mid \mathbf{v})$ over the depth sequence $\mathbf{d}$ conditioned on the input video $\mathbf{v}$. We train the model in three stages, where the spatial or temporal layers of the diffusion model are progressively learned on our compiled realistic or synthetic datasets with variable lengths T. During inference, given an open-world video, it can generate temporally consistent long depth sequences with fine-grained details for the entire video from initialized Gaussian noise, without requiring any supplementary information, such as camera poses or optical flow.

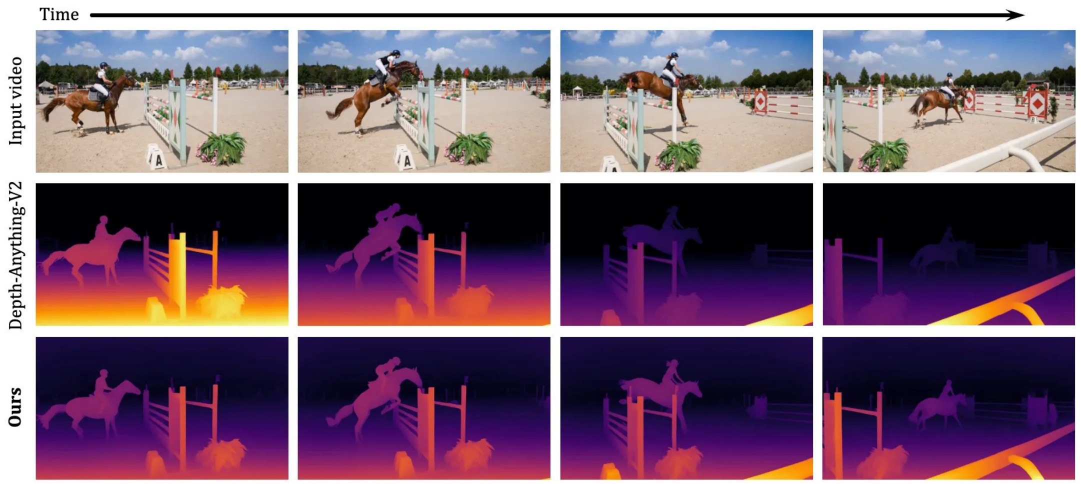
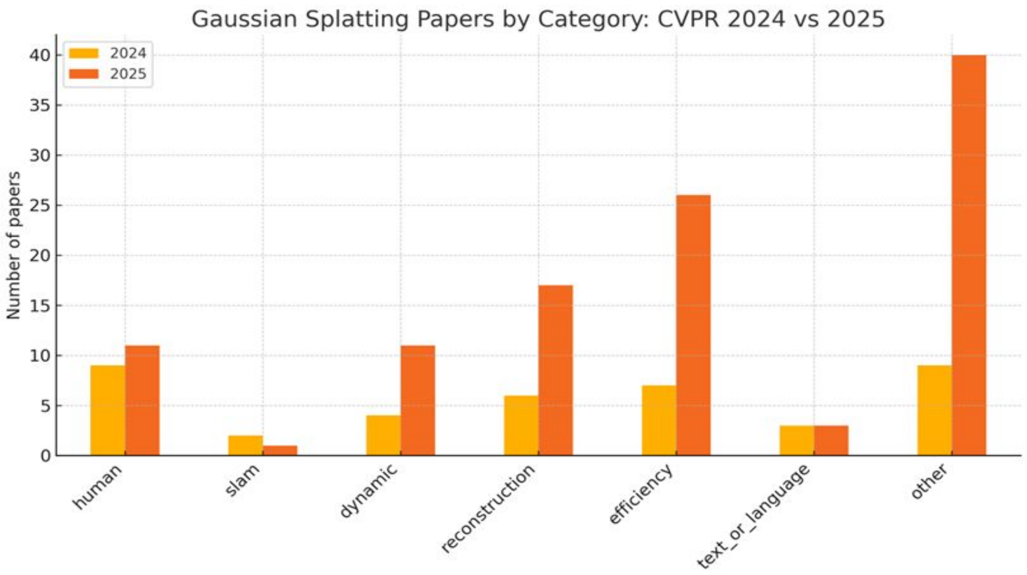


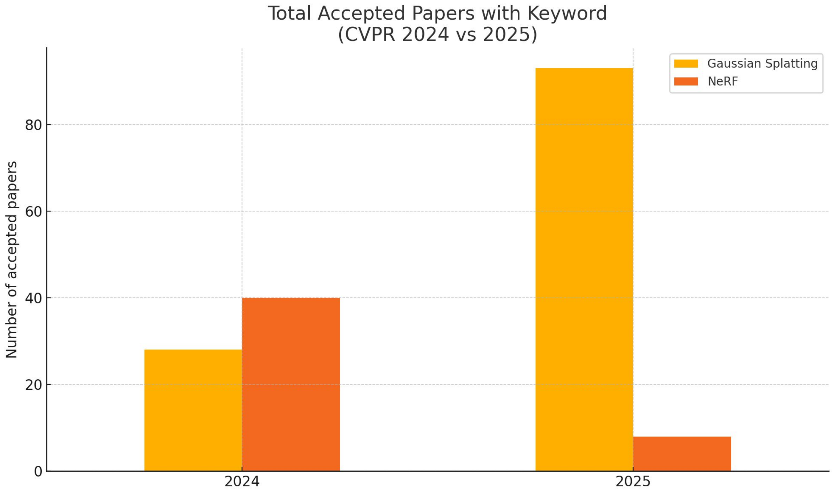
Figure 1. We innovate DepthCrafter, a novel video depth estimation approach, that can generate temporally consistent long depth sequences with fine-grained details for open-world videos, without requiring additional information such as camera poses or optical flow.

CVPR25: Meta Insights into Trends and Tendencies (162/179)

Trends of Gaussian Splatting

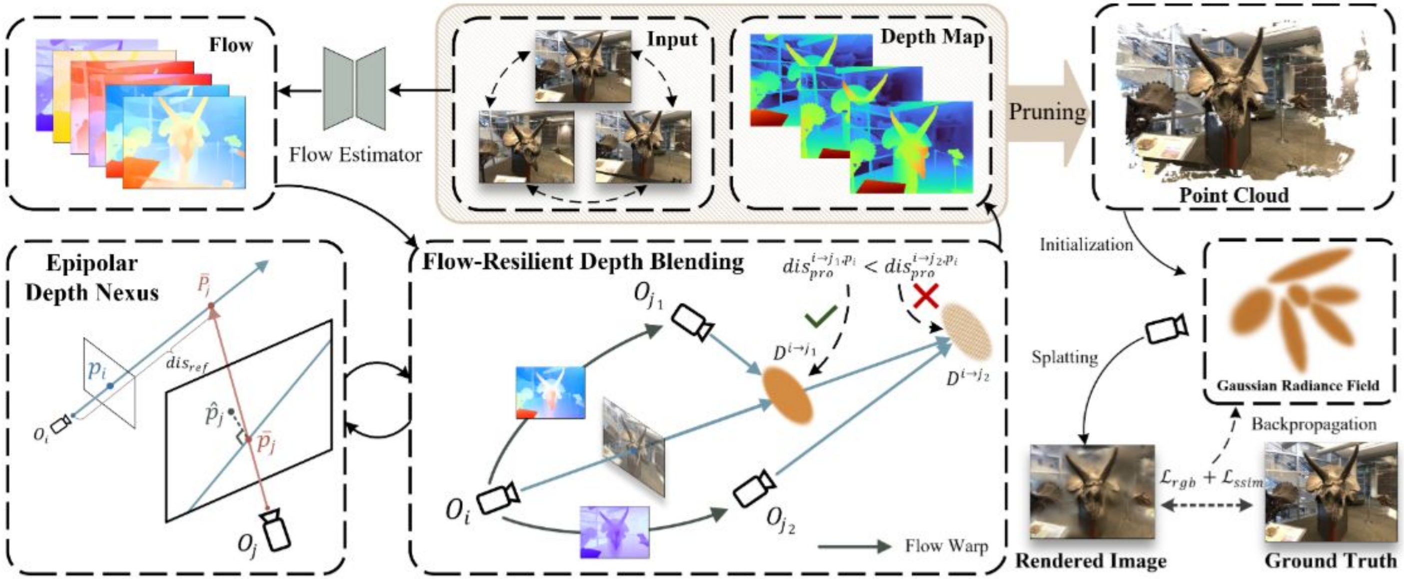
- 93 papers that contain "Gaussian Splatting" in the titles in 2025 vs 28 in 2024 (≈3.3× growth).
  - Gaussian splatting is attracting attention as 3D scene reconstruction is one of the areas of focus at CVPR 2025.
- Comparison with CVPR 2024
  - Efficiency / sparse methods: 26 papers (fastest-growing)
  - Scene reconstruction & dynamic scenes ≈ tripled
  - Human-avatar share shrank from 32% → 12%
- Gaussian Splatting shifts from niche to default for real-time 3D
  - Replacing NeRF as the dominant 3D scene reconstruction method.

