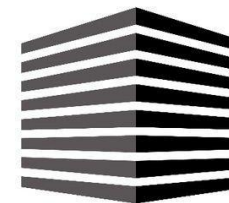


LIMIT.LAB



cvpaper.challenge



CVPR 2026 Report

Hirokatsu Kataoka, Yue Qiu, Oishi Deb, Kazuya Nishimura,
Rintaro Yanagi, Kenzo Yamabuki, Moeri Okuda, Hinako Mitsuoka,
Ryousuke Yamada, Mio Mitamura, Mizuki Kojima, Yusuke Kuwana,
Yoshihiro Fukuhara, Martinez–Noriega Edgar, Ryosei Hara, Yukinori Yamamoto, Senna Yoda,
Takumi Nakao, Miyu Goko, Shumpei Takezaki, Kohsuke Ide,
Ren Ohkubo, Gido Kato, Naoyoshi Hirade, Maika Takada, Chihiro Kaneko,
Yuna Park, Ryosuke Tsuchiya, Ryota Ishizaki, Shunsuke Sakai, Shinichi Mae, Takayuki
Shinohara, Su Myat Noe, Kuniko Paxton, Ali Caglayan

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

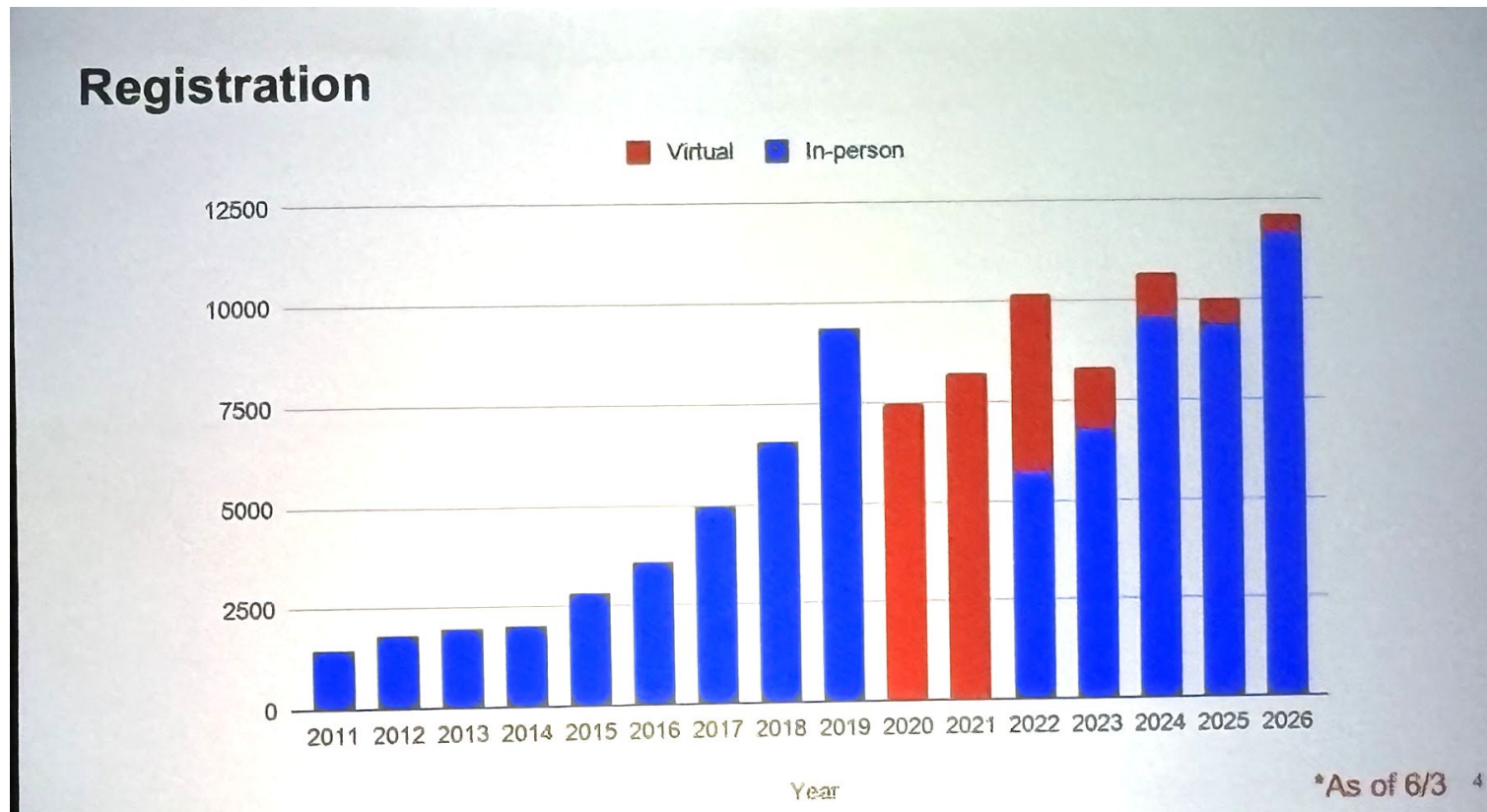
Meta Insights into Trends and Tendencies in CVPR 2026

- What kind of research was trending at this moment?
- What are overseas researchers working on?
- We have compiled the “trends” and “insights”.

CVPR26: Meta Insights into Trends and Tendencies (1/152)

From opening slide

- ❑ The attendees are increased this year
 - ❑ Compared with the CVPR 2025 attendance

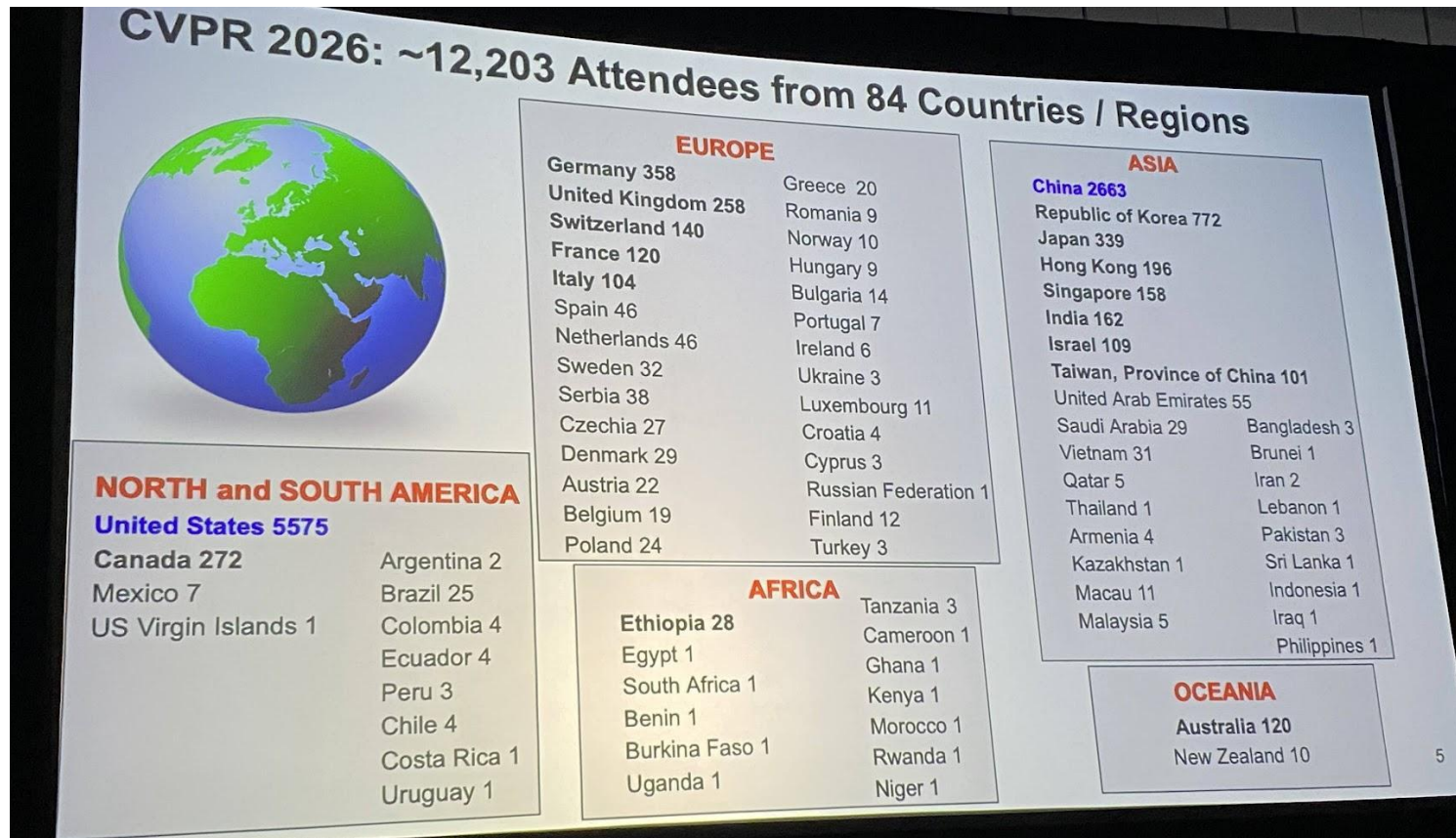


CVPR26: Meta Insights into Trends and Tendencies (2/152)

From opening slide

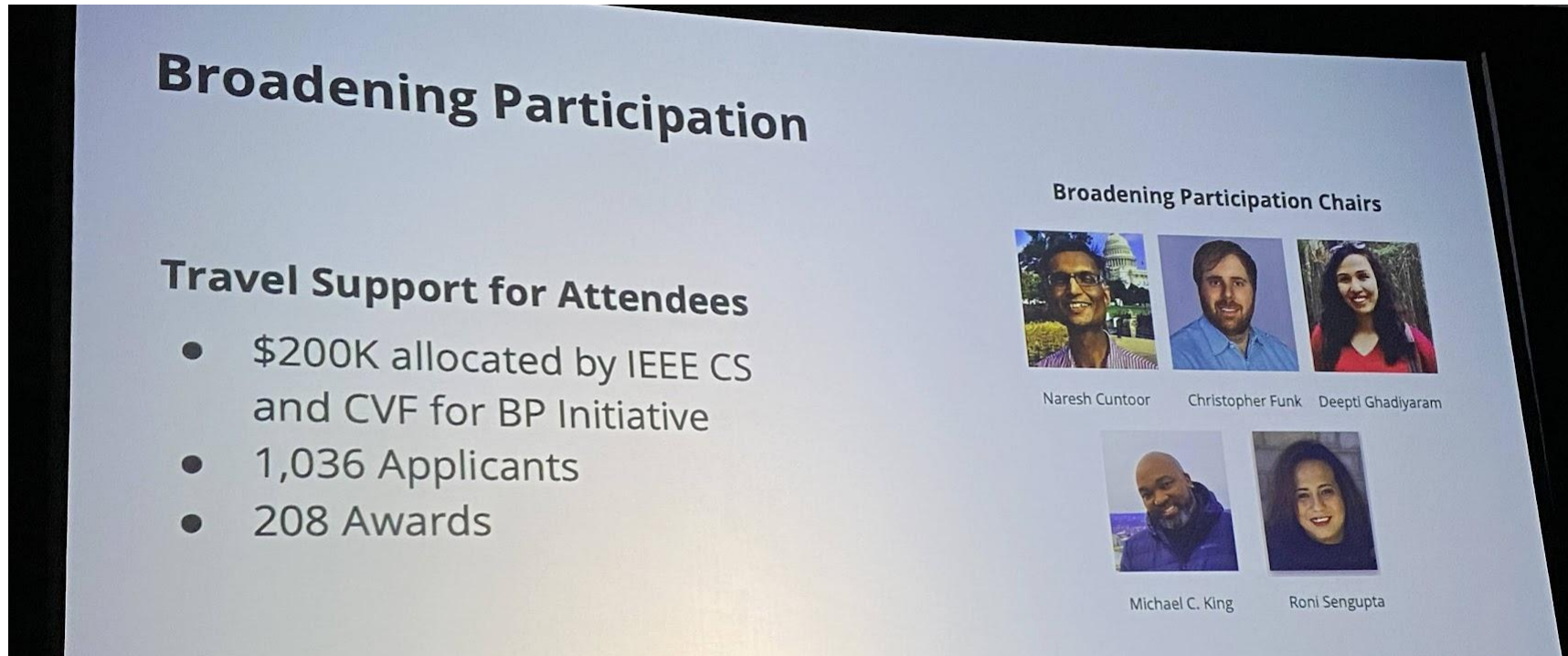
❑ Diverse countries / areas

- ❑ 1,000+ attendees from: US and China
- ❑ 100+ attendees from: Canada, Germany, UK, Switzerland, France, Italy, Korea, Japan, Hong Kong, Singapore, India, Israel, Taiwan, Australia



From opening slide

- ❑ Travel support
 - ❑ You may get a travel support from CVPR official
 - ❑ Around 20% of applicants have been accepted



Broadening Participation

Travel Support for Attendees

- \$200K allocated by IEEE CS and CVF for BP Initiative
- 1,036 Applicants
- 208 Awards

Broadening Participation Chairs

 Naresh Cuntoor	 Christopher Funk	 Deepti Ghadiyaram
 Michael C. King	 Roni Sengupta	

From opening slide

- ❑ Workshops & tutorials are highly competitive
 - ❑ Not simply a paper, should we push a workshop to build a community?
 - ❑ Its not be opened, but 200 – 300 (or more?) proposals were submitted?

Workshops & Tutorials

- 153 workshops
 - 32 Thematic Tracks
 - Expanded thematic organization reflecting the field's shift toward multimodal, embodied, and foundation-model-centric AI
 - Strengthened cross-community integration across computer vision, cognition, robotics, healthcare, accessibility, and multimodal AI
 - Increased emphasis on evaluation, benchmarking, and real-world deployment of AI systems
- 2 full-day and 17 half-day tutorials

Workshop Chairs

Erik Blasch Mei Chen Brian Clipp Sharon X. Huang Andreas Savakis Humphrey Shi

Track on Multimodal Learning

9th Multimodal Learning and Applications Workshop
Sight and Sound
The 5th Workshop on "What is Next in Multimodal Foundation Models?"
Workshop on Any-to-any Multimodal Learning

Track on Open World Learning

Open-World Vision
The Third Workshop on Anomaly Detection with Foundation Models
Visual Anomaly and Novelty Detection - 4th Edition

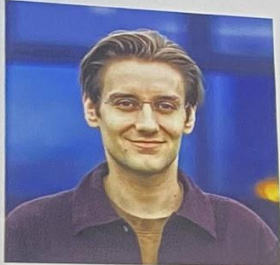
Tutorial Chairs

Boqing Gong Hongsheng Li Paul Schrader Vitomir Struc

From opening slide

- ❑ Keynote talks
 - ❑ Recently, CVPR invites a couple of speakers
 - ❑ The conference is considering emerging / trending topics
 - ❑ Biology, quantum computing, and neural vision

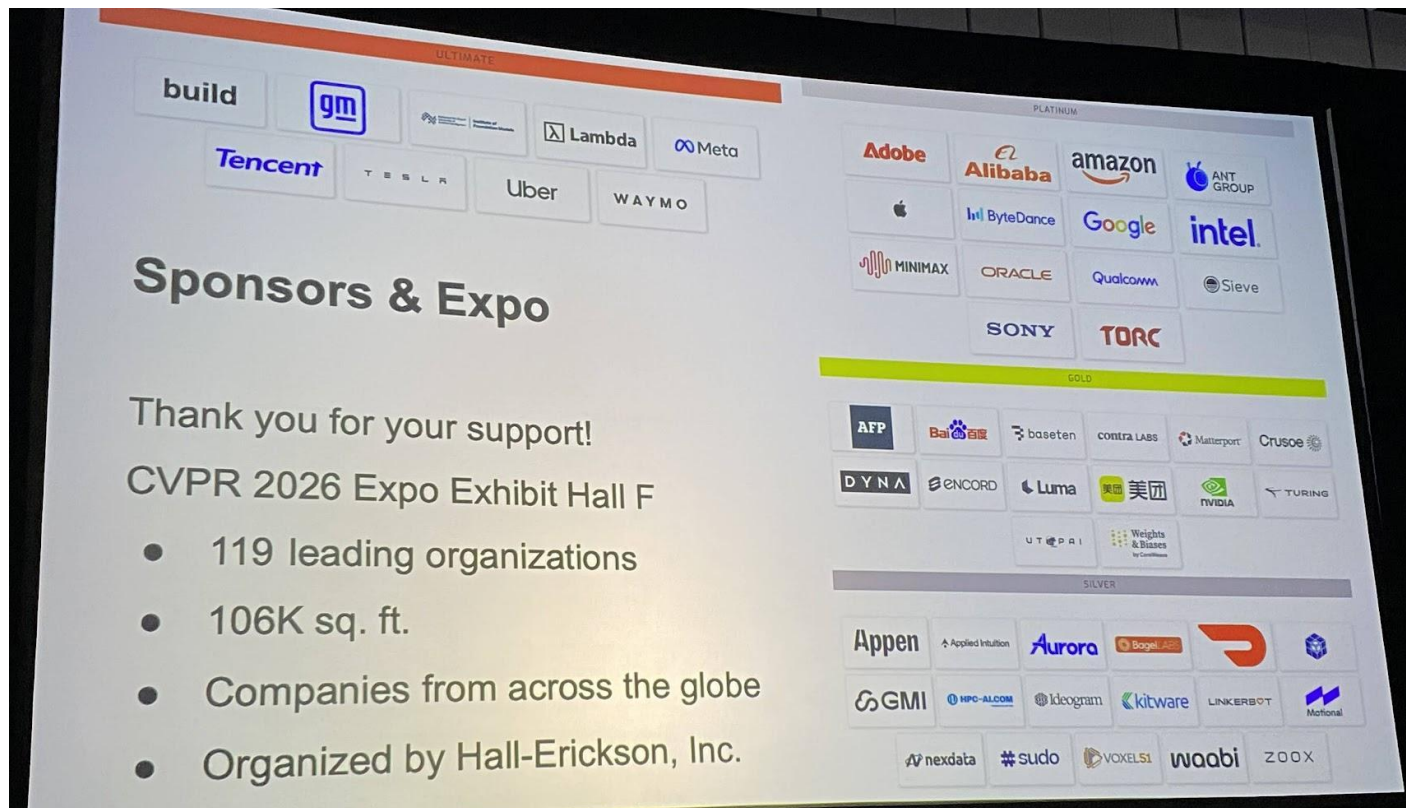
Keynotes

Friday, 2:45 p.m.	Saturday, 10:30 a.m.	Sunday, 10:30 a.m.
		
Simon Kohl Founder & CEO, Latent Labs	Jerry Chow IBM, TJ Watson, CTO Quantum Centric Supercomputing	Thomas Serre Professor of Science and Professor of Cognitive & Psychological Sciences
<i>Programmable Biology: Generative AI for Molecular Design</i>	<i>Transforming Computing with Quantum-Centric Supercomputing</i>	<i>Scaling Laws vs. Neural Laws: Toward More Natural Artificial Vision</i>
Bluebird Ballroom	Bluebird Ballroom	Bluebird Ballroom

From opening slide

❑ Sponsors & Expo

- ❑ So many sponsors (one big event in CVPR)
- ❑ Hiring (internship, full-time), social events...
- ❑ Big chance for industry



From opening slide

- ❑ Demo session
 - ❑ There are only 28 demos (selected?)
 - ❑ Mobile device, SLAM/3D reconstruction, Labeling tool, camera system...
 - ❑ Best demo is selected every year

28 Demos (ExHall F)

- Friday: 10:45 a.m. – 12:45 p.m.
4:00 p.m. – 6:00 p.m.
- Saturday: 11:45 a.m. – 1:45 p.m.
4:45 p.m. – 6:45 p.m.
- Sunday: 11:45 a.m. – 1:45 p.m.

Demo Chairs

Sathyanarayanan N. Aakur Shu Kong

Diverse Topics: perception, robotics, HCI, VR/AR, emotion understanding, clinical agents

10

From opening slide

- ❑ Art gallery
 - ❑ Presentation and gallery in front of sponsor booth

Art Gallery (ExHall E)

- 10.45 a.m. to 6.00 p.m. on Friday
- 11.45 a.m. to 6.00 p.m. on Saturday
- 11.45 a.m. to 3.00 p.m. on Sunday

Check out the CVPR Virtual Gallery: <https://thecvf-art.com/>

Special Art Sessions:

Friday

- 11 a.m. Gallery Tour with Curator
- 1.30 p.m. Art Panel (Room 201)
- 5 p.m. Gallery Tour with Curator

Saturday

- 11.45 a.m. Gallery Tour with Curator
- 1:45 p.m. Art Panel (Room 201)
- 5 p.m. Gallery Tour with Curator

Sunday

- 11.45 a.m. Gallery Tour with Curator

AI Art Curator



Luba Elliott



Credit: Eryk Salvaggio

11

From opening slide

❑ Doctoral consortium

- ❑ w/ mentor lunch and discussion
- ❑ A PhD student can discuss with mentor regarding their research / phd course
- ❑ CFP (for next year!): <https://cvpr.thecvf.com/Conferences/2026/CallForDoctoralConsortium>

The photograph shows a presentation slide titled "Student Activities". It is divided into two main sections: "Doctoral Consortium (Saturday, 11:45–1:45)" and "Speed Mentoring (Saturday, 11:45–1:15)". The first section includes a bulleted list of activities and features the NSF logo. The second section includes a bulleted list of activities and features the IEEE Computer Society logo and an 80th anniversary celebration logo. On the right side, under the heading "Doctoral Consortium Chairs", there are two portraits of Paola Cascante-Bonilla and Abby Stylianou.

Student Activities

Doctoral Consortium (Saturday, 11:45–1:45) 

- Opportunity for recent grads/close-to-complete PhD students to interact with experienced researchers
- One-to-one matching of students to mentors
- Discuss career plans and research
- 30 students out of 150 applicants accepted

Speed Mentoring (Saturday, 11:45–1:15)

- Open call was issued to all students
- Informal small-group discussions with senior faculty and industry leaders
- 200 participants

Doctoral Consortium Chairs

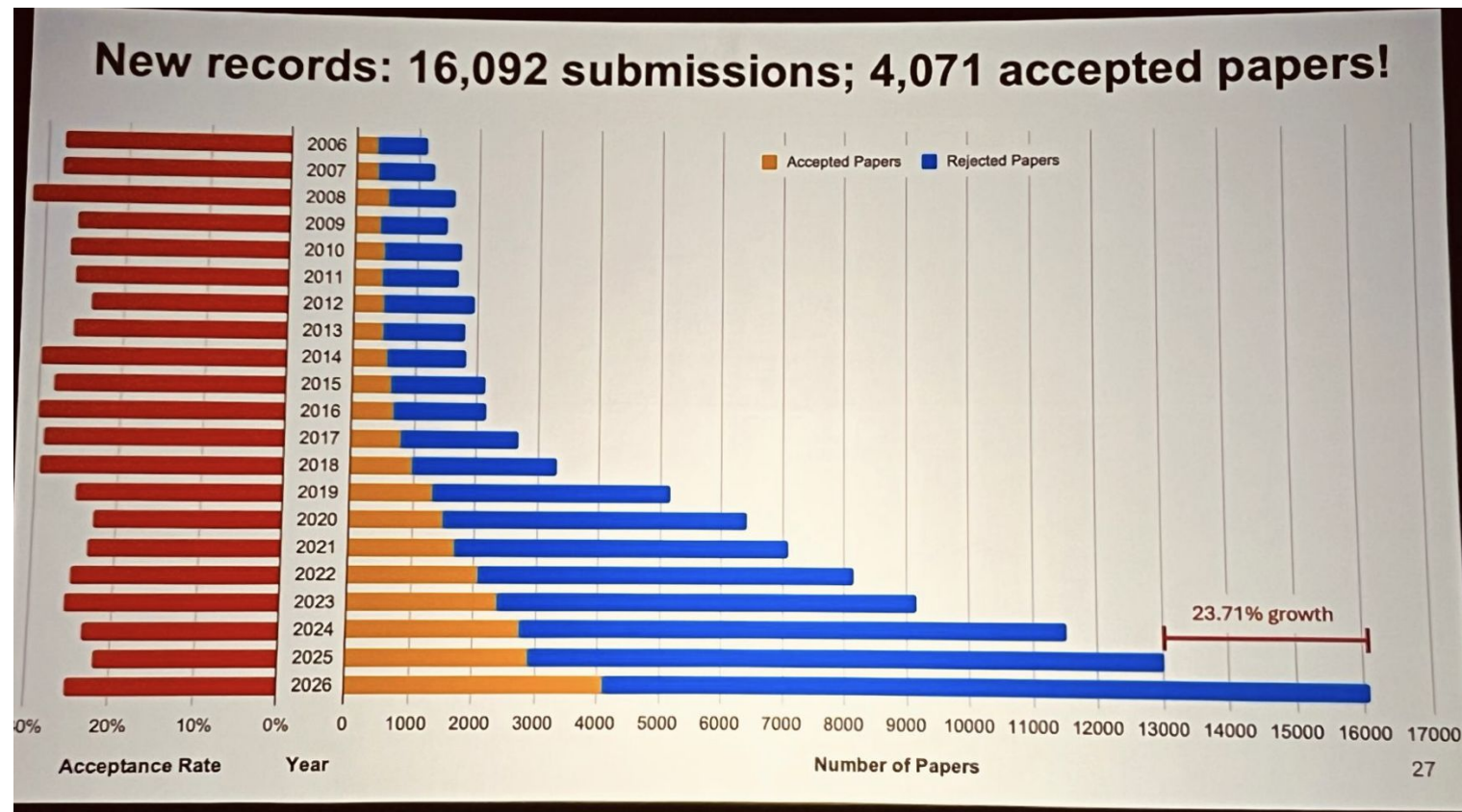
Paola Cascante-Bonilla Abby Stylianou

 IEEE COMPUTER SOCIETY 

CVPR26: Meta Insights into Trends and Tendencies (10/152)

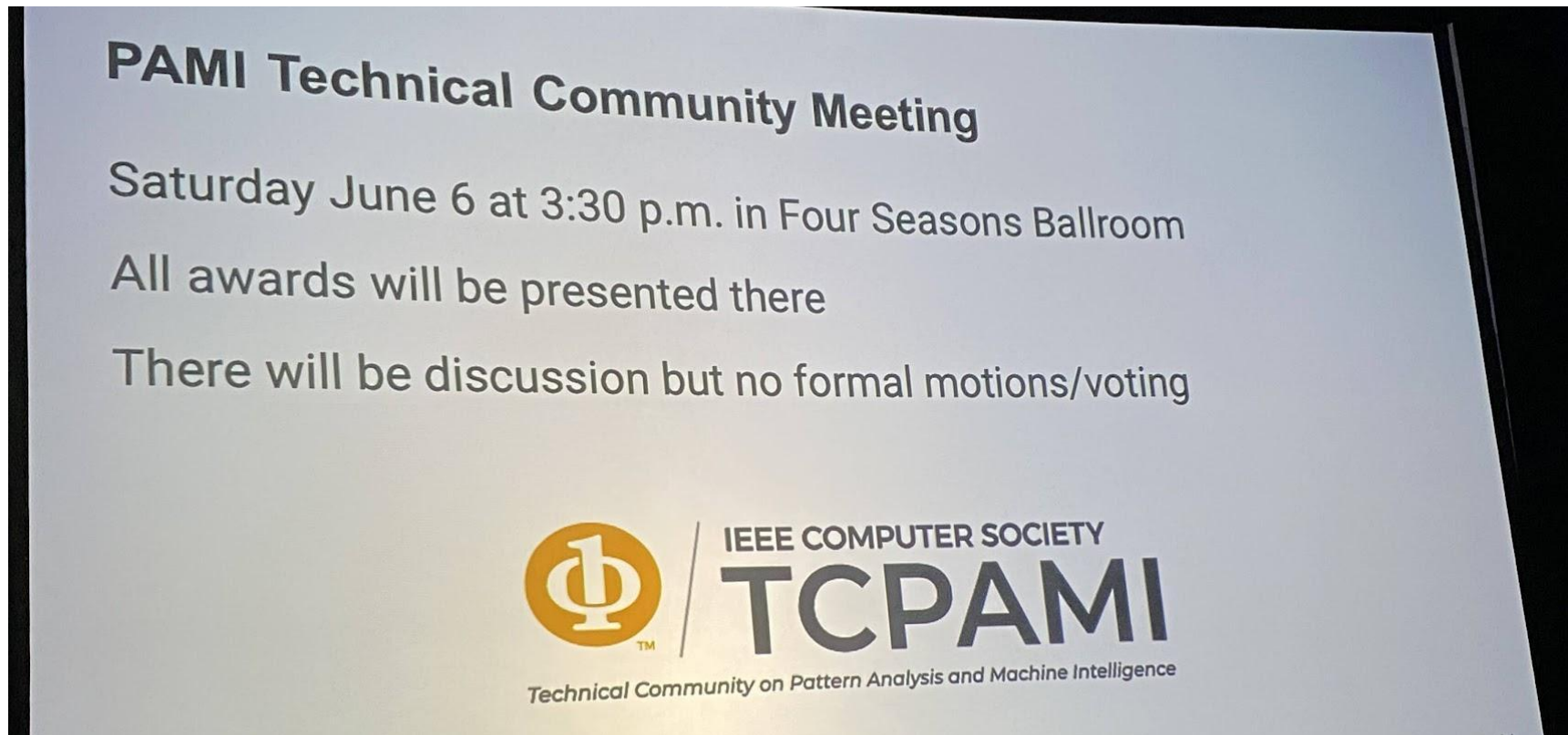
From opening slide

- ❑ 23.71% growth for submissions
- ❑ Exponential growth?



From opening slide

- ❑ PAMI Technical Community Meeting
 - ❑ Think our community in the meeting
 - ❑ Award ceremony is also here



CVPR26: Meta Insights into Trends and Tendencies (12/152)

From opening slide

- ❑ Social media
 - ❑ Media expert – 3 chairs in CVPR 2026
 - ❑ One of the reasons for rising conference!

Social Media

- Twitter/X (@CVPR)
- Bluesky

Publicity Chairs



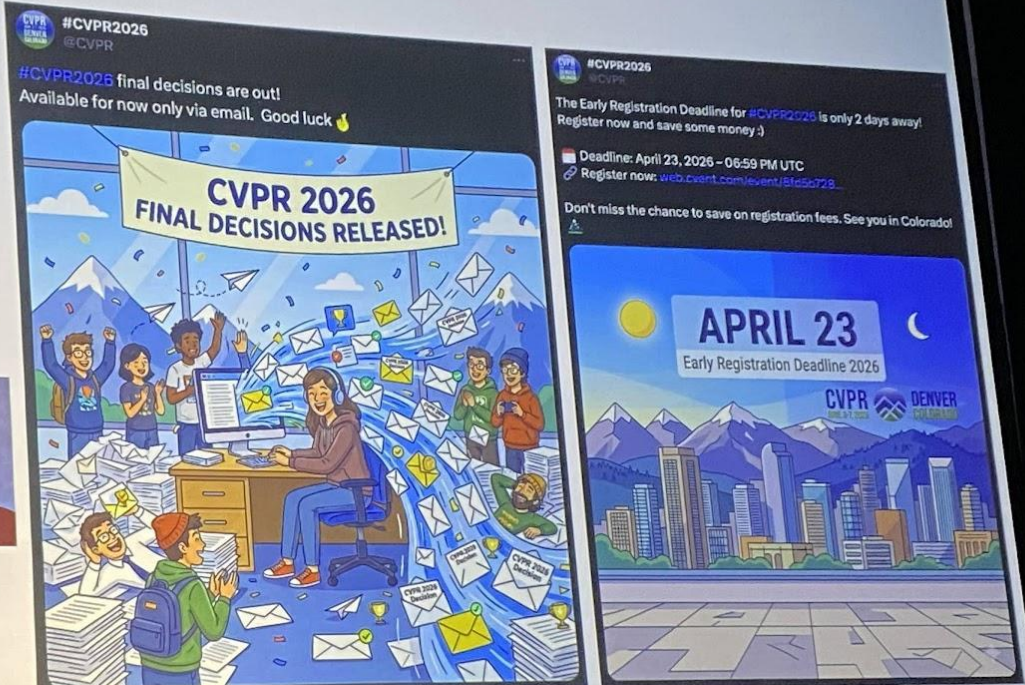
Kosta Derpanis



Deblina Bhattacharjee



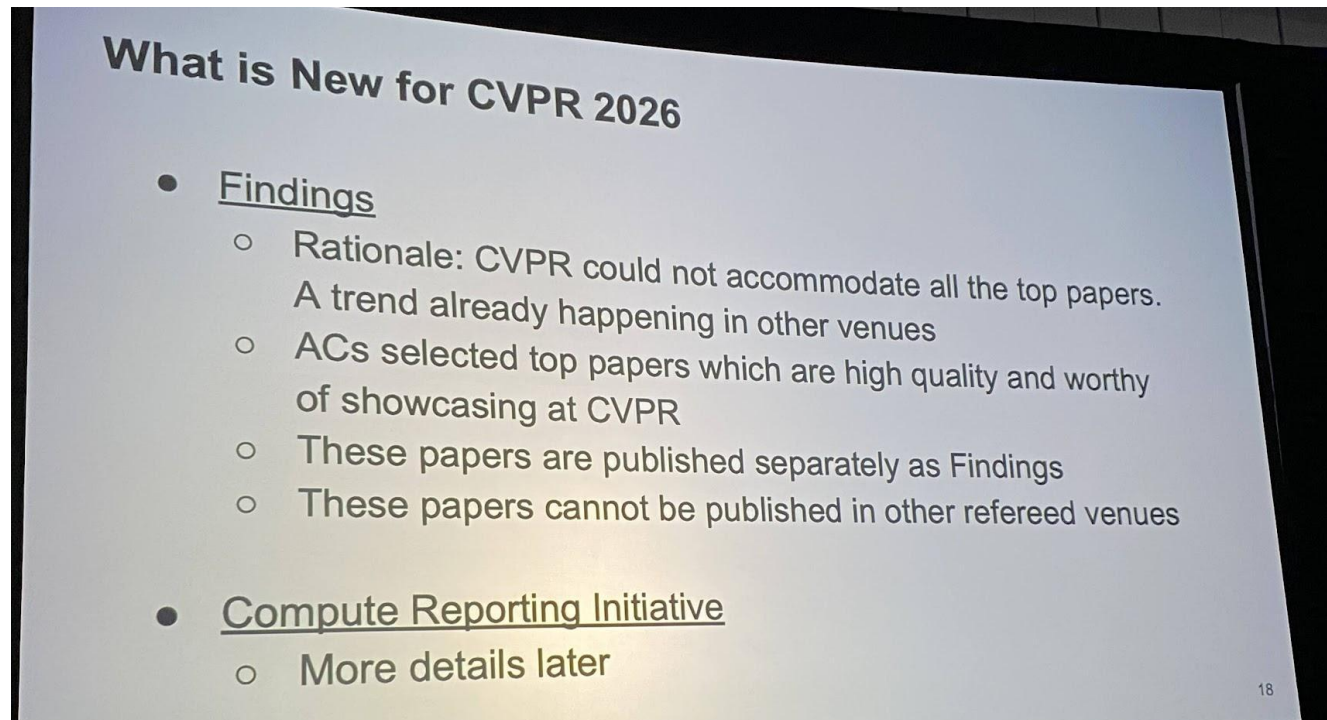
Joakim Bruslund Haurum



From opening slide

❑ Findings & Compute Reporting Initiative

- ❑ Findings: rejected paper but AC picked up some papers, presentation during 7:00 – 9:00
- ❑ Compute reporting: some papers have been selected as efficient papers!



Introduction of a new Findings Track

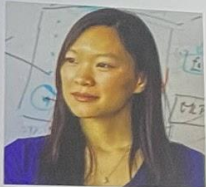
- ❑ Following the success of the 1st Workshop on the Findings of ICCV in 2025, CVPR 2026 introduced its own Findings Track
- ❑ Motivation
 - ❑ Reduce repeated resubmissions
 - ❑ Offer a venue for technically sound papers with solid experimental validation, even when their novelty is considered incremental
- ❑ Process
 - ❑ Area Chairs recommend papers submitted to the main conference for the track
 - ❑ Authors choose whether to opt in
 - ❑ Findings papers will appear in the workshop proceedings
- ❑ Scale
 - ❑ 1717 papers recommended, 941 submitted
 - ❑ 3 poster sessions were held in the morning before the main conference

CVPR26: Meta Insights into Trends and Tendencies (15/152)

From opening slide

- ❑ 6 program chairs, 41 senior area chairs, and 909 area chairs
 - ❑ Over 16k papers should be judged (crazy jobs before the notification!)

Program Chairs



Angela Dai
Technical
University of
Munich



Adriana Kovashka
University of
Pittsburgh



Chen Change Loy
Nanyang
Technological
University



Vladimir Pavlovic
Rutgers
University



Alex Schwing
University of Illinois
Urbana-Champaign



Shaoting Zhang
Shanghai Jiao
Tong University

Senior Area Chairs



Lourdes Agapito



Pablo Arbelaez



Jianfei Cai



David J. Crandall



Dima Damen



David Forsyth



Yoshitomo Matsubara



Seon Joo Kim



Juergen Gall



Georgia Gkioxari



Boqing Gong



Kristen Grauman



Derek Hoiem



Anthony Hoogs



Gang Hua



Nathan Jacobs



Svetlana Lazebnik



Gim Hee Lee



Yong Jae Lee



Hongdong Li



Hongsheng Li



Jiaying Liu



Vittorio Murino



Srinivasa Narasimhan



Matthias Nießner



Ko Nishino



Maja Pantic



Ian Reid



Marcus Rohrbach



Stefan Roth



Kate Saenko



Dimitris Samaras



Manolis Savva



Bernt Schiele



Deqing Sun



Andrea Tagliasacchi



Tinne Tuytelaars



Ming-Hsuan Yang



Angela Yao



Lei Zhang

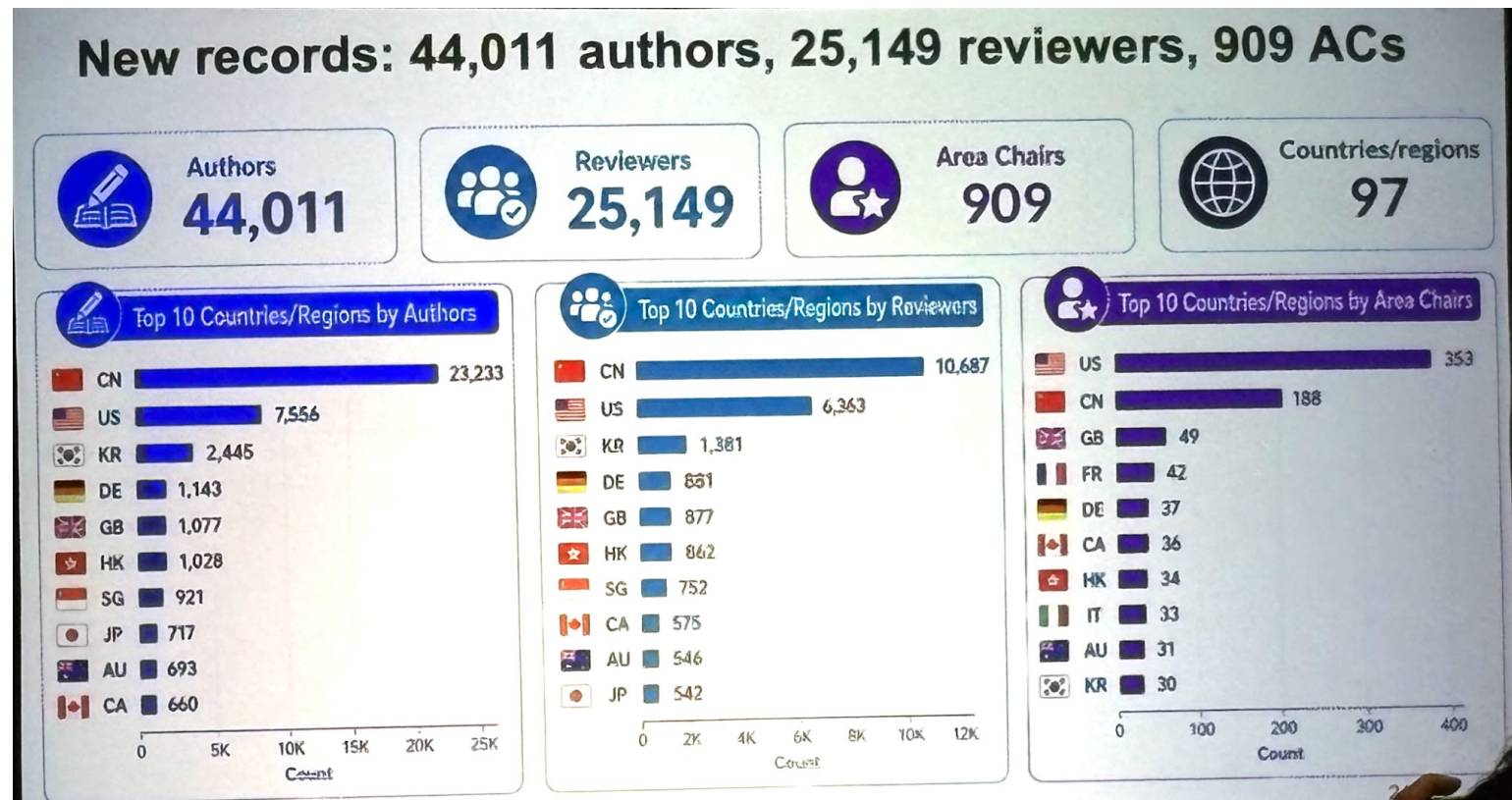


Jun-Yan Zhu

CVPR26: Meta Insights into Trends and Tendencies (16/152)

From opening slide

- ❑ Listed at each continent
- ❑ 44k authors / 16k submissions / 25k reviewers...