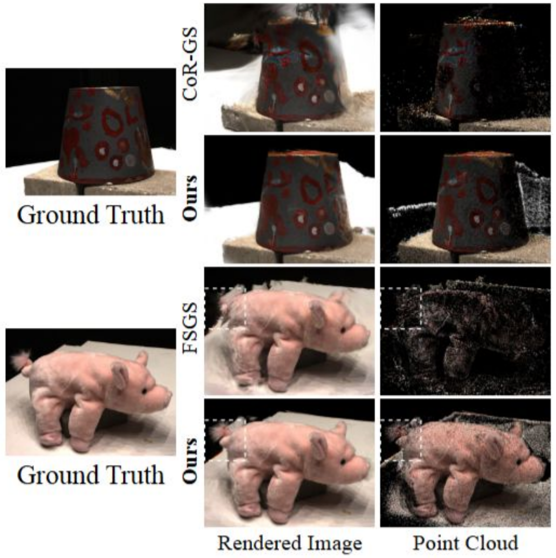




NexusGS: Sparse View Synthesis with Epipolar Depth Priors in 3D Gaussian Splatting

- Summary: This work embeds epipolar depth priors into 3D Gaussian Splatting to improve sparse-view novel view synthesis.
- Novelty: It introduces a point cloud densification strategy initializing dense Gaussians with epipolar depth priors via optical flow, employing Flow-Resilient Depth Blending and Flow-Filtered Depth Pruning to suppress flow errors and produce accurate depth under sparse views.
- Thought: NexusGS cleverly uses epipolar geometry to overcome the limitations of sparse views, providing robust depth initialisation without the need for heavy regularisation. Its effective three-step process relies on optical flow, which may struggle in textureless regions.



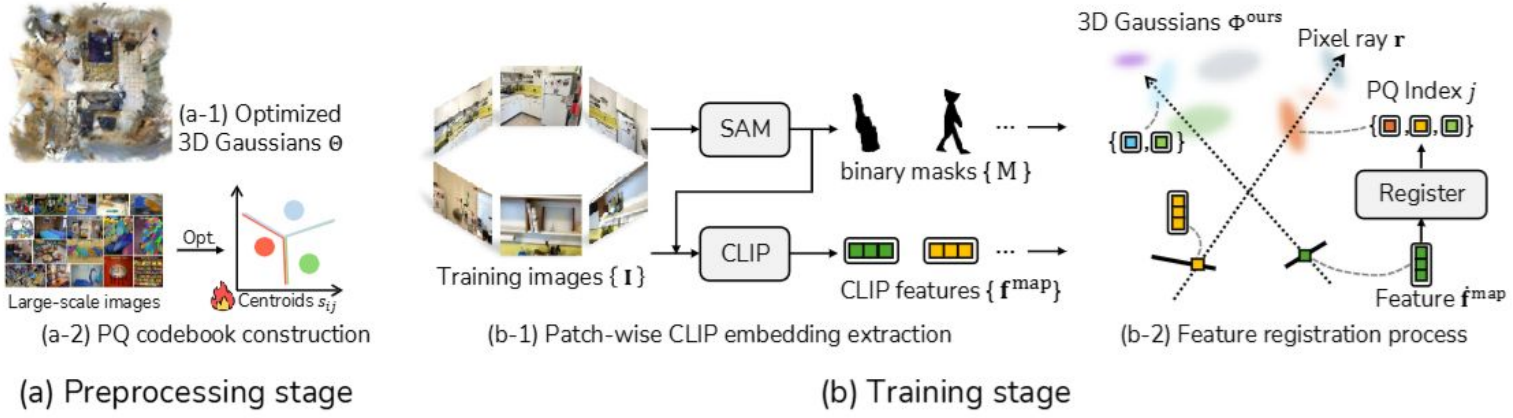


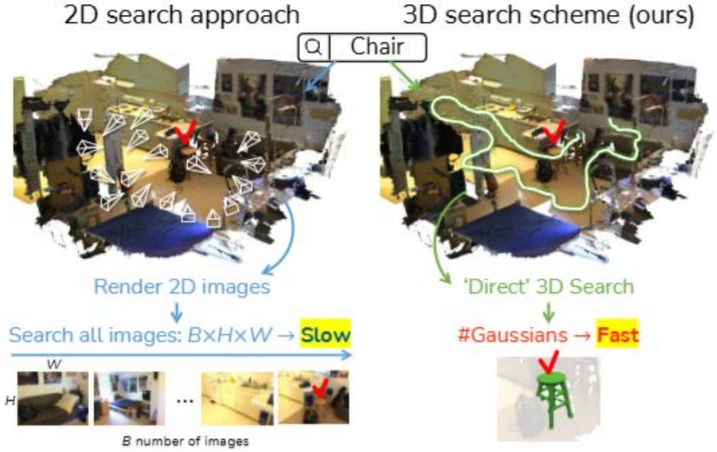
[Y. Zheng et al, CVPR 2025.][Link]

CVPR25: Meta Insights into Trends and Tendencies (164/179)

Dr. Splat: Directly Referring 3D Gaussian Splatting via Direct Language Embedding Registration

- Summary: Dr. Splat directly registers CLIP embeddings to 3D Gaussians, bypassing rendering for efficient open-vocabulary 3D scene understanding.
- Novelty: It introduces a direct feature registration technique that assigns CLIP embeddings to dominant 3D Gaussians and integrates pretrained Product Quantization for compact embedding representation without per-scene optimization.
- Thought: I find the idea of eliminating the rendering stage compelling. It reduces feature distortion and speeds up 3D queries. Pretrained PQ balances memory and accuracy effectively.

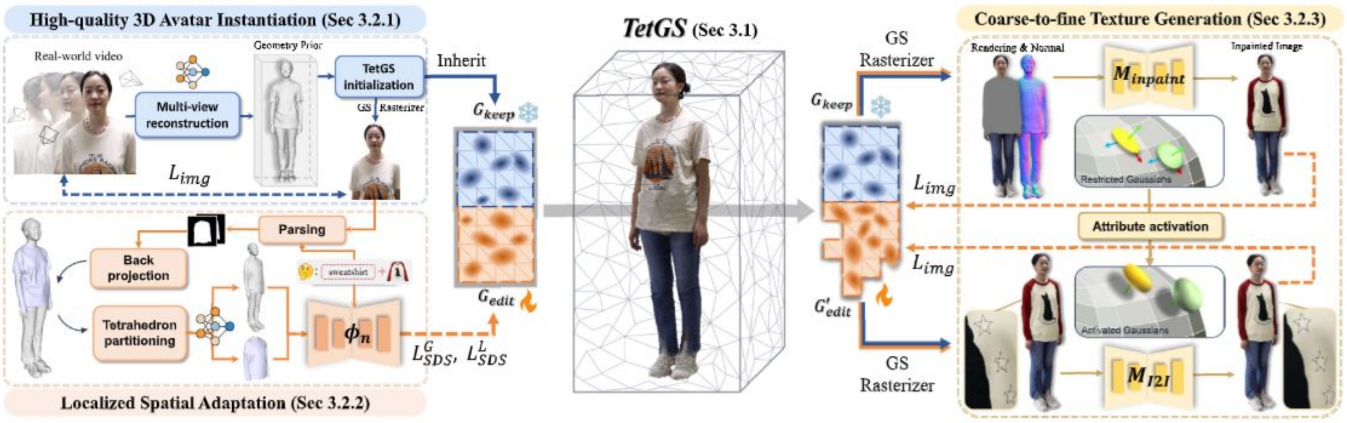

(a-1) Optimized 3D Gaussians Θ
Large-scale images
Opt.
Centroids s_{ij}
(a-2) PQ codebook construction
(a) Preprocessing stage
Training images { I }
SAM
CLIP
binary masks { M }
CLIP features { $\mathbf{f}^{\text{map}}$ }
(b-1) Patch-wise CLIP embedding extraction
3D Gaussians Φ^{ours}
Pixel ray $\mathbf{r}$
PQ Index j
Register
Feature $\mathbf{f}^{\text{map}}$
(b-2) Feature registration process
(b) Training stage

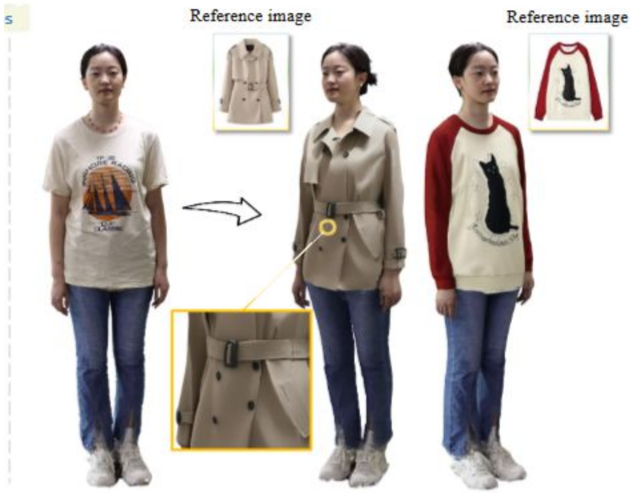


[K. Jun-Seong et al., CVPR 2025.][Link]

Creating Your Editable 3D Photorealistic Avatar with Tetrahedron-constrained Gaussian Splatting

- ❑ Summary: A hybrid TetGS-based pipeline generates editable photorealistic 3D avatars from monocular videos with text/image-guided edits.
- ❑ Novelty: This work proposes TetGS: embedding Gaussian kernels in tetrahedral grids for decoupled spatial adaptation and appearance learning, enabling precise localized geometry edits and photorealistic texture generation with coarse-to-fine and few-shot supervision.
- ❑ Thought: This paper addresses the issue of instability in 3D Gaussian Splatting editing by integrating structured tetrahedral grids. This approach offers clear geometric control and high-fidelity results. Although computationally intensive, the paper's staged pipeline and hybrid representation represent a significant advance in the creation of practical, user-friendly 3D avatars.





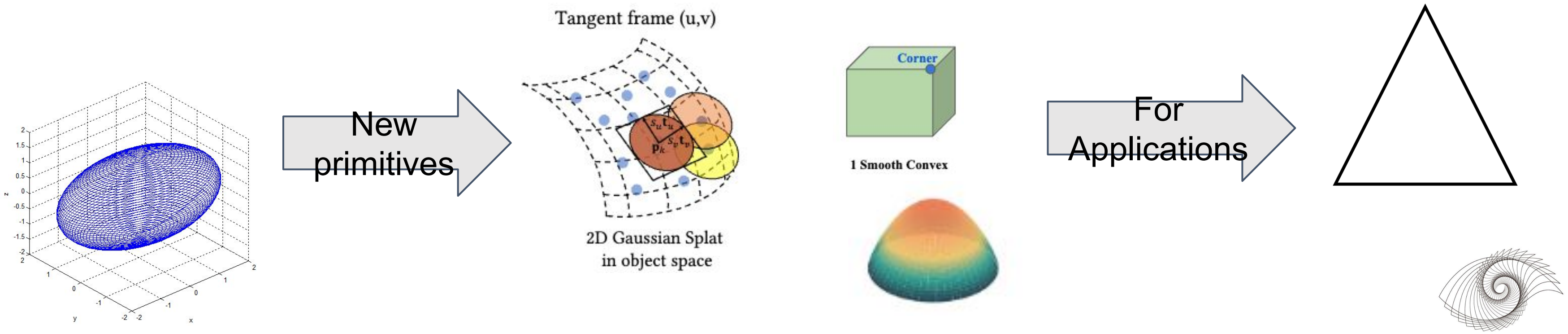
(d) Image-guided editing

[H. Liu et al., CVPR 2025.][Link]

CVPR25: Meta Insights into Trends and Tendencies (166/179)

From Photorealism to Geometric Integrity

- 3DGS weakness: geometrically inaccurate surfaces/blurry edges/multi-view inconsistency.
  - The primary research trend has decisively shifted from simply improving visual quality to correcting these geometric flaws
- The "Splat" is No Longer Just 3D blob (Gaussian).
  - To achieve geometric accuracy, a wave of new primitives has emerged.
    - **2D Surfel Splatting** for surface consistency
    - **3D Convex Splatting** for sharp-edged objects
    - **Deformable Beta Splatting** for higher fidelity with fewer parameters
- Integration with Graphics Pipelines
  - While advanced primitives were introduced, they were still custom representations.The ultimate goal for many applications, such as games and VFX, remains generating standard triangular meshes
    - **Triangle Splatting** represents a potential convergence point for the field, making the fundamental graphics primitive directly optimizable.

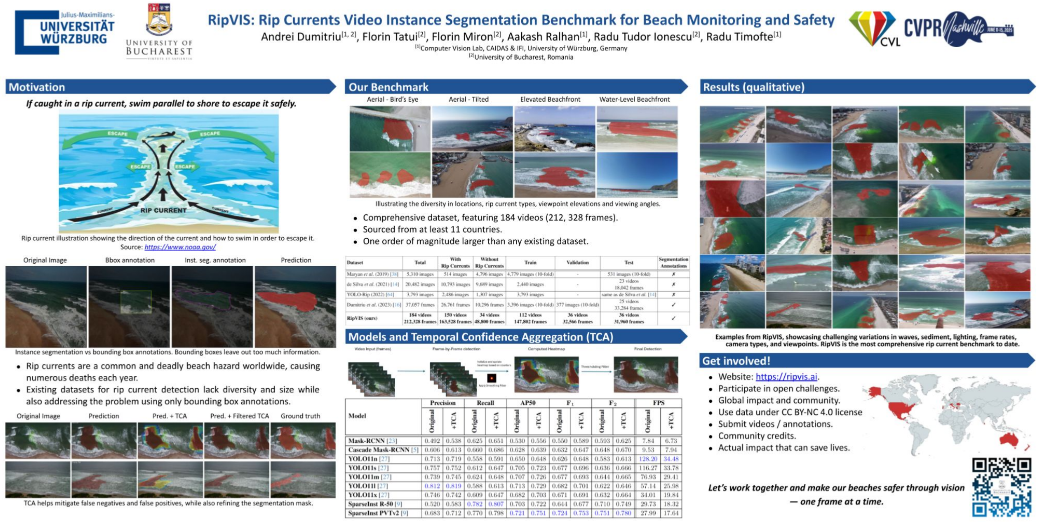

New primitives

For Applications

RipVIS: Rip Currents Video Instance Segmentation Benchmark for Beach Monitoring and Safety

- **Summary** : Creating large-scale video instance segmentation benchmark (RipVIS) for detecting rip currents. (Author said it took around 3 years to complete this)

- **Comment** : Author said Japan already has a nice system to predict rip current but it is closed. So they want to find Japanese collaborator. A competition based on RipVIS will be held in ICCV 2025.



RipVIS: Rip Currents Video Instance Segmentation Benchmark for Beach Monitoring and Safety

Andrei Dumitriu[1, 2], Florin Tatui[2], Florin Miron[2], Aakash Ralhan[1], Radu Tudor Ionescu[2], Radu Timofte[1]

[1]Computer Vision Lab, CAIDAS & IFI, University of Würzburg, Germany

[2]University of Bucharest, Romania

Motivation

If caught in a rip current, swim parallel to shore to escape it safely.

Rip current illustration showing the direction of the current and how to swim in order to escape it. Source: https://www.noaa.gov/

Instance segmentation vs bounding box annotations. Bounding boxes leave out too much information.

- Rip currents are a common and deadly beach hazard worldwide, causing numerous deaths each year.
- Existing datasets for rip current detection lack diversity and size while also addressing the problem using only bounding box annotations.