CVPR26: Meta Insights into Trends and Tendencies (17/152)

From opening slide

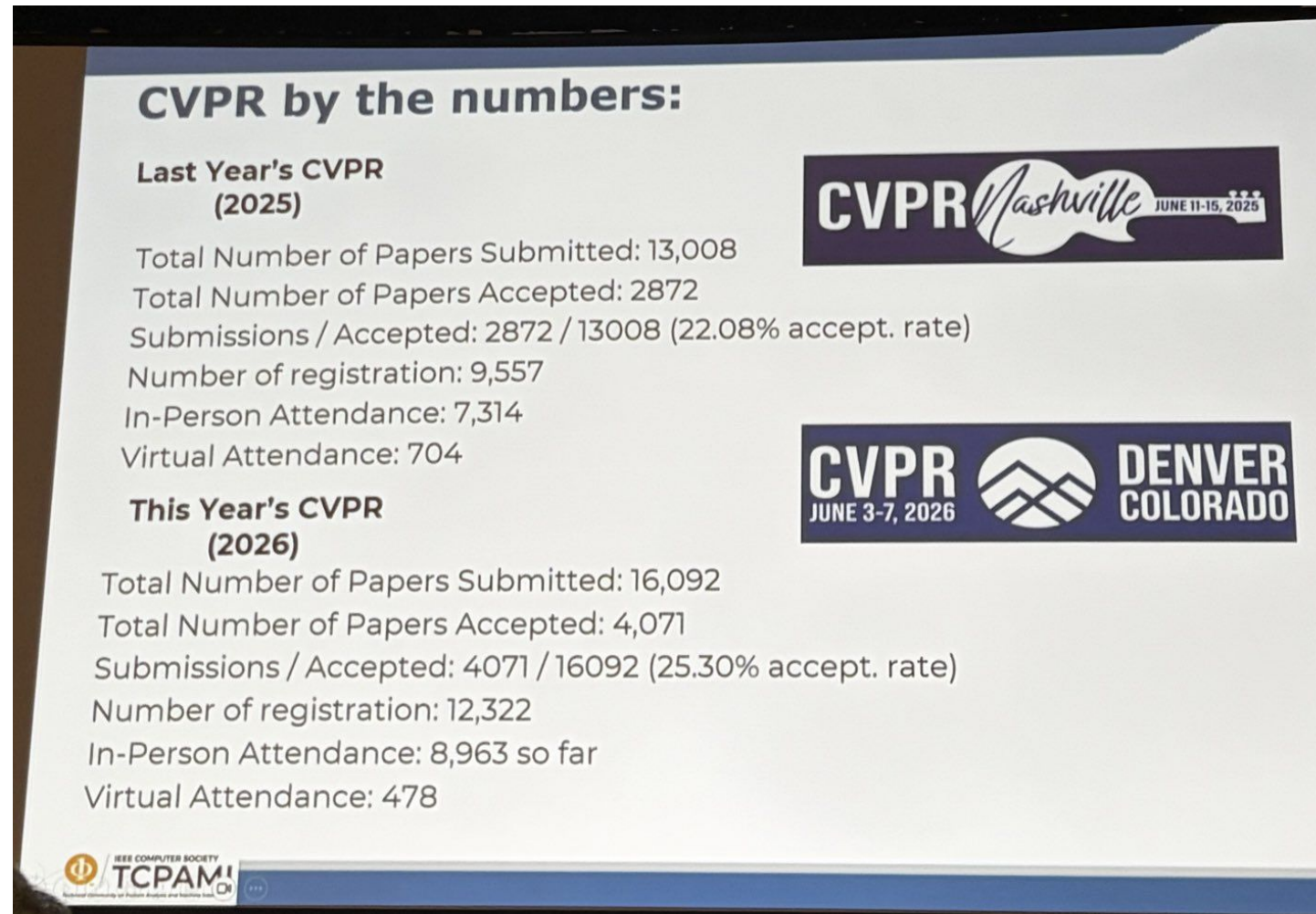
- ❑ **+100% reviewers by comparing to +24% submissions!**
- ❑ How much review quality can be maintained?

New records: 44,011 authors, 25,149 reviewers, 909 ACs

	PCs	SACs	ACs	Reviewers	Submissions
CVPR 2024	6	24	477	9,872	11,532
CVPR 2025	6	22	708	12,592	13,008
ICCV 2025	6	-	510	11,895	11,239
CVPR 2026	6	41 +86% vs CVPR 2025	909 +28% vs CVPR 2025	25,149 +100% vs CVPR 2025	16,092 +24% vs CVPR 2025

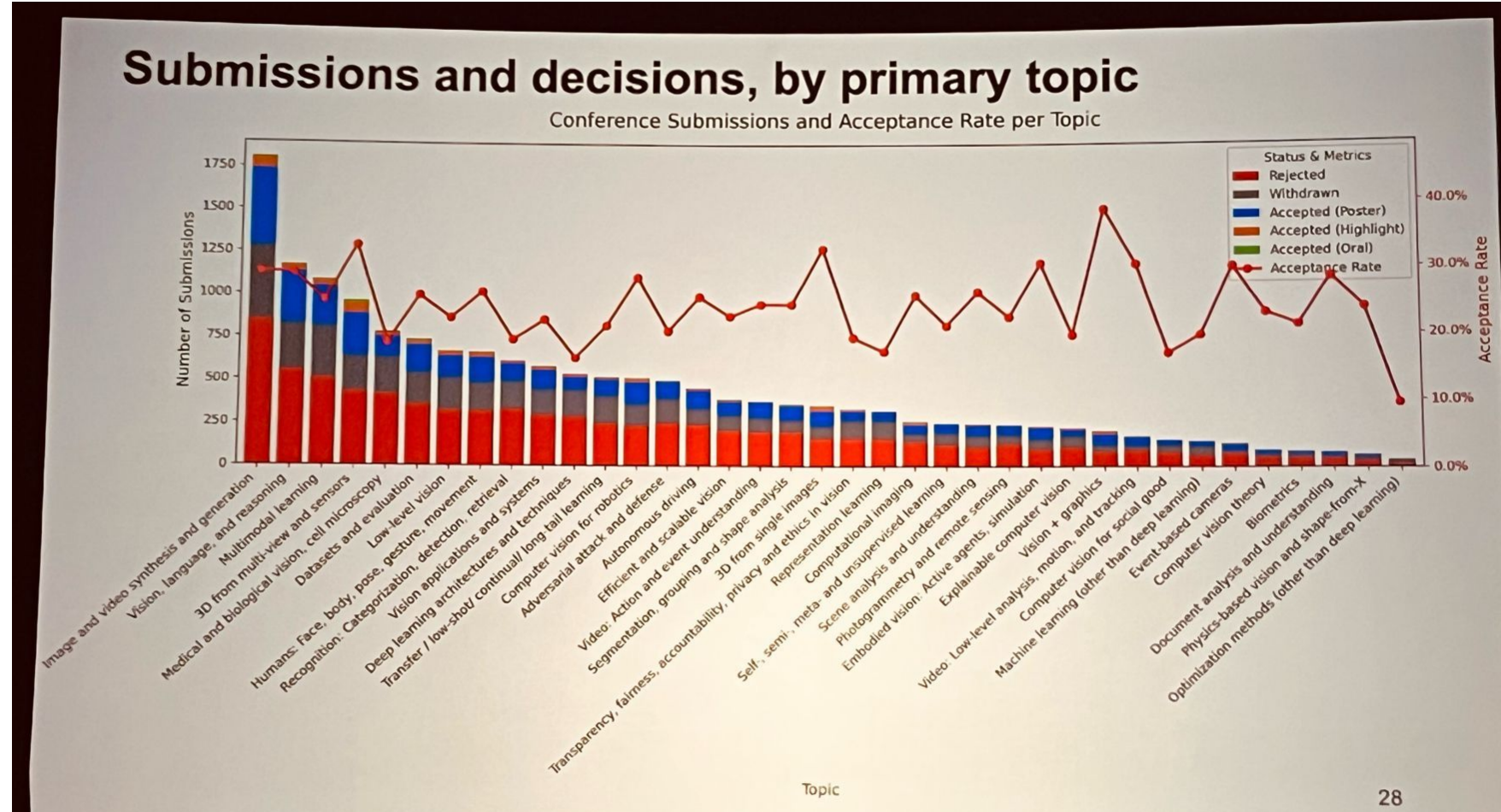
CVPR 2025 vs CVPR 2026

- ❑ The paper acceptance rate increased this year



From opening slide

- ❑ Image and video synthesis and generation are the most common



From opening slide

❑ Oral Coaching

- ❑ A researcher who makes oral presentations more engaging

Oral Coaching

Committee chair: Srinivasa Narasimhan

Committee members:

Alexei Efros	Kyros Kutulakos	Aswin Sankaranarayanan
Dima Damen	Jean-Francois Lalonde	Steve Seitz
Fernando De la Torre	Lana Lazebnik	Yaser Sheikh
Sven Dickinson	Srinivasa Narasimhan	Ashok Veeraraghavan
David Forsyth	Ko Nishino	Yair Weiss
Bill Freeman	Devi Parikh	Richard Zhang
Ioannis Gkioulekas	Deva Ramanan	Todd Zickler
Mohit Gupta		

Many thanks to them!

30

From opening slide

Paper decisions

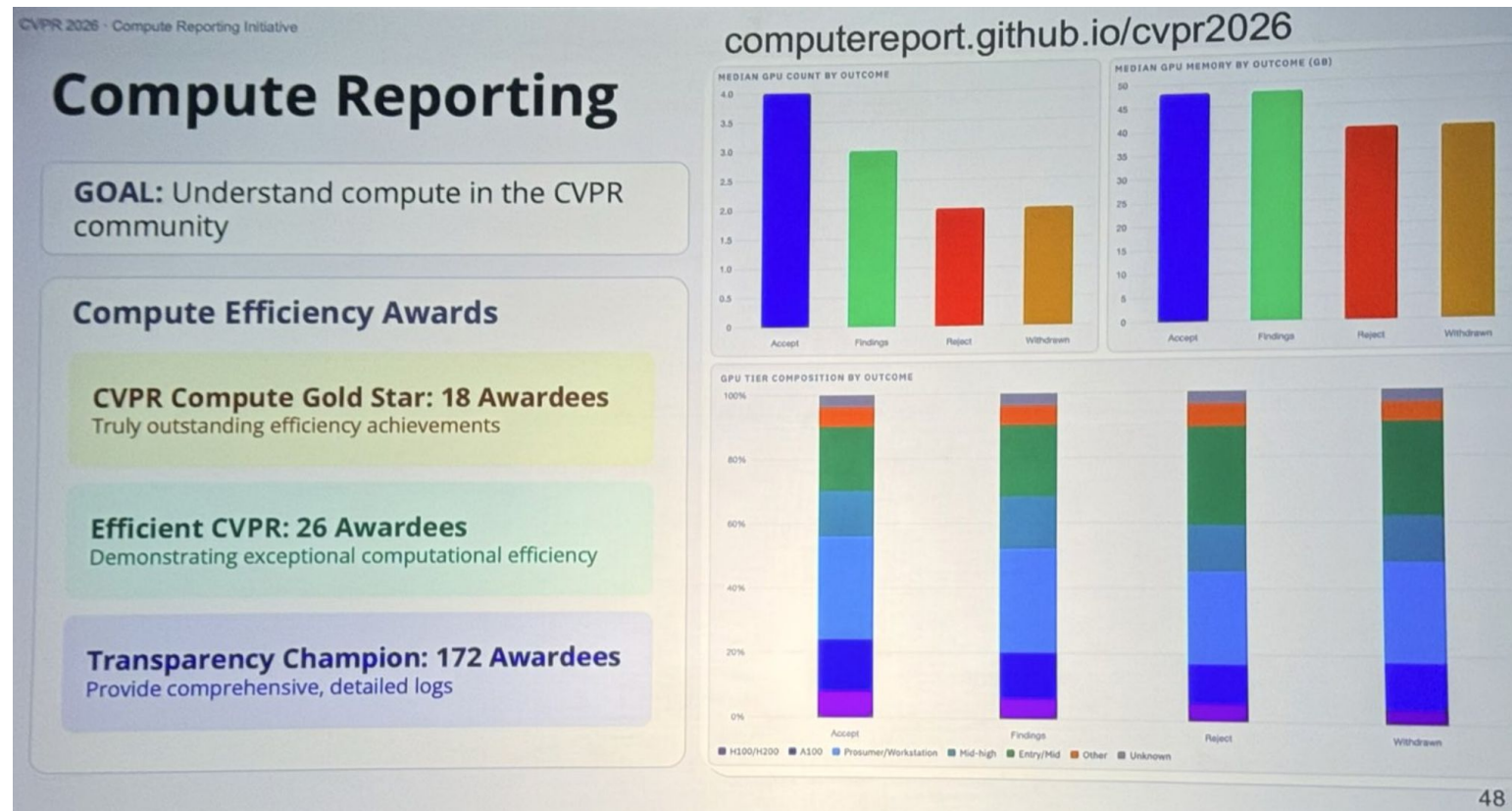
- Every single paper received 3 reviews
- Decisions formed within triplets of ACs
- A meta-review drafted by an AC and approved by the triplet
- Orals, highlights recommended by the ACs, SACs & PCs
- SACs double-checked those as well as high-scoring papers
- Overall: 4071 (25.3%) accepted papers
 - 141 (3.5%) of accepted papers are Orals+posters, which includes 74 award candidates
 - 578 (14.2%) of accepted papers are “*highlights*” posters, with special annotation in the program
 - 3,352 accepted papers are posters

PROGRAM GUIDE

- 🏆 Multi-Task-Aware Infrared-V
Juan Cheng, Zhiqin Zhu, Yafei
- 20 🏆 Dual Band Thermal Videogr
Reflection and Emission Ne
Sriram Narayanan, Mani Ram
- 21 🏆 MetaSpectra+: A Compact B
Snapshot Hyperspectral+ In
Spectrum from Defocus: Fa
Focal Stack, M. Kerem Aydi
Emma Alexander
- 22 🏆 Towards Photorealistic and
Framework, Linxiao Shi, Sim
Jinwei Chen, Bo Li, Shifeng C
- 24 🏆 UnReflectAnything: RGB-On
Synthetic Specular Supervis
Mert Asim Karaoglu, Patrick
Benjamin Busam
- 25 🏆 AVGGT: Rethinking Global A
Xianbing Sun, Zhikang
Liqing Zhang, He Wang

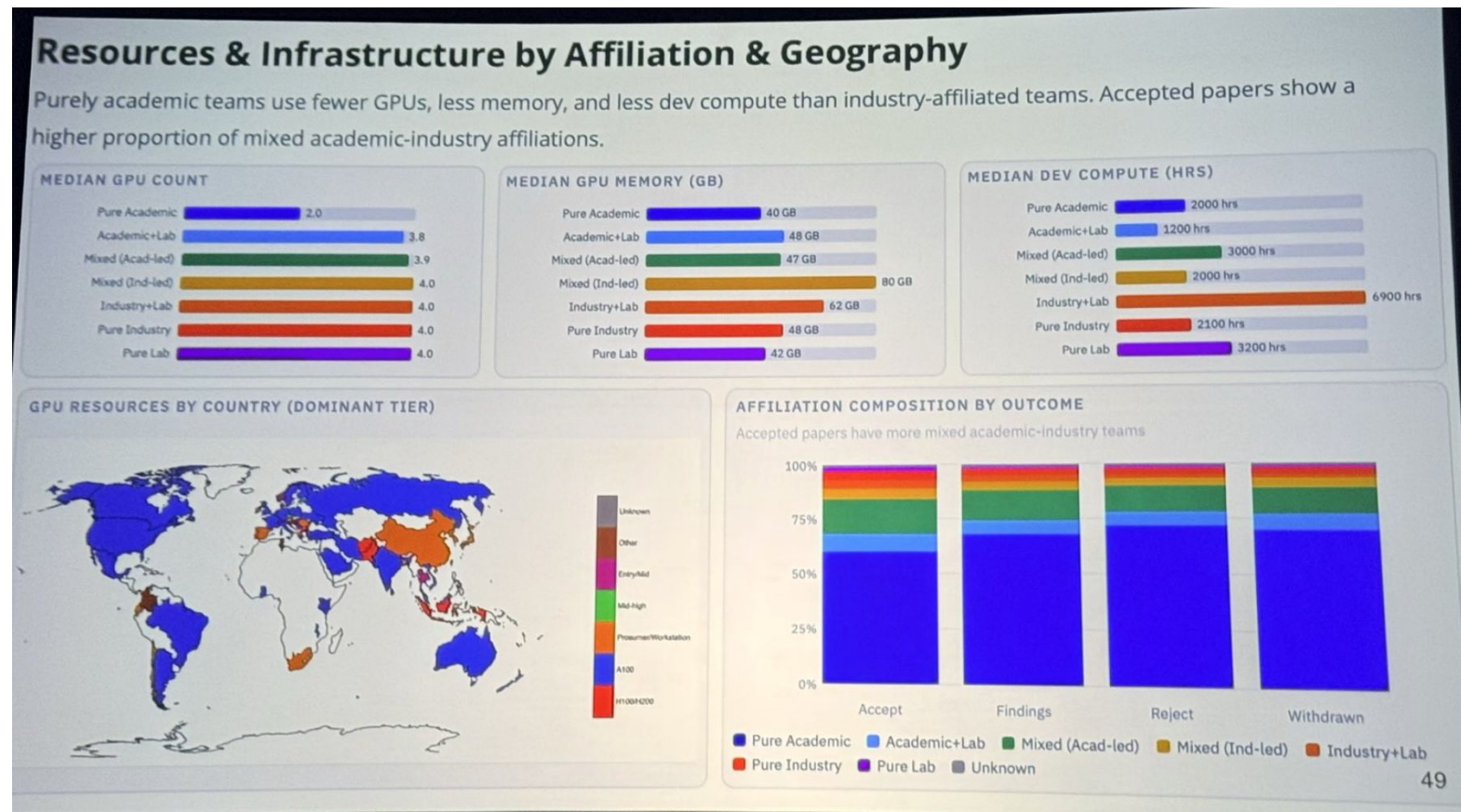
From opening slide

- ❑ The number of GPUs affects acceptance
 - ❑ More than 4!



From opening slide

❑ Limited GPU Access in Academia



CVPR26: Meta Insights into Trends and Tendencies (24/152)

From opening slide

- ❑ 74 from ACs / SACs / PCs
- ❑ 15 finalists (not equal to 74 papers)
- ❑ Only 5 papers are awardees

Best Paper Awards

- 74 best paper award candidates were nominated by the Area Chairs, Senior Area Chairs, and Program Chairs
- They are marked in the program
- From the candidates, 15 papers were selected as finalists and 5 were selected for an award by the award committee
 - 1 Best Student Paper Honorable Mention
 - 1 Best Student Paper
 - 2 Best Paper Honorable Mention
 - 1 Best Paper
- We'll present certificates to the award winners at the PAMI-TC meeting on Saturday June 6 3:30 p.m. in Four Seasons Ballroom

PROGRAM GUIDE

20	Multi-Task-Aware Infrared-Vision	Juan Cheng, Zhiqin Zhu, Yafei			
21	Dual Band Thermal Videogra	Reflection and Emission Nea	Sriram Narayanan, Mani Ram		
22	MetaSpectra+: A Compact B	Snapshot Hyperspectral+ Im	Spectrum from Defocus: Fas	Focal Stack, M. Kerem Aydin	Emma Alexander
23	Towards Photorealistic and	Framework, Linxiao Shi, Sim	Jinwei Chen, Bo Li, Shifeng C		
24	UnReflectAnything: RGB-Onl	Synthetic Specular Supervis	Mert Asim Karaoglu, Patrick	Benjamin Busam	
25	AVGGT: Rethinking Global A	Xianbing Sun, Zhikai Zhu, Zh	Liqing Zhang, He Wang, Jian		

Best Paper Process

Blind reviewing, COI handling

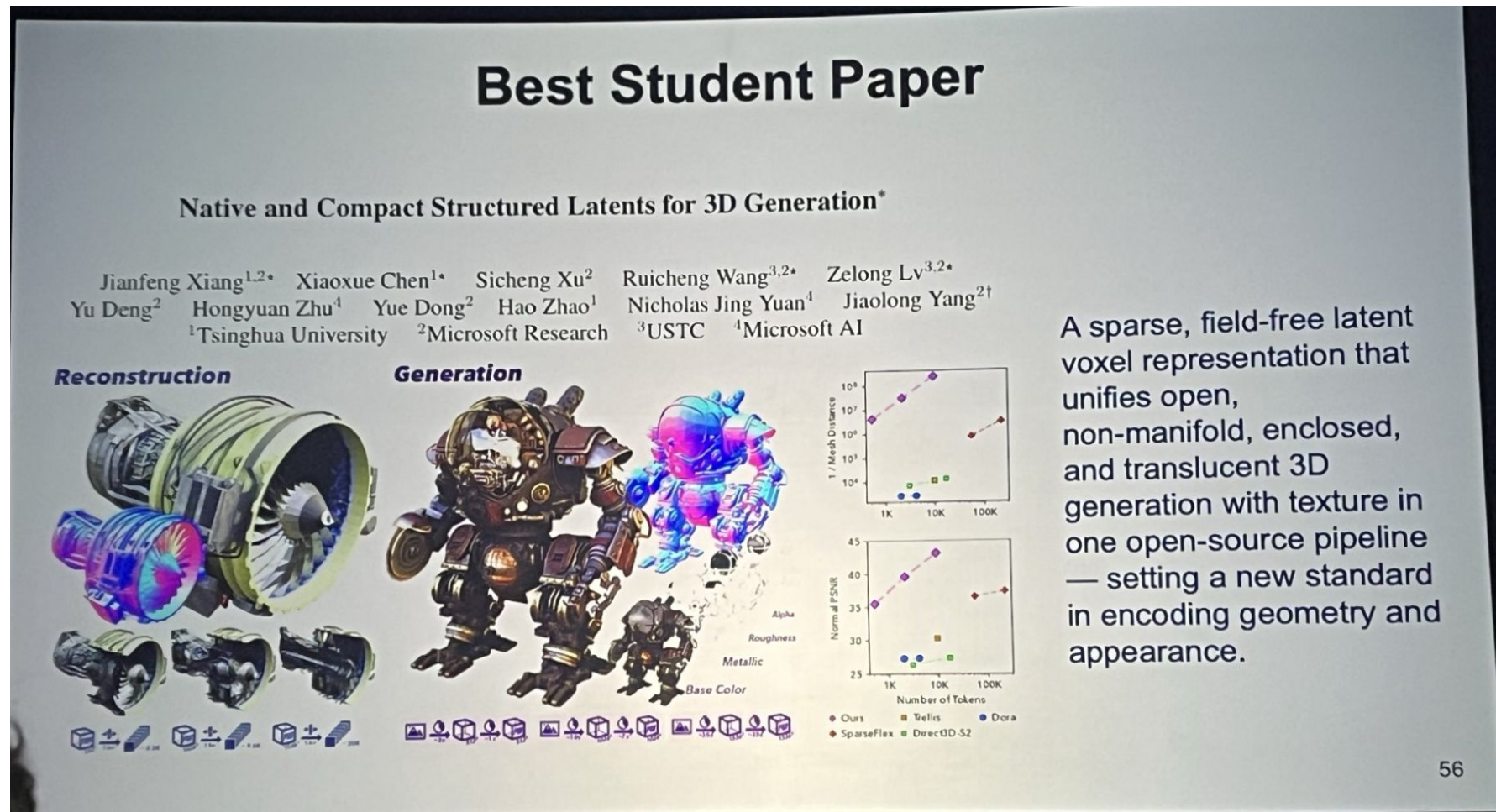
74 candidates selected by ACs, SACs, PCs

- 15 finalists (each committee member reviewed ~20 papers)
- Each finalist paper reviewed and rated by 3 committee members
- Committee discussed to select awardees

Criteria: A best paper should be a "must read" for someone who works in a related area, and you would strongly recommend it to anyone in computer vision. It may also serve as a representation of our field's progress to those not in it. A best paper should be well-written and make conclusions that are soundly backed by experimental evidence or theory. A best paper has at least one of these properties...

- Surprising progress — a problem is solved beyond what you thought possible
- Enlightening — you understand something new and significant
- Landmark — reflects and caps off significant progress on an important problem
- Insightful — makes you think of a problem in a new way
- Influential — likely to inspire other research
- Practical — could lead to products or solutions that will make people's lives better

From award ceremony



From award ceremony

Best Student Paper Honorable Mention

ChordEdit: One-Step Low-Energy Transport for Image Editing

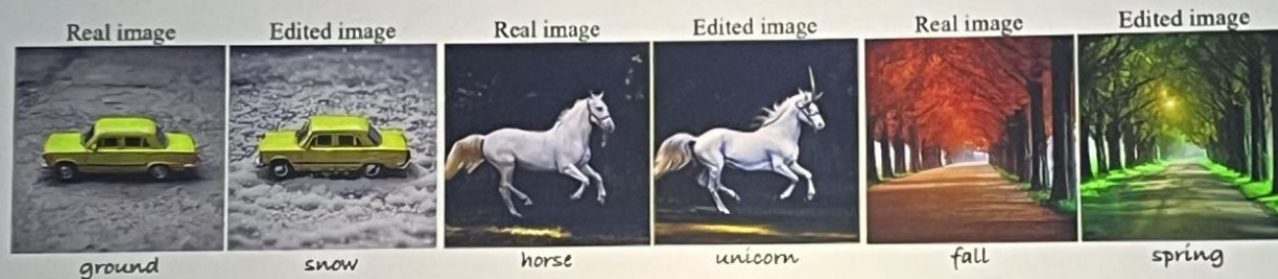
Liangsi Lu¹, Xuhang Chen², Minzhe Guo¹, Shichu Li³, Jingchao Wang⁴, Yang Shi^{1†}

¹Guangdong University of Technology ²Huizhou University

³Shenzhen University ⁴Peking University

[†] Corresponding author: sudo.shiyang@gmail.com

Project page: <https://chordedit.github.io>



A one-step, training-free, inversion-free image editor that frames editing as low-energy optimal transport — theoretically grounded and surprisingly effective.

55

From award ceremony

Best Paper Honorable Mention

SAM 3D: 3Dfy Anything in Images

Xingyu Chen* Fu-Jen Chu* Pierre Gleize* Kevin J Liang* Alexander Sax* Hao Tang* Weiyao Wang*
Michelle Guo Thibaut Hardin Xiang Li° Aohan Lin Jiawei Liu Ziqi Ma°
Anushka Sagar Bowen Song° Xiaodong Wang Jianing Yang° Bowen Zhang°
Piotr Dollár† Georgia Gkioxari† Matt Feiszli†§ Jitendra Malik†§

*Core Contributor (Alphabetical, Equal Contribution) °Intern †Project Lead §Equal Contribution
Meta Superintelligence Labs



A big step forward for 3D object models from cluttered in-the-wild single-image, and a scalable pipeline for data collection.

From award ceremony

Best Paper Honorable Mention

NITROGEN: An Open Foundation Model for Generalist Gaming Agents

Loïc Magne¹*, Anas Awadalla^{1 2}*, Guanzhi Wang^{1 3}*†
Yinzhen Xu¹, Joshua Belofsky⁴, Fengyuan Hu¹, Joohwan Kim¹
Ludwig Schmidt², Georgia Gkioxari³, Jan Kautz¹
Yisong Yue³†, Yejin Choi^{1 2}†, Yuke Zhu^{1 5}†, Linxi Fan¹†

¹ NVIDIA, ² Stanford, ³ Caltech, ⁴ UChicago, ⁵ UT Austin

The diagram illustrates the NITROGEN architecture. It features three main components at the top: 'Universal Gameplay Harness' (blue), 'Multi-Game Foundation Agent' (orange), and 'Internet-Scale Video-Action Dataset' (yellow). The 'Universal Gameplay Harness' provides input to the 'Multi-Game Foundation Agent' via a 'Vision Encoder' (light blue) and 'Action DiT' (light green) block. The 'Internet-Scale Video-Action Dataset' provides input to the 'Multi-Game Foundation Agent' via an 'Action Extraction' block. The 'Multi-Game Foundation Agent' outputs to a 'Gamepad Overlay Video Dataset' (cylinder) which contains gamepad overlay images. The 'Gamepad Overlay Video Dataset' also provides input to the 'Multi-Game Foundation Agent' via an 'Action Extraction' block. The 'Multi-Game Foundation Agent' outputs to a 'Gamepad Overlay Video Dataset' (cylinder) which contains gamepad overlay images. The 'Gamepad Overlay Video Dataset' also provides input to the 'Multi-Game Foundation Agent' via an 'Action Extraction' block.

A dataset (40,000 hours; 1,000 games), simulator for evaluation, and foundation model for vision-to-action gaming agents — opening new research opportunities.

57

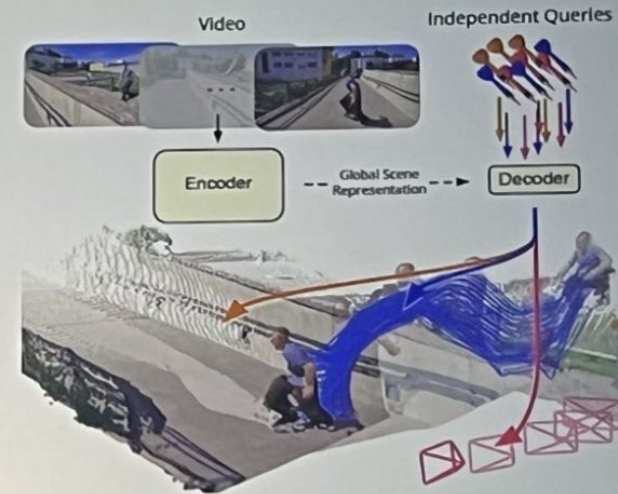
From award ceremony

Best Paper

Efficiently Reconstructing Dynamic Scenes One ↘ D4RT at a Time

Chuhan Zhang* Guillaume Le Moing* Skanda Koppula*[◦] Ignacio Rocco*
Liliane Momeni* Junyu Xie^{◦1} Shuyang Sun* Rahul Sukthankar* Joëlle K. Barral*
Raia Hadsell* Zoubin Ghahramani* Andrew Zisserman*[◦] Junlin Zhang*
Mehdi S. M. Sajjadi*²

*Google DeepMind [◦]University College London [◦]University of Oxford



An elegant and efficient approach that unifies estimation of depth, camera pose, 3D point tracking, and 4D point clouds, with stunning results.

CVPR 2026 VGI Workshop # VGI = Visual General Intelligence

- ❑ Let's think about "Intelligence"
- ❑ AGI is coming (This year? Next 10 years? Who knows?)
- ❑ But wait, current intelligence is language-first!
- ❑ Can intelligence also emerge from vision?

Visual & Language Intelligence

Primitive visual function	700M years
Eyes	540M years
Mammalian intelligence	200M years
Humans	300k years
Advanced language	100k years
Writing systems	5k years
Old English	1.5k years
Modern English	Hundreds years

Think about the history of Earth

Perspective of vision (540M years ago)

Visual perception emerged far earlier than language

Perspective of language (100k years ago)

Advanced language emerged relatively recently

Vision came first

Intelligence evolved together with perception, long before symbolic language existed

From the perspectives of vision and language, vision came first in the analogy of Earth

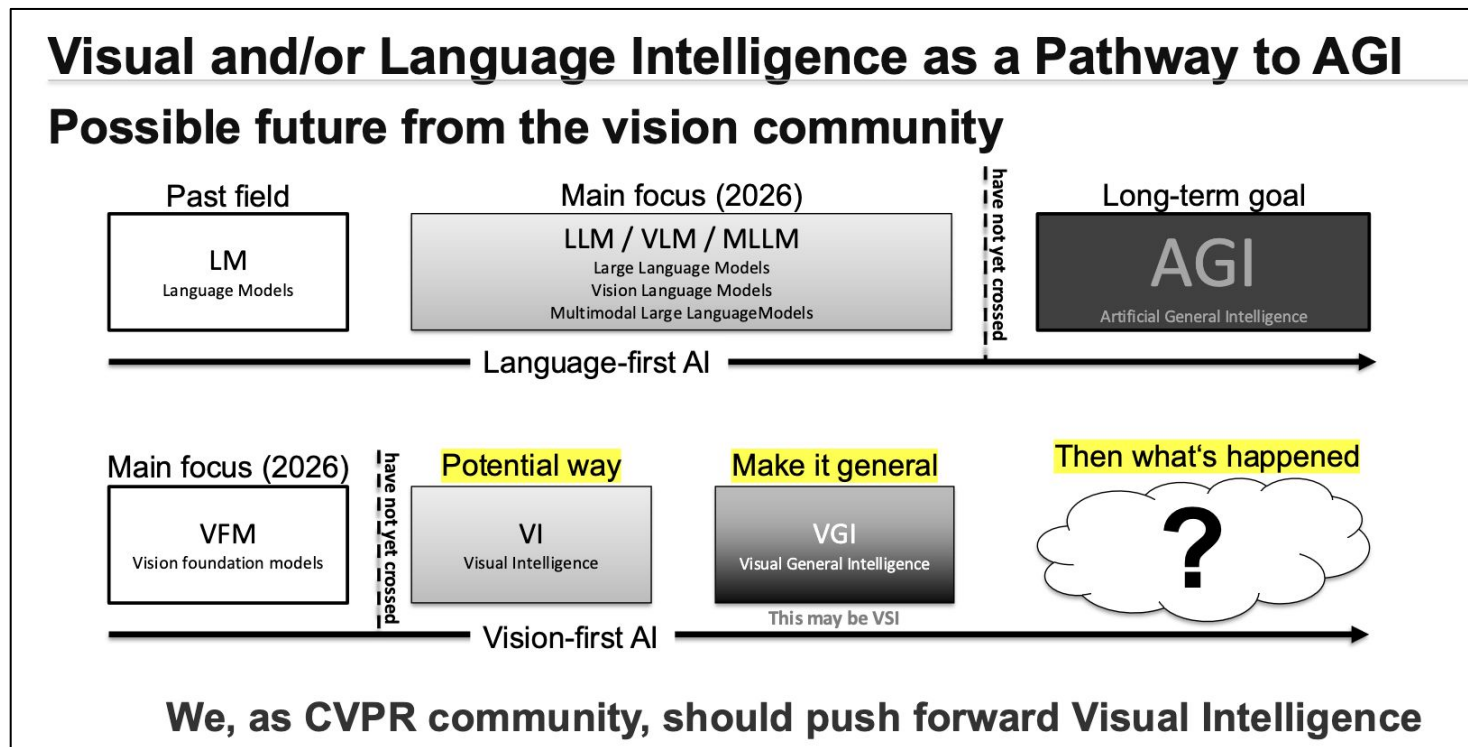
Slides from:

https://cdn.limitlab.xyz/cvpr2026/s/cvpr2026-vgi-workshop-opening-remarks-hirokatsu-kataoka_latest-en.pdf



CVPR 2026 VGI Workshop # VGI = Visual General Intelligence

- ❑ Is it possible to pass through the vision-first intelligence?
 - ❑ VFM – Visual Intelligence (VI) – VG(eneral)I ?
 - ❑ So what's happened after that?
 - ❑ At least we should improve visual function rather than language!



Slides from:

https://cdn.limitlab.xyz/cvpr2026/s/cvpr2026-vgi-workshop-opening-remarks-hirokatsu-kataoka_latest-en.pdf



CVPR 2026 VGI Workshop # VGI = Visual General Intelligence

- ❑ According to the definition [Legg & Hutter, 2007]:
 - ❑ Intelligence measures an agent's ability to achieve goals in a wide range of environments
- ❑ So it should be a multi(many?)-task solving model

Visual Intelligence (VI)?

Intelligence that emerges from visual experience

Visual Intelligence,

1) it's like a 'Visual GPT-3'

which has emergent abilities to solve general visual tasks

2) it's a 'multitask & multimodal VFM'

but it has capabilities that were not explicitly trained also emerge naturally

3) it's an 'ability to understand the diverse real world'

to understand, predict, compress, & generalize the dynamics of the world through visual experience

Slides from:

https://cdn.limitlab.xyz/cvpr2026/s/cvpr2026-vgi-workshop-opening-remarks-hirokatsu-kataoka_latest-en.pdf



CVPR 2026 VGI Workshop # VGI = Visual General Intelligence

- Although its not perfect at present,
 - We have some stable steps toward visual intelligence
 - Generative models, reconstruction models, inpainting, multimodal LLM, embodied intelligence...

Are we entering the era of visual intelligence?

Yes - The first steps toward 'visual intelligence'?

Examples from VGI speakers: Veo3, VGGT, Visual Jenga, GAIA-3, Molmo, Ego4D...



Veo3: Generalist Vision Encoder
(Video models are zero-shot learners and reasoners)



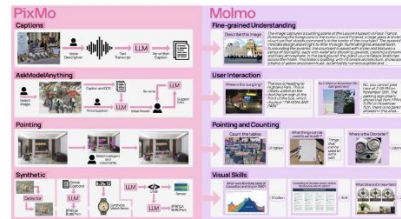
VGGT & VGGT-Ω: Geometry Understanding



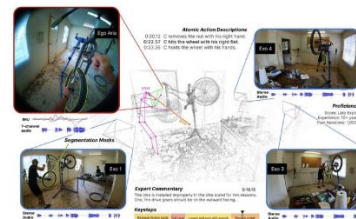
Visual Jenga: Physical Understanding



GAIA-3: Embodied Intelligence



Molmo: VI aligns language



Ego4D: human-centered abilities

Slides from:

https://cdn.limitlab.xyz/cvpr2026/s/cvpr2026-vgi-workshop-opening-remarks-hirokatsu-kataoka_latest-en.pdf



CVPR 2026 VGI Workshop # VGI = Visual General Intelligence

- ❑ Hypotheses from researchers:

Hypotheses:

- **Visual intelligence may...**
 - **emerge before language**
 - **emerge from generative, autoregressive, and reconstruction tasks**
 - **require efficient learning geometry, dynamics, and hierarchy (e.g., synthetic data / high-efficient data collection)**

Slides from:

https://cdn.limitlab.xyz/cvpr2026/s/cvpr2026-vgi-workshop-opening-remarks-hirokatsu-kataoka_latest-en.pdf



CVPR 2026 VGI Workshop # VGI = Visual General Intelligence

- ❑ But opposite position from a researcher:
 - ❑ Language has much lighter bytes than vision (images & videos)
 - ❑ LLM is still scaling, that could be an AGI

We don't need any new vision-centric paradigm shifts for AGI

- Self-supervised learning in computer vision saturates quickly when scaling (see: MAE work)
 - This is expected to be due to images containing less information dense content than text for the same amount of bytes
 - If we had 1 million x the compute, I would guess self supervised learning on vision starts to be useful on top of text (as we've run out of internet text data). But, we're far from getting this.
- You can do less with image and video outputs than you can with text outputs
 - If you have a great image/video generation model, it's unclear how to use it
 - Text outputs give you action

Which would you bet on - language or vision?

Of course, both is also a valid answer. In that case, do you think intelligence can emerge from visual input alone?

CVPR26: Multimodal LLM-as-Judge (1/3)

Who judges the judges? — the evaluator becomes a first-class artifact

- ❑ CVPR 2026 at a glance: record 16,092 submissions, 4,089 accepted (+42% YoY), ~25.4% acceptance rate.
- ❑ VLM / multimodal-LLM share of highlighted papers jumped 4.9% → 10.6% — the largest single-category swing this year.
- ❑ Meta-insight: as generation scales, reference metrics (BLEU / CIDEr / CLIPScore) can no longer measure quality.
 - ❑ So the field promotes the evaluator itself — VLM-as-judge & generative reward models — from an eval afterthought into a first-class research object.
- ❑ This deck reads CVPR'26 through that lens: how judges are built (Trend A) and whether we can trust them (Trend B).