TCA helps mitigate false negatives and false positives, while also refining the segmentation mask.

Our Benchmark

Illustrating the diversity in locations, rip current types, viewpoint elevations and viewing angles.

- Comprehensive dataset, featuring 184 videos (212, 328 frames).
- Sourced from at least 11 countries.
- One order of magnitude larger than any existing dataset.

Dataset	Total	With Rip Currents	Without Rip Currents	Train	Validation	Test	Segmentation Annotations
Maryan *et al.* (2019) [38]	5,310 images	514 images	4,796 images	4,779 images (10-fold)	-	531 images (10-fold)	✗
de Silva *et al.* (2021) [14]	20,482 images	10,793 images	9,689 images	2,440 images	-	23 videos 18,042 frames	✗
YOLO-Rip (2022) [64]	3,793 images	2,486 images	1,307 images	3,793 images	-	same as de Silva *et al.* [14]	✗
Dumitriu *et al.* (2023) [16]	37,057 frames	26,761 frames	10,296 frames	3,396 images (10-fold)	377 images (10-fold)	25 videos 33,284 frames	✓
RipVIS (ours)	**184 videos 212,328 frames**	**150 videos 163,528 frames**	**34 videos 48,800 frames**	**112 videos 147,802 frames**	**36 videos 32,566 frames**	**36 videos 31,960 frames**	✓

Models and Temporal Confidence Aggregation (TCA)

Model	Precision		Recall		AP50		F_1		F_2		FPS	
	Original	+TCA	Original	+TCA	Original	+TCA	Original	+TCA	Original	+TCA	Original	+TCA
Mask-RCNN [23]	0.492	0.538	0.625	0.651	0.530	0.556	0.550	0.589	0.593	0.625	7.84	6.73
Cascade Mask-RCNN [5]	0.606	0.613	0.660	0.686	0.628	0.639	0.632	0.647	0.648	0.670	9.53	7.94
YOLO11n [27]	0.713	0.719	0.558	0.591	0.650	0.648	0.626	0.648	0.583	0.613	128.20	34.48
YOLO11s [27]	0.757	0.752	0.612	0.647	0.705	0.723	0.677	0.696	0.636	0.666	116.27	33.78
YOLO11m [27]	0.739	0.745	0.624	0.648	0.707	0.726	0.677	0.693	0.644	0.665	76.93	29.41
YOLO11l [27]	0.812	0.819	0.588	0.613	0.713	0.729	0.682	0.701	0.622	0.646	57.14	25.98
YOLO11x [27]	0.746	0.742	0.609	0.647	0.682	0.703	0.671	0.691	0.632	0.664	34.01	19.84
SparseInst R-50 [9]	0.520	0.583	0.782	0.807	0.703	0.722	0.644	0.677	0.710	0.749	29.73	18.32
SparseInst PVTv2 [9]	0.683	0.712	0.770	0.798	0.721	0.751	0.724	0.753	0.751	0.780	27.99	17.64

Results (qualitative)

Examples from RipVIS, showcasing challenging variations in waves, sediment, lighting, frame rates, camera types, and viewpoints. RipVIS is the most comprehensive rip current benchmark to date.

Get involved!

- Website: https://ripvis.ai.
- Participate in open challenges.
- Global impact and community.
- Use data under CC BY-NC 4.0 license
- Submit videos / annotations.
- Community credits.
- Actual impact that can save lives.

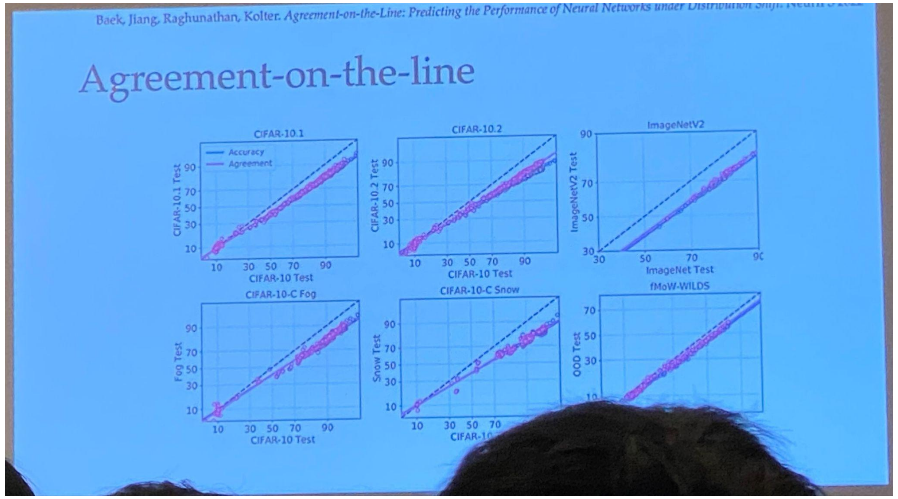
Let's work together and make our beaches safer through vision — one frame at a time.

- [Project page](#)
- [Rip Current Instance Segmentation Challenge](#) (ICCV 2025)

CVPR25: Meta Insights into Trends and Tendencies (168/179)

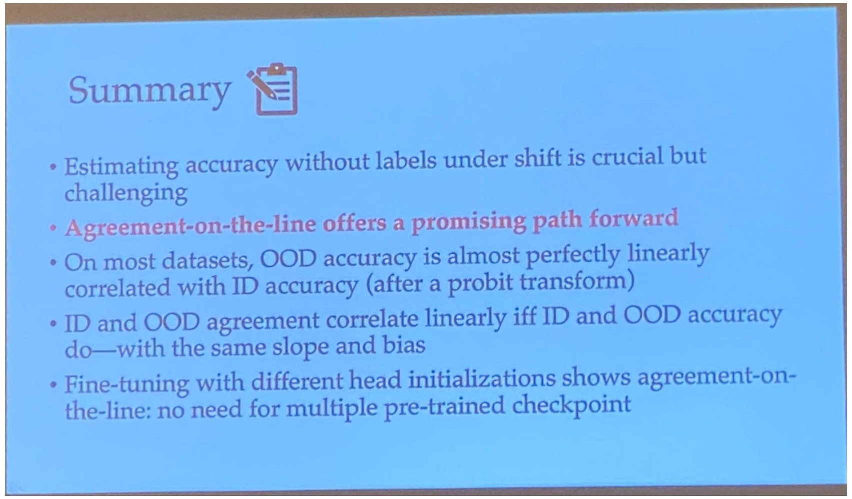
Workshop: Domain Generalization: Evolution, Breakthroughs and Future Horizon

- Speaker: Aditi Raghunathan
  Title: Predicting the Performance of Foundation Models Under Distribution Shift
  - OOD accuracy is almost linearly correlated with ID accuracy.
  - ID and OOD agreement also correlate linearly iff ID and ODD accuracy do.
    - This empirical phenomena will make possible to **estimate FM performance under distribution shift with unlabeled data** . (paper)
  - But computing agreement over multiple FMs is not computationally feasible.
    - **Random initializations of linear heads** works (no need for multiple checkpoint).



Baek, Jiang, Raghunathan, Kolter. *Agreement-on-the-Line: Predicting the Performance of Neural Networks under Distribution Shift.* NeurIPS 2022

Agreement-on-the-line



Summary

- Estimating accuracy without labels under shift is crucial but challenging
- Agreement-on-the-line offers a promising path forward
- On most datasets, OOD accuracy is almost perfectly linearly correlated with ID accuracy (after a probit transform)
- ID and OOD agreement correlate linearly iff ID and OOD accuracy do—with the same slope and bias
- Fine-tuning with different head initializations shows agreement-on-the-line: no need for multiple pre-trained checkpoint

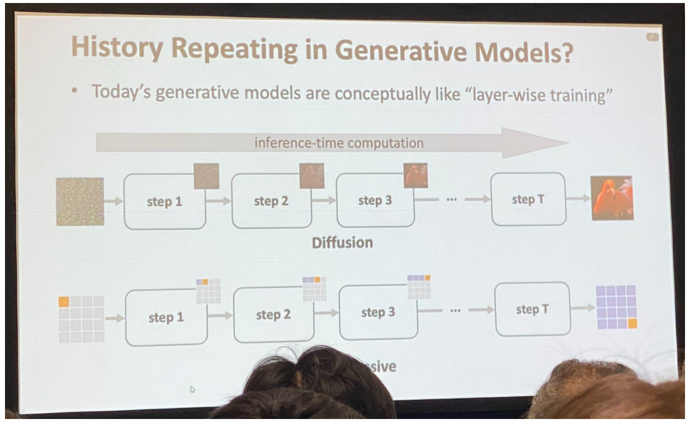
CVPR25: Meta Insights into Trends and Tendencies (169/179)

Workshop: Visual Generative Modeling: What’s After Diffusion?

- Speaker: Kaiming He

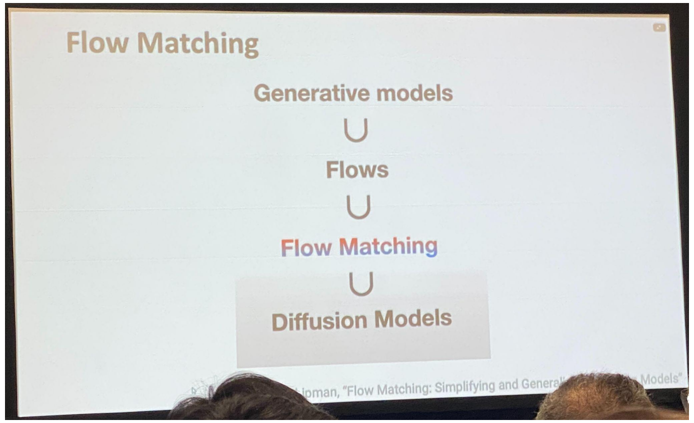
  Title: Towards End-to-End Generative Modeling

  - Before AlexNet layer-wise training was a popular solution in recognition.
  - Today’s generative models are still conceptually like “layer-wise training”.
    - Therefore, there may be significant room for improvement **if the generative model can be made end-to-end** .
  - Approaches based on **Flow** appear to be a promising direction at present.
    - Latest work: Mean Flows for One-step Generative Modeling



History Repeating in Generative Models?

- Today’s generative models are conceptually like “layer-wise training”



Flow Matching

Generative models

∪

Flows

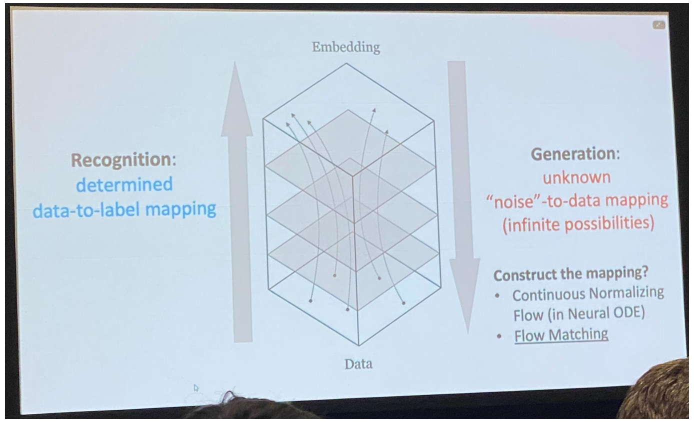
∪

Flow Matching

∪

Diffusion Models

ipman, “Flow Matching: Simplifying and General Models”



Recognition:
determined
data-to-label mapping

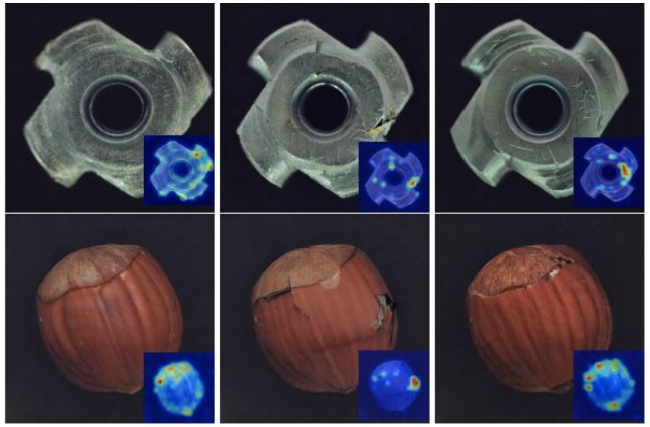
Generation:
unknown
“noise”-to-data mapping
(infinite possibilities)

Construct the mapping?

- Continuous Normalizing Flow (in Neural ODE)
- Flow Matching

Unseen Visual Anomaly Generation

- Summary:
  - Proposed model (AnomalyAny) generates realistic and diverse anomaly images at test time using only normal samples and a pre-trained Stable Diffusion model.
- Novelty:
  - It uniquely leverages cross-attention-guided optimization and prompt refinement to steer anomaly generation without any additional training.
- Thoughts:
  - The zero-training approach is elegant, though incorporating real one-shot anomaly exemplars could further boost generation precision.



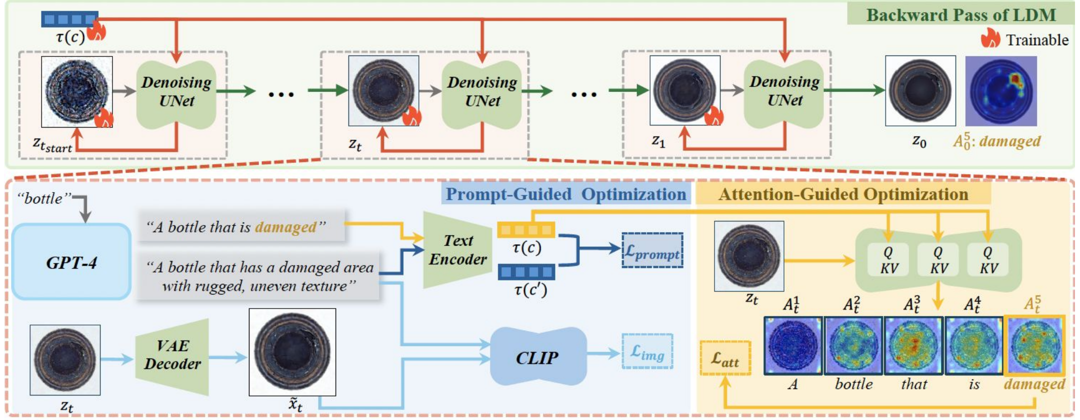


Figure 2. **Illustration of AnomalyAny** with details of the attention-guided & prompt-guided optimization process at time step t.

Workshop on Video Large Language Models

- ❑ Evaluation benchmark
  - ❑ Hallucination and Omission (ARGUS)
  - ❑ Compositional reasoning (VELOCITI)
  - ❑ Road event (RoadSocial)
  - ❑ Temporal understanding (Lost in Time)
- ❑ Efficient handling of long videos
  - ❑ Multi-frame fusion (FiLA-Video)
  - ❑ State-space model (BIMBA)
  - ❑ Frame sampling (Moment Sampling)
- ❑ Video grounding
  - ❑ Weakly Supervised (CoSPaL, STPro)
  - ❑ Low cost (NumPro)
  - ❑ Pixel grounding (PG-Video-LLaVA)
- ❑ Others
  - ❑ Current bottlenecks in training
  - ❑ Zero-shot Action Localization

BIMBA: Selective-Scan Compression for Long-Range Video Question Answering

❑ Summary

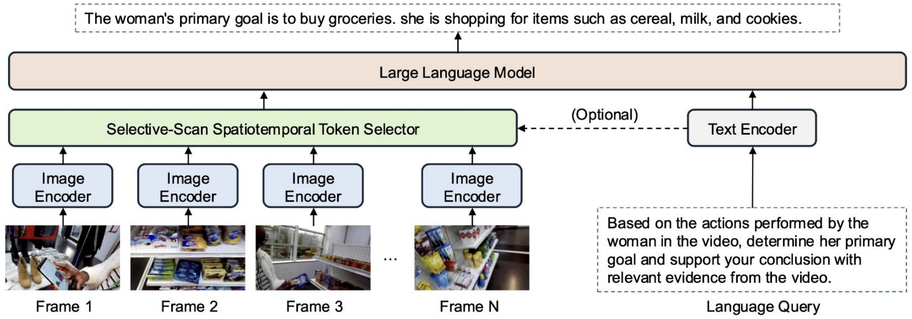
- Video understanding with MLLMs face considerable difficulties when dealing with long videos, primarily due to the prohibitive quadratic computational cost of the self-attention mechanism, which is central to processing vast numbers of spatiotemporal tokens.