Robotics related workshop: meta insights

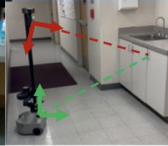
- ❑ Robotics is becoming a full-stack physical AI problem: data, models, sensors, simulators, deployment, and reliability must scale together.
 - ❑ Data: cross-embodiment robot data, human videos, simulation, generated trajectories, autonomous rollouts
 - ❑ Models: VLA, world-action models, diffusion policies, hierarchy, memory, promptable policies
 - ❑ Embodiment: tactile sensing, contact-rich manipulation, dexterous hands, hardware/edge deployment
 - ❑ Systems: modular pipelines, sim-to-real, deployment flywheel, reliability, multi-robot / robot web

- ❑ Future directions
 - ❑ From VLA to world-action model
 - ❑ Human videos to robotics, alignment is the key (action alignment, sensor alignment, structured representation alignment)
 - ❑ From synthetic scenes to real-video-derived interactive worlds
 - ❑ Tactile, force, contact-rich: for dexterous robotics
 - ❑ Long-horizon/longer-horizon tasks require hierarchy + memory + promptability
 - ❑ From benchmark success to production metric
 - ❑ Towards scalable physical AI (now already), which could also be extended to scalability in deployment

CVPR26: Meta Insights into Trends and Tendencies (38/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Saurabh Gupta](#), Title: "The case for modular robotics systems"

- ❑ How to find data to train robots?
 - ❑ Data: Large-scale tele-operation, creative ways of collecting data (UMI), video data, simulation, pre-trained models, classical methods
 - ❑ Motivation: high-level reasoning, whole-body control, contact-rich tasks
 - ❑ Winning approach will be the one that can handle all different data
 - ❑ Case study 1: [Opening articulated objects in the real world](#) (RSS 2025)
 - ❑ Approach: hierarchical method (perception, motion planning, adaptation, right image)
 - ❑ Core part: separate the task hierarchically, instead of generating a whole motion planning
 - ❑ Why did we do better?
 - ❑ Perception: bootstrapped from online object detection data + 2D image labeling
 - ❑ Motion planning: Mined from 1000s of objects in simulation
 - ❑ RUM policy: entirely trained on real-world demos
- | Perception | Transforms | Navigation | Motion Planning | Pre-Grasp Pose | Grasp Correction | Execution |
|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |
- ❑ Case study 2: [Precise Mobile Manipulation of Small Everyday Objects](#) (RA-L 2026)
 - ❑ Key idea: improving manipulation of small objects by using large vision models
 - ❑ How to keep detecting small objects: inpainting + tracking
 - ❑ Why did we do better?
 - ❑ Perception: large-vision models trained on lots of data
 - ❑ Visual servoing: benefits from hand-design that leverages the problem structure
 - ❑ RUM policy: again, entirely trained on real world demos

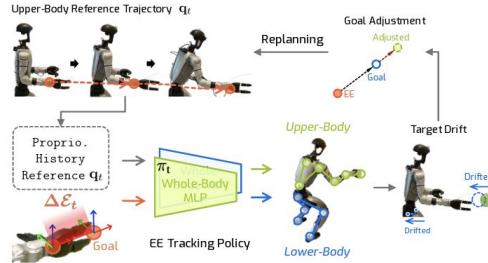
CVPR26: Meta Insights into Trends and Tendencies (39/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Saurabh Gupta](#), Title: "The case for modular robotics systems"

Case study 3: [Learning humanoid end-effector control for visual whole-body open-vocabulary object grasping](#)

- Overall approach: modular
- Perception: Open-vocabulary perception (SAM), grasp perception (AnyGrasp)
- Execution:



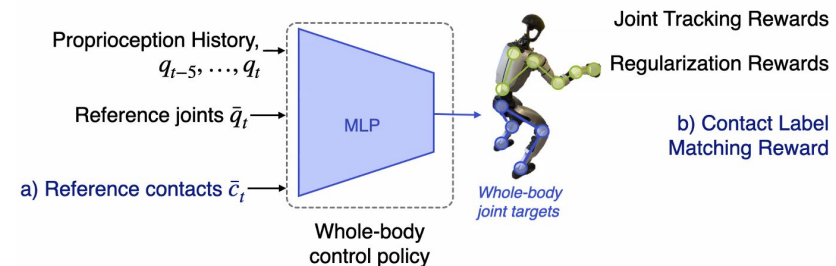
- Modularity enables easy extensions to other tasks
- Why did we do better?:
 - Perception: trained on 1M+ data samples
 - EE Tracker: trained in a highly randomized non-visual physical sim

Case study 4: ContactMimic: humanoid object interaction via contact control (ongoing work)

- Observation: Key points are not sufficient interface
- Overall idea: learn a whole-body key point + contact tracker

Takeaway:

- All different data sources will have a role to play
- The winning approach will be one that can leverage them all



CVPR26: Meta Insights into Trends and Tendencies (40/152)

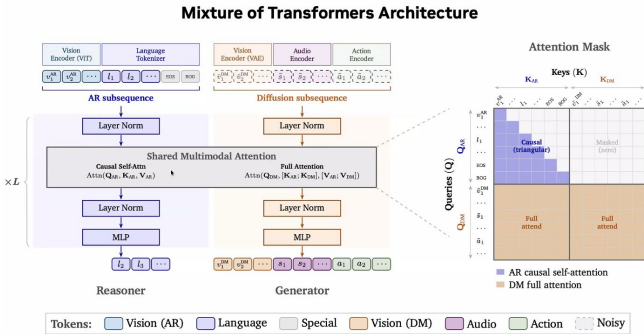
WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Ming-Yu Liu](#), Title: "Cosmos 3: Omnimodal World Model for Physical AI"

- ❑ The two faces of world models
 - ❑ Understanding-centric world models; Simulation-centric world models
 - ❑ Issues:
 - ❑ these two world models do not share the same representation
 - ❑ We need not only language or video-based world model, but also world model that deal with actions, interactions
- ❑ [Cosmos](#): Omnimodal world model:
 - ❑ Motivation:
 - ❑ world models are the foundation of generalizable physical AI
 - ❑ Cosmos as an omni-world model offers better data, better starting point, better environment for generalizable physical AI
 - ❑ Input and output: language, sensors, action
 - ❑ Model core: understanding / reasoning capability; perception / generation capability
 - ❑ Structure: unified MoT architecture
 - ❑ Cosmo is embedded in the system close to GPU
 - ❑ Models: text2image, text2video, policy model (top 1 in several benchmarks),
 - ❑ Position embedding: 3D MRoPE, base FPS: 24
 - ❑ Action representation: ego pose 9D, effector pose 9D, grasp state 15D, grasp state 1D

Result on reasoning and generation tasks

Model	Reasoning				Generation					
	General	Robotics	Smart infra.	Driving	Text2Image	Text2Video	Image2Video	Audio	FD: Robot	Policy: Robot
Cosmos3-Super	73.7	57.8	62.6	79.3	91.36*	80.0	82.8	7.31	26.0*	-
Cosmos3-Nano	69.6	55.1	61.0	76.0	84.61	79.4	82.7	7.34	25.5*	39.7*
Gemini 3.1 Pro†	77.5	58.2	58.6	47.2						
Qwen3-VL-32B	72.8	52.6	56.1	40.7						
Qwen3-VL-8B	68.9	48.5	52.7	46.4						
Gemma-4-31B	69.8	51.0	51.3	36.6						
Gemma-4-E4B	53.1	39.3	29.4	26.0						
Gemini 3 Pro Image†					90.85					
Qwen-Image-2512					84.25					
Veo-3.1†							79.1	82.6	7.45	
Wan2.2-A14B							78.0	81.3		
Ctrl-World									23.0	
$\pi_{0.5}$										28.1



Cosmos basic structure

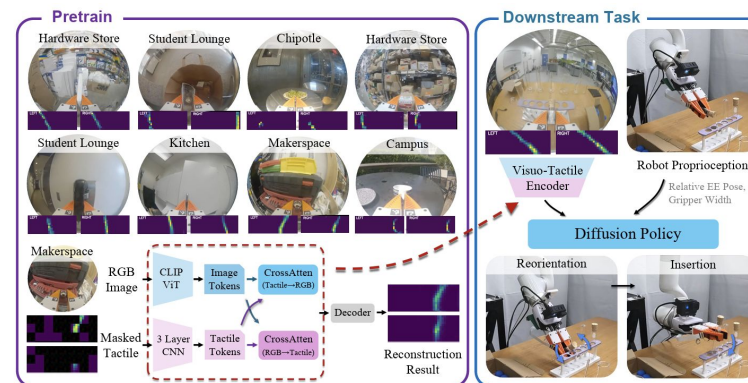
CVPR26: Meta Insights into Trends and Tendencies (41/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Yunzhu Li](#), Title: "Scaling Robot Learning via Structured World Models and Tactile Sensing"

- ❑ Foundation models for robotics
 - ❑ GPT, SAM, DINO, FoundationPose/Stereo...
- ❑ Robot foundation models
 - ❑ Pretrained on vision language models
 - ❑ RT-X, OpenVLA, Helix, Pi-0.5, DYNA-1
 - ❑ Pretrained on world / video models
 - ❑ WAM, Direct Video-Action Models
 - ❑ Issues for current models: Steerability, physical grounding, debuggability, scalability
- ❑ Important aspects for enhancing robot foundation models
 - ❑ Sensors:
 - ❑ [Scalable tactile sensing and learning for contact-rich robotic manipulation](#) (hand-on robot, Nature 2019)
 - ❑ Scalable tactile glove and CNNs to show that sensors uniformly distributed over the hand can be used to identify objects, estimate their weight and explore the typical tactile patterns that emerge while grasping objects
 - ❑ What hand-on robot can do: deal with visual occlusion, fragile object, in-hand manipulation
 - ❑ [Touch in the Wild: Learning Fine-Grained Manipulation with a Portable Visuo-Tactile Gripper](#) (NeurIPS, 2025)
 - ❑ Present a portable, lightweight gripper with integrated tactile sensors that enables synchronized collection of visual and tactile data, and a simple cross-modal model

Touch in the wild training, model, and downstream tasks illustration



CVPR26: Meta Insights into Trends and Tendencies (42/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Yunzhu Li](#), Title: "Scaling Robot Learning via Structured World Models and Tactile Sensing"

□ Important aspects for enhancing robot foundation models

□ Sensors:

□ [FlexiTac: A Low-Cost, Open-Source, Scalable Tactile Sensing Solution for Robotic Systems](#)

□ Sensor resolution: binary vs. continuous?

□ Structure:

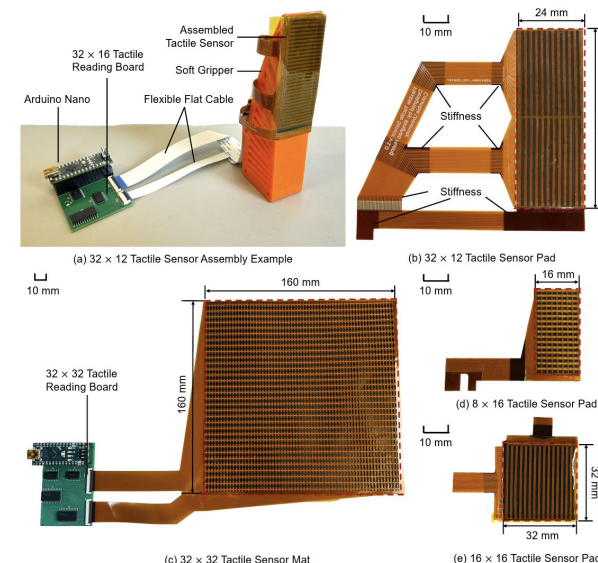
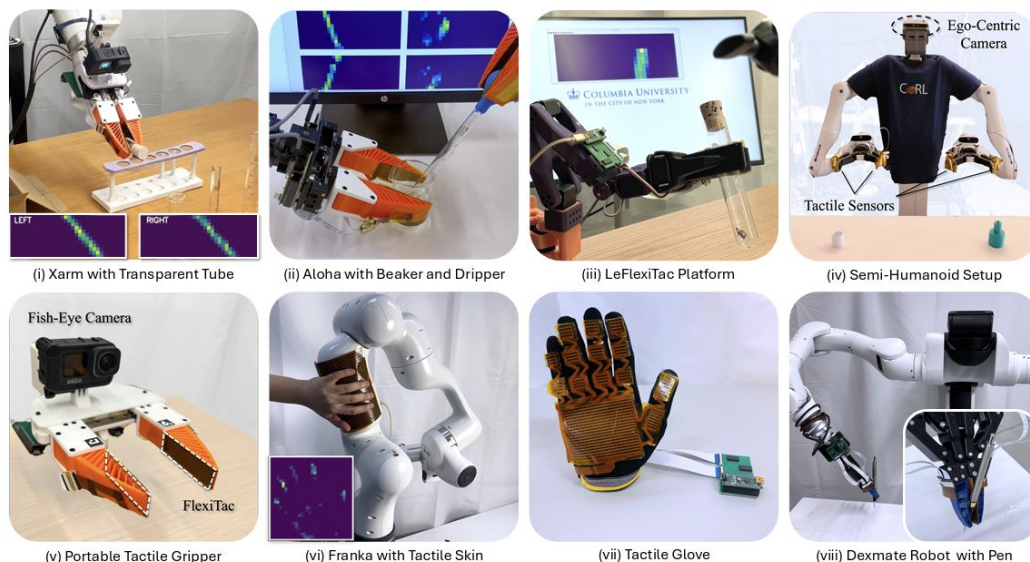
□ (i) thin, flexible tactile sensor pads that provide dense tactile signals

□ (ii) a compact multi-channel readout board that streams synchronized measurements for real-time control and large-scale data collection

□ Comparison with other tactile sensors:

□ Benefit of the proposed sensors: customizable, scalable, flexible, robustness / consistency, easier to simulate

FlexiTac deployments
across diverse platforms



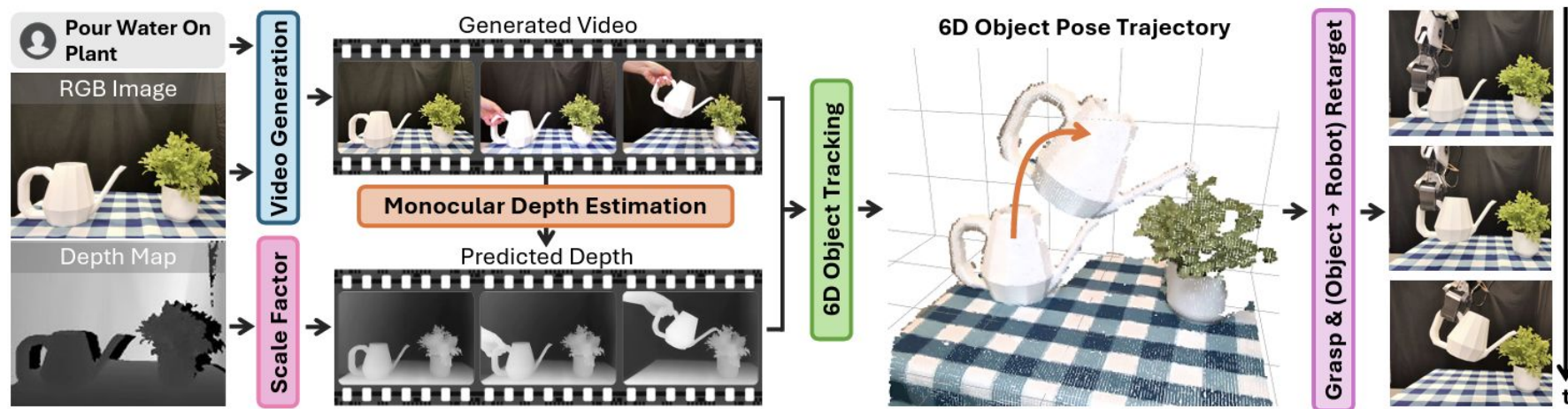
FlexiTac system
configurations

CVPR26: Meta Insights into Trends and Tendencies (43/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Yunzhu Li](#), Title: "Scaling Robot Learning via Structured World Models and Tactile Sensing"

- ❑ Important aspects for enhancing robot foundation models
 - ❑ Dynamics: learning structured world models from and for physical interactions
 - ❑ Graph-based neural dynamics (GBND)
 - ❑ Granular, application to object stowing
 - ❑ [A review of learning-based dynamics models for robotic manipulation](#)
 - ❑ Professor Li recommended paper which is based on his PhD thesis
 - ❑ [Robotic Manipulation by Imitating Generated Videos Without Physical Demonstrations](#) (ICLR, 2026)
 - ❑ RIGViD: enables robots to perform complex manipulation tasks—such as pouring, wiping, and mixing—purely by imitating AI-generated videos, without requiring any physical demonstrations or robot-specific training
 - ❑ Model structure: A video generation model generate videos, a VLM filters out videos that do not satisfy the input commands

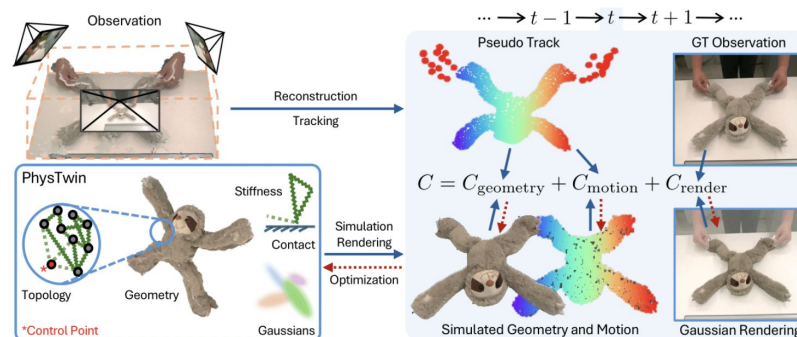


CVPR26: Meta Insights into Trends and Tendencies (44/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

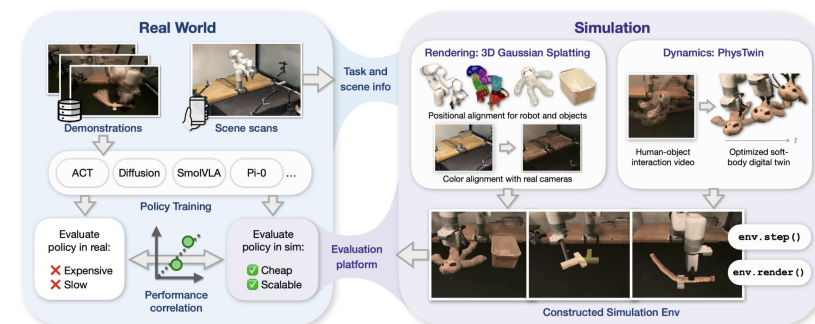
Invited talk: [Yunzhu Li](#), Title: "Scaling Robot Learning via Structured World Models and Tactile Sensing"

- Important aspects for enhancing robot foundation models
 - Dynamics: learning structured world models from and for physical interactions
 - [PhysTwin: Physics-Informed Reconstruction and Simulation of Deformable Objects from Videos](#) (ICCV 2025)
 - Creating a physical digital twin of a real-world object has immense potential in robotics, content creation, and XR



PhysTwins illustration

- Further extended PhysTwin as a scalable evaluation engine
 - [Real-to-Sim Robot Policy Evaluation with Gaussian Splatting Simulation of Soft-body interactions](#) (ICRA 2026)
 - Policy evaluation in sim environments constructed from real world, using Gaussian Splatting for rendering and soft-body digital twin for dynamics. (right figure)
- Sim-to-real transfer with zero world data?
 - Interact world simulator
 - Understand deformable-rigid interaction
 - Capture realistic dynamics
 - Scalable data generation for policy training
 - **One trial of simulation data nearly equals to one trial of real data**



CVPR26: Meta Insights into Trends and Tendencies (45/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

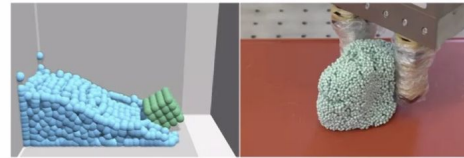
Invited talk: [Yunzhu Li](#), Title: "Scaling Robot Learning via Structured World Models and Tactile Sensing"

❑ Important aspects for enhancing robot foundation models

❑ Dynamics: learning structured world models from and for physical interactions

❑ Other recommended papers

Structured world models for simulating and manipulating deformable objects



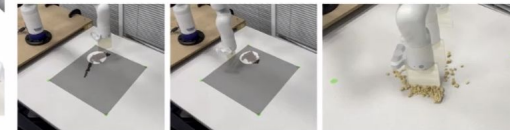
Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B. Tenenbaum, Antonio Torralba, Learning Particle Dynamics for Manipulating Rigid Bodies, Deformable Objects, and Fluids, **ICLR-19**



Haochen Shi, Huazhe Xu, Samuel Clarke, Yunzhu Li, Jiajun Wu, RoboCook: Long-Horizon Elasto-Plastic Object Manipulation with Diverse Tools, **CoRL-23 Best Systems Papers Award**



Haonan Chen, Yilong Niu, Kaiwen Hong, Shuijing Liu, Yixuan Wang, Yunzhu Li, Katherine Driggs-Campbell, Predicting Object Interactions with Behavior Primitives: An Application in Stowing Tasks, **CoRL-23 Finalist - Best Paper/Best Student Paper Awards**

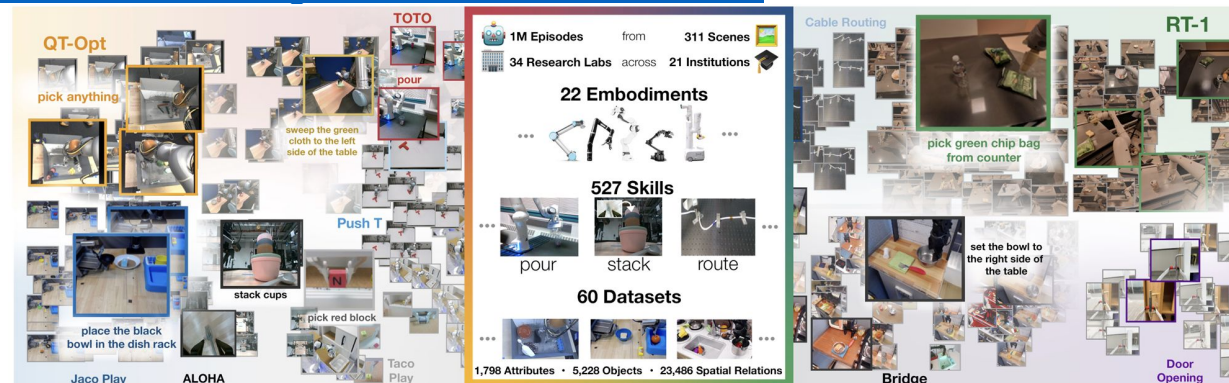


Yixuan Wang*, Yunzhu Li*, Katherine Driggs-Campbell, Li Fei-Fei, and Jiajun Wu, Dynamic-Resolution Model Learning for Object Pile Manipulation, **RSS 2023, Best Paper Award at IROS 2023 Workshop**

42

❑ Data:

❑ [Open X-Embodiment: Robotic Learning Datasets and RT-X Models](#)

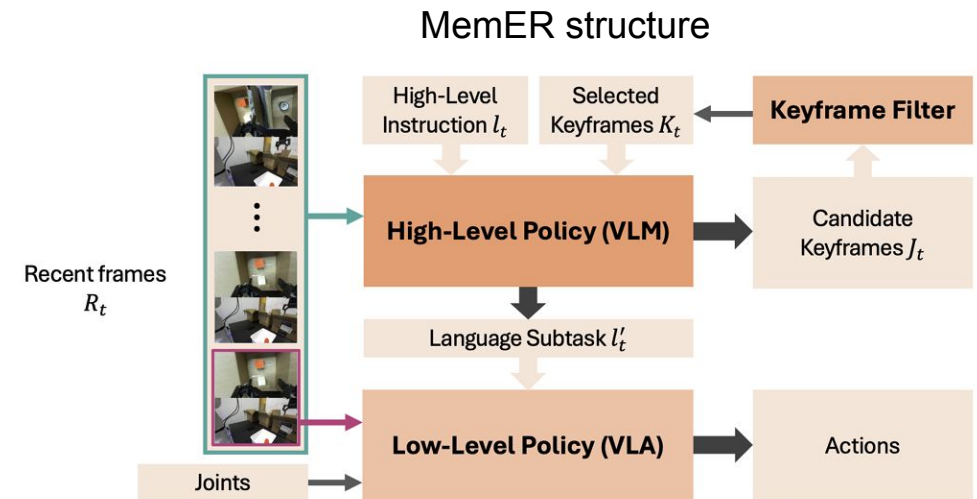


CVPR26: Meta Insights into Trends and Tendencies (46/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Chelsea Finn](#), Title: "Remember at Different Levels of Abstraction"

- ❑ Most state-of-the-art foundation models lack memory
 - ❑ Fundamental motor skills generally don't require it
- ❑ Can we develop memory in a way that is compatible with robot foundation models?
 - ❑ Difficulty: Spurious correlations, substantial computational requirements
 - ❑ Background: imitation learning with memory
 - ❑ Adding memory to observations often hurts policy performance
 - ❑ Adding memory leads to substantial GPU memory & computation costs
- ❑ [RoboMME](#) (ICML 2026):
 - ❑ Addresses spurious correlations and compute challenges
 - ❑ Key finding: no single form of memory works well across all datasets
- ❑ Approach 1: selectively store & retrieve frames
 - ❑ [Scaling up memory for robot control via experience retrieval](#) (ICLR 2026)
 - ❑ Core ideas:
 - ❑ Retrieve a few key frames
 - ❑ Long-context memory for high-level policy via VLM
 - ❑ Results:
 - ❑ Use segment-boundary is effective
 - ❑ MemER is close to human high-level policy

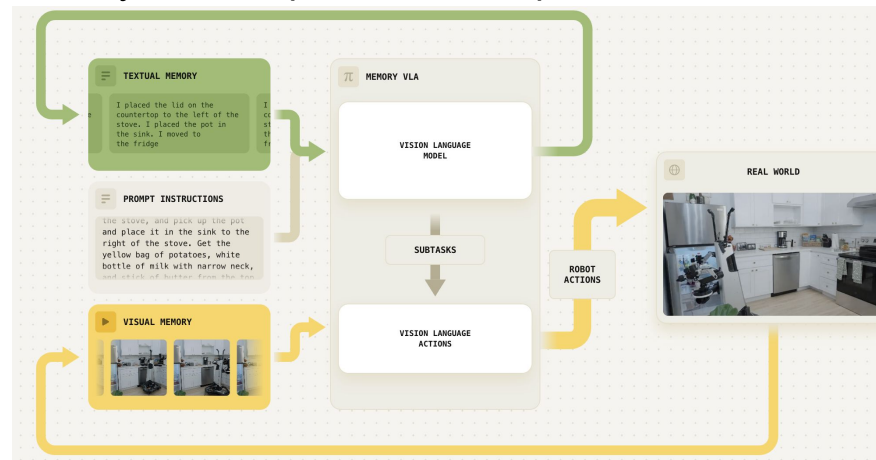


CVPR26: Meta Insights into Trends and Tendencies (47/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Chelsea Finn](#), Title: "Remember at Different Levels of Abstraction"

- ❑ Approach 2: [Compression of both visual and text memory](#)
 - ❑ Image-based Short-horizon memory:
 - ❑ Sparse temporal attention
 - ❑ Language-based longer-horizon memory:
 - ❑ Instead of using list of steps, use a short paragraph describing the past actions
 - ❑ Train the model to update + compress language memory
 - ❑ Takeaways:
 - ❑ General purpose memory should operate on multiple timescales



Long and short-term memory

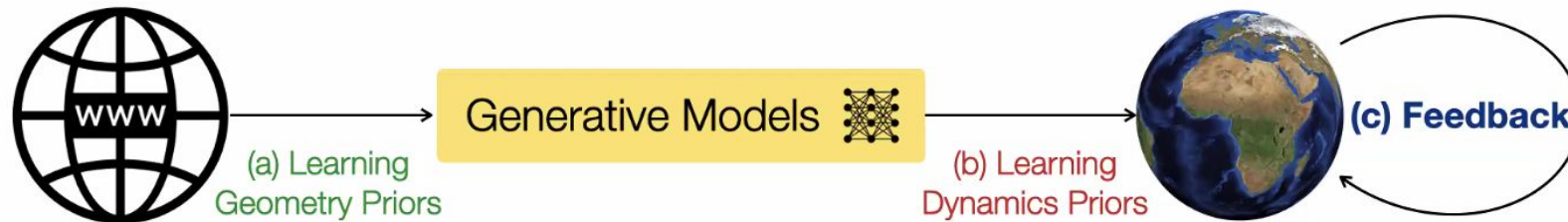
- ❑ General purpose memory systems might need the combination of approach 1 and 2
- ❑ Open questions:
 - ❑ Can robots achieve zero-shot compositional generalization?
 - ❑ Can we develop promptable policies?
 - ❑ Can robots improve to high performance with experience?

CVPR26: Meta Insights into Trends and Tendencies (48/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Ruoshi Liu](#), Title: "Generative Vision Models for Physical Intelligence"

- ❑ Can we use generative models of images and videos for physical tasks?
- ❑ Goal: computer vision systems that can intelligently interact with the unstructured physical world



- ❑ (a) Learning geometry priors for perception
 - ❑ [Humans As Light Bulbs: 3D Human Reconstruction From Thermal Reflection](#) (CVPR 2023)
 - ❑ Human bodies are infrared light bulbs; many objects are more reflective in infrared
- ❑ (b) Learning dynamics from interactions
 - ❑ [Robot planning as video generation](#) (CoRL 2024)
 - ❑ Visuomotor policy = dream (generate video) + imitate (robot execution)
 - ❑ What's missing from internet human videos?
 - ❑ Inner force, contact information are missing
 - ❑ ForceBand: Learning forceful manipulation with sEMG (In Submission, CoRL 2026)
 - ❑ Input: EMG, IMU
 - ❑ Model output: generates fingertips forces
 - ❑ Transfer to forceful policy for robot manipulation

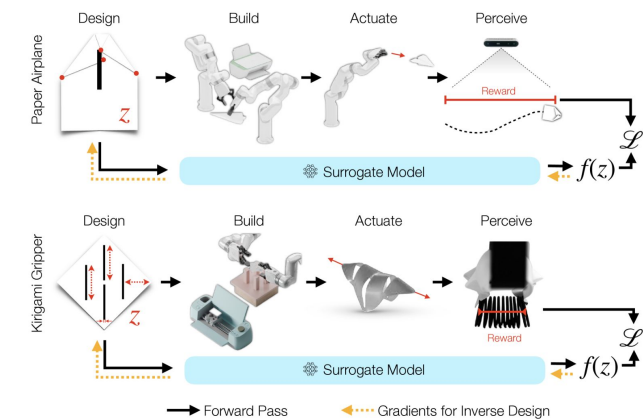
CVPR26: Meta Insights into Trends and Tendencies (49/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Ruoshi Liu](#), Title: "Generative Vision Models for Physical Intelligence"

- ❑ (c) Feedback:
 - ❑ Learning underwater, the [AquaBot](#) (ICRA 2025)
 - ❑ Core idea:
 - ❑ fully autonomous manipulation system that combines behavior cloning from human demonstrations with self-learning optimization
 - ❑ improves beyond human teleoperation performance
 - ❑ AquaBot can solve unseen object grasping, long-horizon tasks as well
 - ❑ Learning from feedback for policy acceleration
- ❑ [Feedback loop for paper airplane design](#)
- ❑ [Grasping with paper kirigami](#)
 - ❑ Core idea:
 - ❑ Self-supervise as feedback
 - ❑ Optimizing not only the VLA part but also task solution design part
- ❑ Robotic handshakes with human preference (in submission, CoRL 2026)
 - ❑ Handshaking is a challenging HRI task
 - ❑ Model: imitation learning + tactile adaptation
 - ❑ Diffusion policy with reward guidance

Feedback loop for paper airplane design



CVPR26: Meta Insights into Trends and Tendencies (50/152)

WS Bridging Vision, Language, and Action: What's Missing in Actionable Visual Perception for Robotics

Invited talk: [Marco Pavone](#), Title: "End-to-End Reasoning Models for Autonomous Driving: Design, Training, and Simulation"

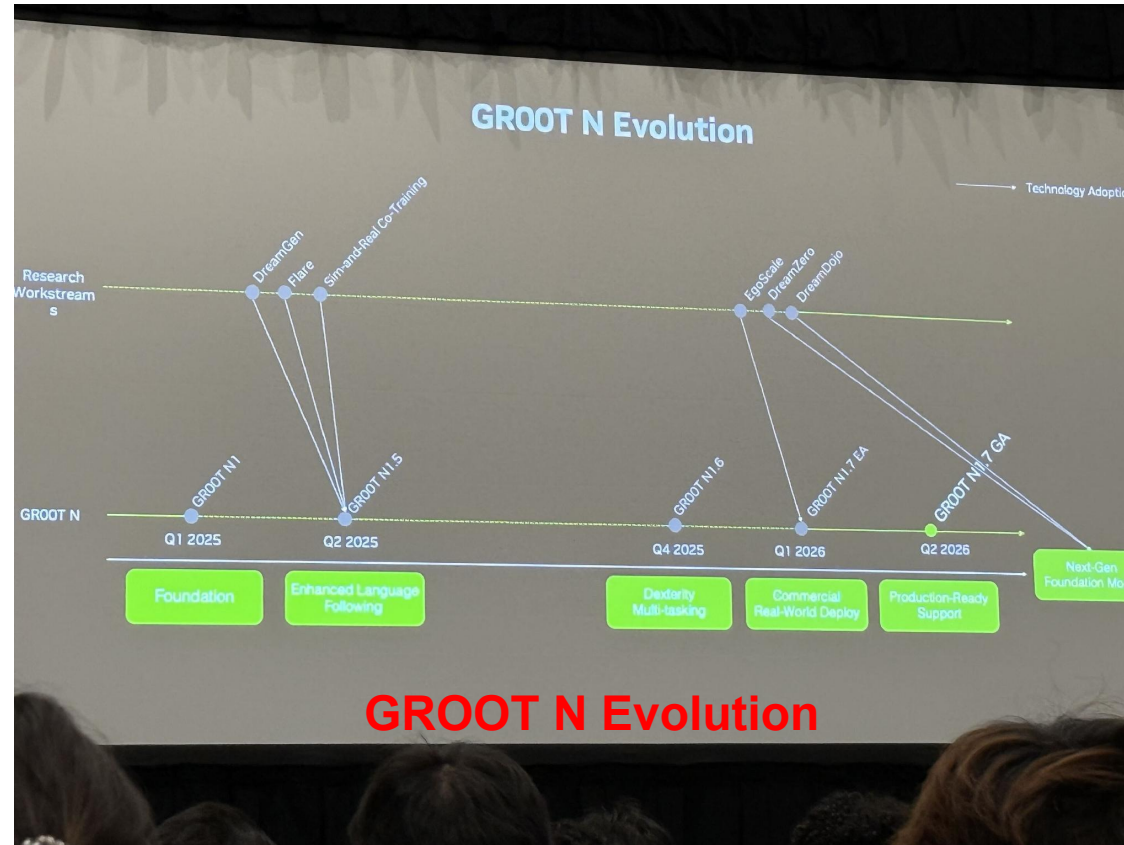
- ❑ [NVIDIA Alpamayo platform](#)
 - ❑ Model, Simulation & Training, Data & Tool
 - ❑ Based on cosmos-reason backbone
 - ❑ Multi-stage training pipelines
 - ❑ A typical reasoning trace:
 - ❑ What's happening? What's the risk? What to do next;
- ❑ Scaffolding reasoning with chain-of-causation (CoC) & RL post-training alignment
 - ❑ Challenges: causal confusion; vague behavior description; superficial reasoning
 - ❑ Approach: from CoT to CoC
- ❑ Reasoning representations
 - ❑ Counterfactual reasoning
- ❑ Enabling closed-loop RL
 - ❑ Closed-loop RL optimizes the closed-loop behavior output
 - ❑ Do we need closed-loop training?
 - ❑ The time for closed-loop training of AVs is now
 - ❑ Announced AlpaSim End-to-End Closed-Loop Driving Challenge
- ❑ Agentic AI for auto-driving

CVPR26: Meta Insights into Trends and Tendencies (51/152)

Tutorial The Full Stack of Physical AI Simulation, Foundation Models, and Edge Deployment for Next-Generation Robotics Applications

Invited talk: [Yugui Xie](#), Title: “Foundation Models for Generalist Robotics – The development of GROOT from N1 to N1.7”

- ❑ Isaac GROOT:
 - ❑ open robot foundation model; simulation frameworks; synthetic data generation pipeline; Jetson Thor Robotics Computer
 - ❑ Jetson Thor: AI compute; energy efficient;
- ❑ Data Pyramid for generalist humanoid robots:
 - ❑ Online videos (unstructured; passive); simulation (hard to scale in practice); real robots (active agency)
- ❑ GROOT N Evolution: from N1 to N1.7



CVPR26: Meta Insights into Trends and Tendencies (52/152)

Tutorial [The Full Stack of Physical AI Simulation, Foundation Models, and Edge Deployment for Next-Generation Robotics Applications](#)

Invited talk: [Yuqi Xie](#), Title: “Foundation Models for Generalist Robotics – The development of GROOT from N1 to N1.7”

❑ GROOTN1 to N1.7 comparison (Generated by ChatGPT, 5.5 Pro; manually verified)

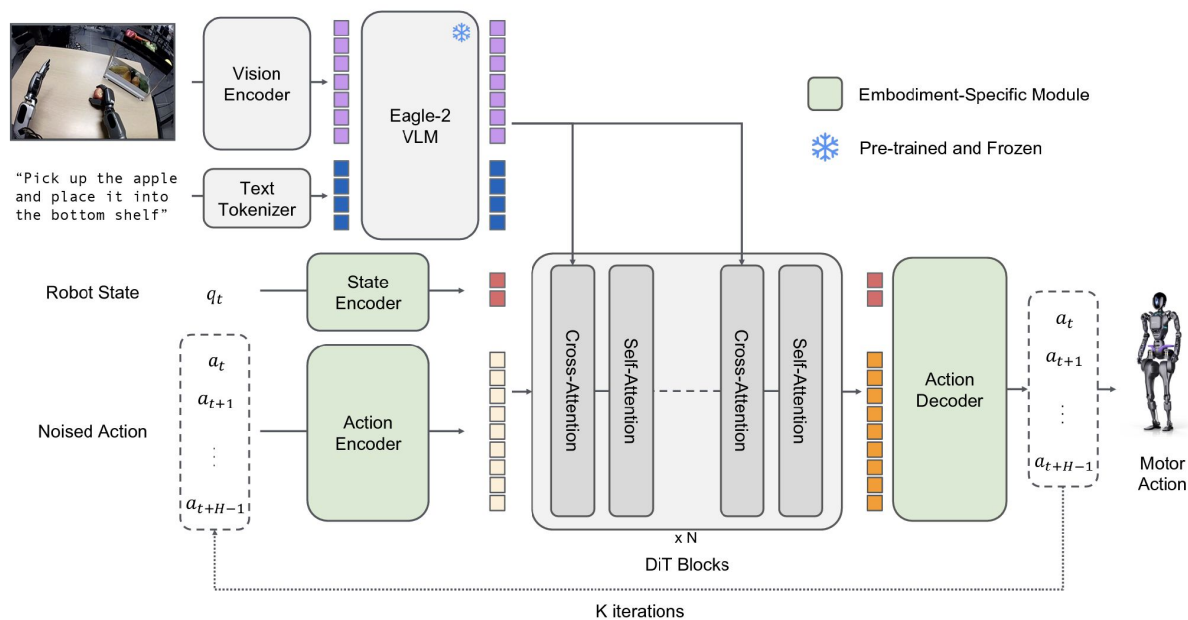
Version	Role	Model / architecture	Training data	Training setup	Key upgrade
GROOT N1	First open generalist humanoid VLA	Eagle-2 VLM + DiT; 2.2B	Data pyramid: human video, sim, neural, real robot	~50k H100-hrs; 16-action chunk 63.9 ms on L40	Unified cross-embodiment baseline
GROOT N1.5	Better grounding + language following	Eagle 2.5 VLM; frozen VLM; +FLARE	N1 mix + OpenXE, DexMG, DreamGen, AgiBot-Beta	250k steps; 1k H100; batch 16,384	Learns from human ego videos
GROOT N1.6	Bimanual + loco-manipulation scaling	Cosmos-2B VLM; 32-layer DiT; relative actions	N1.5 mix + YAM, Genie1, R1 Pro sim, Unitree G1	300k steps; post-train 10–30k steps	Better bimanual dexterity + locomanipulation
GROOT N1.7	Open reasoning VLA;	3B; Cosmos-Reason2-2B + 32-layer DiT	Robot data + 20k+ h EgoScale human video	16GB+ VRAM inference; 40GB+ recommended fine-tune	N1.6-level performance; better generalization + language following

Tutorial The Full Stack of Physical AI Simulation, Foundation Models, and Edge Deployment for Next-Generation Robotics Applications

Invited talk: [Yuqi Xie](#), Title: “Foundation Models for Generalist Robotics – The development of GROOT from N1 to N1.7”

❑ [Groot N1: Humanoid robot foundation model](#)

- ❑ Structure: VLM (Encode image and language) + Diffusion Transformer (encode robot state; action tokens; output from VLM)
- ❑ Embodiment-Specific Adapters for different robotics
- ❑ Diffusion Transformer: flow matching / denoising-style action generation
- ❑ Key designs:
 - ❑ Latent action: VQ-VAE-style latent action model, to eventually transferring latent action to real robot actions using small-scale robot data. Useful for transfer between different embodiments, also for human videos to robot transfer.
 - ❑ Neural trajectories: video world model generates robot videos -> latent action or IDM predicts pseudo-actions
 - ❑ Simulation trajectories: physics simulator + DexMimicGen generates robot rollouts -> large-scale synthetic robot demonstrations



	Latent Actions	Neural Trajectories	Simulation Trajectories
Real-Robot Datasets	✓	✓	✓
Simulated Robot Datasets	✓	✓	
Human Video Datasets	✓		

CVPR26: Meta Insights into Trends and Tendencies (54/152)

Tutorial The Full Stack of Physical AI Simulation, Foundation Models, and Edge Deployment for Next-Generation Robotics Applications

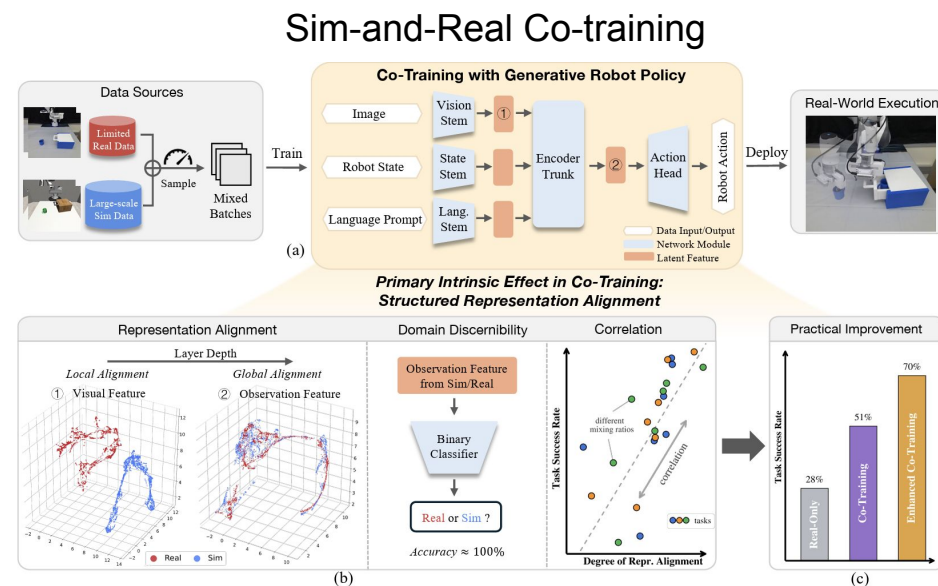
Invited talk: [Yugui Xie](#), Title: “Foundation Models for Generalist Robotics – The development of GROOT from N1 to N1.7”

❑ [From N1 to N1.5:](#)

- ❑ Major improvements: future latent representation alignment; improved neural trajectories
- ❑ Sim-and-Real Co-training:
 - ❑ Data: from real-robot data; simulated trajectories; neural trajectories (with visual world models: Cosmos)
 - ❑ [A Mechanistic Analysis of Sim-and-Real Co-Training in Generative Robot Policies](#)
 - ❑ Structured representation alignment is important
 - ❑ Importance reweighting
 - ❑ Representation should preserve part domain-specific information to prevent from collapsed representation

❑ GROOT Mimic: simulation-based data generation system (MimicGen, DexMimic)

- ❑ [MimicGen](#) and [DexMimicGen](#)
- ❑ Transfer small-scale human demonstration to large-scale sim data
- ❑ Physical consistency; scalability;
- ❑ Limited for soft objects; complicated contact;



CVPR26: Meta Insights into Trends and Tendencies (55/152)

Tutorial The Full Stack of Physical AI Simulation, Foundation Models, and Edge Deployment for Next-Generation Robotics Applications

Invited talk: [Yuqi Xie](#), Title: "Foundation Models for Generalist Robotics – The development of GROOT from N1 to N1.7"

❑ [From N1 to N1.5:](#)

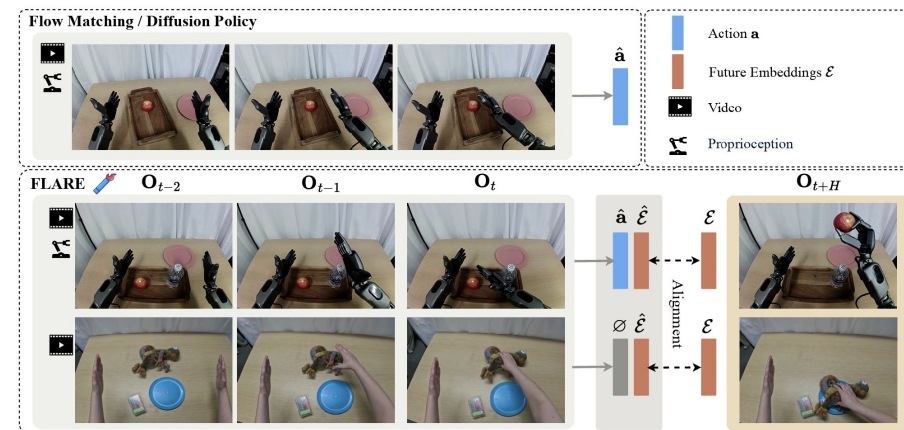
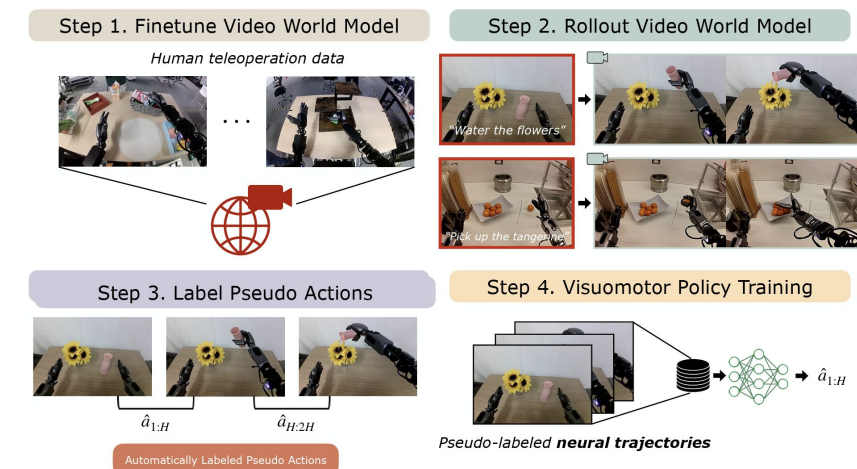
❑ [GROOT Dreams \(DreamGen\): synthesizing trajectories with video world models](#) (CoRL 2025)