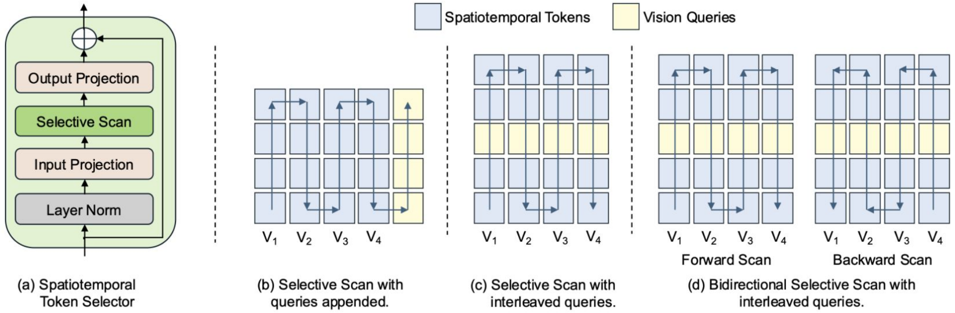
❑ Novelty

- BIMBA uses Mamba's efficient selective scan to extract key information from high-dimensional video, reducing it to a compact, information-rich token sequence.

❑ Thoughts

- The "Vanilla" (linear, Q-former) method quickly runs into GPU memory issues, and pooling struggles to capture long-range dependencies. Therefore, SSM-based methods, which can efficiently model temporal information, are expected to advance further for video understanding.


The woman's primary goal is to buy groceries. she is shopping for items such as cereal, milk, and cookies.
Large Language Model
Selective-Scan Spatiotemporal Token Selector
(Optional)
Text Encoder
Image Encoder
Image Encoder
Image Encoder
Image Encoder
...
Frame 1
Frame 2
Frame 3
Frame N
Based on the actions performed by the woman in the video, determine her primary goal and support your conclusion with relevant evidence from the video.
Language Query



(a) Spatiotemporal Token Selector

(b) Selective Scan with queries appended.

(c) Selective Scan with interleaved queries.

(d) Bidirectional Selective Scan with interleaved queries.

CVPR25: Meta Insights into Trends and Tendencies (173/179)

Watermark

- ❑ For generative model
  - ❑ Robust Watermark against Fine-Tuning
  - ❑ Box-Free Watermark Removal
  - ❑ Black-Box Forgery Attacks on Semantic Watermarks for Diffusion Models
- ❑ Localized watermark/Partial theft
  - ❑ Robust Watermarking Scheme Against Partial Image Theft
  - ❑ Manipulation Localization
- ❑ For 3D Gaussian Splatting
  - ❑ GuardSplat, 3D-GSW
- ❑ Others
  - ❑ Robust Message Steganography
  - ❑ Open-source Dataset Copyright

CVPR25: Meta Insights into Trends and Tendencies (174/179)

Video LLM

- Frame/Token Selection and Compression
  - Dynamic Compression of Tokens
  - Adaptive Keyframe Sampling, Flexible Frame Selection, M-LLM Based Video Frame Selection
- Temporal understanding
  - Sequential Knowledge Transfer
  - Consistency of Video Large Language Models in Temporal Comprehension
  - Fine-Grained Compositional and Temporal Alignment
- Spatial-Temporal Grounding
  - VideoRefer Suite, LLaVA-ST, VideoGLaMM
- Others
  - Mitigating Action-Scene Hallucination
  - Real-time interactive streaming

CVPR25: Meta Insights into Trends and Tendencies (175/179)

Hand–Object Interaction

- ❏ Dataset
  - ❏ HD-EPIC: A Highly-Detailed Egocentric Video Dataset
  - ❏ EgoPressure: A Dataset for Hand Pressure and Pose Estimation in Egocentric Vision
- ❏ 3D Generation
  - ❏ EasyHOI: Unleashing the Power of Large Models for Reconstructing Hand-Object Interactions in the Wild
  - ❏ LatentHOI: On the Generalizable Hand Object Motion Generation with Latent Hand Diffusion
  - ❏ HOIGPT: Learning Long-Sequence Hand-Object Interaction with Language Models

CVPR25: Meta Insights into Trends and Tendencies (176/179)

Workshop： 4D vision Modeling the Dynamic World

1. 4D Gaussian Splatting
   4D LangSplat: 4D Language Gaussian Splatting via Multimodal Large Language Models 4D LangSplat
   TRL-GS: Cascaded Temporal Residue Learning for 4D GS TRL-GS

2. 3D Rigging, 4D object intrinsics generation
   CVPR Poster Category-Agnostic Neural Object Rigging
   Birth and Death of a Rose
   Anymate: A Dataset and Baselines for Learning 3D Object Rigging

3. 4D Reconstruction
   St4RTrack: Simultaneous 4D Reconstruction and Tracking in the World St4rtrack

CVPR25: Meta Insights into Trends and Tendencies (177/179)

Insight : 4D vision Modeling the Dynamic World

・Gaussian Splatting is also powerful in 4D [4D LangSplat]

- 4D LangSplat embeds multimodal-LLM-generated, temporally consistent object captions into a dynamic 4D Gaussian-Splatting scene representation, enabling open-vocabulary and time-aware semantic querying.
-

・Temporal object intrinsics generation

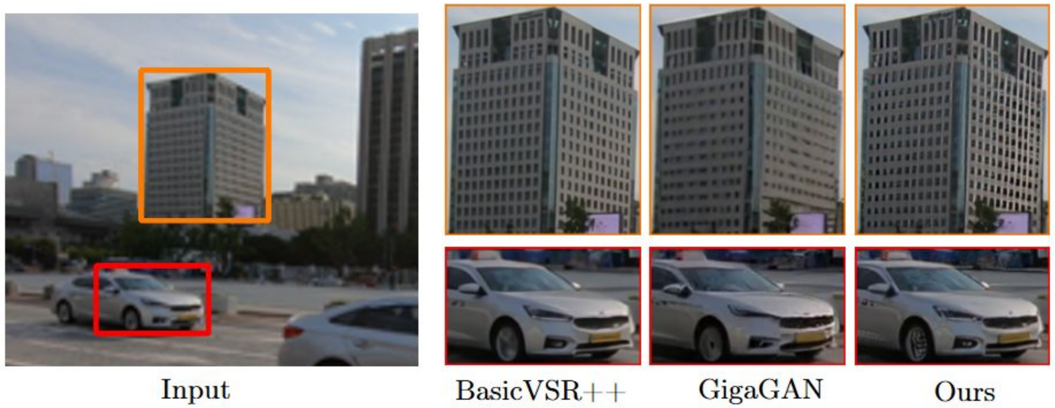
- The rose of Death learn temporal object intrinsics by distilled it's 4D prior from a 2d diffusion.

  [Birth and Death of a Rose]

CVPR25: Meta Insights into Trends and Tendencies (178/179)

VideoGigaGAN: Towards Detail-rich Video Super-Resolution

- Summery
  - The proposed model (Video-GAN) achieves both high quality and temporal consistency across output frames, delivering higher detail than existing models.
- Novelty
  - VideoGigaGAN can produce video results with both high-frequency details and temporal consistency while artifacts like aliasing are significantly mitigated.
- Thoughts
  - In video super-resolution, the issue of achieving both frame-to-frame consistency and quality is likely to continue to be debated.

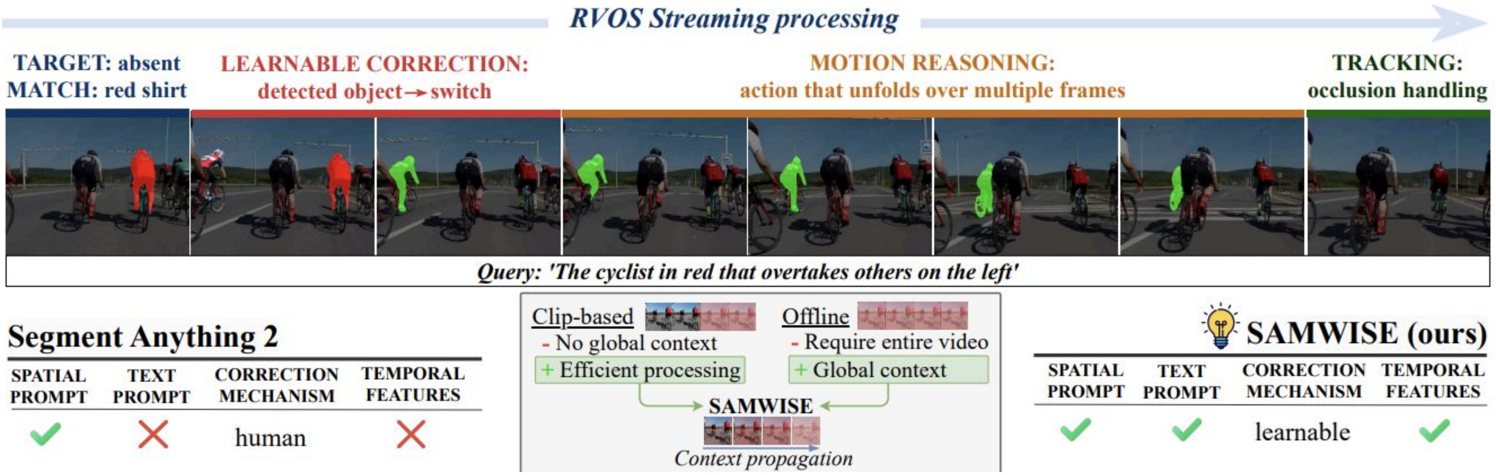


Yiran Xu et al. “VideoGigaGAN: Towards Detail-rich Video Super-Resolution”, in CVPR 2025.

CVPR25: Meta Insights into Trends and Tendencies (179/179)

SAMWISE: Infusing Wisdom in SAM2 for Text-Driven Video Segmentation(Highlight)

- ❑ **Summary:**
  - ❑ SAMWISE leverages SAM2 for text-driven video segmentation by integrating natural language understanding and temporal modeling without fine-tuning, while also introducing a novel module to adjust SAM2's inherent tracking bias for improved performance on dynamic objects.
- ❑ **Novelty:**
  - ❑ Efficiently achieves text-driven video segmentation without re-training the foundation model, while overcoming SAM2's tracking limitations.
- ❑ **Thoughts:**
  - ❑ Similar to last year, we're still seeing a large number of derivative studies building upon SAM and SAM2. I think the integration of text prompts without fine-tuning is particularly impressive.