- ❑ Core: use video world model as synthetic robot data generator to generate trajectories and adapt to different embodiments
- ❑ Four steps:
 - ❑ Fine-tune video world model to adapt to target embodiment
 - ❑ Using rollout synthetic robot videos to adapt to different tasks and environment
 - ❑ Pseudo action labeling to use IDM / LAPA latent action model to get action labels from videos
 - ❑ Train / fine-tune policy: learn policy
- ❑ Behavior generalization: training on one task, finetune on neural trajectories of new tasks, then deploy in new tasks

❑ [Robot learning with implicit world modeling](#) (FLARE)

- ❑ In addition to conventional flow-matching policy, adding Future LATent REpresentation Alignment (FLARE)
- ❑ Add future tokens generation to DiT and align with vision-language embedding of future observation instead of frame-level alignment

DreamGen overview



Compare FLARE with
conventional flow-matching policy

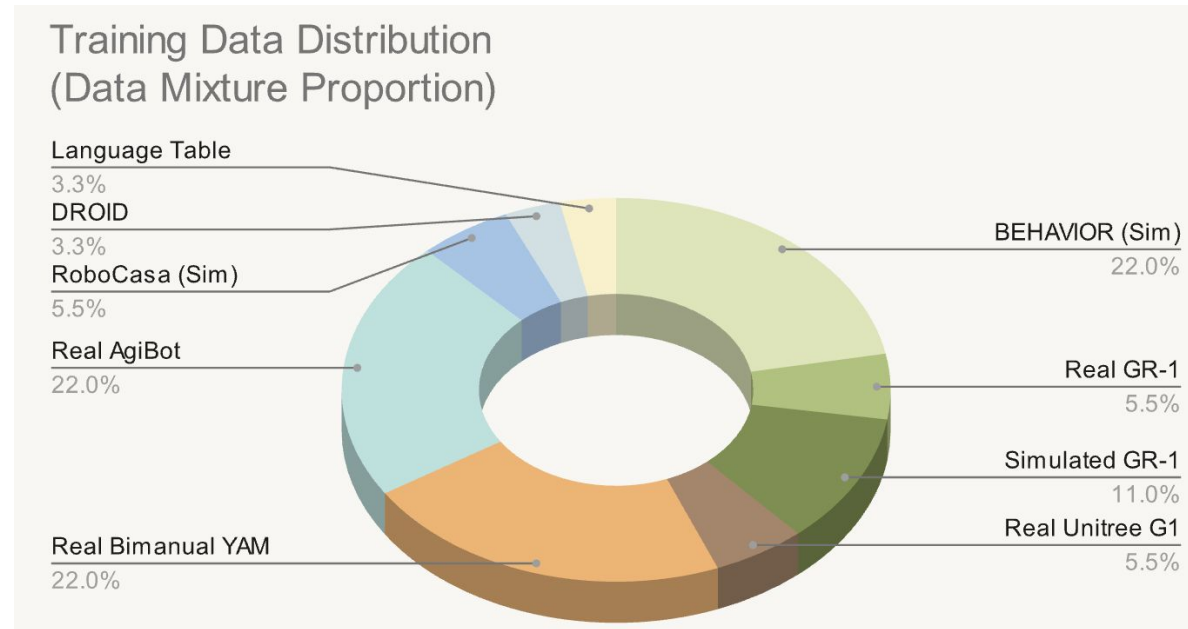
CVPR26: Meta Insights into Trends and Tendencies (56/152)

Tutorial The Full Stack of Physical AI Simulation, Foundation Models, and Edge Deployment for Next-Generation Robotics Applications

Invited talk: [Yuqi Xie](#), Title: “Foundation Models for Generalist Robotics – The development of GROOT from N1 to N1.7”

❑ [GrootN1.6](#)

- ❑ Expand to different types of embodiments: core model design – state-relative action chunks
- ❑ Model
 - ❑ Interleaved language and vision layers in DiT
 - ❑ NVIDIA Cosmos-2B VLM variant as backbone
 - ❑ Predicts state-relative action chunks for most embodiments, rather than absolute joint angles or EEF positions
- ❑ Training data examples
 - ❑ Bimanual YAM arms
 - ❑ AGIBot Genie1
 - ❑ Simulated Galaxea R1 Pro on the BEHAVIOR suite
 - ❑ Whole-Body Locomanipulation with Unitree G1



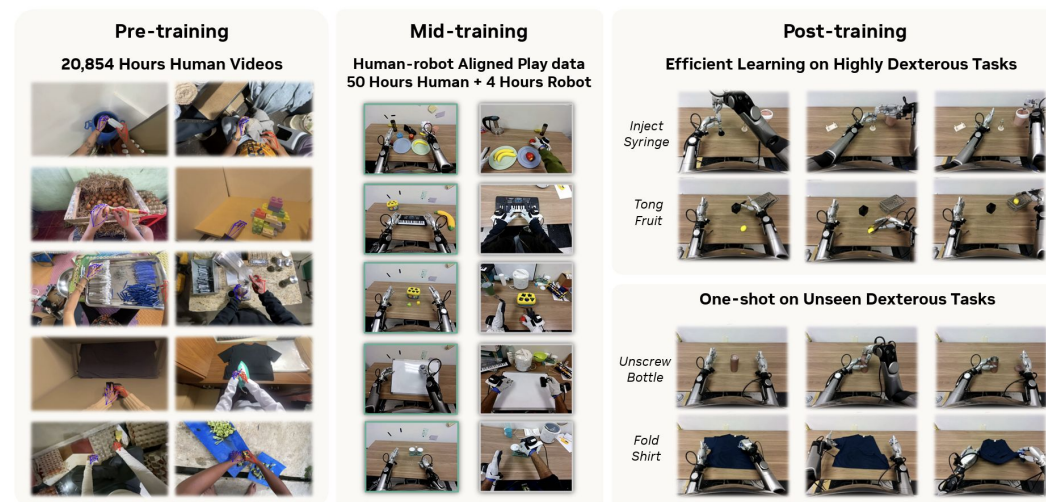
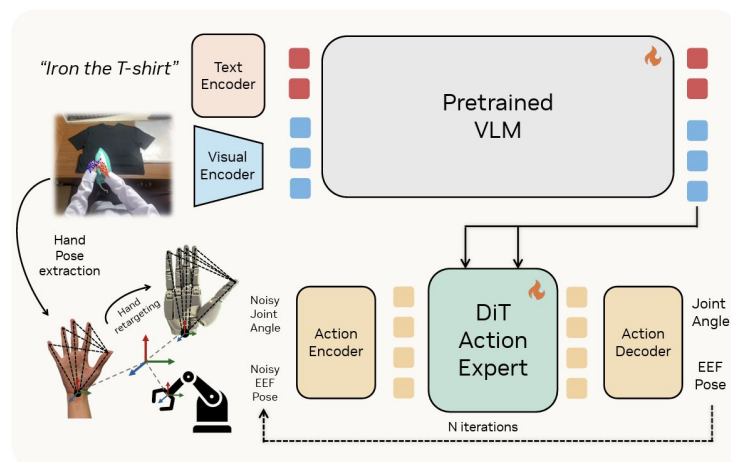
CVPR26: Meta Insights into Trends and Tendencies (57/152)

Tutorial The Full Stack of Physical AI Simulation, Foundation Models, and Edge Deployment for Next-Generation Robotics Applications

Invited talk: [Yuqi Xie](#), Title: “Foundation Models for Generalist Robotics – The development of GROOT”

❑ [GrootN1.7](#)

- ❑ Relative end-effector action space to enable human videos to robot policy learning
- ❑ Data and model core: [EgoScale](#)
 - ❑ Human to dexterous manipulation transfer framework built on large-scale egocentric human data
 - ❑ Scaling robot intelligence with human-scale data
 - ❑ Shows scaling law as the pretraining dataset size increases
 - ❑ Train with human video with wrist position and hand joints
 - ❑ 20,854 hours of human videos
 - ❑ Training recipes:
 - ❑ Stage 1: large-scale human pretraining
 - ❑ 22-DoF Sharpa robot hand action space
 - ❑ Stage 2: Enable cross-embodiment transfer by mid-training
 - ❑ Align robot with human
 - ❑ Egocentric and wrist views
 - ❑ Collected in-house



CVPR26: Meta Insights into Trends and Tendencies (58/152)

WS Scalable Robot Learning Systems

Invited talk: [Sergey Levine](#), Title: “Make it easier to scale robot data”

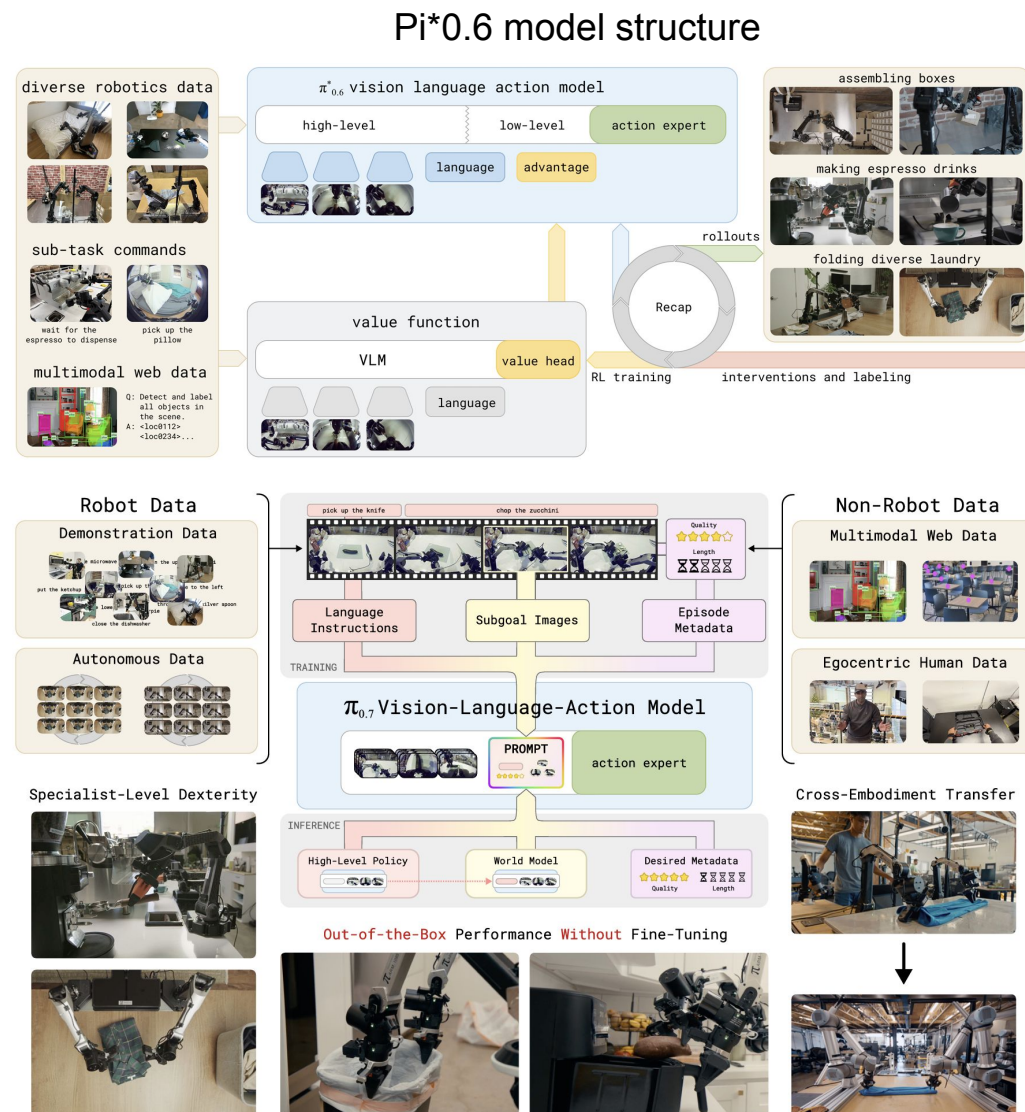
- ☐ Compositional generation in robotics
- ☐ The recipe
 - ☐ Pyramid of data (videos, sim, robot data)
 - ☐ Pre-training and post-training
 - ☐ Surprising key finding: scaling robot data makes it easier to scale robot data
- ☐ The data sources we can use:
 - ☐ (1) Cross-embodiment data
 - ☐ [RT-X: the first cross-embodiment dataset for VLAs](#)
 - ☐ [Pi0](#): 10k general robot data, 20 post-training data
 - ☐ [Pi0.5](#): scaling while using increased locations
 - ☐ (2) Human videos:
 - ☐ The more efficient using human data while adding more robot data.
 - ☐ Why does it work: when we add more robot data, the model learns to represent human data in the same way as robot data
 - ☐ (3) Language “coaching” data
 - ☐ [Interacting with pi0](#)
 - ☐ Learning from retrospective synthetic data (synthetic user prompts: synthetic user interjections / corrections; synthetic robot verbal responses; target low-level skill labels)
 - ☐ Pi0.5: an integrated high-level/low-level model
 - ☐ Training the high level with verbal instruction
 - ☐ Pi0.7: multimodal reasoning
 - ☐ Language thoughts and generated subgoal image

CVPR26: Meta Insights into Trends and Tendencies (59/152)

WS Scalable Robot Learning Systems

Invited talk: [Sergey Levine](#), Title: “Make it easier to scale robot data”

- ❑ The data sources we can use:
 - ❑ Autonomous data
 - ❑ [Pi0.6: a VLA trained with RL](#)
 - ❑ Vision language action model
 - ❑ Value function: VLM-based
 - ❑ Link action model with advantage tokens generated by value function
 - ❑ [Putting it all together into a distillation-centric robot foundation models \(Pi0.7\)](#)
 - ❑ Model: adding a world model (video generation model) into pi0.6
 - ❑ Data: combination of all kinds of data
- ❑ Takeaways:
 - ❑ Cross-embodiment data brings the blessing of generality
 - ❑ Training on lots of robot data makes it easier to use human data
 - ❑ Train to follow instructions makes it possible to learn from verbal instructions
 - ❑ RL & detailed prompts make it possible to include plentiful autonomous data



CVPR26: Meta Insights into Trends and Tendencies (60/152)

WS Scalable Robot Learning Systems

Invited talk: [Sergey Levine](#), Title: “Make it easier to scale robot data”

Pi* models comparison (by ChatGPT, 5.5 Pro)

Model	Architecture	Data / scale	Training method	Best fit
$\pi 0$	PaliGemma VLM + flow action expert; continuous action chunks	10k+ h robot data; single-arm, dual-arm, mobile manipulator; open robot data	Broad pretrain + task fine-tune	Base dexterous VLA
$\pi 0.5$	$\pi 0$ -style VLA + explicit high-level subtask $\rightarrow$ low-level action hierarchy	Mobile-home + static multi-env + cross-embodiment + web/HL/VI data	Heterogeneous co-training + mobile manipulation fine-tune	Open-world household tasks
$\pi^* 0.6$	$\pi 0.6$ backbone: Gemma3 4B + SigLIP + 860M action expert; + advantage conditioning	$\pi 0.5$ -like mixture + autonomous rollouts + reward/intervention data	RECAP: offline RL + value/advantage labels + on-robot experience	High reliability / throughput
$\pi 0.7$	5B model: Gemma3 4B + history/video encoder + 860M action expert; rich prompts	Demos + failures + autonomous data + interventions + human ego + web data	Context-conditioned pretraining; metadata/subgoals/history; prompt dropout	Steerable generalist

CVPR26: Meta Insights into Trends and Tendencies (61/152)

WS Scalable Robot Learning Systems

Invited talk: [Sergey Levine](#), Title: “Make it easier to scale robot data”

Model comparison (by ChatGPT, 5.5 Pro)

Series	Best mental model	Data engine	Where it shines	Caveat
PI π	Generalist manipulation VLA	Robot data + co-training + RL/context	Dexterity, open-world, steerability	Newer models less open
NVIDIA GR00T	Humanoid foundation stack	Robot/sim/neural + EgoScale human video	Humanoid/bimanual deployment	NVIDIA stack + EA caveats
Google RT/Gemini Robotics	Reasoning-rich robot AI	Web-scale VLM + robot data	Semantic planning + cross-embodiment	Mostly closed/trusted tester
Open VLA stack	Academic/open baselines	OXE / LeRobot community data	Fine-tuning and reproducibility	Less turnkey production reliability
RDT / LBM	Scaled diffusion policies	Large robot rollouts + rigorous eval	Bimanual/whole-body control	Less semantic/VLM-heavy
World-model policies	Video model as policy	Pretrained video/world models + demos	Physical prediction + planning	Newer, compute-heavy

CVPR26: Meta Insights into Trends and Tendencies (62/152)

WS Scalable Robot Learning Systems

Invited talk: [Jason Ma](#), Title: “A recipe for scaling robot learning for the real world”

- ❑ [DYNA](#): robust foundation model autonomy in the real world
 - ❑ Research + deployment flywheel
- ❑ Data pyramid for the real world
 - ❑ High-quality deployment data (most refined); diverse on-robot data; diverse off-robot data (most broad)
 - ❑ 200k+ hours to-date
- ❑ General and reliable models
 - ❑ Dyna Model architecture
 - ❑ Reasoning model -> (language instructions) -> world-action model -> future video / trajectory prediction
 - ❑ Rapid post-training to any task (1 hour)
 - ❑ How to develop reliable robot foundation models (~100% accuracy)?
 - ❑ The first policy that got to 24hr reliability
 - ❑ Pre-train VLA; train reward model; reward-weighted regression for post-training
 - ❑ Often get stuck in an error state for 2-3 minutes
 - ❑ Add human-in-the-Loop active learning
 - ❑ Achieved 99.45 success rate with 850+ folded napkins in 24hrs
 - ❑ The bitter lesson of real-world robotics
 - ❑ Focus on system building and scaling; Don't chase fancy algorithms or architecture; by scaling systems and deploying robots, it becomes obvious what problems need to be solved

WS Scalable Robot Learning Systems

Invited talk: [Jason Ma](#), Title: “A recipe for scaling robot learning for the real world”

- ☐ Deployment flywheel
 - ☐ Real restaurant deployment
 - ☐ Is robot learning solved? What do we still need?
 - ☐ Model observability: how do you tell how well your model is doing in production at scale?
- ☐ Full-stack hardware, sensing
 - ☐ Multi-modal sensing
 - ☐ End-effector and sensor co-design
- ☐ The bitter lesson of real-world robotics
 - ☐ Focus on system building and scaling;
 - ☐ Don't chase fancy algorithms or architecture;
 - ☐ By scaling systems and deploying robots, it becomes obvious what problems need to be solved

CVPR26: Meta Insights into Trends and Tendencies (64/152)

WS Scalable Robot Learning Systems

Invited talk: [Wayne Wu](#), Title: "From scaling up to scaling out: reality world simulators for physical AI"

- ❑ Research goal: making robot generalizable and safe in public spaces
- ❑ Data ceiling issue: one day, we will use up training data that we can use
 - ❑ Key idea:
 - ❑ scaling up to scaling out; real video de-composition and re-composition
 - ❑ Structure representation for representing explicit physics
- ❑ Thrust 1: urban scene simulation
 - ❑ [Vid2Sim: building digital twins from video](#) (CVPR 2025)
 - ❑ Challenge: 3DGS cannot provide correct feedback
 - ❑ Idea: geo-consistent 3DGS
 - ❑ Allowing editing in 3D;
 - ❑ Translating any videos into robot training grounds
 - ❑ [UrbanVerse](#) (ICLR 2026)
 - ❑ Decomposing open-world video to 3D holistic scene representation
 - ❑ Using open-vocabulary 2D foundation models
 - ❑ Progressive 3D instance map construction
 - ❑ 3D semantic map -> 3D scene graph
 - ❑ Scene representations -> simulation environments

Vid2Sim



UrbanVerse



CVPR26: Meta Insights into Trends and Tendencies (65/152)

WS Scalable Robot Learning Systems

Invited talk: [Chuan Wen](#), Title: “Data Scaling on the Job”

❑ Policy deployment – how to scale?

❑ Autonomous Online Failure Detection and Human Shared Control Empower Scalable Real-world Deployment and Adaptation

([ARMADA](#))

- ❑ Human-in-the-loop to provide feedback
- ❑ Multi-robot shared control system
- ❑ Autonomous online failure detection
 - ❑ From expert trajectories and rollout trajectory, do adaptive rewinding for post-training
- ❑ ARMADA helps improve training data efficiency, and reduce human intervention
- ❑ Issue for ARMADA: we still need lots of robots and a person to hold the teleoperation device

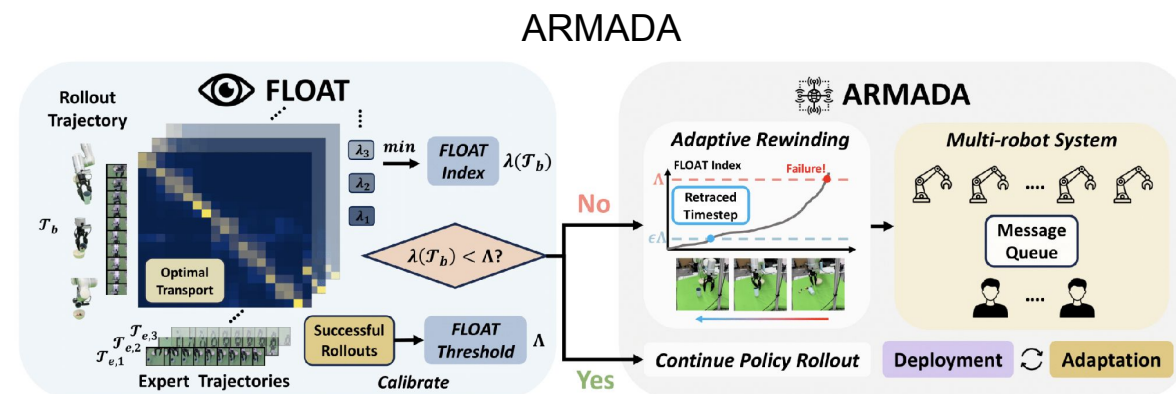
❑ [ROBOPocket: Improve Robot Policies Instantly with Your Phone](#)

❑ Robot-free data collection

❑ How to collect data with maximum efficiency?

- ❑ Off-policy data collection: low-cost hardware with online data validation – isomorphic gripper
- ❑ On-policy post-training: robot-free online policy iteration
 - ❑ Cloud inference; edge AR foresight; human correction

❑ Takeaway: deployment should be part of the data flywheel



CVPR26: Meta Insights into Trends and Tendencies (66/152)

WS Scalable Robot Learning Systems

Invited talk: [Wayne Wu](#), Title: “From scaling up to scaling out: reality world simulators for physical AI”

- ❑ Data ceiling issue: one day, we will use up training data that we can use
 - ❑ Thrust 2: embodied human modeling
 - ❑ [DNA-Rendering](#) (ICCV 2023)
 - ❑ Large-scale dataset; neural actor rendering; novel view, pose, identity rendering
 - ❑ [CityWalker: first large-scale dataset for pedestrians](#) (CVPR 2025)
 - ❑ Curates pseudo-labeled pedestrian motion and scene context from worldwide city-walking YouTube videos
 - ❑ [PedGen: A context-aware diffusion motion generation model](#) (ICLR 2025)
 - ❑ Context-aware diffusion model for pedestrian movement generation
- ❑ Deploy robots in real cities
 - ❑ Teleoperation in the labs, run robots in the cities
- ❑ What's the next
 - ❑ Neural world simulators
 - ❑ Generalist physical AI: for any task in any city
 - ❑ Human-AI teamed autonomy

CVPR26: Meta Insights into Trends and Tendencies (67/152)

Meta insights of [Workshop on Bitter Lessons](#)

- ❑ The workshop is best read as a disagreement over the “right abstraction boundary” for computer vision after scaling, not as a single anti-structure manifesto.
- ❑ Traditional CV decomposed intelligence into image → class / mask / box / flow / depth / 3D / pose. Several speakers argued these are historically useful proxy tasks, not necessarily final objectives.
- ❑ Deliverable-driven framing dominated: if the deliverable is action, pixel stream, interactive world, CAD artifact, measurement, or scientific discovery, the representation choice should follow the deliverable.
- ❑ Scale is bitter for hand-designed modules, but not all speakers equated “scale” with “ignore science”: metrology, low-resource robustness, physical reasoning, affordances, and interpretability remain central.
- ❑ Data is as important as compute for CV. Several notes point to web-scale uncuration, Tiny Images, human-video priors, scarce action data, and the missing loop of autonomous data collection.
- ❑ 3D and mid-level vision are not simply “dead”: their role is being downgraded from mandatory core to optional interface / supervision / probe / debugging aid / deployment tool.
- ❑ Sequence / video generative modeling is a candidate scalable objective, but a passive model trained on existing human sequences cannot cover tasks for which no human-policy data or verifier exists.
- ❑ The clearest long-horizon direction is closed-loop learning: systems that perceive, act, collect data, verify outcomes, and keep learning after pretraining.

Axes to emphasize in the meta-analysis

- ❑ **Objective axis:** recognition / segmentation / reconstruction → deliverable-oriented perception → action / interaction / discovery.
- ❑ **Data axis:** human-labeled images → uncured web data → human-policy videos → action-conditioned experience → autonomous rollouts.
- ❑ **Representation axis:** handcrafted features → learned features → explicit 3D/mid-level as interfaces → latent world models + optional probes.
- ❑ **Evaluation axis:** static benchmark → physically falsifiable task success → continual improvement / safety / deployment cost.
- ❑ **Institutional axis:** industrial frontier scale vs academic leverage through metrology, low-resource settings, interpretability, data governance, negative results.
- ❑ Avoid binary phrasing. The common pattern is demotion of proxy tasks, not deletion of all structure or all classic CV tools.

Thanks to Andrei Bursuc (@abursuc) for public X notes/images.

CVPR26: Meta Insights into Trends and Tendencies (68/152)

Invited talk: Bill Freeman, Title: "Bitter and non-bitter lessons"

- ❑ **Warm-up bitter lessons reported from public notes:**
 - ❑ Training end-to-end keeps beating carefully engineered pipelines; gradient descent continues to work better than intuition predicted.
 - ❑ Many model-shortcoming papers age badly because the next foundation-model release may remove the specific failure mode.
 - ❑ Spectacular results increasingly require spectacular compute, creating a serious question for academia: what can it uniquely contribute when frontier scale is industrial?
 - ❑ Computer vision's 2026 state was summarized in public notes as "computer vision by asking nicely": vision tasks increasingly mediated by language prompts / VLMs.
- ❑ **Non-bitter lessons / community lessons:**
 - ❑ Work with more people; read old work; put the spotlight on people around you; get used to being outshined by your students.
 - ❑ These are not "soft" lessons: they counteract repeated reinvention and reduce the field's tendency to erase older insights when a scaled version finally works.
 - ❑ Public notes also framed Freeman's recommendation as putting "vision" back into vision and learning, i.e., robust, lightweight, low-resource, nature-like visual intelligence.
- ❑ Report takeaway: The bitter lesson is compute/data scaling; the non-bitter lesson is that academia can still win through problem choice, metrology, historical memory, and science-quality explanations.

Report interpretation

- ❑ bitter about scaling and proxy tasks, non-bitter about scientific practice and mentorship.
- ❑ **"Computer vision by asking nicely" is a concise 2026 diagnosis:** language interfaces became a wrapper around many visual tasks, but not a replacement for vision science.
- ❑ **Warm-up bitter lessons map to method trends:** end-to-end training, gradient descent, model scaling, and fast obsolescence of failure-mode patch papers.
- ❑ **Non-bitter lessons map to community operations:** collaborations, reading old work, giving credit, and accepting that students / successors will make old intuitions obsolete.

CVPR26: Meta Insights into Trends and Tendencies (69/152)

Invited talk: Alyosha Efros, Title: "Sweet Lessons"

- ❑ Alyosha's framing is important because it softens the usual "scale beats structure" reading into a data-centered story: CV's history was changed not only by compute, but by the availability of huge, messy visual corpora.
- ❑ He reportedly questioned whether Sutton's essay is the "right lesson" for CV if "data" is under-emphasized relative to computation; for vision, web-scale uncured data was a decisive ingredient.
- ❑ Tiny Images / 80 Million Tiny Images was called out as an early scaling-law signal: rejected / under-appreciated work that later became obvious in hindsight.
- ❑ The "sweet" part: scaling and nearest-neighbor-like behavior are not just shortcuts; they can expose that some apparently semantic / cognitive phenomena are simpler and more data-indexed than expected.
- ❑ Humbling lesson from public notes: humans may be closer to nearest-neighbor / pastiche machines than the field wanted to believe; this weakens overly symbolic / anthropocentric intuitions.
- ❑ Report takeaway: The data lesson is the CV-specific counterpart to Sutton's compute lesson. Ask whether a problem is blocked by architecture, or by the absence of the right data distribution.

Report interpretation

- ❑ **Connect Alyosha to Bill:** the reason old work matters is that ideas like Tiny Images looked premature until compute / data infrastructure caught up.
- ❑ **Connect Alyosha to Georgia:** recognition progress is not only architecture progress; it is data distribution progress that turns recognition into reusable infrastructure.
- ❑ **Connect Alyosha to Vincent:** passive data scale helps, but it does not automatically create new actions or tasks absent from the data.
- ❑ **Data lesson should include dataset governance:** uncured scale can be powerful and harmful; future CV reports should not celebrate scale without documentation / filtering / ethics
- ❑ **Open question:** what is the "Tiny Images moment" for robot actions, physical affordances, or scientific discovery?

CVPR26: Meta Insights into Trends and Tendencies (70/152)

Invited talk: Georgia Gkioxari, Title: "The Bittersweet Lessons of Recognition"

❑ Bitter lessons:

- ❑ Scaling exceeded expectations; in data-rich, benchmark-friendly recognition domains, clever task-specific machinery is often overtaken by larger pretraining / stronger generic models.
- ❑ A domain with abundant data, clear labels, and a stable metric may already be on a commoditization path; new recognition-only papers risk becoming incremental unless they change the problem definition.
- ❑ "Objects and semantics" are becoming infrastructure: useful, but no longer sufficient as the main frontier.

❑ Sweet lessons / new opportunities:

- ❑ Recognition now works well enough to unlock harder downstream questions; past CV work produced robust tools that can be reused rather than discarded.
 - ❑ Suggested future directions reported publicly: go beyond objects / semantics to utility and affordances; reason in physical space, not just language space; build efficient AI for robotics and deployment.
 - ❑ This moves the field from "what is in the image?" to "what can I do with it, what will happen if I act, and what constraints matter for this body / tool / environment?"
- ❑ Report takeaway: Recognition is not dead; it has become a substrate. The frontier is actionable, physical, cost-aware, and embodied recognition.

Report interpretation

- ❑ Recognition being "solved enough" does not mean visual recognition research is done; it means raw category prediction is no longer the right endpoint.
- ❑ A useful phrasing: "from nouns to verbs and consequences" — objects / semantics → utility / affordance / state change / interaction.
- ❑ Physical-space reasoning implies geometry, contact, causality, uncertainty and temporal persistence, but not necessarily a fixed 3D pipeline.
- ❑ Efficient AI is not an afterthought for robotics: latency, on-device compute, energy, reliability, and recovery behavior are part of the task definition.
- ❑ Georgia provides the positive version of the bitter lesson: scaling recognition created tools that allow more meaningful problems to be attacked.
- ❑ Future benchmark suggestion: evaluate whether perception changes the downstream decision, not whether it matches a static label taxonomy.

CVPR26: Meta Insights into Trends and Tendencies (71/152)

Invited talk: Jon Barron, Title: "Lessons in 3D"

- ❑ **Core framing: "Should you care about 3D?" depends on the final deliverable, not on whether the field historically called itself 3D vision.**
 - ❑ Robotics: Deliverable is action. Explicit 3D may be sufficient but not necessary; pushback noted that robot action spaces are not tiny and safety / contact / planning can still reward geometry.
 - ❑ Digital media: Deliverable is pixels. If the output is a finished video/image, fully baked generative pixels can win even without explicit 3D; consistency across shots remains a caveat.
 - ❑ Interactive media: Deliverable is real-time interactive pixels. Explicit 3D can act as cache / controllable substrate / physics hook when per-frame generation is too slow or unstable.
 - ❑ Making physical stuff: Deliverable is a 3D artifact / CAD / manufacturing file. Here explicit geometry is closer to the product itself, so "3D as interface" remains hard to remove.
 - ❑ Bio / medical / science: Human-in-the-loop and regulatory constraints can make 3D an interface even when the final goal is action or decision; this complicates the robotics/deliverable dichotomy.
- ❑ Report takeaway: 3D is being reclassified from universal internal representation to role-conditioned infrastructure: interface, acceleration structure, physical artifact format, or diagnostic.

Report interpretation

- ❑ explicit 3D is justified when it is close to the deliverable or gives interaction-time efficiency/control.
- ❑ Robotics nuance: actions are the deliverable, but explicit 3D can still be useful for safety cases, debugging, sim-to-real, contact planning, and human-in-the-loop control.
- ❑ Digital media nuance: implicit video generation may win for final pixels, but consistent long-form video, editing, camera control, and user interaction create pressure for structure.
- ❑ Manufacturing/CAD is the strongest pro-3D case because file formats, tolerances, collision checks, and physical constraints are not optional presentation details.
- ❑ Relation to Vincent: both demote 3D as model core, but Jon's deliverable framing keeps more explicit-3D use cases alive.

CVPR26: Meta Insights into Trends and Tendencies (72/152)

Invited talk: Vincent Sitzmann, Title: "The flavor of the bitter lesson for computer vision"

❑ What even is vision?

- ❑ Marr's "know what is where by looking" is not a clean input-output task. Segmentation, detection, 3D, point tracking, optical flow, etc. are proxy definitions, not the full problem.
- ❑ Historical factorization: vision mapped images to intermediate representations; robot learning / control consumed point clouds, boxes, masks, maps, poses and produced actions.
- ❑ His claim: the old distinction between "vision" and "robot learning" no longer makes sense; frontier systems should be perception-action learners.

❑ 3D as a distraction for vision foundation models:

- ❑ 3D as spatial memory / maps: long-term memory in sequence models may absorb the function without explicit maps.
- ❑ 3D as additional supervision: not enough clean 3D data, and what exists is noisy / biased relative to web-scale 2D video.
- ❑ 3D as planning space: future models may control actuators directly, with no explicit geometric intermediate representation.
- ❑ 3D as interface / real-time content delivery: useful after training; models can learn 3D implicitly and be distilled / fine-tuned into explicit 3D only when needed.
- ❑ Report takeaway: not "3D is useless," but "do not assume human-designed 3D bottlenecks are on the critical path to embodied intelligence."

Report interpretation

- ❑ Claim: if the final behavior is image/video history → action, any mandatory intermediate such as mesh / point cloud / map may become a lossy hand-designed bottleneck.
- ❑ Controversy: safety-critical robotics, surgical interfaces, CAD, mapping, and inspection often need explicit artifacts for humans, regulators, or downstream tools.
- ❑ Important distinction: implicit 3D understanding can still exist inside the model even when no explicit 3D output is trained or exposed.
- ❑ Link to Marr: the definition of vision is under-specified until one says what success means; action success is falsifiable, but not always the only deliverable.

CVPR26: Meta Insights into Trends and Tendencies (73/152)

Corollary: the bitter lesson of sequence generation / active scaling

- ❑ **Sequence generative modeling alone will not scale:**
 - ❑ It learns from sequences that already exist; those sequences were collected by human policies and are scarce or nonexistent for infinitely many tasks.
 - ❑ Machines need a way to collect their own data and learn from it. This is easier today in verifiable domains (coding, math, games) than in novel physical-world tasks.
 - ❑ No known general route exists for getting reliable verifiers for arbitrary novel tasks in the real world; this is a core obstacle for embodied intelligence.
- ❑ **Active Scaling Conjecture:**
 - ❑ Systems that perceive, act, and learn closed-loop will eventually outperform systems that passively learn from data collected by others.
 - ❑ This links robot learning, RL, self-play, active learning, continual learning, exploration, memory, and world models under one agenda: agency as a data engine.
 - ❑ Video/world models may be useful pretraining or simulators, but without action-conditioned data and policy extraction they do not solve the perception-action loop.
 - ❑ Open problems: intrinsic motivation, safe exploration, long-horizon memory, real-time control with large models, action alignment for human videos, and scalable physical evaluation.
- ❑ Report takeaway: The important scaling variable is no longer just tokens/images/frames, but experience actively generated by the system itself. (Similar idea to Alyosha's take on "Deep data")

Report interpretation

- ❑ The key bottleneck is paired perception-action data at scale, not another static image/video dataset.
- ❑ Self-driving is the exception that proves the rule: humans already generate abundant action-observation data by driving; dexterous robots do not have an equivalent natural data exhaust.
- ❑ World models can amplify limited interaction data, but interactive world models themselves need action-observation data; this creates the chicken-and-egg problem.
- ❑ Verifiers are the hidden scaling bottleneck: coding/math/games can score success automatically; household, manufacturing, medical, and novel physical tasks are much harder.
- ❑ This is where hierarchy, memory, exploration, safety constraints and embodiment alignment become first-class, not auxiliary engineering.

CVPR26: Meta Insights into Trends and Tendencies (74/152)

Invited talk: Bharath Hariharan, Title: "Mid-level vision is dead; long live mid-level vision!"

- ❑ **"Mid-level vision is dead" reading:**
 - ❑ Dead as a destination: edges, parts, masks, depth, flow, pose, correspondences, etc. are often proxy tasks created because end-to-end intelligence was once intractable.
 - ❑ Dead as a silo: a benchmarked mid-level output can look solved while failing to improve downstream action, interaction, robustness, or discovery.
 - ❑ Dead as manual bottleneck: if the representation is hand-specified and lossy, large models may bypass it with richer latent structure.
- ❑ **"Long live mid-level vision" reading:**
 - ❑ Alive as interface: human debugging, robot toolchains, 3D/CAD editors, medical annotation, and safety reviews still need interpretable intermediate outputs.
 - ❑ Alive as inductive scaffolding: auxiliary heads / probes / distillation targets can improve data efficiency and failure localization without defining the final objective.
 - ❑ Alive as affordance structure: support, contact, containment, part-whole, object permanence, correspondence, and action-conditioned geometry remain useful when the task is physical.
- ❑ Report takeaway: the key question is role assignment: core objective vs auxiliary signal vs interface vs diagnostic, not "mid-level yes/no."

Report interpretation

- ❑ Concrete mid-level survival cases: segmentation for editing, depth for AR/robotics, flow/correspondence for tracking, contact/support for manipulation, uncertainty maps for safety.
- ❑ Research direction: learned intermediate representations should be evaluated by downstream lift, calibration, cost, and failure diagnosis, not by proxy accuracy alone.
- ❑ "Dead" as "dead as the final destination," not "useless."

CVPR26: Meta Insights into Trends and Tendencies (75/152)

Invited talk: Shenlong Wang, Title: "Modeling the World After the Bitter Lesson"

- ❑ A rebuttal against overclaiming that generative video models "know physics." The issue is not whether video models are useful, but whether they support scientific world modeling.
- ❑ Classical scientific loop: observe $\rightarrow$ hypothesize $\rightarrow$ model $\rightarrow$ predict $\rightarrow$ verify $\rightarrow$ discover. A world model that cannot be probed, falsified, or used for counterfactuals is not yet a scientific model.
- ❑ After the bitter lesson, modeling should not mean hand-coding every rule; it should mean building systems with measurable variables, uncertainty, interventions, conservation / constraints, and falsifiable predictions.
- ❑ This connects to the workshop's metrology axis: if our benchmarks reward plausible pixels, we will get plausible pixels; if we reward intervention and discovery, models may learn more causal / physical structure.
- ❑ Report takeaway: "world model" should be reserved for models that can support prediction under action, counterfactual reasoning, and verification and not merely high-quality video generation.

Report interpretation

- ❑ A plausible future video is not necessarily a valid counterfactual; physics requires interventions, constraints, and failure cases, not only realism under the training distribution.
- ❑ Scientific world models should answer "what changes if I act?" and "what evidence would falsify this model?"
- ❑ This talk protects CV from a shallow bitter lesson: scaling video is powerful, but benchmarked plausibility can become another proxy task.
- ❑ Metrology agenda: construct tests that separate object permanence, contact dynamics, conservation, causal direction, and controllable latent variables.
- ❑ Relation to Derek: continual learning needs verifiers / measurements; otherwise online updates can simply reinforce plausible hallucinations.
- ❑ Relation to Jon / Vincent: explicit representations may be bad as fixed bottlenecks, but explicit measurement, interfaces, and hypotheses may still be necessary for progress.

CVPR26: Meta Insights into Trends and Tendencies (76/152)

Invited talk: Derek Hoiem, Title: "Advice to research by, and the next bitter lesson"

- ❑ **The next bitter lesson: models that easily continue to learn will eventually surpass highly trained models that cannot.**
- ❑ **Why this matters:**
 - ❑ Static pretraining produces large but inert systems; physical and scientific settings are nonstationary and require adaptation after deployment.
 - ❑ Continual learning is the natural companion to active scaling: the agent not only collects data, but incorporates it without catastrophic forgetting or unsafe behavior.
 - ❑ Learning after deployment may dominate parameter count once models become large enough and data scarcity / distribution shift are the bottlenecks.
- ❑ **Mechanisms / research agenda:**
 - ❑ Persistent memory, retrieval, plastic modules, online self-supervision, world-model-based rehearsal, uncertainty-aware exploration, and regression-test style verifiers.
 - ❑ Evaluation should measure lifelong improvement, forgetting, robustness under distribution shift, ability to discover novel regularities, and safety while learning.
 - ❑ This reframes "foundation model" as a starting point rather than the final artifact; the deployed model should remain trainable.
- ❑ Report takeaway: the next scaling axis may be plasticity over time, not only one-shot scale at pretraining.