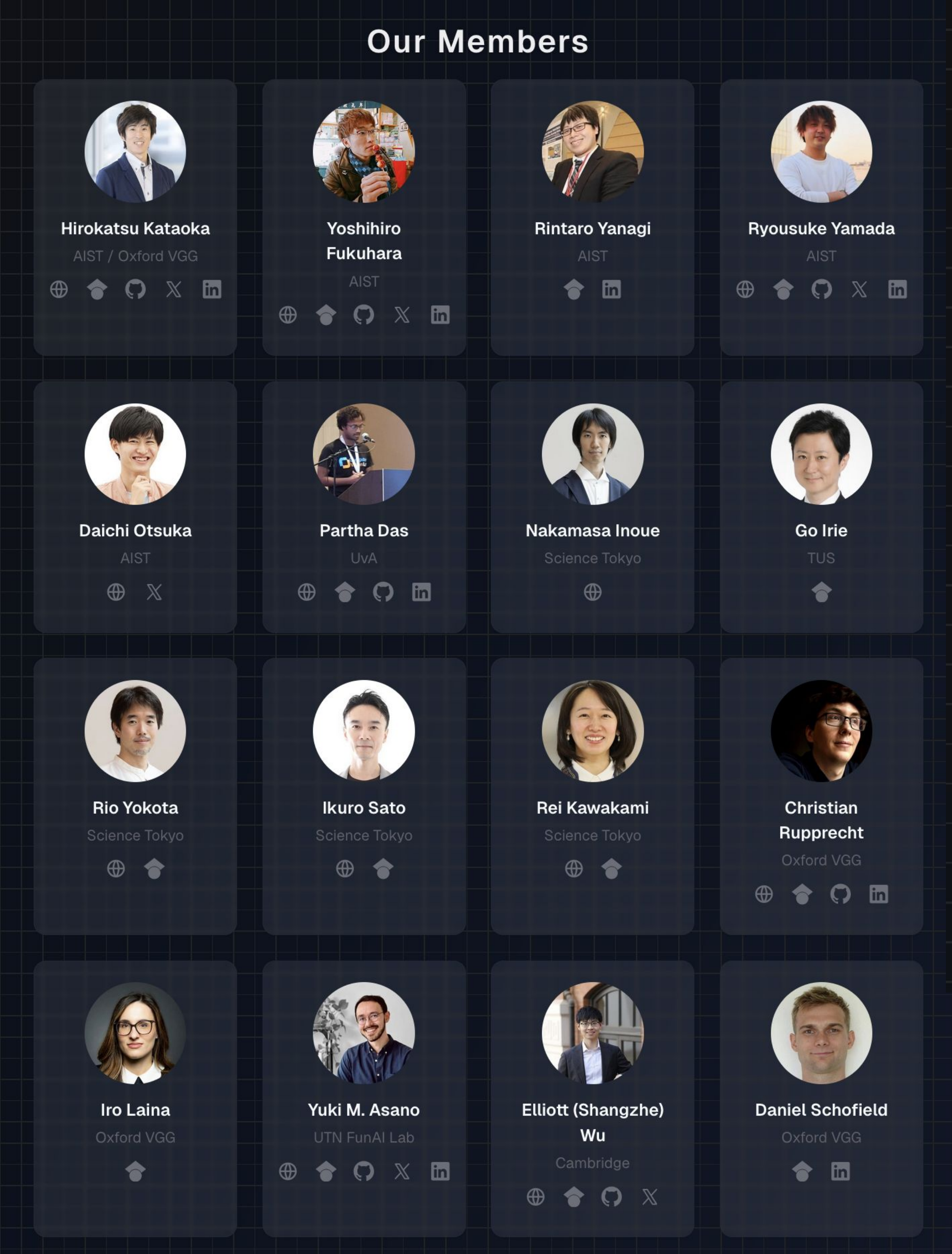


LIMIT.LAB

LIMIT.Lab

A collaboration hub

for building multimodal AI models under limited resources



Our Members

Hirokatsu Kataoka
AIST / Oxford VGG

Yoshihiro Fukuhara
AIST

Rintaro Yanagi
AIST

Ryousuke Yamada
AIST

Daichi Otsuka
AIST

Partha Das
UvA

Nakamasa Inoue
Science Tokyo

Go Irie
TUS

Rio Yokota
Science Tokyo

Ikuro Sato
Science Tokyo

Rei Kawakami
Science Tokyo

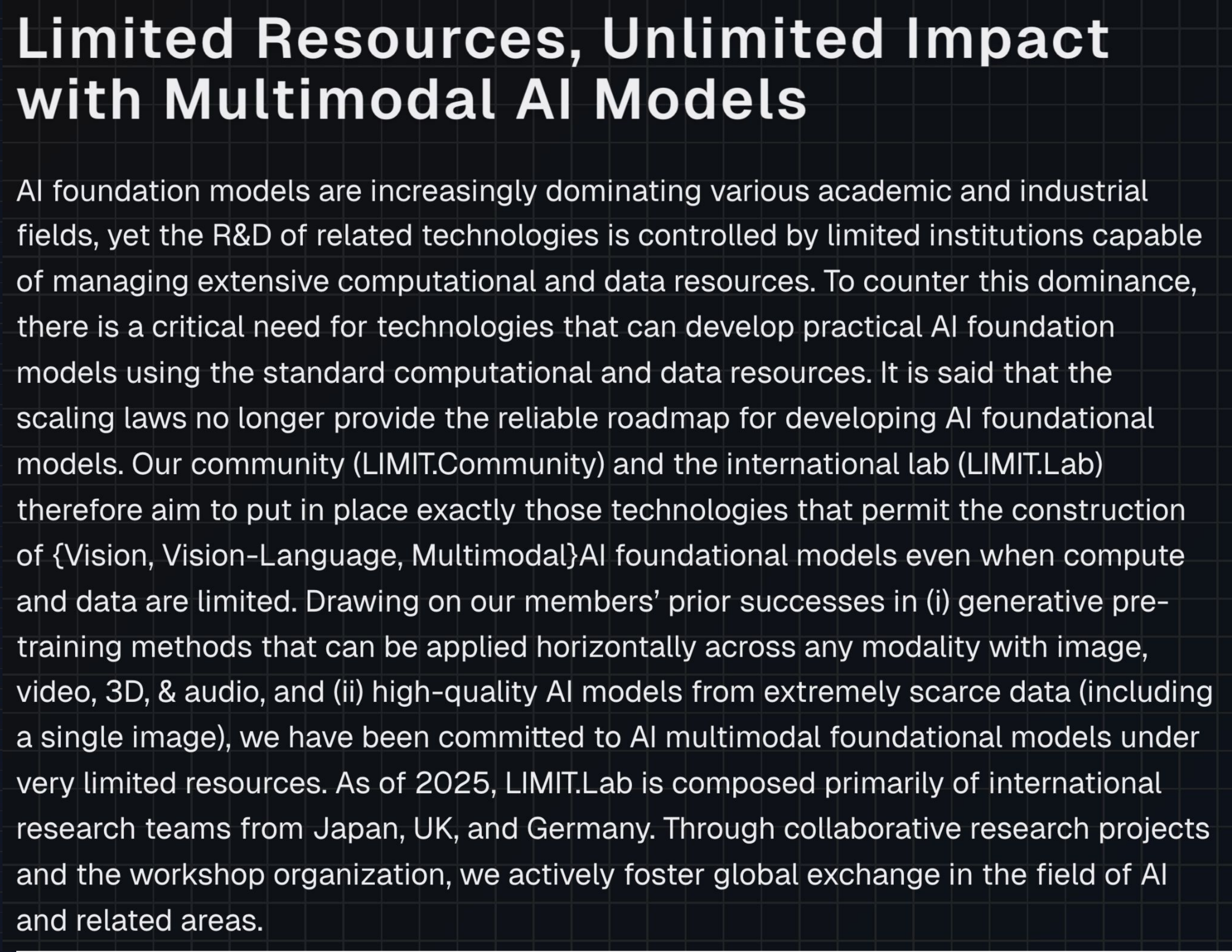
Christian Rupprecht
Oxford VGG

Iro Laina
Oxford VGG

Yuki M. Asano
UTN FunAI Lab

Elliott (Shangzhe) Wu
Cambridge

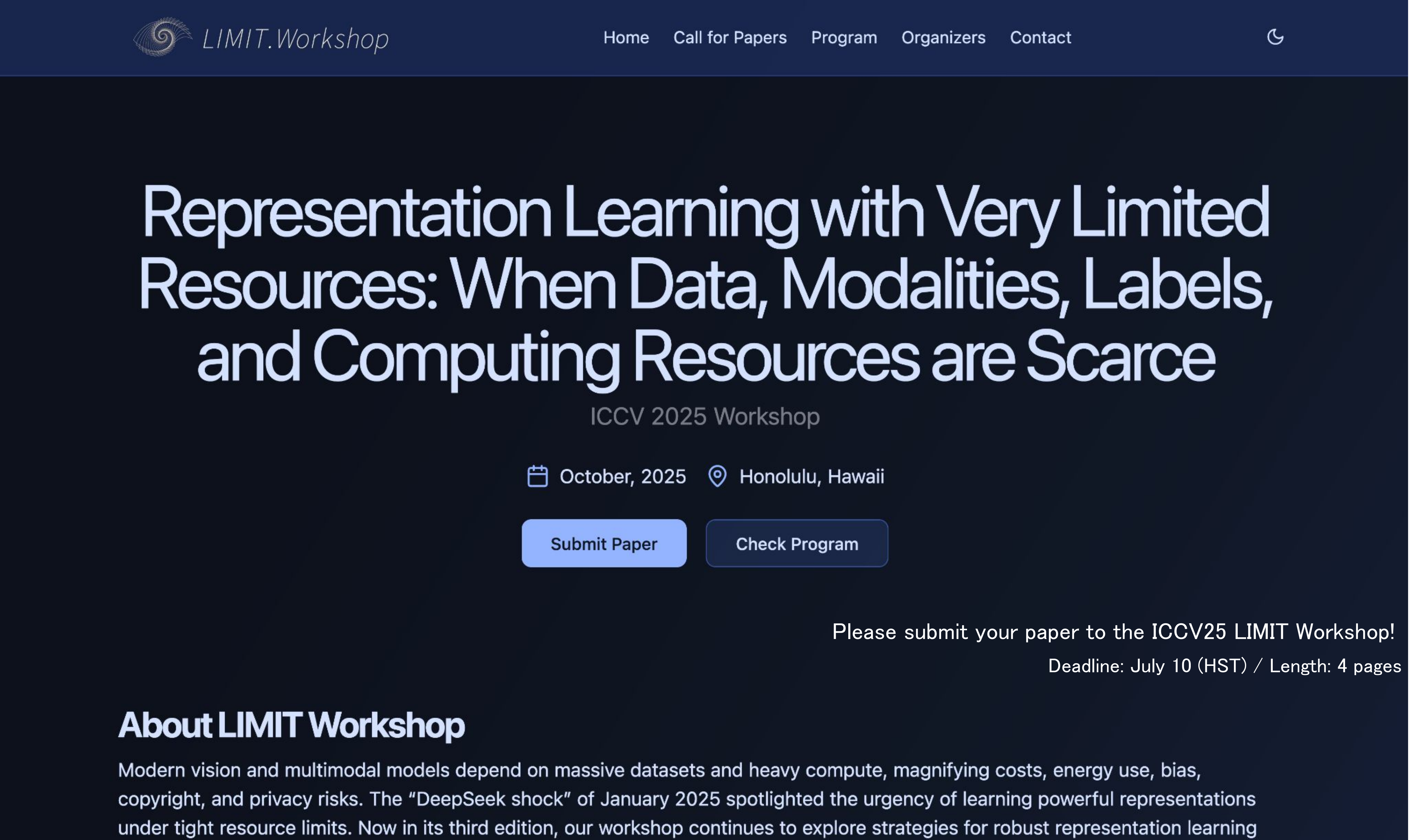
Daniel Schofield
Oxford VGG



Limited Resources, Unlimited Impact with Multimodal AI Models

AI foundation models are increasingly dominating various academic and industrial fields, yet the R&D of related technologies is controlled by limited institutions capable of managing extensive computational and data resources. To counter this dominance, there is a critical need for technologies that can develop practical AI foundation models using the standard computational and data resources. It is said that the scaling laws no longer provide the reliable roadmap for developing AI foundational models. Our community (LIMIT.Community) and the international lab (LIMIT.Lab) therefore aim to put in place exactly those technologies that permit the construction of {Vision, Vision-Language, Multimodal}AI foundational models even when compute and data are limited. Drawing on our members' prior successes in (i) generative pre-training methods that can be applied horizontally across any modality with image, video, 3D, & audio, and (ii) high-quality AI models from extremely scarce data (including a single image), we have been committed to AI multimodal foundational models under very limited resources. As of 2025, LIMIT.Lab is composed primarily of international research teams from Japan, UK, and Germany. Through collaborative research projects and the workshop organization, we actively foster global exchange in the field of AI and related areas.

Left: Core members / Right: Our mission



Representation Learning with Very Limited Resources: When Data, Modalities, Labels, and Computing Resources are Scarce

ICCV 2025 Workshop

October, 2025 Honolulu, Hawaii

Please submit your paper to the ICCV25 LIMIT Workshop!

Deadline: July 10 (HST) / Length: 4 pages

About LIMIT Workshop

Modern vision and multimodal models depend on massive datasets and heavy compute, magnifying costs, energy use, bias, copyright, and privacy risks. The "DeepSeek shock" of January 2025 spotlighted the urgency of learning powerful representations under tight resource limits. Now in its third edition, our workshop continues to explore strategies for robust representation learning

https://iccv2025-limit-workshop.limitlab.xyz/

LIMIT.LAB

https://limitlab.xyz/

Join us! -> Slack invitation [Link]