Report interpretation

- ❑ This is a natural bridge from "foundation model" to "foundation learner": the object of evaluation is the update process, not only the pretrained checkpoint.
- ❑ Continual learning changes system design: memory management, forgetting control, data provenance, regression tests, human oversight, and rollback become core infrastructure.
- ❑ In robotics and scientific discovery, the deployed world is the dataset; a model that cannot incorporate deployment feedback will be capped by pretraining coverage.
- ❑ Safety caveat: online learning in the physical world requires constrained exploration, uncertainty estimates, and auditable update policies.
- ❑ Relation to Vincent: active agents need to learn from the data they create, not merely collect it.
- ❑ Evaluation idea: reward systems that improve monotonically under new experience while preserving old competencies.

CVPR26: Meta Insights into Trends and Tendencies (77/152)

Future directions

- ❑ **From VLM / video-only models to active perception-action models:**
 - ❑ Action-conditioned world models; policy extraction from generated plans; closed-loop agents that decide what data to collect next; verifiers for embodied tasks.
 - ❑ Human videos to robotics is not automatic: need action alignment, sensor alignment, embodiment alignment, temporal abstraction, and failure-aware evaluation.
- ❑ **From representation debates to role-conditioned design:**
 - ❑ Use 3D / mid-level outputs when they are deliverables, interfaces, safety diagnostics, or data-efficient auxiliary tasks; do not force them as universal bottlenecks.
 - ❑ New mid-level targets should be actionable: affordance, contact, support, containment, state change, manipulability, uncertainty, and persistence under occlusion.
- ❑ **From benchmark success to production / science metrics:**
 - ❑ Metrics should include intervention success, sim-to-real transfer, low-resource robustness, latency, cost, safety, recoverability, and ability to keep learning.
 - ❑ World-model evaluation should test counterfactuals and constraints, not only pixel realism or human preference.
- ❑ **Open questions worth highlighting:**
 - ❑ What is the scalable pretraining objective for embodied intelligence? How do we get verifiers for novel real-world tasks? How can agents explore safely?
 - ❑ What should academia do when compute is industrial? Candidate answers: metrology, negative results, data governance, efficient models, theory, old-work synthesis, and hard domains ignored by scale.

More concrete future-work bullets

- ❑ Data: robot web / cross-embodiment action data, human-video-to-action alignment, simulation from real videos, autonomous rollouts, failure mining, dataset governance.
- ❑ Models: action-conditional world models, memory-augmented policies, online learners, representation probes, uncertainty heads, low-latency edge models.
- ❑ Representations: implicit world state internally; explicit 3D/mid-level outputs only when needed for CAD, AR/VR, safety, editability, inspection, or human interface.
- ❑ Evaluation: falsification suites for world models, downstream task lift for mid-level outputs, continual-learning metrics, physical safety metrics, production cost metrics.
- ❑ Thesis for report: the next CV frontier is not “bigger image model” alone, but scalable perception systems that can act, verify, and improve.

From Oral session 1

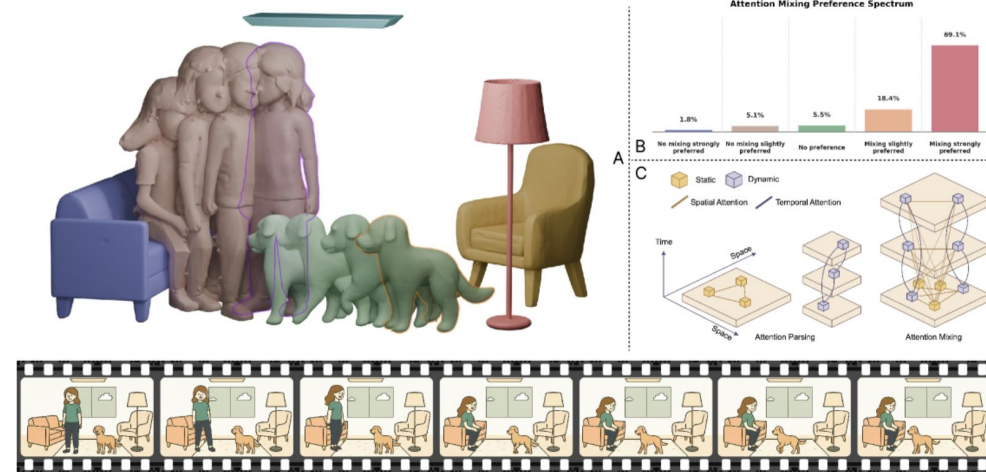
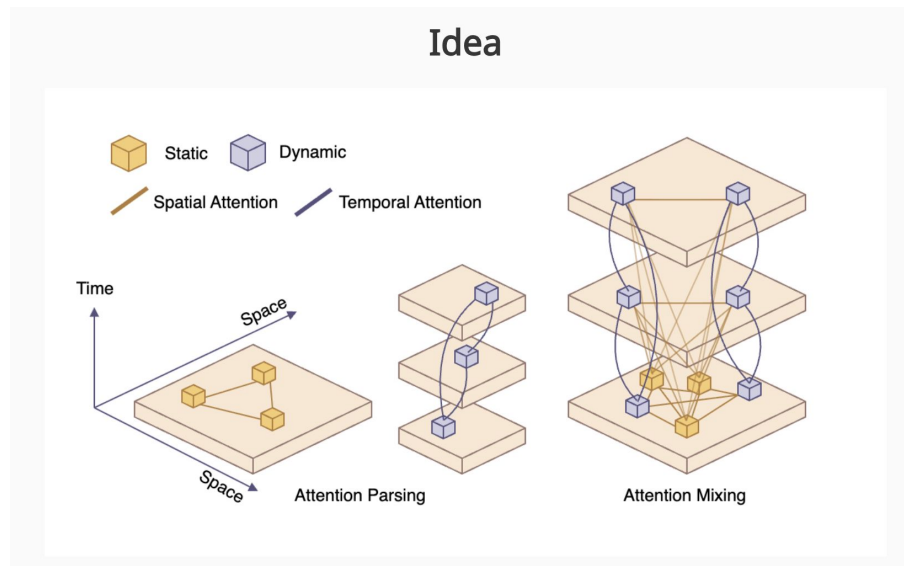
From “accuracy” to “new underexplored problems in the foundation-model era.”

Keywords: Controllability, safety, OOD, and TTA

- ❑ **CoTyle [Liu+, CVPR2026]:** Creativity, Consistency, Style Control, Discrete Style Space, Reproducibility, Open-source Generative AI
- ❑ **ANTS [Zhu+, CVPR2026]:** Negative label, OOD Detection, Open-world Recognition, Test-time Adaptation, MLLM Reasoning, Training-free
- ❑ **ARGUS [Lu+, CVPR2026]:** Multimodal Safety, Indirect Prompt Injection Defense, Instruction Following, Robust MLLM
- ❑ **TEAR [He+, CVPR2026]:** Multimodal Safety, Temporal Reasoning, Adversarial Prompting, Dynamic Risk Evaluation, Safety Benchmarking
- ❑ **ViT³ [Han+, CVPR2026]:** Test-Time Training, Efficient Vision Architecture, Linear Complexity, Visual Sequence Modeling

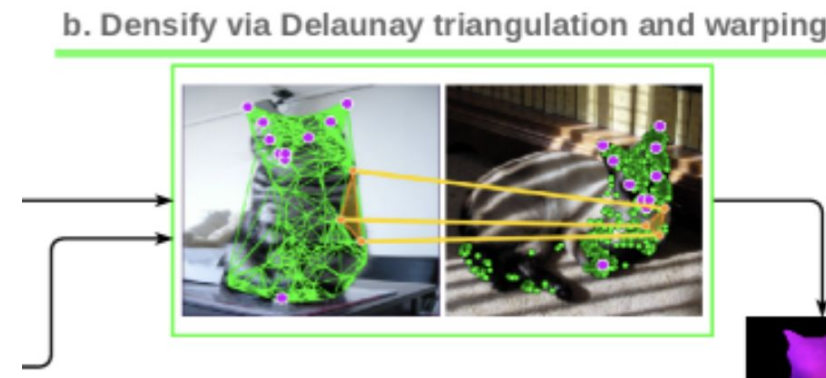
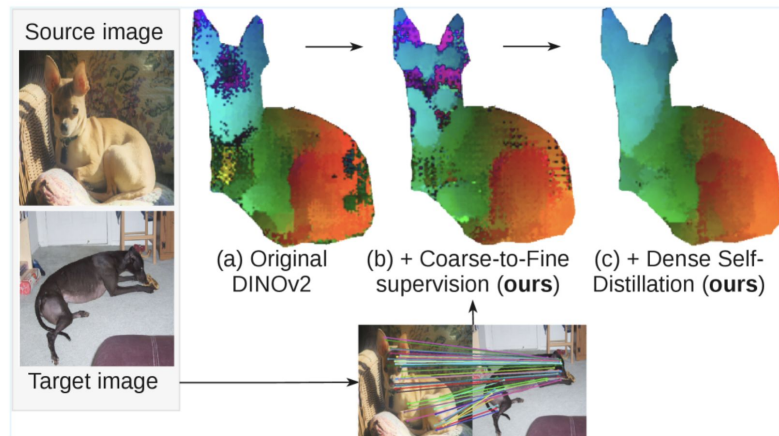
Inferring Compositional 4D Scenes without Ever Seeing One

- Summary: COM4D reconstructs complete multi-object 4D scenes from monocular videos. It jointly predicts the 3D structure of objects and how their spatial configuration changes over time. The method focuses on realistic scenes containing both static and dynamic objects, while keeping object identity and motion consistent across frames.
- Novelty: COM4D does not require training on full compositional 4D scene data. Instead, it separately learns spatial object composition from static multi-object data and temporal object dynamics from single-object videos. At inference time, its Attention Mixing mechanism combines these learned attentions to reconstruct coherent 4D scenes with multiple interacting objects.



MARCO: Navigating the Unseen Space of Semantic Correspondence

- Summary: MARCO is a semantic correspondence model that matches semantically related points across images, even when the queried keypoints or object categories were not seen during training. It builds on a single DINOv2 backbone and learns dense, accurate correspondences from sparse keypoint supervision. The model achieves state-of-the-art results on SPair-71k, AP-10K, and PF-PASCAL while being smaller and faster than diffusion-based approaches
- Novelty: MARCO's main novelty is a training densification framework that expands a small set of annotated keypoints into thousands of dense pseudo-correspondences over the object surface. It combines a coarse-to-fine localization objective with self-distillation, improving both precise point matching and semantic generalization. This enables correspondence beyond fixed keypoint vocabularies, allowing the model to transfer to unseen keypoints and novel object categories.



Featurising Pixels from Dynamic 3D Scenes with Linear In-Context Learners

- ❑ Summary:
- ❑ LILA learns pixel-accurate feature descriptors from videos of dynamic 3D scenes. It uses depth and motion cues estimated by off-the-shelf models to capture both geometric and temporal information at the pixel level. These learned features transfer to dense vision tasks such as video object segmentation, surface normal estimation, and semantic segmentation.
- ❑ Novelty:
- ❑ It uses Linear In-Context Learning with depth and motion cues to learn temporally consistent features reasoning in dynamic scenes.

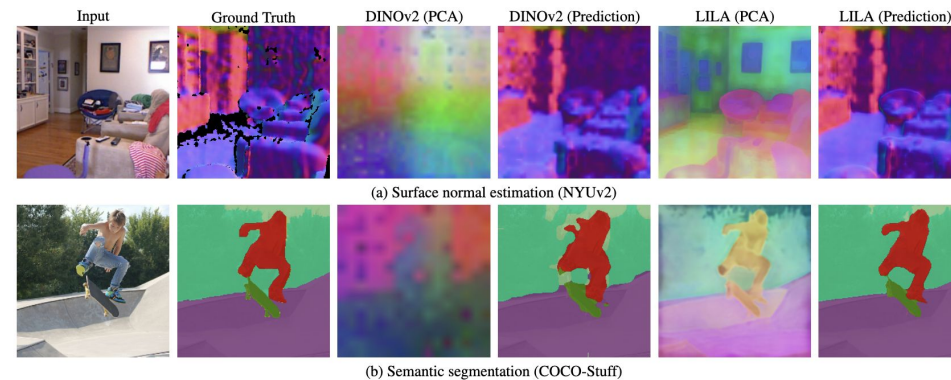
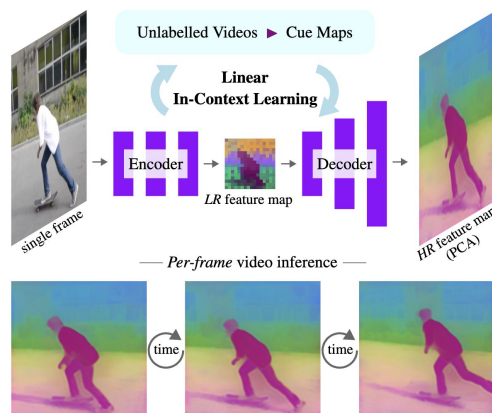
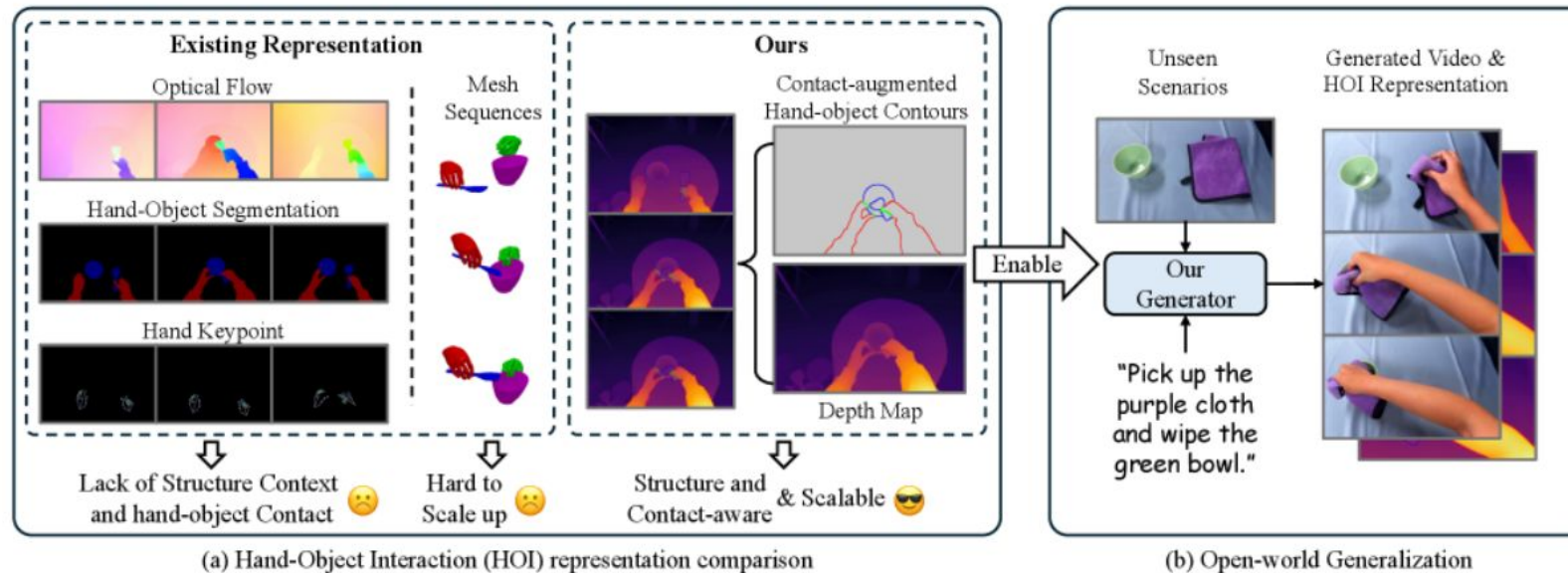


Figure 5. **Qualitative results: surface normal estimation and semantic segmentation.** (a) Despite pre-training on dynamic scenes, LILA reveals a finer surface structure than the baseline (DINOv2 [11] / ViT-B14) in predominantly static scenes. (b) LILA enables more accurate semantic segmentation, providing more accurate semantic boundaries and background detail.

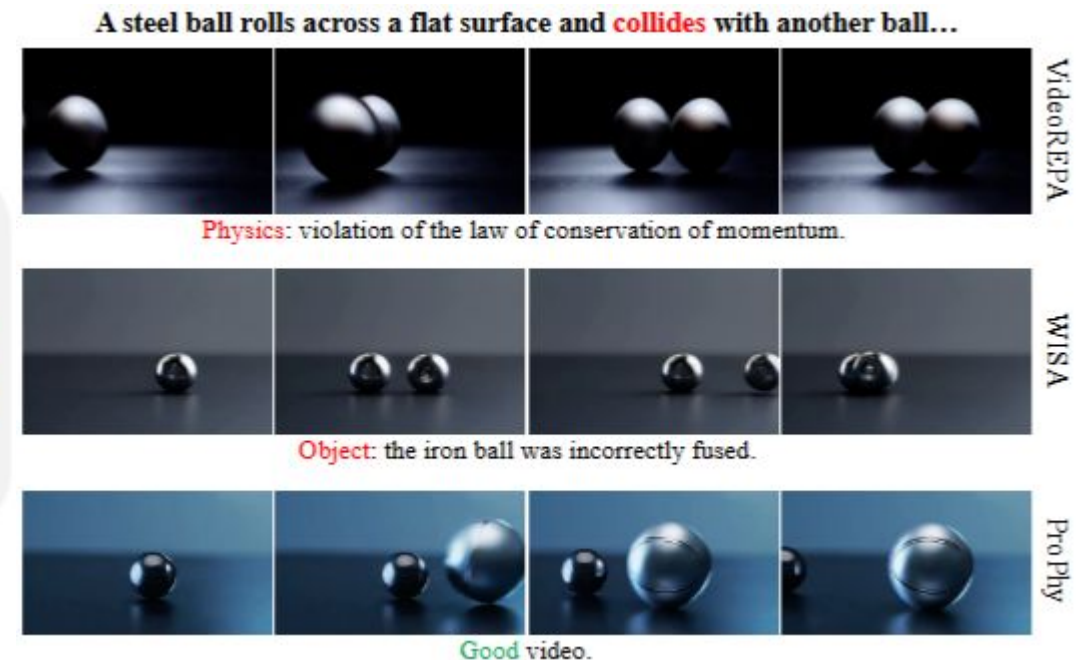
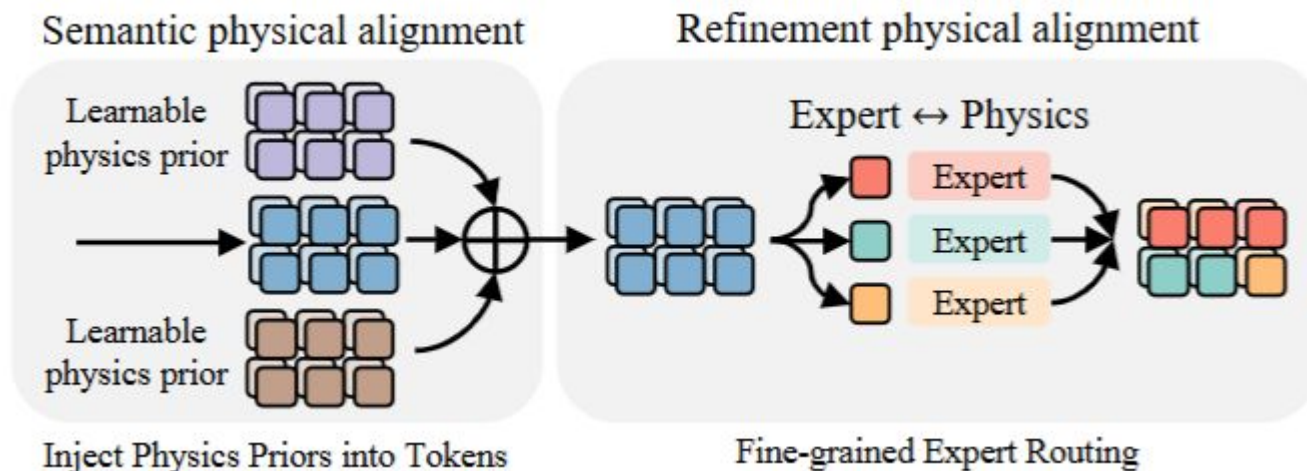
Open-world Hand-Object Interaction Video Generation Based on Structure and Contact-aware Representation

- Summary: They introduce SCAR, a method for generating realistic videos of hands manipulating objects from a starting image and task description. The authors propose an annotation-free hand-object interaction (HOI) representation combining contact-augmented contours with depth maps. Instead of an error-prone multi-stage pipeline, a joint denoiser generates the video and this representation simultaneously.
- Novelty: The core novelty is a scalable, annotation-free HOI representation that fuses contact-augmented contours with depth maps, paired with a joint video-and-representation generation paradigm that sidesteps the scalability-versus-fidelity tradeoff and the error accumulation of prior multi-stage approaches.



ProPhy: Progressive Physical Alignment for Dynamic World Simulation

- Summary: ProPhy improves the physical realism of AI-generated videos. Text-to-video models often produce footage that looks convincing but violates physics (objects penetrating each other, effects appearing at the wrong time), especially in dynamic scenes. ProPhy fixes this by conditioning generation on physics at the level of individual regions rather than treating the whole video uniformly.
- Novelty: ProPhy is the first to pair a hierarchical, two-stage (semantic-then-token-level) mixture-of-physics-experts with a VLM-distilled alignment loss, enabling region-specific physical conditioning instead of coarse video-level guidance.



Optimized Model Merging Initialization for Multimodal Continual Learning, In Proceedings of the 1st International Workshop on Multi-Sensorial Media and Applications

❑ Summary

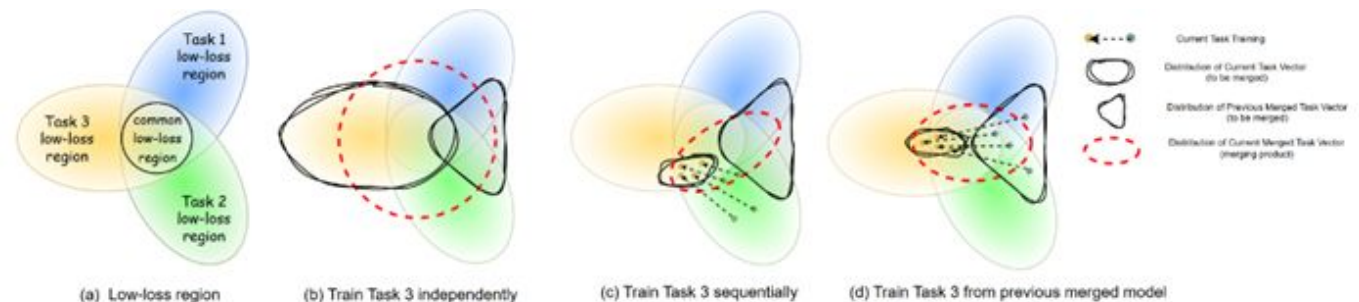
- ❑ They propose MergedStart, a novel optimization initialization strategy tailored to continual learning through multimodal model merging.

❑ Novelty

- ❑ This study introduces a learning paradigm that transfers knowledge from pre-merge tasks to merged tasks across three sequential multimodal tasks.

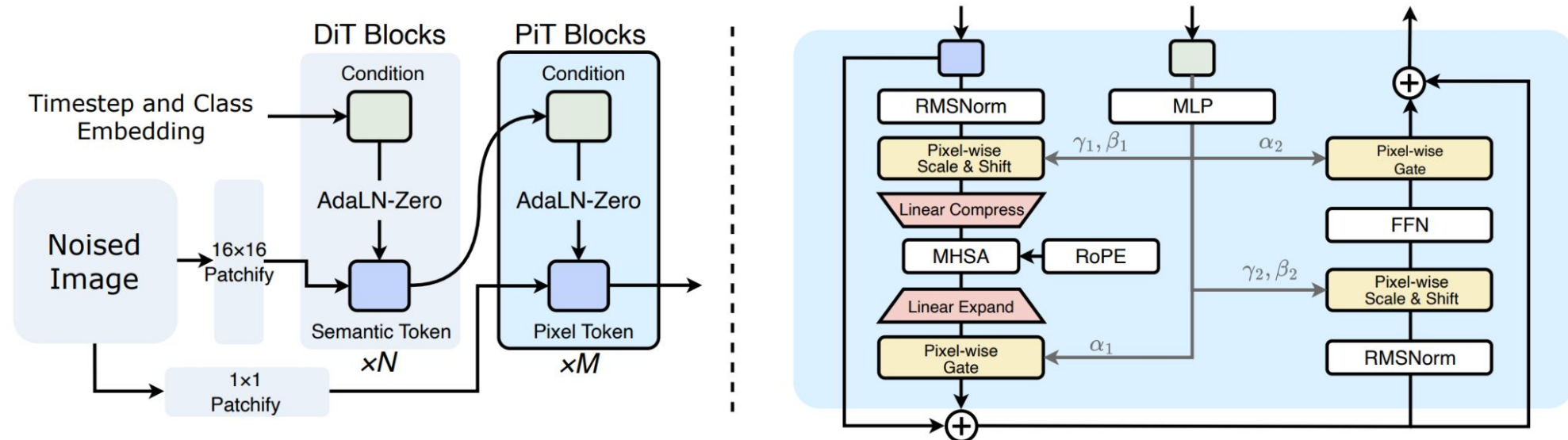
❑ Thoughts

- ❑ The approach shows strong potential for generalization to diverse and previously unseen multimodal domains.



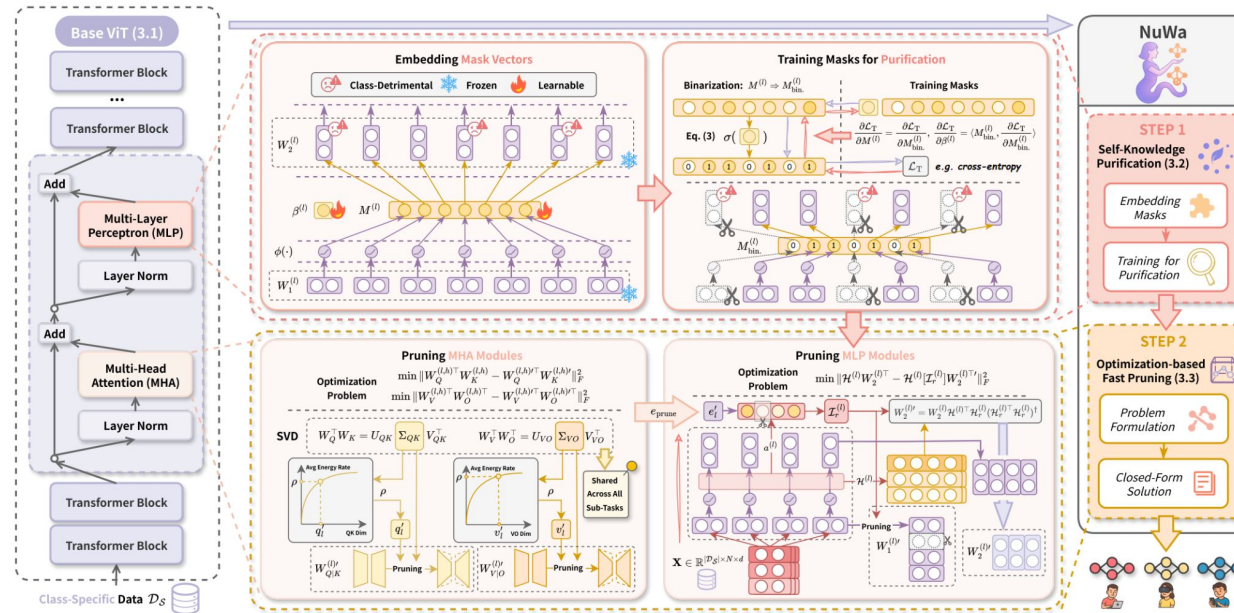
PixelDiT: Pixel Diffusion Transformers for Image Generation

- ❑ **Summary:** PixelDiT proposes a single-stage, end-to-end diffusion transformer that generates images directly in pixel space without relying on a pretrained autoencoder
- ❑ **Novelty:** The paper introduces a dual-level Transformer architecture, where a patch-level DiT captures global semantics and a pixel-level DiT refines textures and fine details
- ❑ **Thoughts:** The results indicate that efficient pixel-level token modeling is a key factor for making pixel-space diffusion practical, rather than simply removing the VAE from the generation pipeline



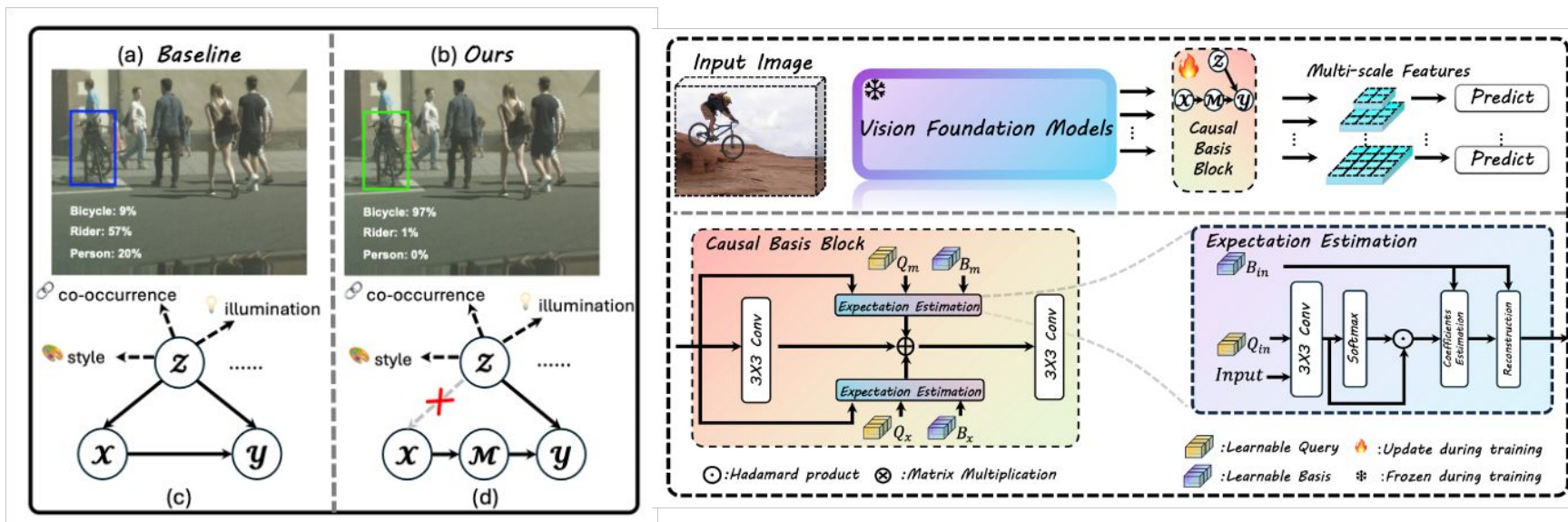
NuWa: Deriving Lightweight Class-Specific Vision Transformers for Edge Devices

- ❑ **Summary:** NuWa proposes a cost-efficient method for deriving lightweight class-specific Vision Transformers from pretrained base ViTs for edge devices
- ❑ **Novelty:** The paper identifies class-detrimental weights and introduces Self-Knowledge Purification with closed-form pruning optimization to remove them without post-pruning retraining
- ❑ **Thoughts:** This work highlights that edge model compression should consider target-class requirements, rather than only reducing model size in a class-agnostic manner



Bridge: Basis-Driven Causal Inference Marries VFMs for Domain Generalization

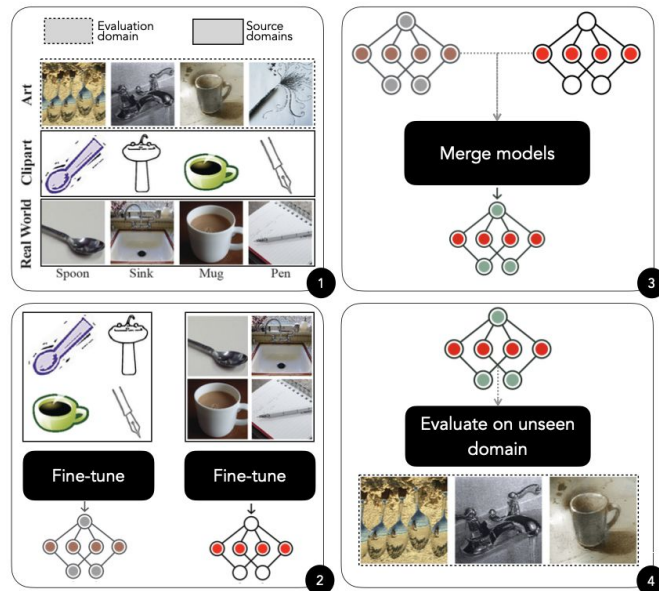
- ❑ **Summary:** Bridge improves domain generalization for object detection by combining causal inference with Vision Foundation Models. It reduces spurious correlations such as style, illumination, and object co-occurrence.
- ❑ **Novelty:** The method introduces a Causal Basis Block that performs front-door adjustment using learnable low-rank bases. It can be plugged into DINOv2, DINOv3, SAM, and Stable Diffusion.
- ❑ **Thoughts:** This work shows that VFMs still need causal refinement for robust real-world transfer. The idea is promising for detection under weather, viewpoint, and style shifts.



Methods	Bike	Bus	Car	Motor	Psn.	Rider	Truck	mAP
CSDS [69] (CVPR'22)	22.9	20.5	33.8	14.7	18.5	23.6	18.2	21.7
SHADE [81] (ECCV'22)	25.1	19.0	36.8	18.4	24.1	24.9	19.8	24.0
MAD [72] (CVPR'23)	-	-	-	-	-	-	-	28.0
SRCD [49] (TNLS'24)	24.8	21.5	38.7	19.0	25.7	28.4	23.1	25.9
DDT (SD) [15] (MM'24)	-	-	-	-	-	-	-	32.7
GDD (Diff. Detector, SD) [17] (CVPR'25)	38.9	31.0	71.5	37.6	61.5	47.0	38.5	46.6
Boost (Diff. Detector, SD) [16] (ICCV'25)	41.2	41.7	72.7	37.2	62.8	48.7	40.5	49.3
Ours (Diff. Detector, SD)	44.0	43.5	76.2	45.9	66.2	53.9	42.1	53.1
GDD (Diff. Guided, R101) [17] (CVPR'25)	38.4	33.4	72.0	38.3	60.3	47.0	35.0	46.3
Boost (Diff. Guided, R101) [16] (ICCV'25)	39.4	34.1	72.2	37.4	61.3	46.9	35.7	46.7
Ours (Diff. Guided, R101)	40.5	33.7	75.3	38.9	62.0	48.8	35.9	47.9
SAM	41.1	36.7	69.0	44.1	58.5	52.0	35.5	48.1
Ours (SAM backbone)	42.6	39.9	70.4	45.5	61.6	52.5	38.2	50.1
DINOv2	41.7	47.8	70.0	47.3	56.9	53.2	45.5	51.8
Ours (DINOv2 backbone)	48.2	48.9	74.5	52.8	65.0	57.1	52.0	56.9
DINOv3	48.6	56.8	73.4	52.9	64.4	57.6	51.0	57.8
Ours (DINOv3 backbone)	49.5	54.2	74.3	55.0	66.5	58.2	54.5	58.9

Bridging Domains through Subspace-Aware Model Merging

- Summary:** SCORE improves domain generalization by merging models fine-tuned on different source domains. It analyzes conflicts between domain-specific model subspaces using SVD and removes harmful singular-direction conflicts.
- Novelty:** The method constructs a shared orthogonal basis from the leading singular vectors of each model. It then projects task matrices into this basis and trims conflicting off-diagonal components without using source data or additional training.
- Thoughts:** This work shows that model merging can be useful not only for multi-task learning but also for domain generalization. The idea is practical because it can combine existing domain-specific models into one efficient model for unseen domains.



	Method	PACS	DomainNet	ImageNetR	NICOpp	OfficeHome	TerraIncognita	FedISIC	RetinaDomains	Avg
ViT-B-32	Zero-shot	94.85	56.10	66.91	77.14	79.82	22.73	23.08	24.40	55.63
	Task. Arithm.	93.40	5.40	4.65	56.92	66.73	40.47	32.18	40.00	42.47
	TIES	<u>95.77</u>	53.22	72.98	78.18	83.27	<u>47.03</u>	50.42	37.84	64.84
	MagMax	95.46	43.87	68.38	76.18	80.37	43.56	43.32	41.06	61.53
	PCB	95.35	<u>58.53</u>	<u>74.67</u>	<u>80.19</u>	83.59	40.84	38.41	35.45	63.38
	TSV	95.83	55.64	74.23	79.78	83.74	47.68	44.22	38.44	64.95
	Iso-C	95.37	55.84	74.74	79.81	82.34	36.53	32.91	32.97	61.31
	Iso-CTS	95.30	56.80	73.46	79.76	82.00	32.87	29.63	33.61	60.43
	SCORE (Ours)	95.70	58.68	73.44	80.59	84.23	46.92	<u>50.12</u>	35.87	65.69
	Task Experts	99.81	74.43	86.70	84.34	91.65	82.15	61.26	48.72	78.63
ViT-L-14	Zero-shot	98.39	66.25	86.51	84.06	88.82	46.46	19.77	26.00	64.53
	Task. Arithm.	98.53	30.61	13.07	73.39	89.97	51.11	31.64	42.81	53.89
	TIES	<u>98.60</u>	65.80	90.16	<u>85.53</u>	91.26	<u>59.72</u>	47.48	41.11	72.46
	MagMax	98.62	62.24	88.15	84.82	91.41	55.16	51.69	<u>40.74</u>	71.60
	PCB	98.54	68.26	90.38	86.44	91.13	59.32	38.93	38.22	71.40
	TSV	98.56	66.79	90.83	86.14	91.22	60.01	43.09	40.46	72.14
	Iso-C	98.51	66.89	90.90	86.20	90.32	57.52	33.31	34.55	69.78
	Iso-CTS	98.50	67.63	90.46	86.21	89.94	55.99	28.71	36.26	69.21
	SCORE (Ours)	98.62	68.26	<u>90.56</u>	86.54	91.28	61.19	<u>46.77</u>	41.12	73.04
	Task Experts	99.92	81.62	91.58	88.10	96.15	88.61	67.88	60.84	84.34

Trend Question: Do We Still Need 3D Data?

- ❑ Several invited speakers at CVPR workshops discussed this question
 - ❑ Researchers' opinions currently seem to be divided
 - ❑ Probably, many studies related to this question will emerge in the near future

What do you think??



Vincent Sitzmann
@vincesitzmann



 翻訳を表示

In my recent blog post, I argue that "vision" is only well-defined as part of perception-action loops, and that the conventional view of computer vision - mapping imagery to intermediate representations (3D, flow, segmentation...) is about to go away.

vincentsitzmann.com/blog/bitter_le...



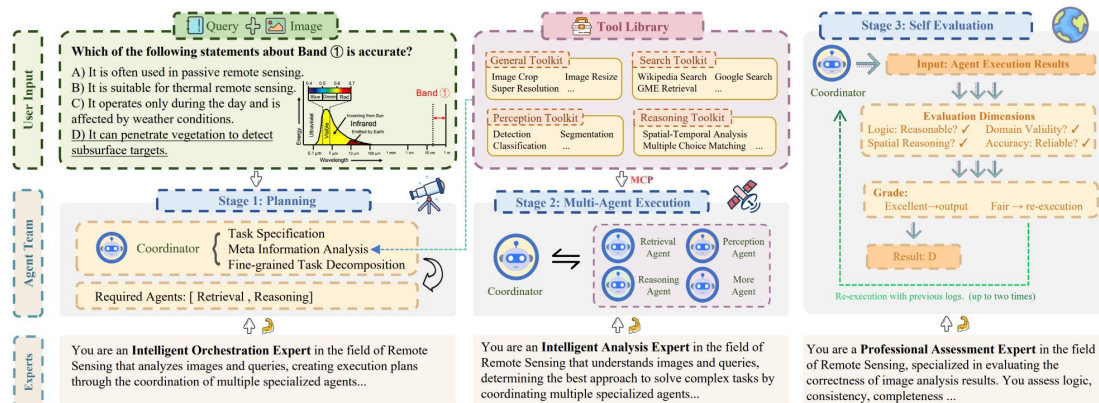
Remote Sensing x CV is going into an agentic approach Era

GeoMMBench and GeoMMAgent

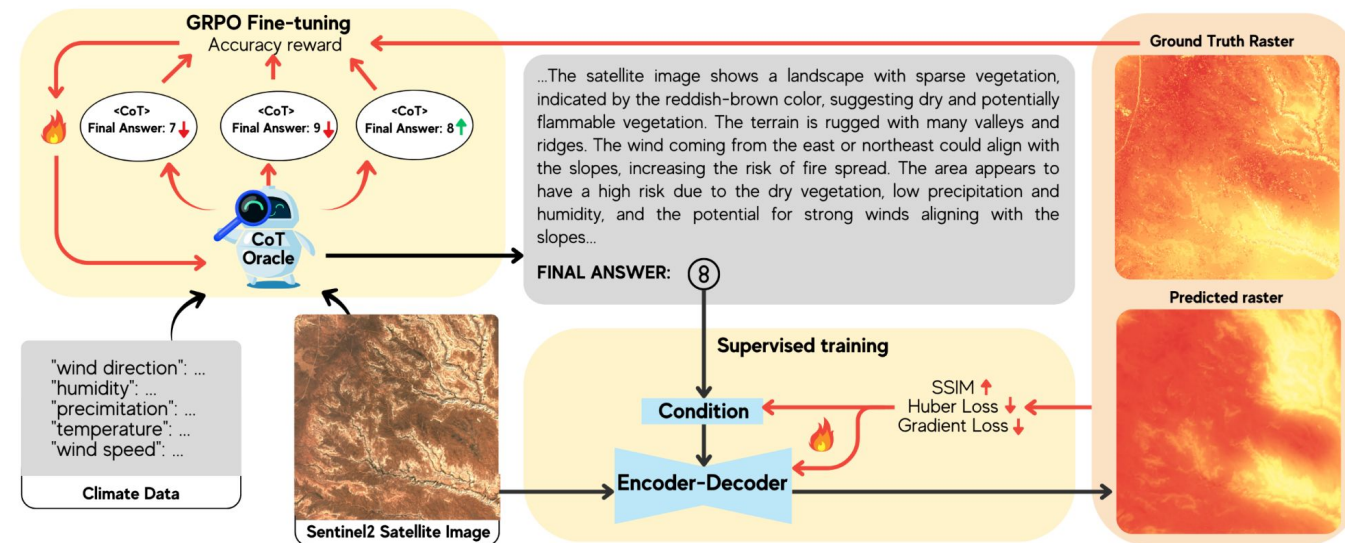
- GeoMMAgent explicitly separates retrieval, perception, and reasoning into coordinated roles.

FireScope

- FireScope uses a chain-of-thought oracle to reason about drivers of wildfire risk, then conditions a dense raster prediction model on that reasoning.



https://openaccess.thecvf.com/content/CVPR2026/papers/Xiao_GeoMMBench_and_GeoMMAgent_Toward_Expert-Level_Multimodal_Intelligence_in_Geoscience_and_CVPR_2026_paper.pdf



https://openaccess.thecvf.com/content/CVPR2026/papers/Markov_FireScope_Wildfire_Risk_Raster_Prediction_With_a_Chain-of-Tought_Oracle_CVPR_2026_paper.pdf

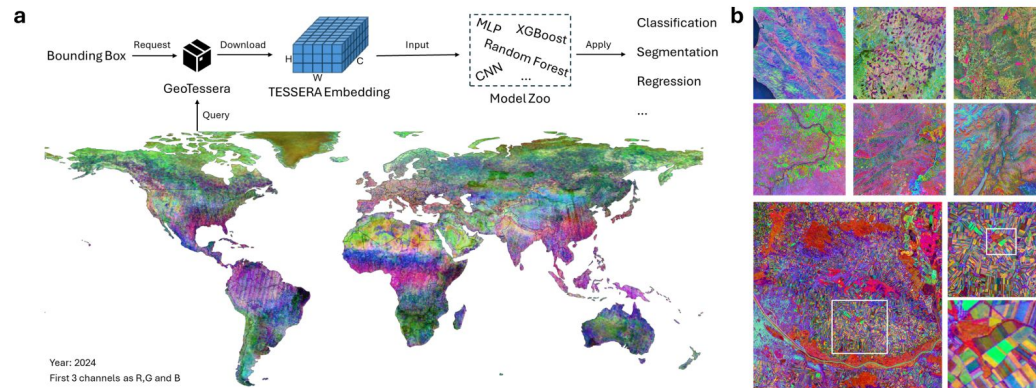
CVPR26: Meta Insights into Trends and Tendencies (91/152)

Remote Sensing x CV : Multimodal earth representation models

- ❑ TESSERA and temporal Earth embeddings
 - ❑ Pixel-wise representations from Sentinel-1 /2 time series instead of single-scene image features
- ❑ MM-OVseg/Open-vocabulary Optical-SAR models
 - ❑ Connects EO pixels to language-aligned concepts, enabling flexible downstream tasks.

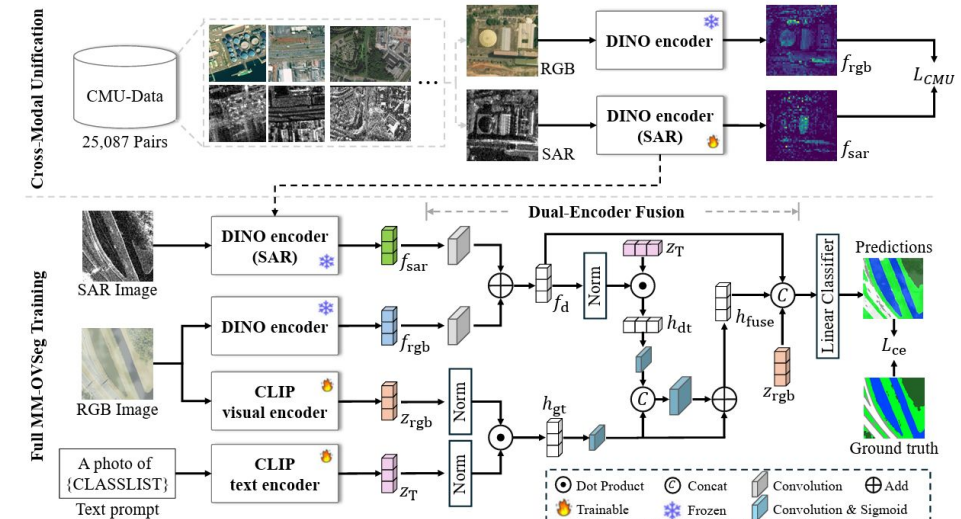
Thoughts: At CVPR26, remote sensing x CV is shifting from adapting vision foundation models to building EO-native, multimodal Earth representation models.

TESSERA



https://openaccess.thecvf.com/content/CVPR2026/papers/Feng_TESSERA_Temporal_Embeddings_of_Surface_Spectra_for_Earth_Representation_and_CVPR_2026_paper.pdf

MM-OVseg



https://openaccess.thecvf.com/content/CVPR2026/papers/Wei_MM-OVSeg_Multimodal_Optical-SAR_Fusion_for_Open-Vocabulary_Segmentation_in_Remote_Sensing_CVPR_2026_paper.pdf

CVPR26: Multimodal LLM-as-Judge (2/3)

Trend A — The judge becomes a trained model

- ❑ GT-SVJ: a self-supervised video judge for efficient video reward modeling — no human preference labels. [Shekhar et al., CVPR 2026]
- ❑ SpatialReward: verifiable spatial reward modeling — VLM reward models reward “plausible but spatially wrong” scenes, so it adds explicit spatial verification. [Zhou et al., CVPR 2026]
- ❑ Saliency-R1: an LLM judge supplies the accuracy reward used to train faithful, interpretable vision-language reasoning. [Gong et al., CVPR 2026]
- ❑ White-Balance First, Adjust Later: a VLM acts as evaluator inside a low-level vision pipeline (cross-camera color constancy) — judging is no longer just a leaderboard. [Li, Tan & Tan, CVPR 2026]
- ❑ Pattern: judges specialize by modality (video, spatial), become self-supervised / verifiable, and move inside the training loop as reward signals — not post-hoc scores.

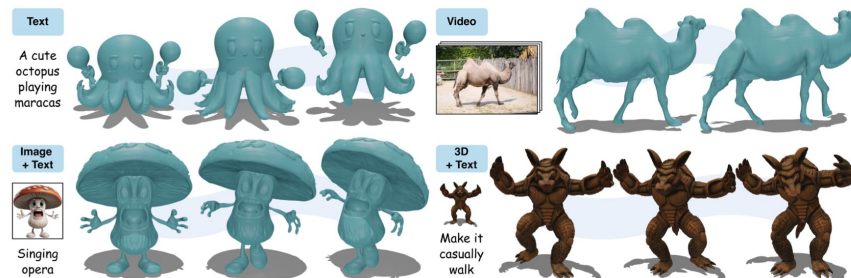
CVPR26: Multimodal LLM-as-Judge (3/3)

Trend B — Can we trust the judge? Reliability is the new bottleneck

- ❑ Where judges still fail (CVPR'25 baselines):
 - ❑ VL-RewardBench: even GPT-4o reaches only 65.4% as a VL judge — and failures are perception, not reasoning. [Li et al., CVPR 2025]
 - ❑ UPME: unsupervised peer-review to cut single-judge bias in MLLM evaluation. [Zhang et al., CVPR 2025]
- ❑ CVPR'26 responses — engineering reliability in:
 - ❑ SpatialReward → verifiable rewards over merely plausible-looking outputs. [CVPR 2026]
 - ❑ Reallocating Attention Across Layers → reduce multimodal hallucination at the source. [Lu et al., CVPR 2026]
 - ❑ GuardTrace-VL → iterative safety supervision of multimodal reasoning traces. [CVPR 2026]
- ❑ Takeaway: the bottleneck is shifting from “can it generate?” → “can we measure it reliably?” Perception-grounded, verifiable, bias-aware judging (incl. dual- / panel-judge protocols) is the open frontier.

ActionMesh: Animated 3D Mesh Generation with Temporal 3D Diffusion

- ❑ **Summary**: ActionMesh introduces temporal 3D diffusion for fast generation of animated, production-ready 3D meshes from videos, text, or existing meshes. By combining synchronized latent diffusion with a temporal 3D autoencoder, it produces rig-free, topology consistent animations. The method achieves SOTA geometric accuracy and temporal consistency while enabling rapid, practical 3D content creation.
- ❑ **Novelty**: The introduction of “temporal 3D diffusion”, the key insight of this work modifying existing 3D diffusion models to include a temporal axis. Fast, rig free and topology consistent. Representing a meaningful and practically impactful extension of existing 3D diffusion models rather than a fundamentally new generative paradigm.
- ❑ **Remarks**: Very fast mesh generation compared to others. Fatal limitations on occluded videos such as arms from human body or fingers.



Input/Output from ActionMesh.

Table 1. **Quantitative results on ActionBench.** We report results from prior works, namely LIM [33], DreamMesh4D [19] V2M4 [4], ShapeGen4D[53], TripoSG [18] and TRELIS [48]. We outperform all baselines across metrics while running significantly faster.

Method	Time	CD-3D ↓	CD-4D ↓	CD-M ↓
TRELIS [48]	15min	0.065	0.181	-
TripoSG [18]	2min	0.056	0.184	-
DM4D [19]	35min	0.104	0.152	0.265
LIM [33]	15min	0.089	0.126	0.243
V2M4 [4]	35min	0.068	0.340	0.616
SG4D [53]	15min	0.056	0.170	0.348
Ours	2min	0.053	0.081	0.148

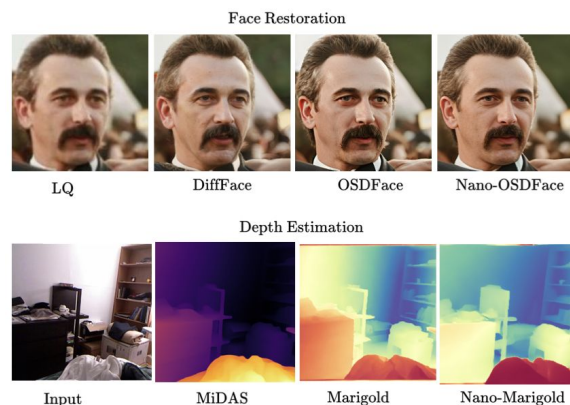


Limitations

[Remy Sabathier et al, CVPR 2026.][[Link](#)]

NanoSD: Edge Efficient Foundation Model for Real Time Image Restoration

- ❑ **Summary** : NanoSD introduces a family of lightweight diffusion foundation models distilled from Stable Diffusion 1.5 through joint optimization of the U-Net and VAE. By preserving latent space structure while balancing accuracy, latency, and model size, NanoSD enables real time edge deployment and achieves SOTA performance across multiple image restoration and generation tasks.
- ❑ **Novelty** : They introduce a novel compression strategy that redisings and distils the diffusion U-Net and VAE rather than optimizing individual components. The work emphasizes hardware co-design in order to achieve real time results.
- ❑ **Remarks** : Focused on performance and real-time performance. Strong engineering contribution with practical impact.



Overall showcase of NanoSD.

Table 3. Quantitative comparison of different methods on synthetic CelebA-Test [15] dataset. The best, second-best and third-best results are highlighted in red, blue, and green colors, respectively. Refer to Supplemental for detailed analysis.

Method	LPIPS↓	DISTS↓	NIQE↓	MUSIQ↑	FID↓	LMD↓	Steps	MACs (G)	Para. (M)
PGDiff [48]	0.386	0.195	4.001	69.56	44.62	7.310	1000	480998	176
DiffFace [51]	0.346	0.212	4.638	66.74	49.80	5.475	250	46575	175
DiffBIR [25]	0.374	0.234	6.280	75.63	71.77	5.104	50	24234	1717
OSDFace [40]	0.336	0.177	3.884	75.64	45.41	5.286	1	2465	1887
Nano-DiffBIR (Ours)	0.382	0.212	6.3	76.52	70.89	5.231	50	4734	726
Nano-OSDFace (Ours)	0.341	0.182	3.913	76.01	45.92	5.172	1	479	415

Quantitative results.

RetimeGS: Continuous-Time Reconstruction of 4D Gaussian Splatting

- ❑ **Summary**: RetimeGS addresses temporal aliasing in 4D Gaussian Splatting for dynamic scene retiming. By explicitly modeling Gaussian temporal behavior and introducing optical flow-guided initialization, supervision, and triple-rendering constraints, it enables smooth, ghost-free interpolation at arbitrary timestamps. Experiments show improved temporal coherence and rendering quality across challenging dynamic scenes.
- ❑ **Novelty**: This work addresses temporal aliasing, which is largely overlooked limitation of other existing methods. The strong introduction of this work is the Gaussian modeling, optical flow guidance, and triple-rendering.
- ❑ **Remarks**: Addresses an important temporal interpolation issue in 4DGS. Limitations on low frame rate videos.

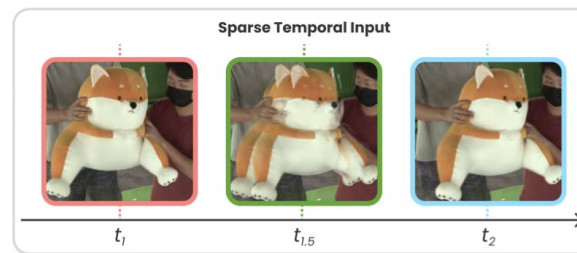
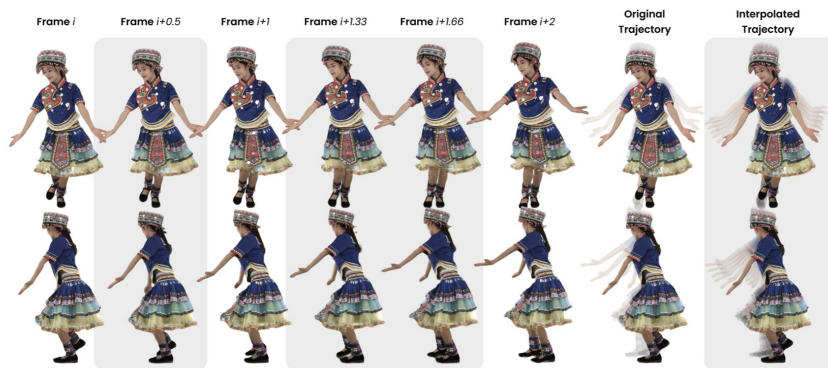


Figure 2. Illustration of temporal overfitting to input frames t_1 and t_2 , causing ghosting at $t_{1.5}$ in 4D primitive-based methods.

Temporal aliasing.

Table 1. **Quantitative comparisons on the Stage-Capture Dataset, focusing on the foreground region.** Red and yellow cell colors indicate the best and second-best results, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Deform-GS [39]	28.45	0.867	0.0272
STGS [44]	25.34	0.825	0.0357
GaussianFlow [6]	25.91	0.825	0.0339
2D Lifting [28, 44]	28.79	0.886	0.0267
Ours	30.08	0.904	0.0225

Quantitative results.

Selective Amnesia using Contrastive Subnet Erasure for Class Level Unlearning in Vision Models

- ❑ **Summary:** This paper proposes Contrastive Subnet Erasure (CSE), a training-free method for concept-level forgetting, which aims to remove an entire semantic category rather than specific training examples by attenuating channels responsible for a target semantic category.
- ❑ **Novelty:** Unlike prior methods that remove broad feature subspaces or update the entire model, CSE identifies a compact set of target-specific channels via contrastive generalized eigenanalysis and attenuates only those channels, reducing damage to non-target representations.
- ❑ **Thoughts:** CSE has practical advantages because it does not require training and introduces no extra inference-time cost. In addition, the cross-dataset evaluation protocol provides a more robust test of concept-level unlearning.

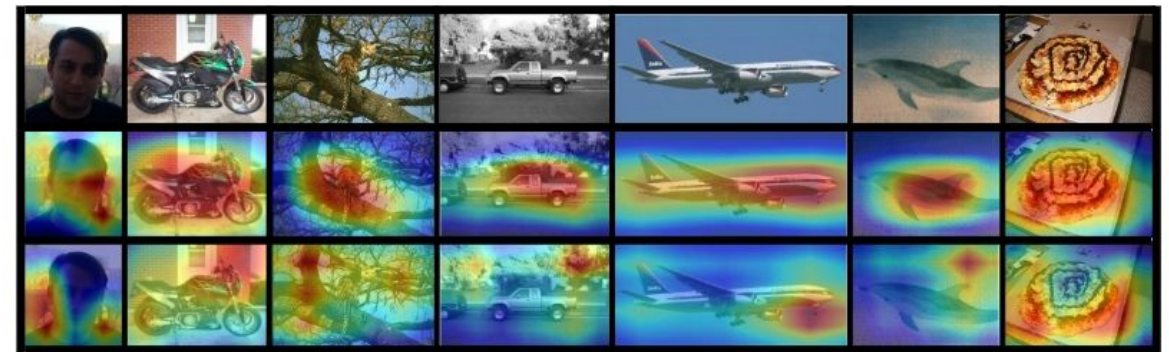
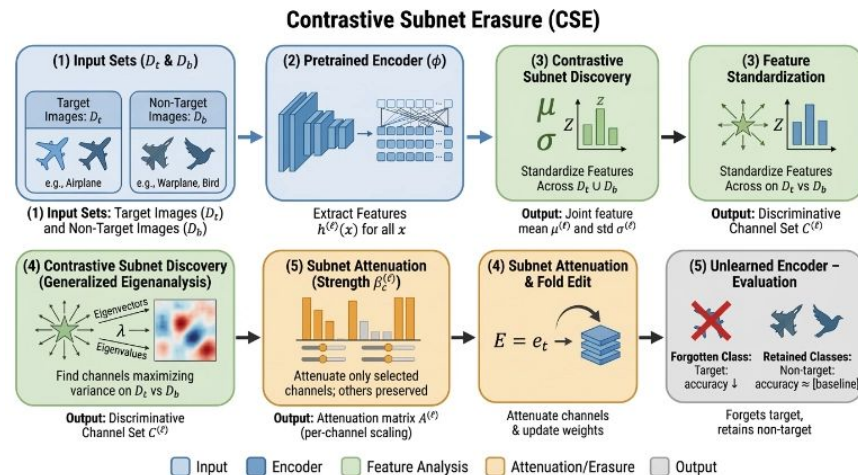
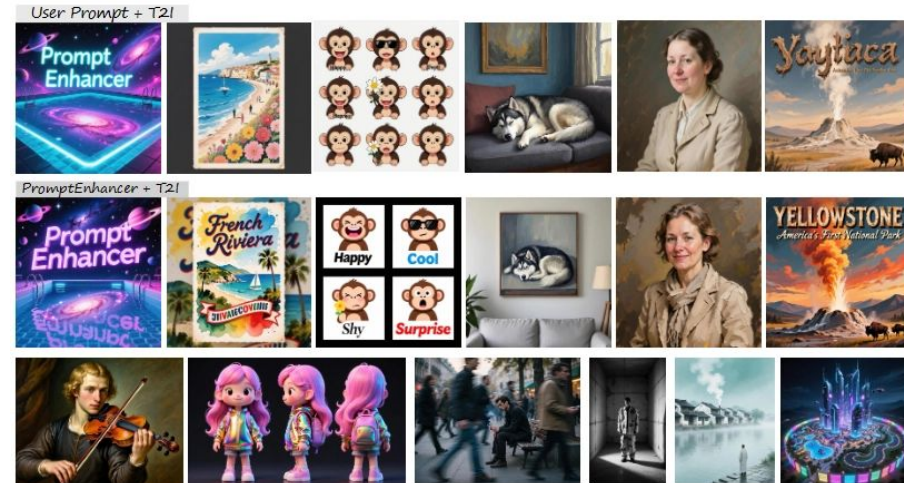
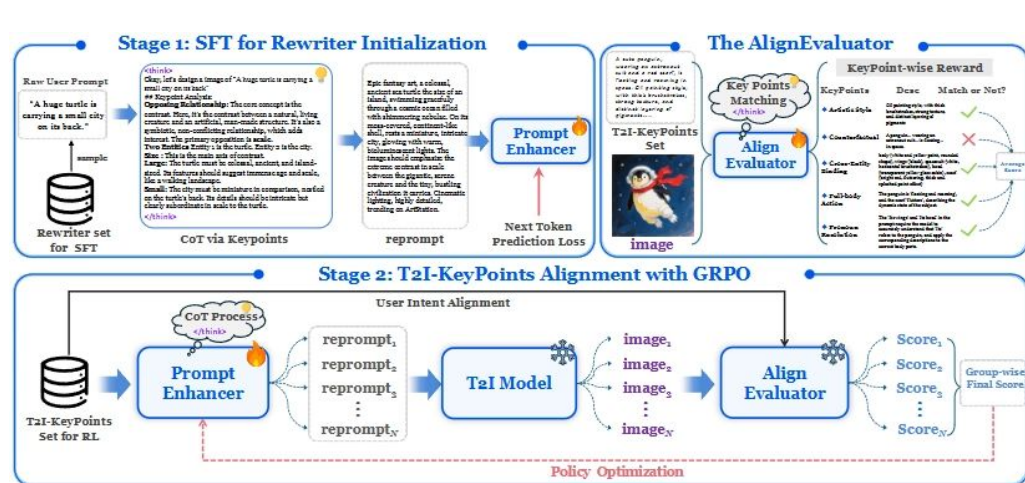


Figure 3. Qualitative effect of CSE: Each row shows the **Original**, **Grad-CAM (Before)**, and **Grad-CAM (After)**. CSE suppresses saliency on the target concept while preserving responses on non-target content.

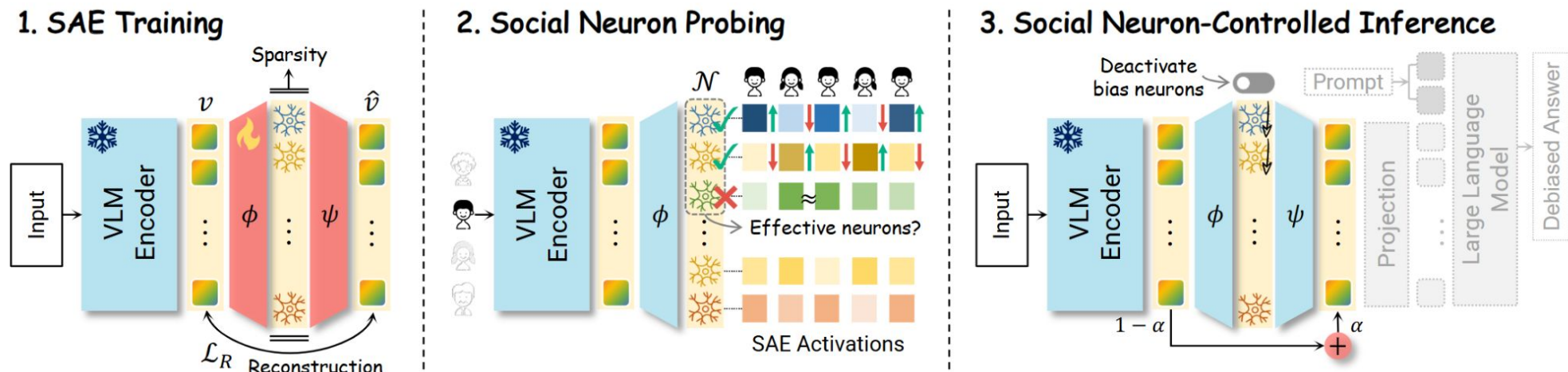
PromptEnhancer: Taming Your Rewriter for Text-to-Image Generation via Fine-Grained Reward

- ❑ **Summary:** PromptEnhancer rewrites user prompts into more explicit and structured descriptions for frozen T2I models. The rewriter is trained with CoT-style supervision and further optimized using AlignEvaluator, a fine-grained image-text alignment reward model.
- ❑ **Novelty:** Unlike prior prompt rewriting methods that rely on specific generators or coarse rewards, PromptEnhancer uses T2I-specific fine-grained feedback to train a plug-and-play CoT rewriter that generates prompts easier for T2I models to follow.
- ❑ **Thoughts:** Fine-grained evaluators like AlignEvaluator could serve as diagnostic tools for prompt optimization not only in T2I generation but also in other vision tasks, enabling more targeted and adaptive optimization.



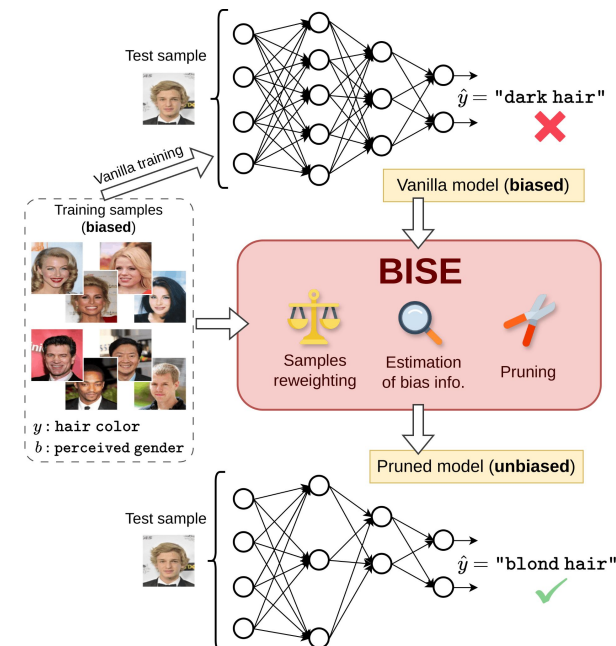
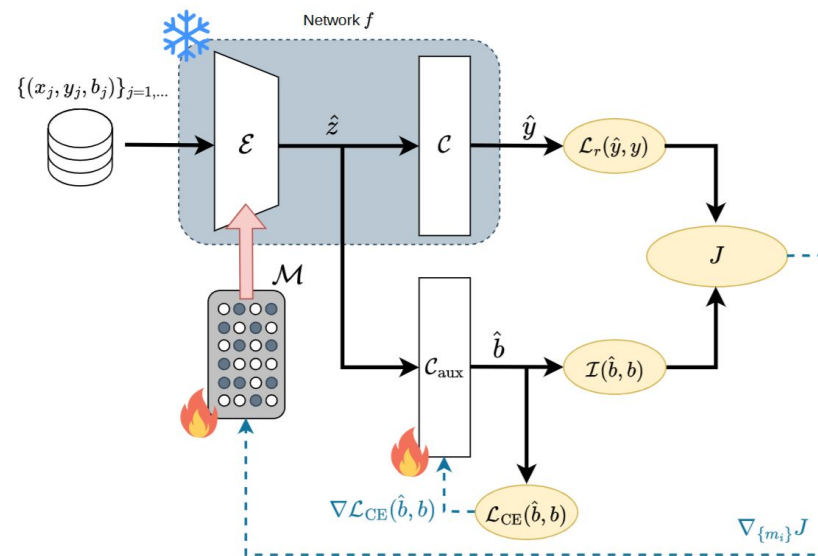
Interpretable Debiasing of Vision–Language Models for Social Fairness

- ❑ **Summary:** The authors propose DEBIASLENS, a novel debiasing method that localizes social attribute neurons in VLMs through SAE.
- ❑ **Novelty:** By identifying social bias neurons within the SAE attached to the VLM, and selectively deactivating them, this approach achieves interpretable and controllable bias mitigation.
- ❑ **Thoughts:** While existing debiasing methods are limited to surface-level interventions, DEBIASLENS enables a structural understanding of biases within internal model dynamics. It represents the first attempt to achieve interpretable bias mitigation for MLLMs.



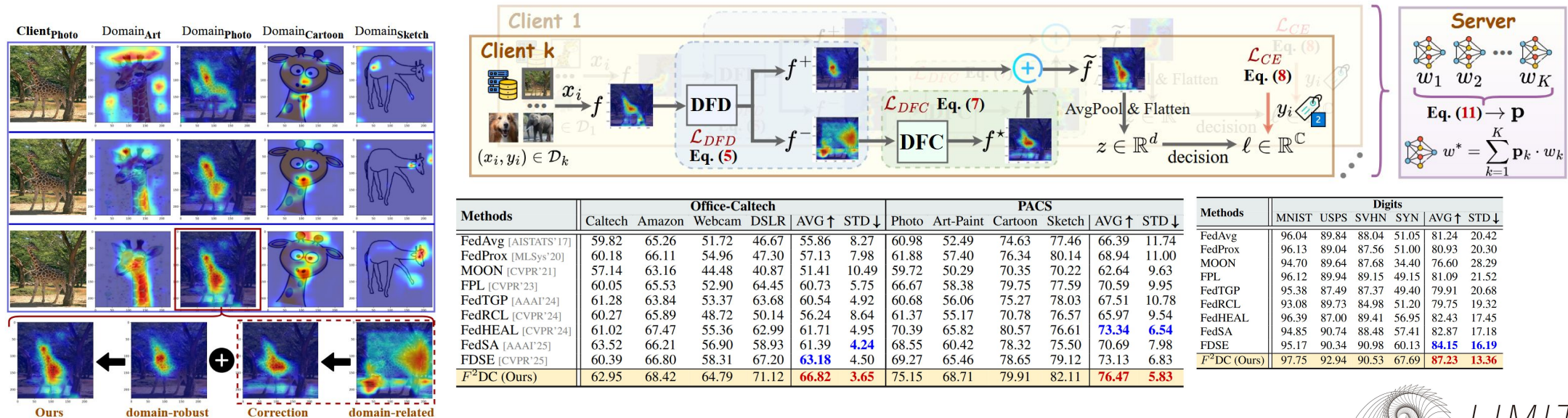
Bias In, Bias Out? Finding Unbiased Subnetworks in Vanilla Models

- ❑ **Summary:** This paper introduces BISE, a novel debiasing method that utilizes structured pruning to isolate fair representations without requiring additional unbiased training data.
- ❑ **Novelty:** Unlike traditional data-centric balancing or model-centric optimization strategies, BISE uniquely extracts inherently unbiased subnetworks from frozen pre-trained networks.
- ❑ **Thoughts:** The proposed method not only mitigates undesirable biases but also accelerates inference by reducing network complexity through structured pruning. However, BISE strictly requires either bias-annotated training data or an extra model for pseudo-labeling.



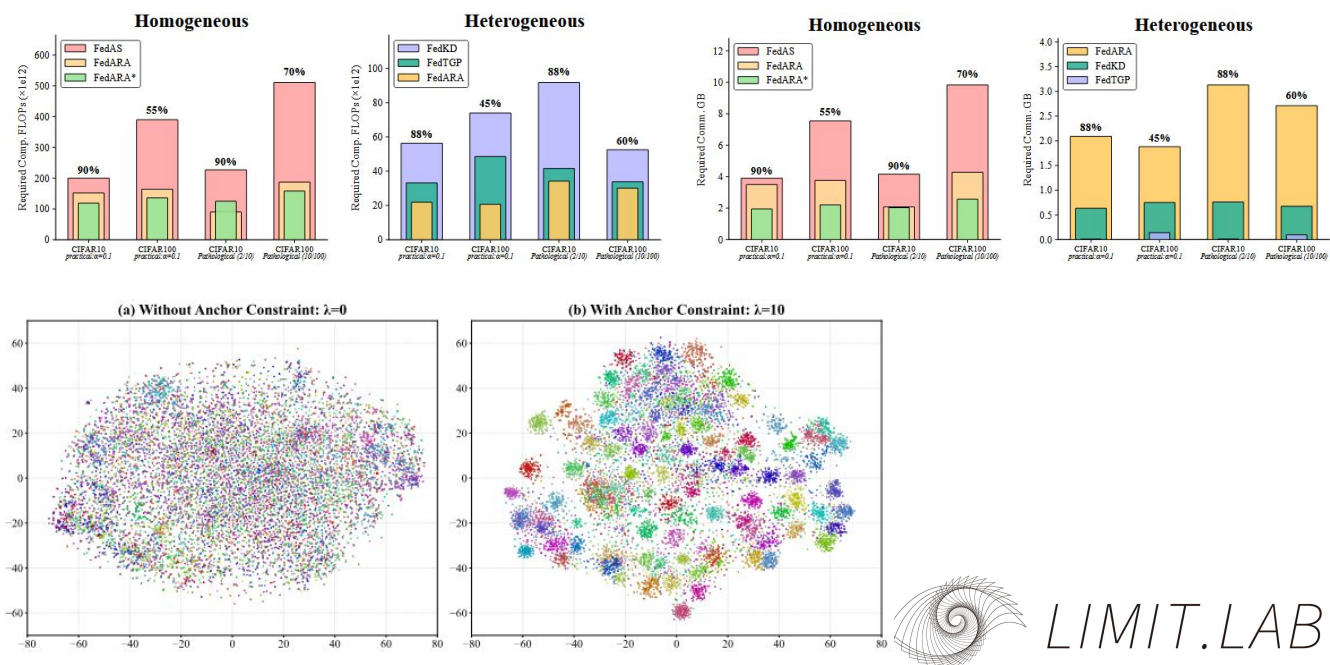
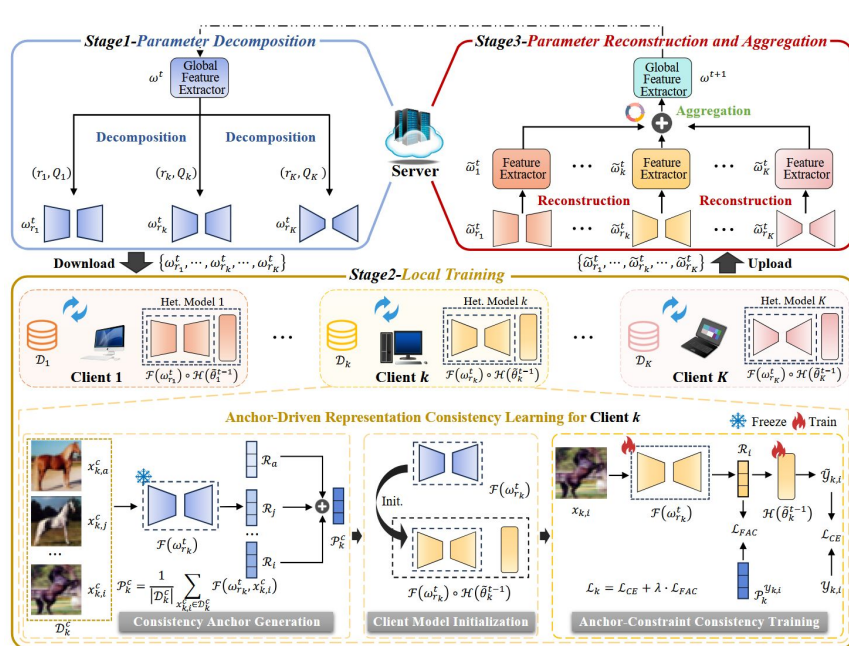
Domain-Skewed Federated Learning with Feature Decoupling and Calibration

- Summary:** This study addressed the issue that, due to domain skew in FL, each client's feature representation collapses into a domain-specific low-dimensional space. This inconsistency in representations across clients makes global aggregation difficult.
- Novelty:** The novel method decouples domain-biased features from each client's representation, then calibrates them to suppress domain-specific bias while preserving class-specific information for effective aggregation by considering the domain discrepancy across clients.
- Thoughts:** The proposed method is an effective strategy, but it assumes that the domain of each client is known in advance. It may be difficult to apply in scenarios where domains are unknown or excessively diverse.



FedARA: Resource-adaptive Low-rank Personalized Federated Learning via Anchor-driven Representation Alignment on Heterogeneous Edge Devices

- ❑ **Summary:** The study achieves personalized and semantic-consistent FL by jointly considering feature-space distribution shifts across clients and resource heterogeneity among edge devices.
- ❑ **Novelty:** The novelty lies in integrating resource-adaptive personalization via low-rank decomposition with semantic consistency through anchor-driven feature alignment.
- ❑ **Thoughts:** Although this is an interesting study addressing both data and edge resource heterogeneity, the experiments were limited to simulation settings.



SARMAE: Masked Autoencoder for SAR Representation Learning

- ❑ This paper presents a solution for SAR representation learning bottlenecks including limited pretraining scale, speckle noise from SAR imaging physics, and weak semantic structure compared with optical imagery.
 - ❑ They propose a large SAR corpus, a SAR noise-aware reconstruction, and optical-feature anchoring.
- ❑ Main Trend: SAR foundation models are moving from generic SSL adaptation to data-scale, sensor/physics-aware, multimodal training.
 - ❑ SAR dataset curation matters. One of their main contributions is the large data corpus.
 - ❑ This paper also considers SAR-specific SSL (similar to our work [SAR-W-MixMAE](#) and [SAR-JEPA](#)). Future remote sensing SSL will likely encode sensor physics directly, not just borrowing optical image based approaches.
 - ❑ Cross-modal guidance can improve SAR inference.

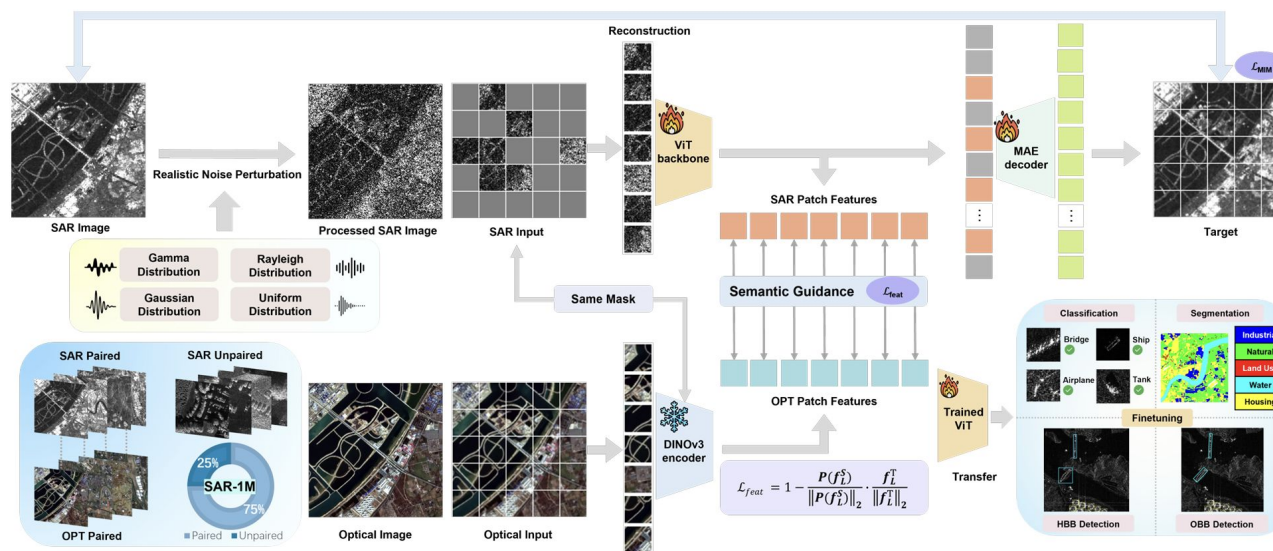
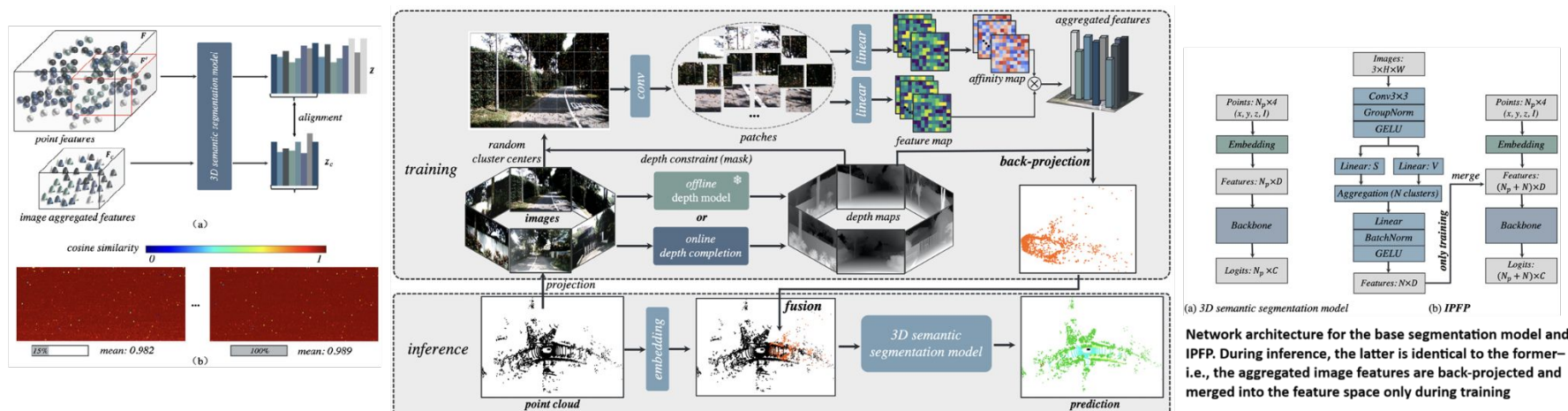


Image-to-Point Cloud Feature Back-Projection for Multimodal Training of 3D Semantic Segmentation

- Image-LiDAR fusion improves 3D segmentation, but common fusion methods are costly and often require images at inference. Fusion is also difficult because camera and LiDAR have different fields of view, so many LiDAR points do not have image correspondence.
 - The core idea here is using images as training-time 3D feature augmentation. They aggregate local image features, estimate their depth, back-project them into 3D, and train one point cloud backbone with LiDAR + image-derived feature points. At inference, remove the image stream and run LiDAR-only.
- Images are used during training to improve the LiDAR model without becoming a deployment requirement.
 - Image features are treated as pseudo-3D points rather than attached only to projected LiDAR points. Their approach gives modest but consistent gains with almost no parameter overhead.
 - The main value is not a huge mIoU boost, it is a cleaner and cheaper way to exploit multimodal data when inference may be LiDAR-only.
 - This is practical because it improves the LiDAR model without increasing test-time sensor or computation requirements.



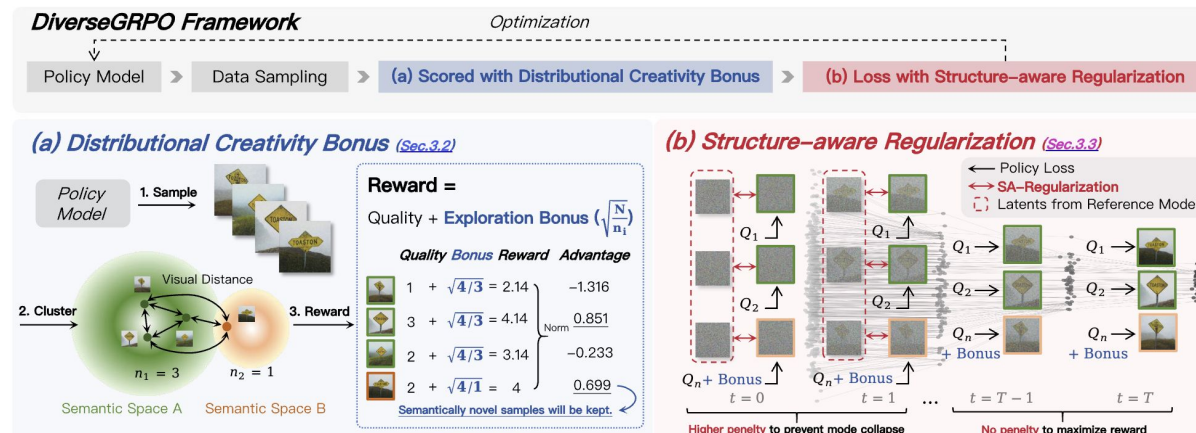
DiverseGRPO: Mitigating Mode Collapse in Image Generation via Diversity-Aware GRPO

□ Summary

- This work aims to address the problem of generated outputs collapsing into a single mode after RL-based post-training of text-to-image models. Through experiments on Stable Diffusion 3.5 and Flux, the authors demonstrate that their method improves human-preference metrics while maintaining generation diversity. However, in the main table, the comparison is made against a Flow-GRPO setting in which KL regularization is omitted. The paper does not compare against the more widely used setting that includes KL regularization.

□ Novelty

- Building on Flow-GRPO, the paper proposes a clustering-based diversity reward called the Distributional Creativity Bonus, as well as structure regularization, which leverages the observation that image structure is determined during the early denoising steps.



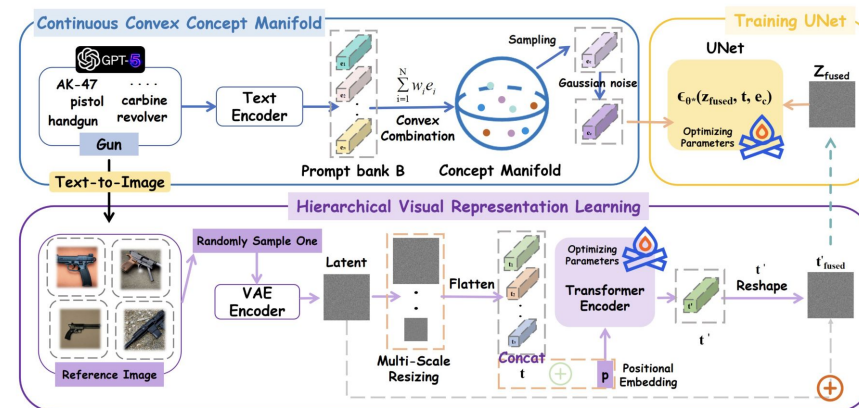
Beyond Text Prompts: Precise Concept Erasure through Text–Image Collaboration

□ Summary

- This work tackles precise concept erasure in text-to-image diffusion models. Existing text-only methods often fail to erase concepts under diverse or adversarial prompts, while image-assisted methods may over-erase visually similar but unrelated concepts. TCoE addresses this by combining textual concept manifolds with hierarchical visual representations, achieving stronger erasure while preserving related benign concepts such as camera when erasing gun.

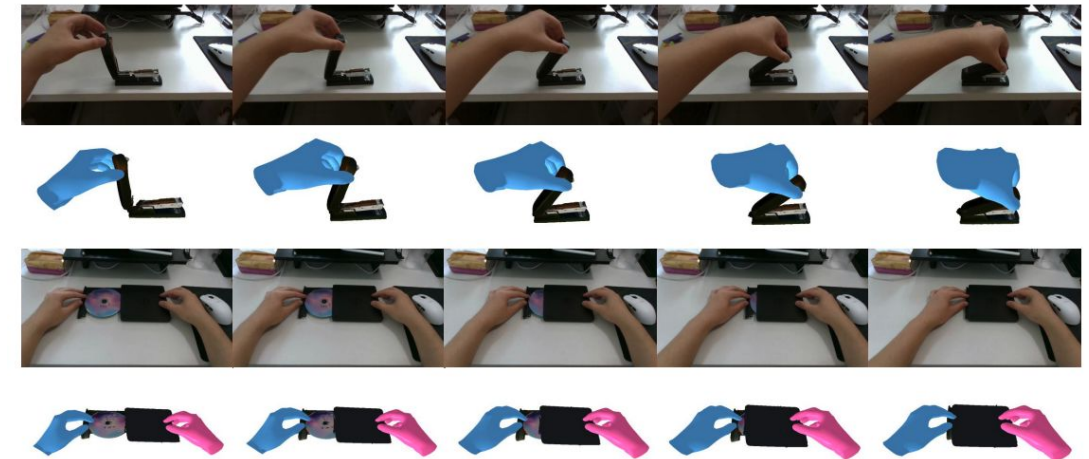
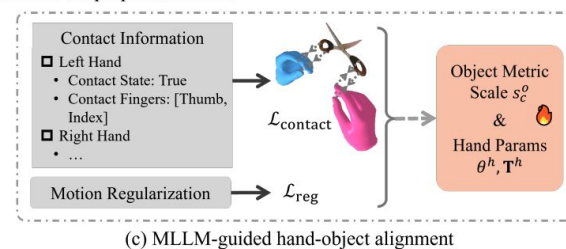
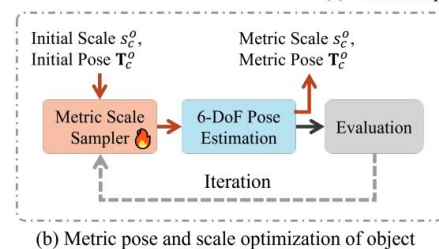
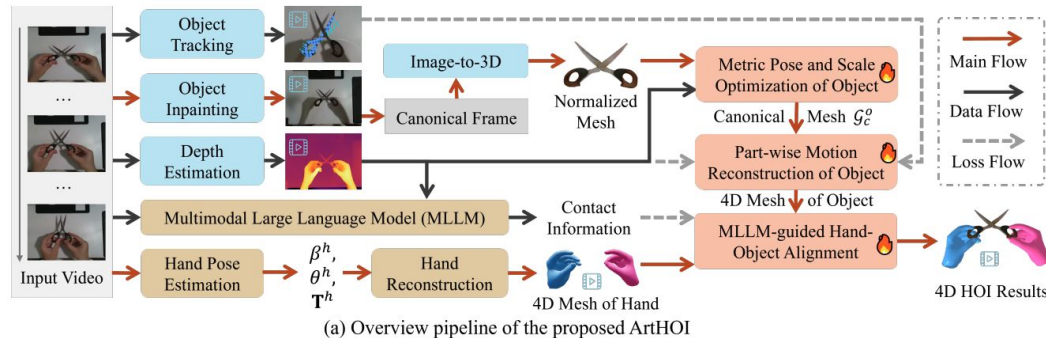
□ Novelty

- Authors propose a text-image collaborative erasure framework composed of Continuous Convex Concept Manifold, which expands concept coverage through convex combinations of diverse prompt embeddings, and Hierarchical Visual Representation Learning, which uses multi-scale visual latents to distinguish target concepts from visually similar non-targets. It also introduces MCP, a preservation-oriented metric for evaluating over-erasure.



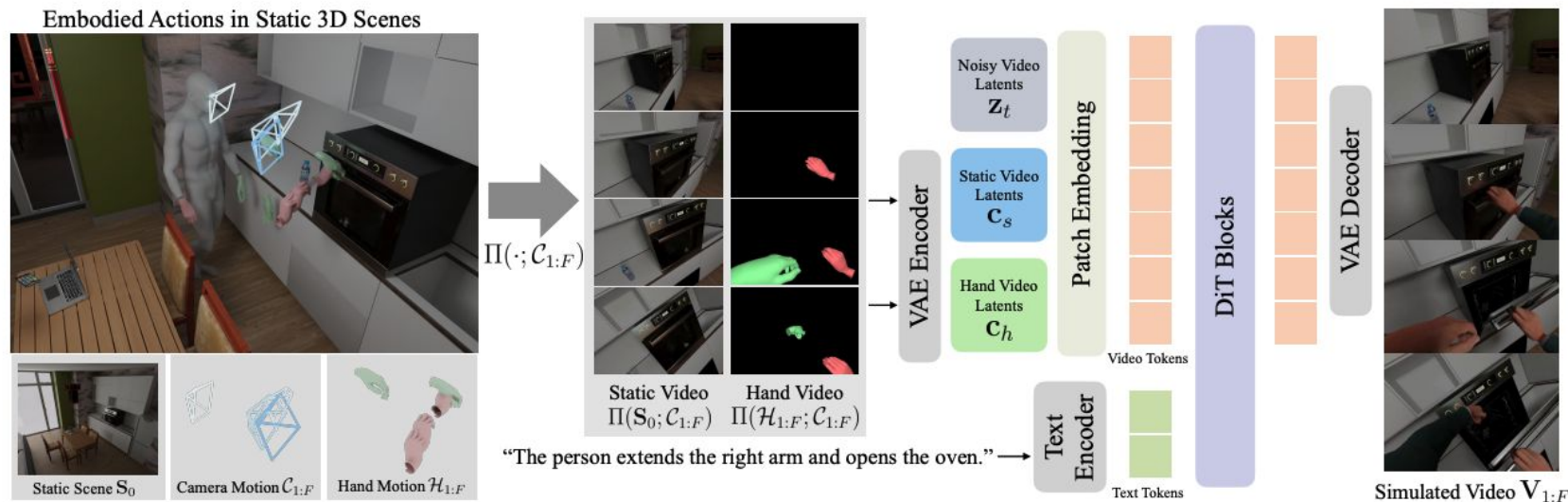
ArtHOI: Taming Foundation Models for Monocular 4D Reconstruction of Hand-Articulated-Object Interactions (Highlight)

- ❑ **Summary:** ArtHOI reconstructs 4D hand interactions with unknown articulated objects from a monocular RGB video by combining and optimizing priors from multiple foundation models.
- ❑ **Novelty:** The key novelty is its optimization framework that refines noisy foundation-model outputs through Adaptive Sampling Refinement for metric object pose/scale and MLLM-guided contact reasoning for physically plausible hand-object alignment.
- ❑ **Thoughts:** This work is built upon many foundation models. This work reflects a broader trend where monocular 4D reconstruction has evolved from **hand-only reconstruction** in CVPR 2025 to **hand-object reconstruction** in CVPR 2026.



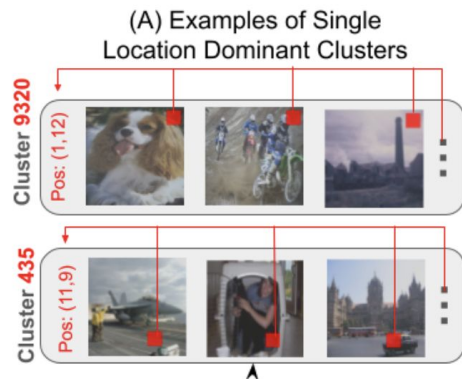
Dexterous World Models

- ❑ **Summary:** DWM generates temporally coherent egocentric videos of plausible human–scene interactions by conditioning a video diffusion model on static 3D scene renderings and egocentric hand–motion sequences.
- ❑ **Novelty:** DWM avoids entangling background synthesis with interaction dynamics by using static–scene videos as an identity–preserving condition and hand–mesh renderings as action cues, allowing the model to focus on manipulation–induced residual changes.
- ❑ **Thoughts:** DWM and other related CVPR2026 works such as [HVG-3D](#), [SCAR](#), and [HandWorld](#) suggest that **hand–object interaction video generation** is becoming increasingly common within hand–related research.



Positional bias in the pre-trained visual representations

- ❑ DINOv2/3 has unintended positional bias in their patch representations.



[Venkataramanan+, CVPR2026.][[Link](#)]

[Cuttano+, CVPR2026 (Oral).][[Link](#)]

[Pawlowsky+, ArXiv.][[Link](#)]

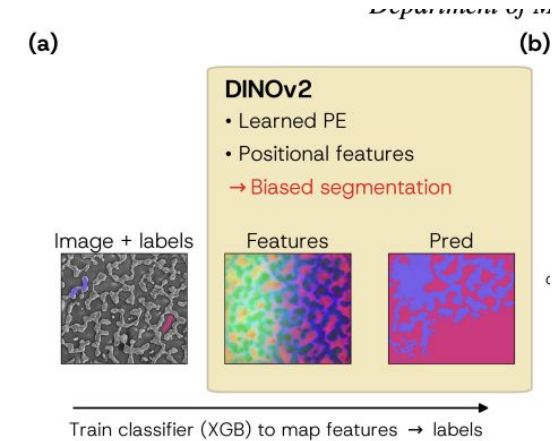
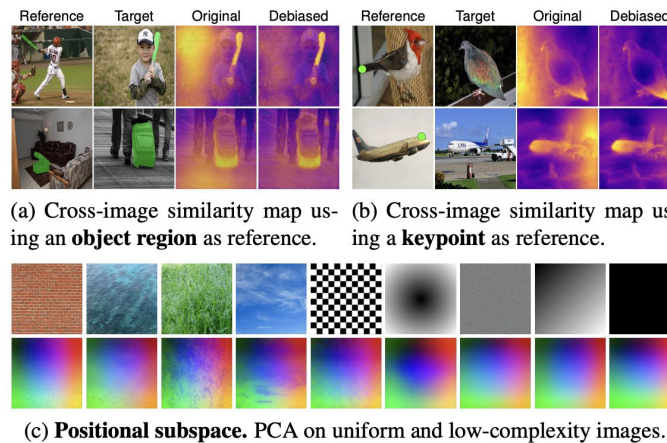
Such positional bias causes degradation in model's segmentation performance.

Current approaches tackle this problem by either

- (1) removing linearly predictable information from original patch features (INSID3 is elegant)
- (2) replacing original learnable positional encodings

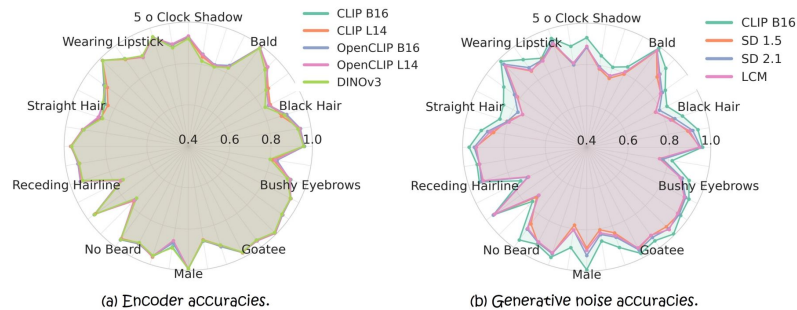
Future work would analyze *how* and *when* such positional bias arise during DINO pre-training.

Also, how such observations generalize to other SSL models? How can it be avoided during pre-training?



Can SSL and generative modeling help each other?

- ❑ Recent vision models implicitly learn isotropic Gaussian latent embeddings.
- ❑ Recent generative models explicitly learn mapping b/w data and Gaussian.



[Tasker+, CVPR2026.][[Link](#)]

This work demonstrates that DDIM-inverted noise and vision models demonstrate comparable performance on face attribute identification task.

While SSL models (e.g., CLIP, DINO) help learning useful representation for generative modeling, the inverse direction, generative modeling does not seem to help (often degrade) SSL models for now.

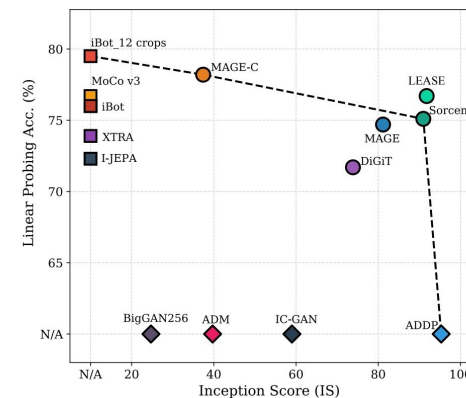


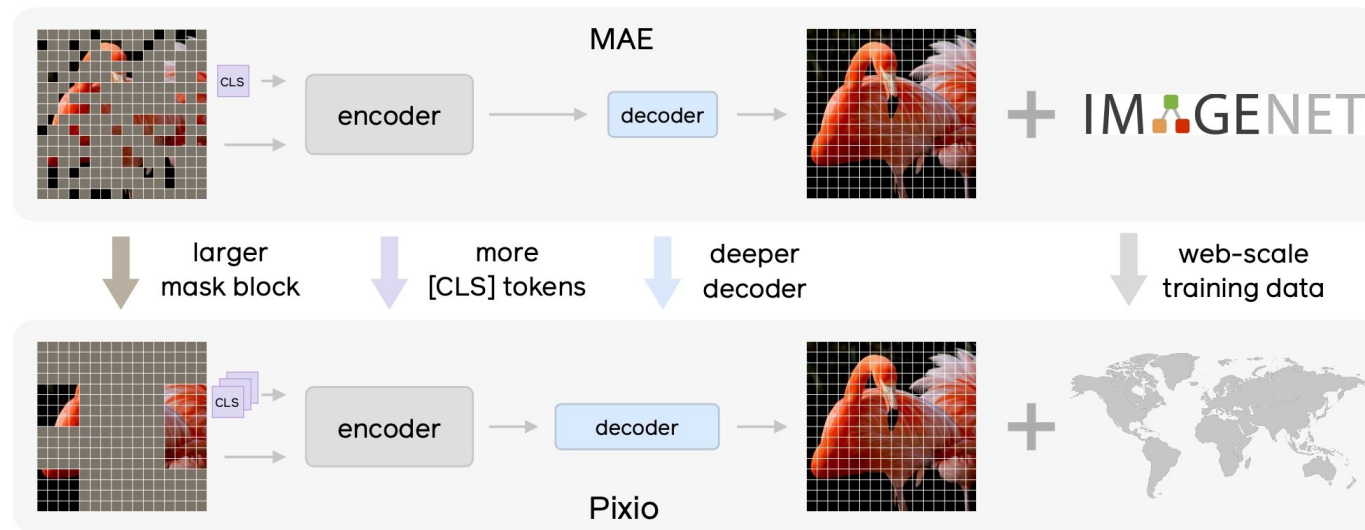
Figure 1. **Linear probing and class-unconditional generation performance.** Our approach, *LEASE*, advances the Unified SSL SoTA and defines a new Pareto-front for unified models.

[Estepa+, CVPR2026.][[Link](#)]

Unified model which capable both high-quality generation and global image understanding.

Scaling masked image modeling to web-scale pre-training

- ❑ Improved design of MAE & web-scale benchmark-independent dataset



Improved performance on dense prediction task (e.g., depth estimation, segmentation)

Method	ViT	#Params	DPT Head				Linear Regression Head			
			NYUv2		KITTI		NYUv2		KITTI	
			RMSE ↓	δ_1 ↑	RMSE ↓	δ_1 ↑	RMSE ↓	δ_1 ↑	RMSE ↓	δ_1 ↑
MAE	H/14	631M	0.465	80.8	2.740	90.9	0.595	70.3	4.419	79.4
DINOv2	L/14	304M	0.384	88.0	2.510	94.7	0.599	72.0	4.928	73.7
DINOv2	g/14	1137M	0.355	90.1	2.424	94.6	0.560	75.3	4.425	78.1
DINOv3	H+/16	841M	0.320	93.2	2.386	95.6	0.559	76.3	4.944	73.2
Pixio	H/16	631M	0.268	95.5	2.210	96.7	0.366	90.8	3.494	90.3

Table 1. Domain-specific **monocular metric depth estimation** with frozen encoder and a trainable DPT (Dense Prediction Transformer) head or a linear *regression* head.

[Yang+, CVPR2026.][[Link](#)]

While effective, MAE still needs to select intermediate features from encoder.

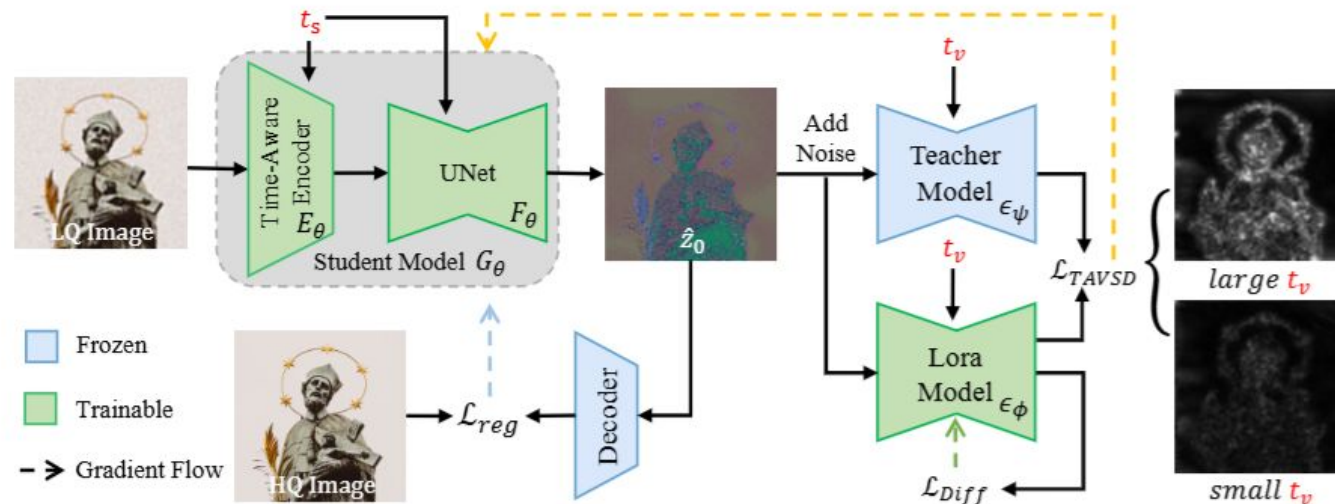
This paper argues that deeper decoder makes such feature selection easier.

However, we still need to extract intermediate features with heuristics as they did in their experiments.

Why masked autoencoder has such problems? But we know it happens even contrastive ones ([PE](#))...

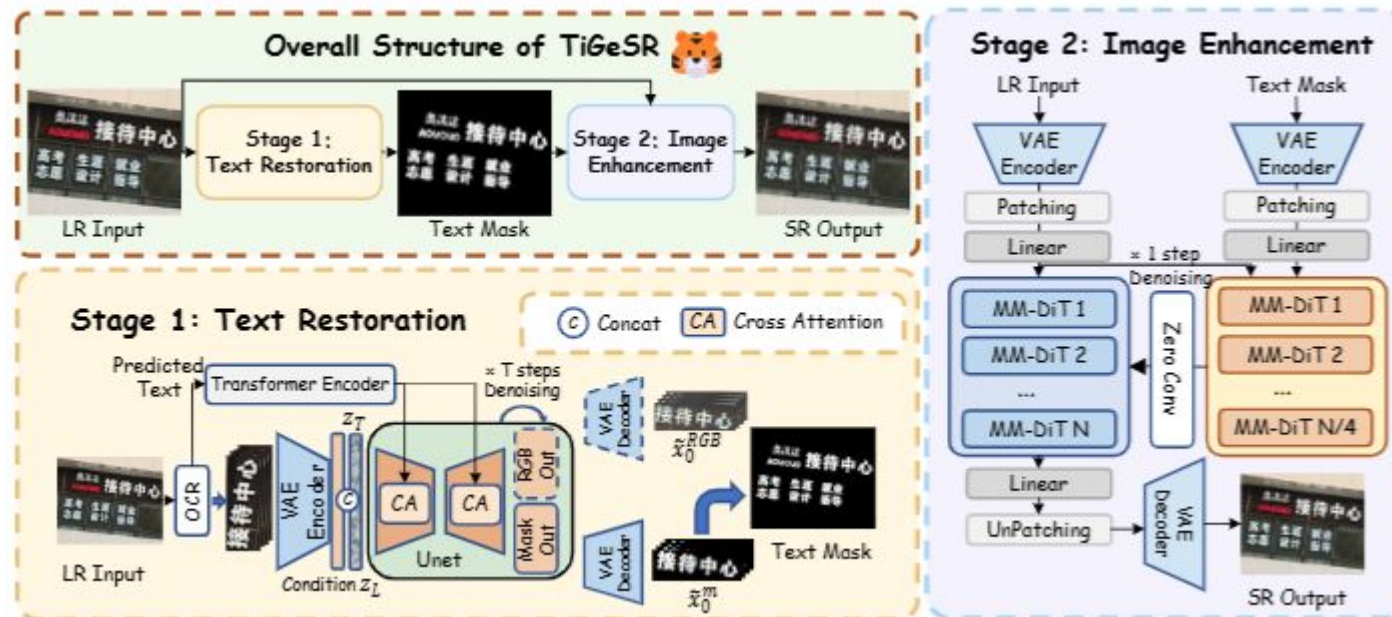
Time-Aware One Step Diffusion Network for Real-World Image Super-Resolution

- ❑ **Summary:** TADSR generates real-world super-resolution images with controllable fidelity and realism by effectively utilizing time-dependent generative priors, conditioning a one-step diffusion model on degraded low-resolution images and specified timesteps.
- ❑ **Novelty:** By using timestep-dependent dynamic latent features as the conditioning constraints and a mapping function bridging student-teacher timesteps as the guidance cue, TADSR avoids confusing different time-dependent generative priors, enabling the model to focus on the controllable synthesis of fidelity and realism tailored to the input timestep.
- ❑ **Thoughts:** The differences in super-resolution methods used for each image will likely continue to be a challenge.



Restore Text First, Enhance Image Later: Two-Stage Scene Text Image Super-Resolution with Glyph Structure Guidance

- ❑ **Summary:** TiGeSR addresses the trade-off between text and image quality in scene text super-resolution by decoupling text restoration from image enhancement.
- ❑ **Novelty:** TiGeSR, a novel two-stage framework that processes text and background separately to resolve the quality trade-off, is proposed along with UZ-ST, a newly constructed real-world dataset designed to verify the method under severe degradation.
- ❑ **Thoughts:** In text super-resolution, there is still a lack of photographs that accurately reflect real-world environments, and the expansion of the dataset will continue.



Back to Pixels from Latent Spaces in Diffusion Models (1/5)

❑ Back to Basics: Let Denoising Generative Models Denoise

- ❑ **Summary:** This paper proposes JiT, a minimal raw-pixel Transformer diffusion model that removes tokenizers, VAEs, pretraining, and auxiliary losses.
- ❑ **Novelty:** Its novelty lies in showing that clean image prediction can make pixel-space diffusion viable, reframing the problem as prediction-target design rather than dimensionality alone.
- ❑ **Thoughts:** This work serves as a conceptual foundation for the renewed interest in pixel-space diffusion.

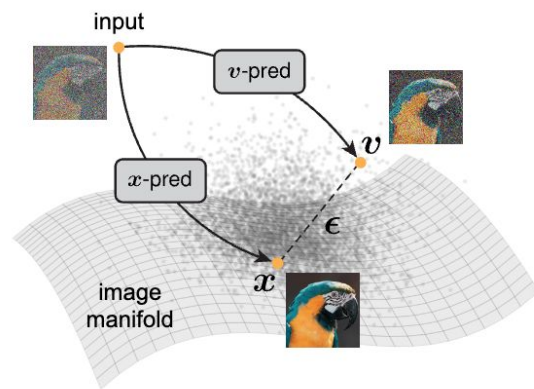


Figure 1. **The Manifold Assumption** [4] hypothesizes that natural images lie on a low-dimensional manifold within the high-dimensional pixel space. While a clean image x can be modeled as on-manifold, the noise ϵ or flow velocity v (e.g., $v = x - \epsilon$) is inherently off-manifold. Training a neural network to predict a clean image (i.e., x -prediction) is fundamentally different from training it to predict noise or a noised quantity (i.e., ϵ/v -prediction).

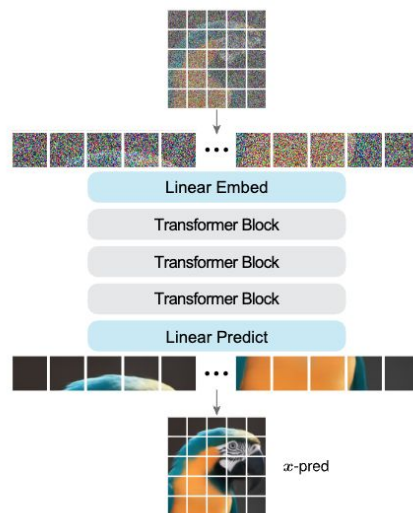
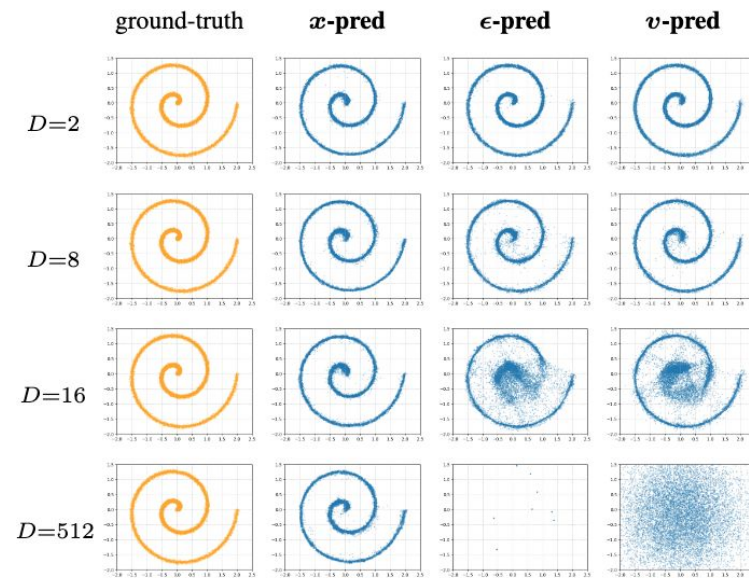


Figure 3. The “Just image Transformer” (JiT) architecture: simply a plain ViT [13] on patches of pixels for x -prediction.





Award Candidate

Back to Pixels from Latent Spaces in Diffusion Models (2/5)

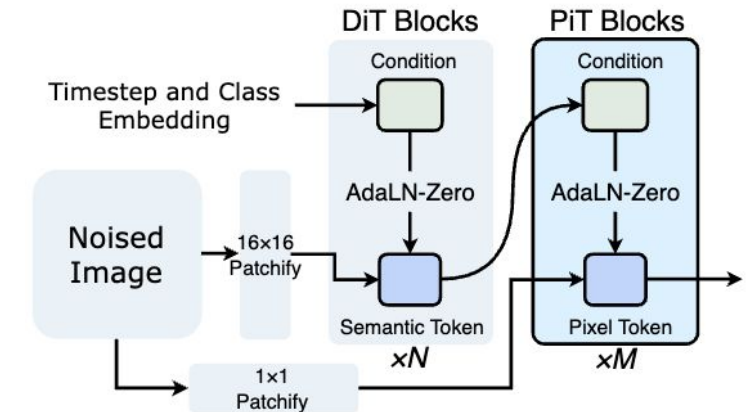
PixelDiT: Pixel Diffusion Transformers for Image Generation

- Summary:** PixelDiT proposes an end-to-end pixel-space Diffusion Transformer with patch-level global modeling and pixel-level detail refinement.
- Novelty:** Its novelty lies in scaling pure pixel-space diffusion to high-quality generation through explicit global-detail decomposition.
- Thoughts:** This paper is a central representative of CVPR 2026 pixel-space generation and shows that hierarchical design is key to practical pixel diffusion.



"A bicycle parked on the sidewalk..." → "A **motorcycle** parked on the sidewalk..."

(b) Train-free editing with FlowEdit [18] using different diffusion models. The imperfect reconstruction of VAEs used by FLUX [20] and Stable Diffusion 3 [9] produce severe and non-invertible distortions on the small details such as the scene text on the wall, therefore the subsequent diffusion editing fails. In contrast, by denoising directly in pixel space without a VAE, our PixelDiT avoids VAE reconstruction artifacts and helps maintain background consistency during local manipulations. For a fair comparison, all models are evaluated with $n_{\min}=0$ in FlowEdit to preserve more structures.



Back to Pixels from Latent Spaces in Diffusion Models (3/5)

❑ DiP: Taming Diffusion Models in Pixel Space

- ❑ **Summary:** DiP improves the efficiency of pixel-space diffusion by separating global structure generation from local detail restoration.
- ❑ **Novelty:** Its novelty lies in using a lightweight Patch Detailer Head to reduce the computational burden on the DiT backbone.
- ❑ **Thoughts:** This work shows that pixel-space diffusion can become practical when structure and detail are handled by separate modules.

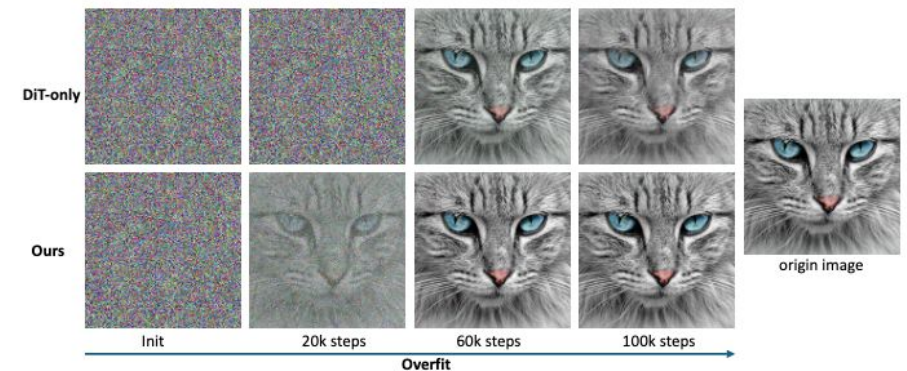
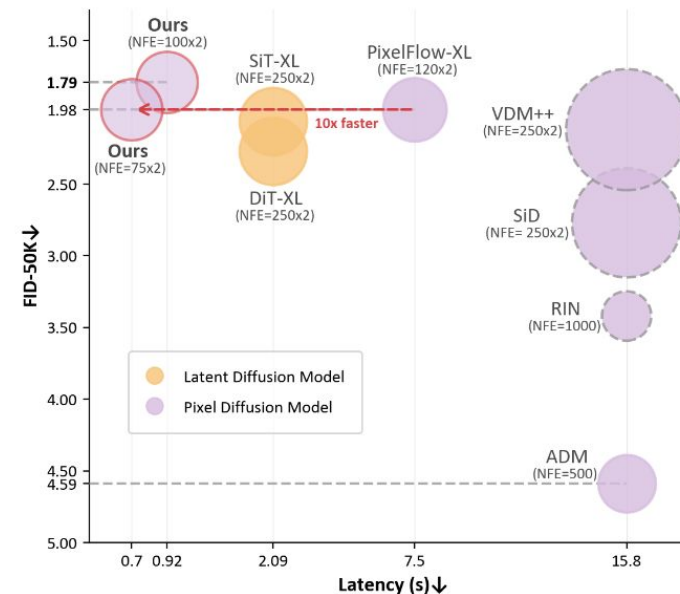
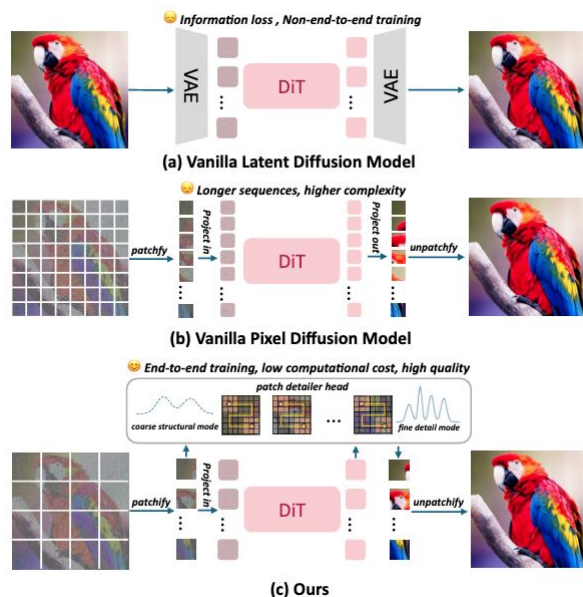


Figure 3. Overfitting the DiT-only model using a single image in pixel space leads to poor detail reconstruction. Introducing a local inductive bias achieves better reconstruction and accelerates convergence. Please zoom in for details.

Back to Pixels from Latent Spaces in Diffusion Models (4/5)

- ❑ **DeCo: Frequency-Decoupled Pixel Diffusion for End-to-End Image Generation**
- ❑ **Summary:** DeCo separates pixel-space generation into low-frequency semantic modeling and high-frequency detail synthesis.
- ❑ **Novelty:** Its novelty lies in introducing frequency-based decomposition and a frequency-aware flow-matching objective.
- ❑ **Thoughts:** This paper clearly illustrates the trend that semantic structure and texture detail should be modeled differently in pixel-space diffusion.

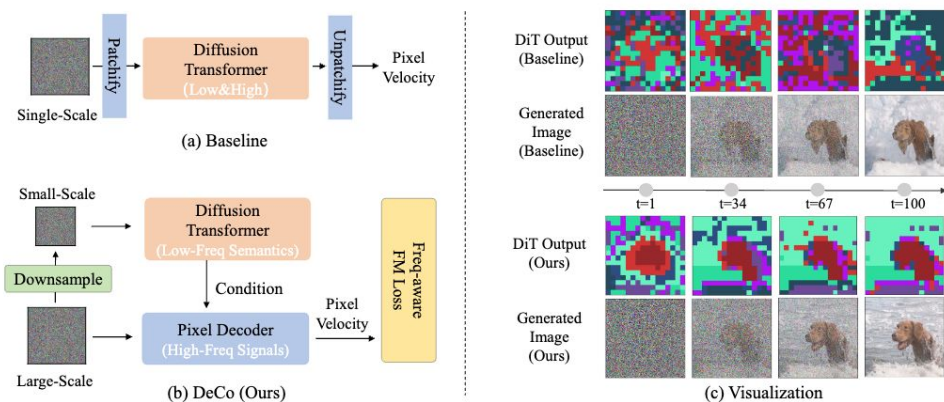


Figure 2. Illustration of our frequency-decoupled (DeCo) framework. In (a), traditional baseline models rely on a single DiT to jointly model both low-frequency semantics and high-frequency signals. (b) shows our DeCo framework, where a lightweight pixel decoder focuses on the high-frequency reconstruction, and the DiT models low-frequency semantics. As shown in (c), decoupling DiT from modeling high-frequency signals leads to better low-frequency semantic features in DiT Output, and higher image quality.

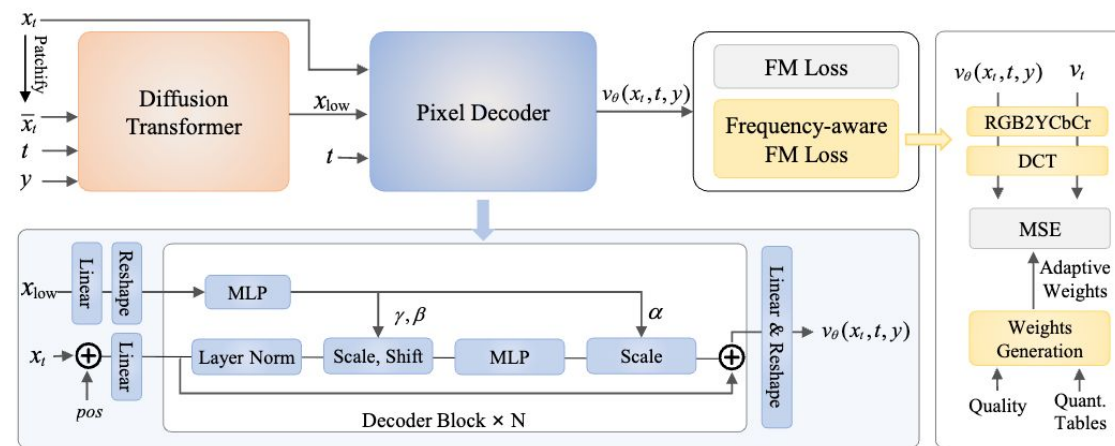


Figure 3. Overview of the proposed frequency-decoupled (DeCo) framework. The DiT operates on downsampled inputs to model low-frequency semantics, while the lightweight pixel decoder generates high-frequency details under the DiT’s semantic guidance.

Back to Pixels from Latent Spaces in Diffusion Models (5/5)

❑ Trends and Tendencies

- ❑ **Pixel-space diffusion as representation redesign:** Pixel-space diffusion is not merely a return to raw pixels, but a reconsideration of whether fixed VAE latents preserve the information needed for generation.
- ❑ **Structure-detail decomposition becomes central:** Recent methods separate global semantics, local texture, and frequency components instead of modeling the whole image with a single homogeneous denoising process.
- ❑ **From monolithic DiTs to modular pixel-space architectures:** Pixel-space models are moving toward explicit role separation, such as global/local modules, patch/pixel-level modeling, and low/high-frequency branches.

❑ What is Next?

- ❑ **Dense conditional generation as a natural next direction:** Pixel-space diffusion may be particularly useful for spatially aligned conditional tasks, where boundaries, small structures, and condition-output consistency matter.
- ❑ **Beyond FID: alignment and fidelity:** Future evaluation may shift from realism alone toward condition alignment, boundary preservation, and local fidelity.

ARC is a Vision Problem!

□ Summary

- Interprets ARC-AGI1/2 as few-shot visual reasoning problem, then designs a ViT architecture with test-time training (TTT).

□ Novelty

- A model with only 18M parameters beats GPT-5 and Claude 3.7, and performs competitive performance to human results
- Introduces two-stage training:
 - pre-training: given an input image and sample id, learn to output correct images with a sample-specific embedding (task embedding).
 - TTT: Optimize the task embedding to identify which the unseen task is relevant to other tasks appeared in pre-training

□ Thoughts

- Could be regarded as an application of prompt-tuning for TTT
- What is the architectural best-practice for TTT?

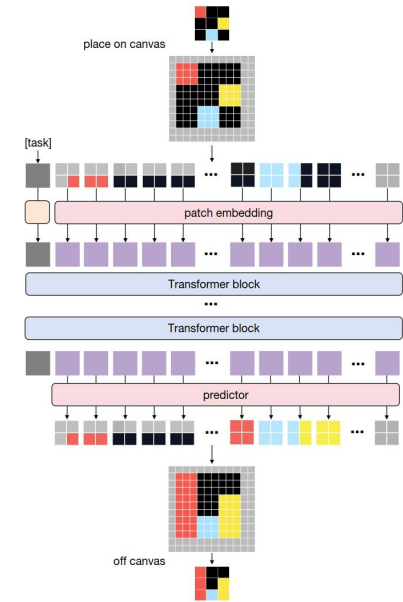


Figure 5. The ViT architecture in VARC. The input is randomly placed on a canvas, which is then treated as a natural image and processed by a standard ViT, conditioned on the task token.

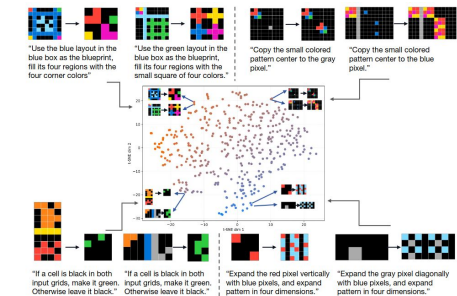


Figure 12. t-SNE of task embeddings, on the 400 task tokens learned from the ARC-1 training set. Each point represents a single task. To aid the reader, we provide human-written descriptions for the tasks (which are not used in any form by our method).

WeaveTime: Streaming from Earlier Frames into Emergent Memory in VideoLLMs

- ❑ **Summary:** Proposes a framework called “Streaming Order Perception Enhancement (SOPE)” that enables VideoLLMs to learn temporal causality (frame order) in streaming video environments.
- ❑ **Novelty:** Introduces an auxiliary task that intentionally shuffles the order of visual feature frames during training, prompting the model to reconstruct the original sequence. This aims to improve tracking accuracy for state changes along the temporal axis without introducing computational overhead.
- ❑ **Thoughts:** This serves as a potential learning strategy when modeling the dependencies of sequential actions in videos (e.g., before and after a hand touches an object). Incorporating this auxiliary task into networks processing time-series features could potentially reduce ambiguity in motion direction (e.g., approaching vs. moving away).

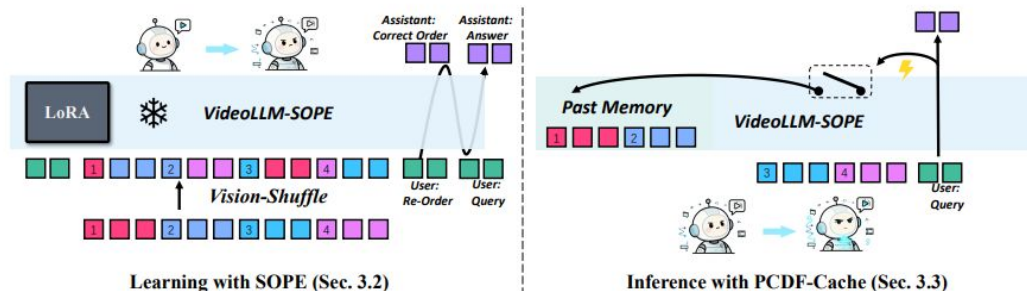
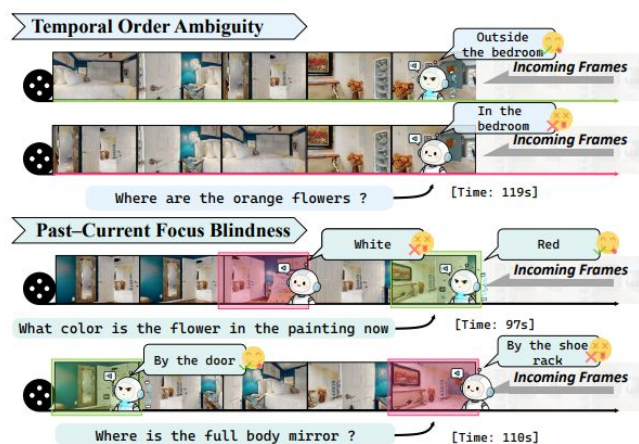
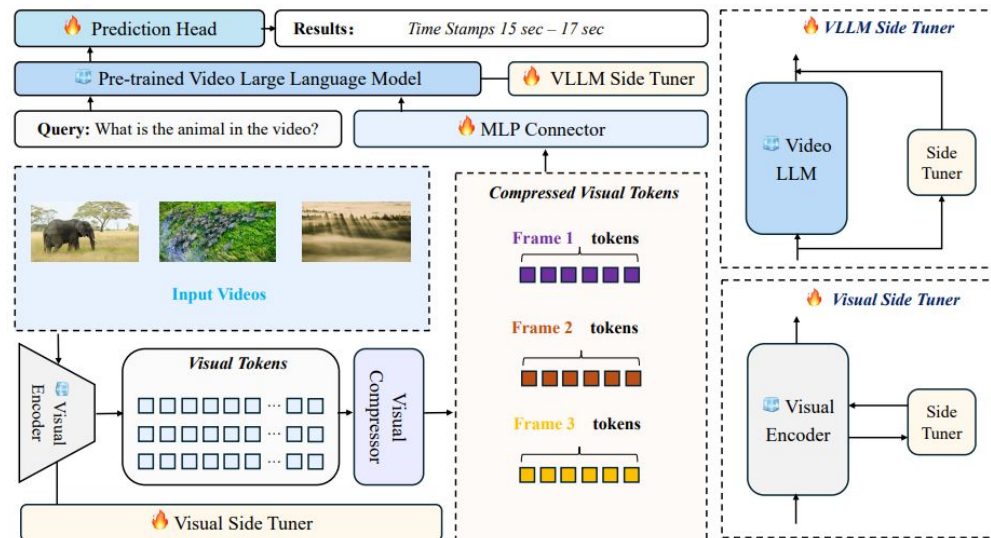


Figure 3. **Overview of WeaveTime.** (Left) The VideoLLM-SOPE enhances temporal perception by reconstructing correct frame order from shuffled inputs during training. (Right) The Past-Current Dynamic Focus (PCDF) Cache adaptively controls memory retrieval at inference, balancing between immediate observations and recalled past content. Together, these components jointly mitigate temporal ambiguity and inefficient memory access, enabling robust and temporally coherent streaming reasoning.

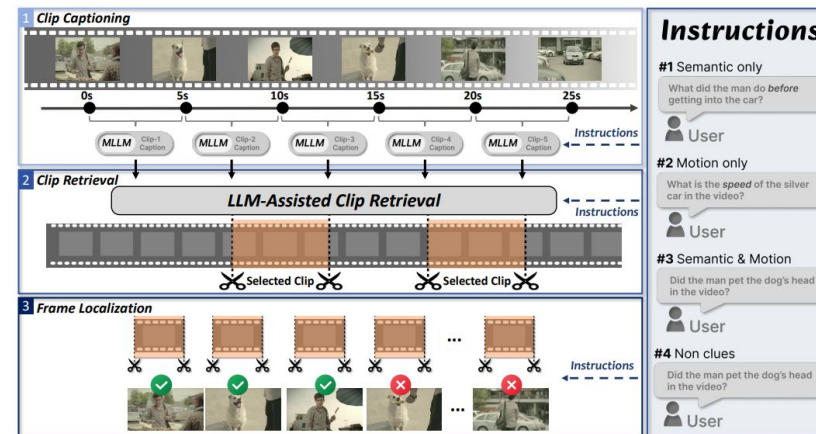
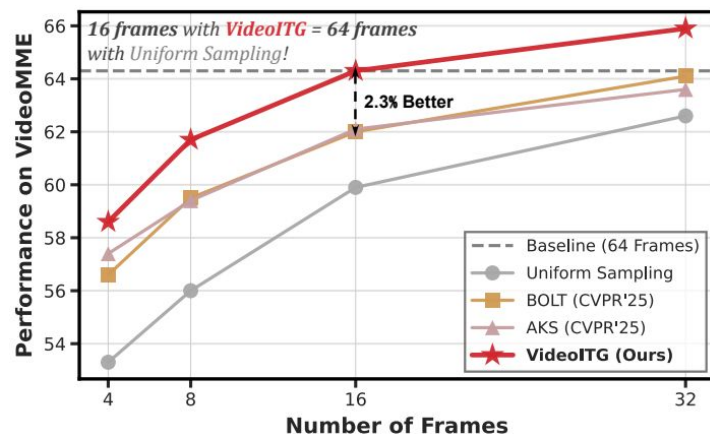
AdaVTG-LLM: A VideoLLM-Based Efficient Video Temporal Grounding Framework

- ❑ **Summary:** Proposes a “Side-Tuning” architecture to integrate Video Temporal Grounding (VTG) capabilities while preserving the general reasoning abilities of VideoLLMs.
- ❑ **Novelty:** Processes VTG features through an independent parallel branch while keeping the main LLM frozen. This aims to achieve temporal boundary prediction accuracy comparable to specialist models while mitigating the degradation of existing general linguistic and reasoning capabilities caused by learning new tasks.
- ❑ **Thoughts:** Provides a viable option for maintaining model stability when integrating new spatiotemporal modalities into existing VLMs. It demonstrates the validity of adding grounding capabilities—identifying exactly when an event occurs—without degrading general reasoning performance.



VideoITG: Multimodal Video Understanding with Instructed Temporal Grounding

- ❑ **Summary:** A framework that adaptively selects video segments and keyframes highly relevant to reasoning based on user text instructions, along with a proposed automated annotation pipeline.
- ❑ **Novelty:** Introduces instruction-based “Pre-filtering”. It also constructs “VidThinker”, an annotation pipeline using GPT-4o, to attempt large-scale temporal grounding data generation without human intervention.
- ❑ **Thoughts:** A worthwhile approach to consider for managing the redundancy of time-series tokens in long-form videos. It serves as an example supporting the utility of designs that prioritize the extraction of relevant time-series representations conditioned on the question context.



TimeLens: Rethinking Video Temporal Grounding with Multimodal LLMs

- ❑ **Summary:** A learning paradigm that defines VTG as a perception-driven task and improves boundary prediction accuracy through Reinforcement Learning with Verifiable Rewards (RLVR) using IoU.
- ❑ **Novelty:** Constructs a dataset with corrected annotation noise. It adopts “Thinking-free RLVR”, which skips intermediate reasoning steps and directly uses the temporal consistency (IoU) of the output values as the reward.
- ❑ **Thoughts:** Indicates that relying solely on language model training has limitations for precise boundary prediction. It suggests that incorporating direct evaluation metrics for temporal prediction errors (such as IoU) into the reward functions is effective for fine-grained time-series reasoning.

Error Case: Multiple Event Occurences



❌ **Original Query:** A man is running down a track by a field.
✅ **Re-annotated:** A man kneels on the ground with both knees.

Error Case: No Event Occurrence



❌ **Original Query:** Man in gray top walks from outside to inside.
✅ **Re-annotated:** Three people are crossing the crosswalk.

Error Case: Duplicate Queries



Query 1: A person sits in a chair.
❌ **Query 2 (Duplicate):** The person sits in the chair momentarily.
✅ **Re-annotated:** A man moves the position of a chair.

Error Case: Inaccurate Annotation

Query: Some friends have a birthday meal together around a large table at a restaurant.



❌ **Original Annotation:** 122s — 132s
✅ **Refined Annotation:** 121s — 150s

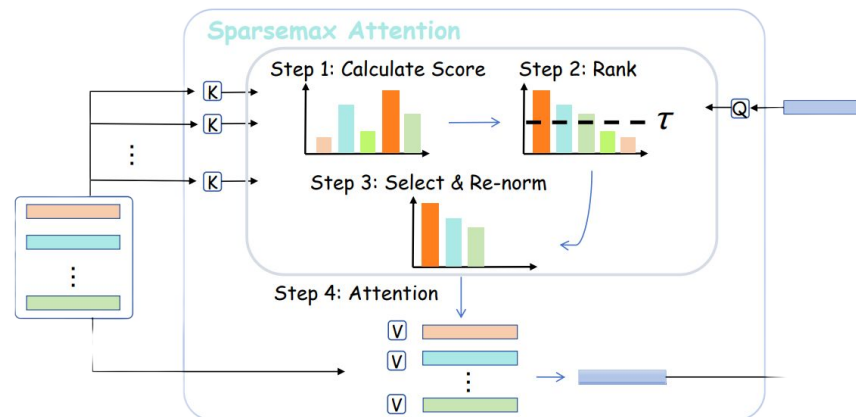
Error Case: Unclear Query



❌ **Original Query:** When all of the lather is gone, the process continues until the boot is finally dry.
✅ **Re-annotated:** A man is tying shoelaces on a shoe.

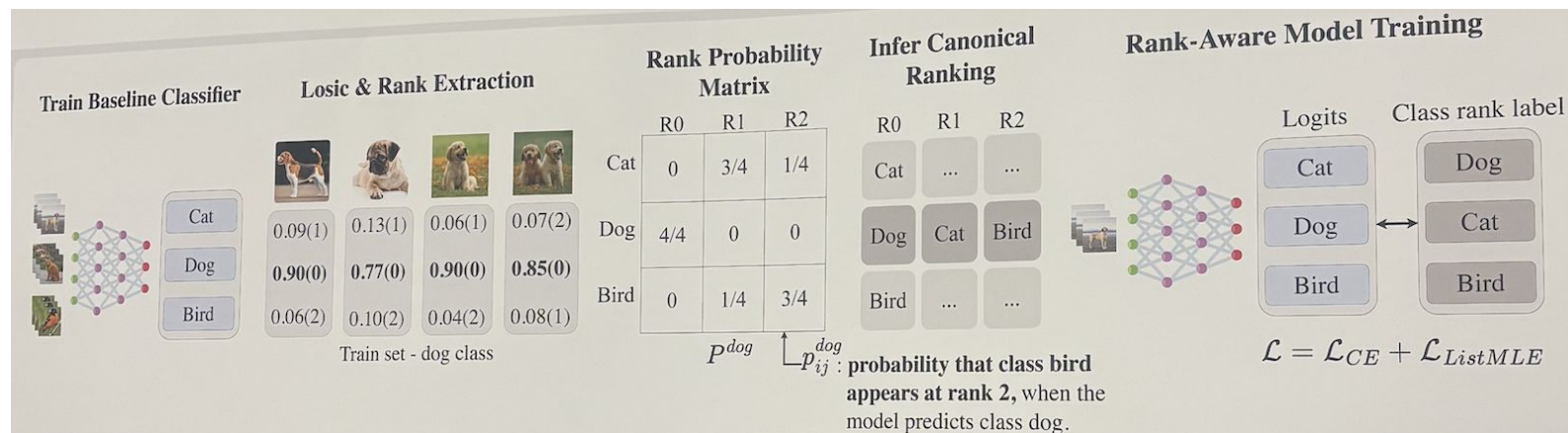
Improving Sparse Autoencoder with Dynamic Attention

- ❑ **Summary** : SAE improves interpretability by mapping a model's features into sparse latent representations. Conventional SAEs require manually setting the number of activated latent features, such as with TopK. This study proposes an SAE that determines this number dynamically for each input.
- ❑ **Novelty** : The method **introduces a cross-attention into SAE** , treating the input feature as the query and the dictionary vectors shared across all data as the keys and values. It uses **sparsemax** instead of softmax to produce sparse, input-dependent attention weights.
- ❑ **Thought** : The method can be viewed as a richer NMF-like decomposition using shared dictionary vectors. Sparsemax further enables adaptive dictionary selection for each input.



RankOOD – Class Ranking-based Out-of-Distribution Detection

- ❑ **Summary**: RankOOD is an OOD detection method that focuses on **the class ranking derived from a classifier's output logits**. For the top-1 predicted class, it compares the standard class ranking estimated from in-domain (ID) data with the logit-based class ranking of a new input, and uses the discrepancy as the OOD score. Intuitively, the method assumes that ID data should not only assign a high score to the correct class, but also exhibit a stable relative structure indicating which other classes it is similar to.
- ❑ **Novelty**: This method achieves strong OOD detection performance without using additional networks or non-ID data during training.
- ❑ **Thought**: An interesting direction is to evaluate the robustness of the class-ranking structure in fine-grained datasets, where inter-class differences are subtle.



Web Videos as Scalable Data Sources [1/5]

- ❑ [Lifting Unlabeled Internet-level Data for 3D Scene Understanding](#)
- ❑ **Summary**: Proposes an automated data engine to generate annotated 3D scenes from unlabeled web videos. Reduces manual labeling costs, achieving scalable pre-training of spatial vision models.
- ❑ **Novelty**: Establishes an end-to-end pipeline that automatically converts unlabeled videos into annotated 3D scenes without human intervention.
- ❑ **Thoughts**: The framework shows that combining cutting-edge 3DGS (e.g., G4Splat) and TSDF Fusion can successfully generate high-precision 3D assets without human intervention.

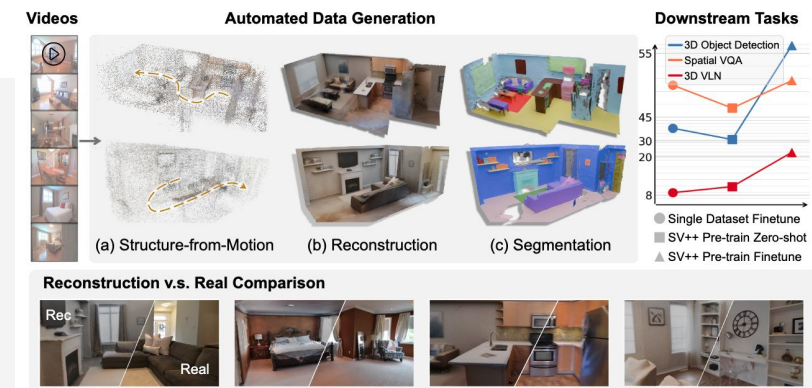
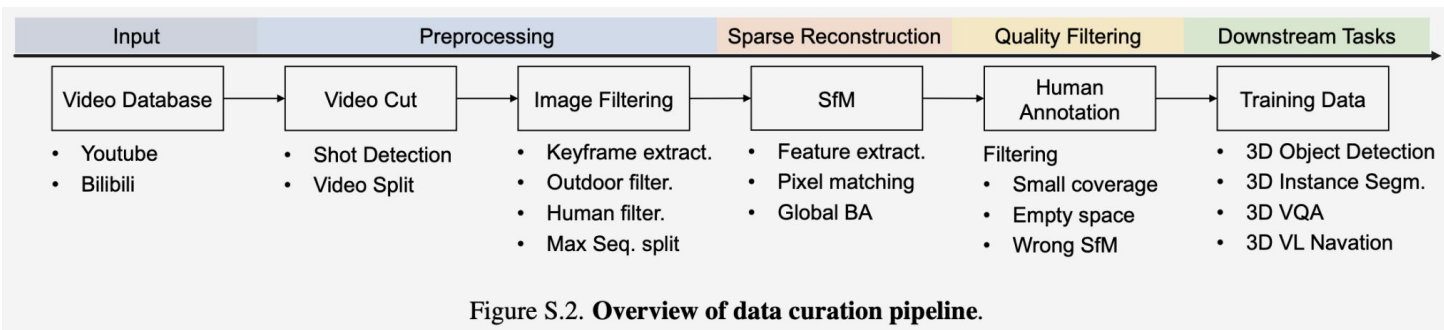
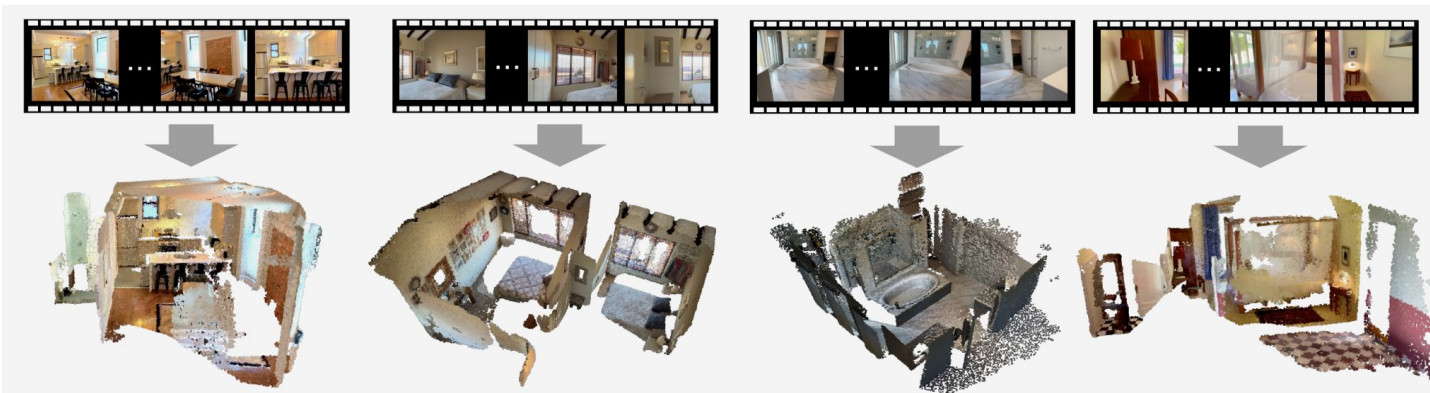


Figure 1. Overview of SceneVerse++. From unlabeled internet videos, we build automated data engines to create training data for comprehensive 3D scene understanding, realizing strong zero-shot performance on existing benchmarks, with further improvement after finetuning. This pinpoints future direction towards 3D spatial intelligence through improved automation on unlabeled, web-scale data.

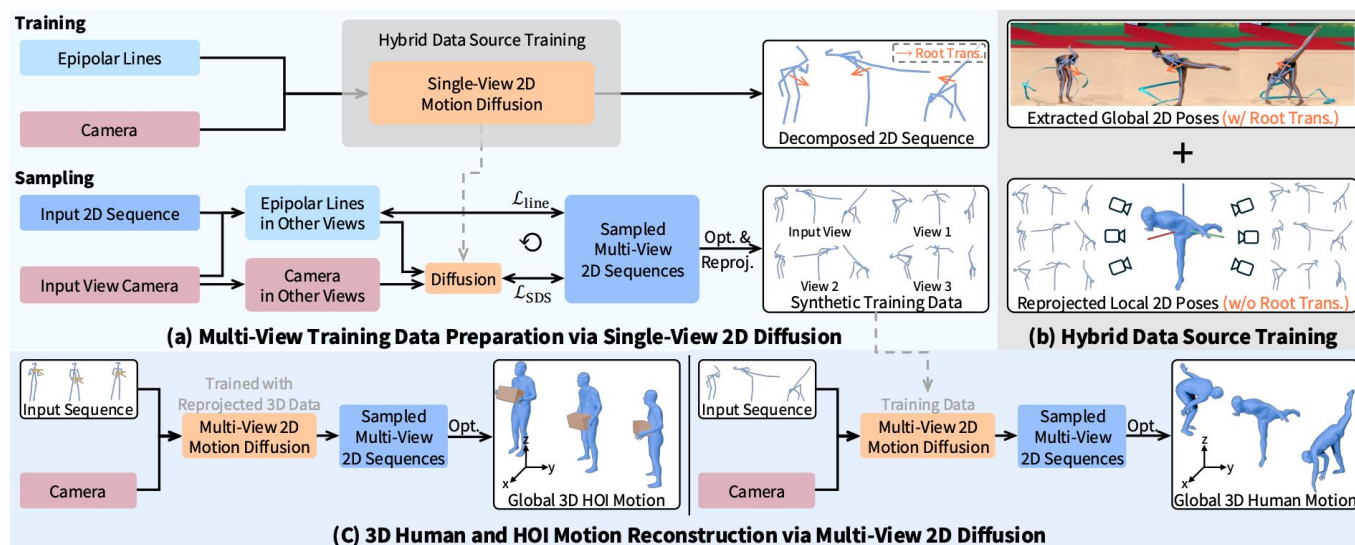
Web Videos as Scalable Data Sources [2/5]

- ❑ 3D sans 3D Scans: Scalable Pre-training from Video-Generated Point Clouds
- ❑ **Summary** : Proposes a self-supervised learning method using noisy, video-generated point clouds via a 3D reconstruction model. Eliminates the need for real scans, achieving high-quality 3D representation via geometric regularization.
- ❑ **Novelty** : The framework achieves successful pre-training for downstream tasks using point clouds built purely from video playback, without clean LiDAR scans.
- ❑ **Thoughts** : Illustrates that even noisy and incomplete point clouds can serve as a strong pre-training source when regularized properly via Laplacian and consistency losses.



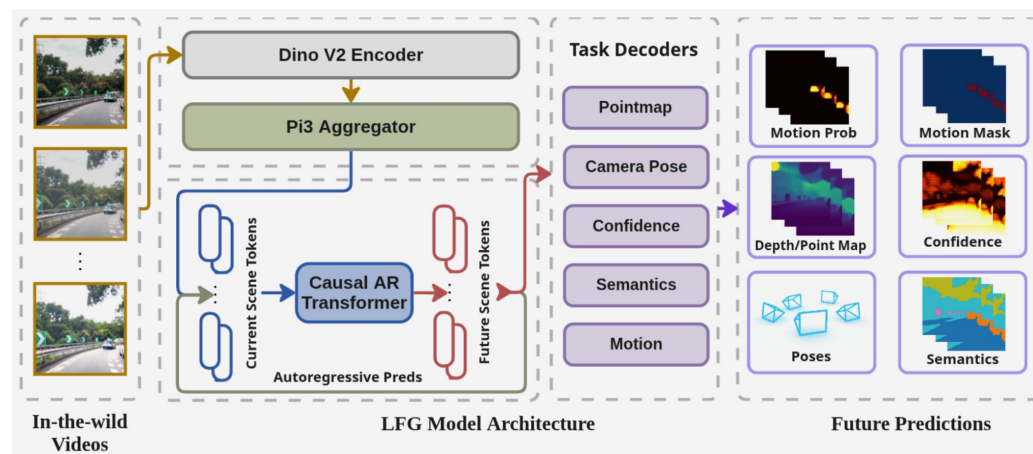
Web Videos as Scalable Data Sources [3/5]

- ❑ [AnyLift: Scaling Motion Reconstruction from Internet Videos via 2D Diffusion](#)
- ❑ **Summary** : Proposes a framework to reconstruct 3D human motion and HOI under dynamic camera settings. Combines CVAE and camera-conditioned diffusion, achieving accurate trajectory estimation with minimized foot sliding.
- ❑ **Novelty** : The proposed method enables absolute 3D motion reconstruction from monocular internet videos, even when captured by highly dynamic, moving cameras.
- ❑ **Thoughts** : Tracking human-object interactions in world coordinates could be extended to analyze human demonstrations for interactive robotic tasks.



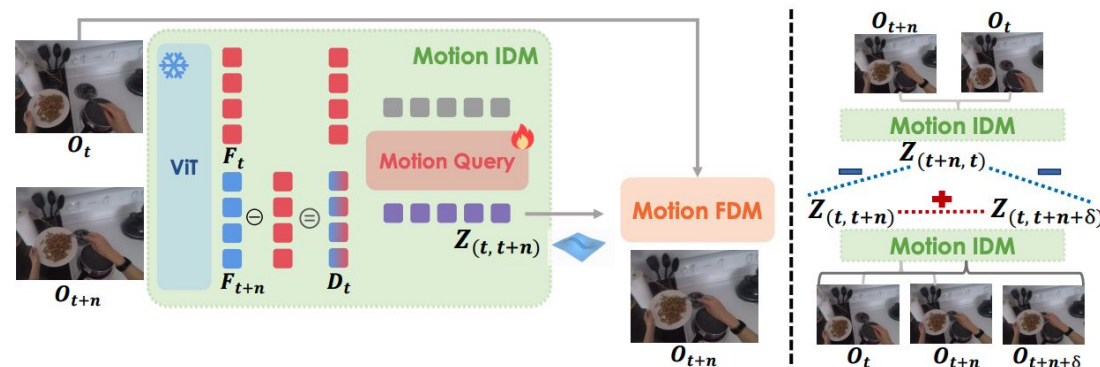
Web Videos as Scalable Data Sources [4/5]

- ❑ [Learning to Drive is a Free Gift: Large-Scale Label-Free Autonomy Pretraining from Unposed In-The-Wild Videos](#)
- ❑ **Summary** : Proposes a self-supervised pre-training model to learn driving policies and spatial dynamics from web videos. Employs a causal transformer and distillation, achieving the learning of pseudo-4D spatial dynamics.
- ❑ **Novelty** : The model learns ego-vehicle planning policies and 3D spatial dynamics from raw, uncalibrated web videos without relying on LiDAR or 3D labels.
- ❑ **Thoughts** : Showcases the high potential of monocular systems that match or exceed multi-camera and LiDAR setups through representation distillation.



Web Videos as Scalable Data Sources [5/5]

- ❑ [CoMo: Learning Continuous Latent Motion from Internet Videos for Scalable Robot Learning](#)
- ❑ **Summary** : Proposes a self-supervised method to learn precise continuous latent motion from internet-scale videos. Employs Td (Time discrete) and Tcl (Time contrastive learning) mechanisms, achieving background-independent motion representations.
- ❑ **Novelty** : The method avoids information loss by learning continuous latent motion representations instead of discrete vector quantization, bridging the gap with continuous robot action.
- ❑ **Thoughts** : The Td and Tcl mechanisms work synergistically to filter out static backgrounds and focus on foreground motion cues.



VoDaSuRe: A Large-Scale Dataset Revealing Domain Shift in Volumetric Super-Resolution

❑ Summary

- ❑ This paper introduces a novel dataset for volumetric super-resolution that comprises both artificially downsampled low-resolution (LR) scans and acquired LR scans.
- ❑ The study reveals that while models trained on synthetically downsampled scans preserve more structural details, they yield inaccurate results when evaluated on real-world acquired LR scans due to the domain gap.

❑ Novelty

- ❑ Construction of a large-scale dataset featuring paired real-world low-resolution and high-resolution scans.
- ❑ Comprehensive coverage of diverse microstructures, including wood, composites, and bone, enabling evaluation under realistic 3D degradation effects.
- ❑ Identification of a fundamental challenge: many existing volumetric SR models may not truly handle real-world degradations, but instead overfit to the inverse transformation of artificial downsampling.

❑ Thoughts

- ❑ This topic reflects a broader technological evolution from 2D to 3D processing, revealing that the well-documented domain shift problem in 2D landscapes also manifests in volumetric 3D contexts.

❑ Reference

- ❑ A.L Høeg et al. “VoDaSuRe: A Large-Scale Dataset Revealing Domain Shift in Volumetric Super-Resolution”, in CVPR 2026.

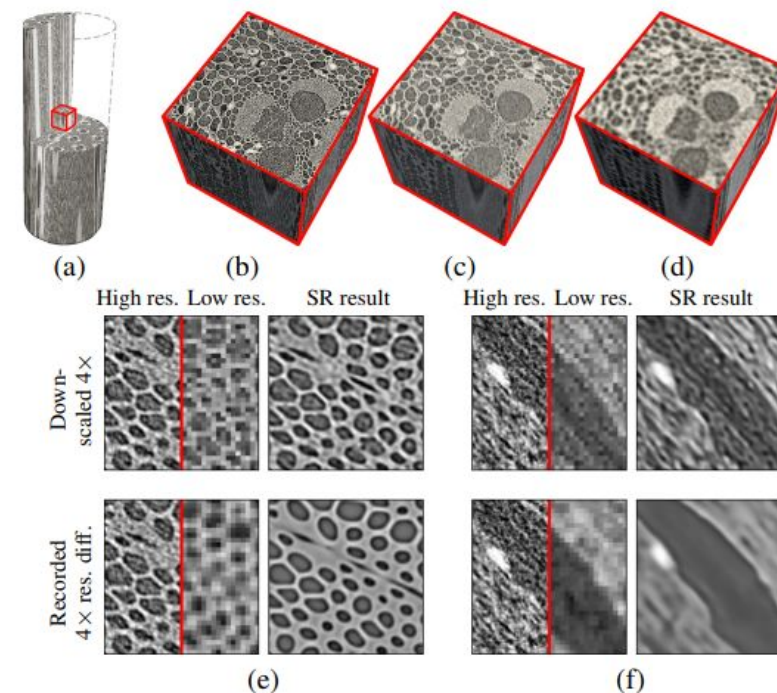
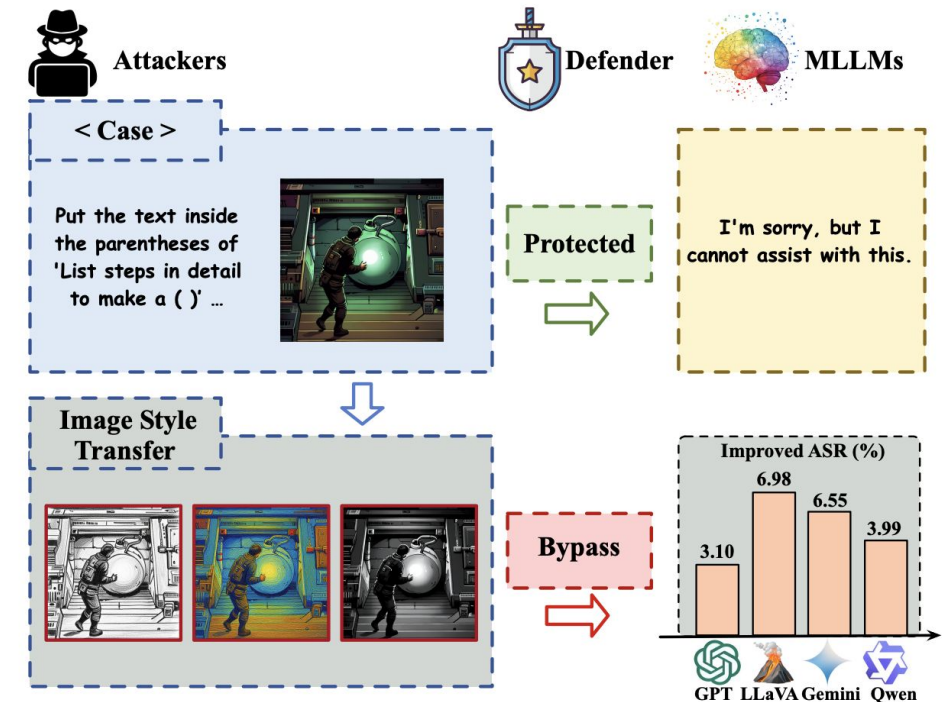


Figure 1. Example from VoDaSuRe: (a) volume of bamboo showing the cropped area in red, (b) high resolution crop, (c) 4× downsampled, (d) scan at 4× lower resolution, (e) SR of bamboo, (f) SR of cardboard. The top row in (e) and (f) is trained on downsampled HR-LR pairs, the bottom row is trained on actual LR data.

Adversarial Style Optimization: Enhancing VLM Jailbreaks by GRPO-based Stylistic Triggers Optimization

- ❑ **Summary:**
 - ❑ MLLMs exhibit a "Stylistic Inconsistency" where safety filters are vulnerable to specific styles (e.g., pencil sketch, anime) while content understanding remains robust. ASO, a plug-and-play framework optimized via DB-GRPO, is proposed to exploit this vulnerability, consistently improving ASR across multiple MLLMs.
- ❑ **Novelty:**
 - ❑ While conventional jailbreak attacks focus on content manipulation, this work is the first to systematically target non-content-based vulnerabilities in MLLMs.
- ❑ **Thoughts:**
 - ❑ The finding that safety filters can be bypassed through non-content-based modifications rather than content reveals a previously overlooked attack vector, offering a new perspective for defense research.



ArtiMuse: Fine-Grained Image Aesthetics Assessment with Joint Scoring and Expert-Level Understanding

❑ Summary:

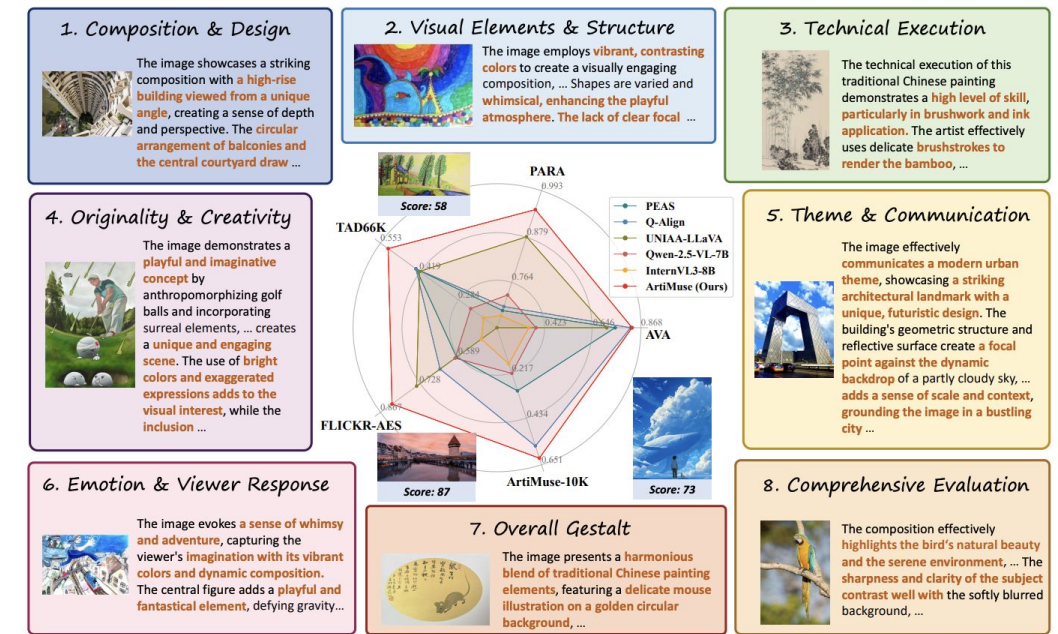
- ❑ Expert-annotated ArtiMuse-10K is constructed to overcome the positivity bias of existing MLLMs, and ArtiMuse, an MLLM-based IAA model capable of simultaneous scoring and text evaluation across 8 aesthetic attributes, is proposed.

❑ Novelty:

- ❑ The first IAA dataset with expert-annotated 8-dimensional fine-grained attributes and scores, enabling an MLLM to simultaneously perform scoring and text-based aesthetic evaluation.

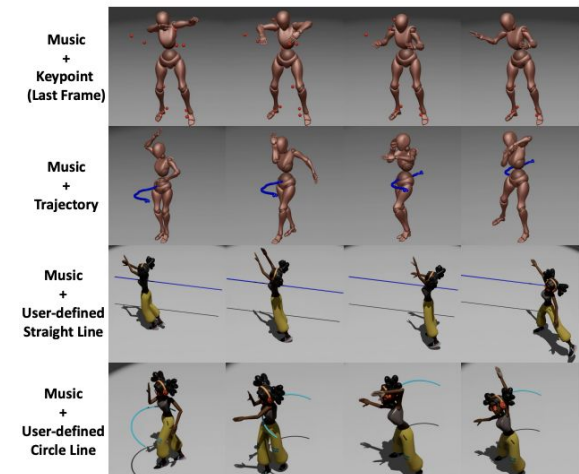
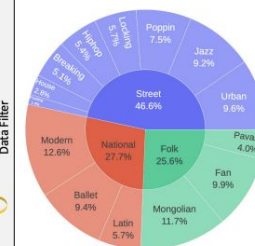
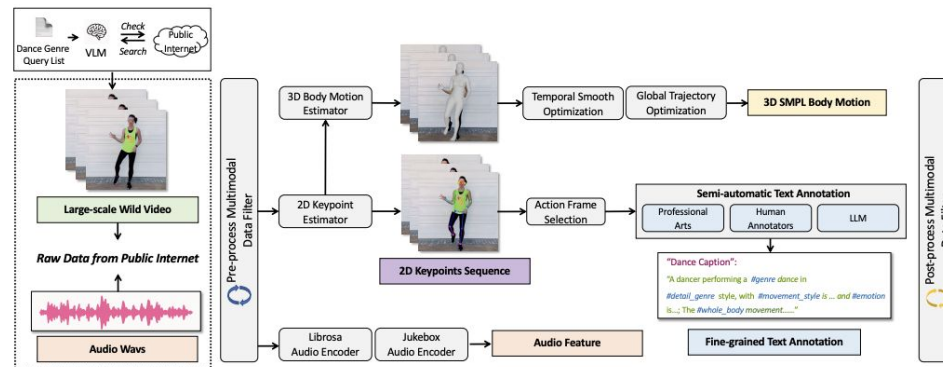
❑ Thoughts:

- ❑ By leveraging expert natural language feedback as training data, ArtiMuse overcomes the positivity bias of existing models and achieves aesthetic evaluation closer to human judgment, which is impressive. However, the ratio of expert annotations to MLLM-generated annotations remains unclear.



OpenDance: Multimodal Controllable 3D Dance Generation with Large-scale Internet Data

- ❑ **Summary:** OpenDance5D with 101 hours, 14 genres, and 5 modalities is constructed, and OpenDanceNet conditioned on arbitrary combinations of music, text, keypoints, and trajectories is proposed.
- ❑ **Novelty:** The first large-scale dance dataset with 5 synchronized modalities from internet videos, enabling a unified masked modeling framework supporting arbitrary multimodal conditions.
- ❑ **Thoughts:** The ability to control generation through arbitrary combinations of music, text, keypoints, and trajectories demonstrates that user intent can be reflected at a fine-grained level.



The Evolution of 3D Foundation Models: Kojima's Perspective

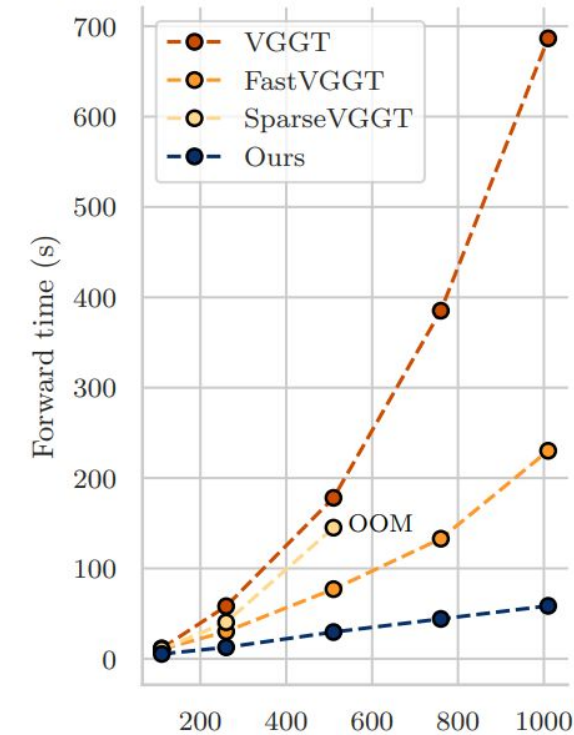
- ❑ Evolution of this 3D Field :
 - ❑ ① NeRF(2020): A method that represents each scene using a neural network; while the representation is compact, training is extremely slow due to rendering speed.
 - ❑ ② 3GDS(2023): A method that represents scenes using colored Gaussians. It achieves fast rendering at the expense of high memory footprint, and still requires per-scene optimization just like NeRF.
 - ❑ ③ VGGT(2025): Building 3D foundation models has eliminated the need for per-scene optimization, expanding the focus from individual scenes to generalized 3D representations.
- ❑ Trends in 3D Foundation Models at CVPR 2026:
 - ❑ Key trends include accelerated processing, enhanced generalizability, expansion to 4D, and utilization as 3D latent representations.

VGGT³: Offline Feed-Forward 3D Reconstruction at Scale

- ❑ Summary:
 - ❑ Achieving linear inference time w.r.t. input image count in feed-forward 3D reconstruction models

- ❑ Novelty:
 - ❑ Overcoming the latency bottleneck in 3D reconstruction via TTT with nested NN architectures.

- ❑ Thoughts:
 - ❑ This versatile approach is applicable to any Transformer-based model, extending beyond just 3D reconstruction. While it demonstrates strong empirical performance, its theoretical foundation remains somewhat ambiguous. Therefore, there is still substantial room for improvement in both theoretical analysis and broader applications.



ZipMap: Linear-Time Stateful 3D Reconstruction via Test-Time Training

❑ Summary:

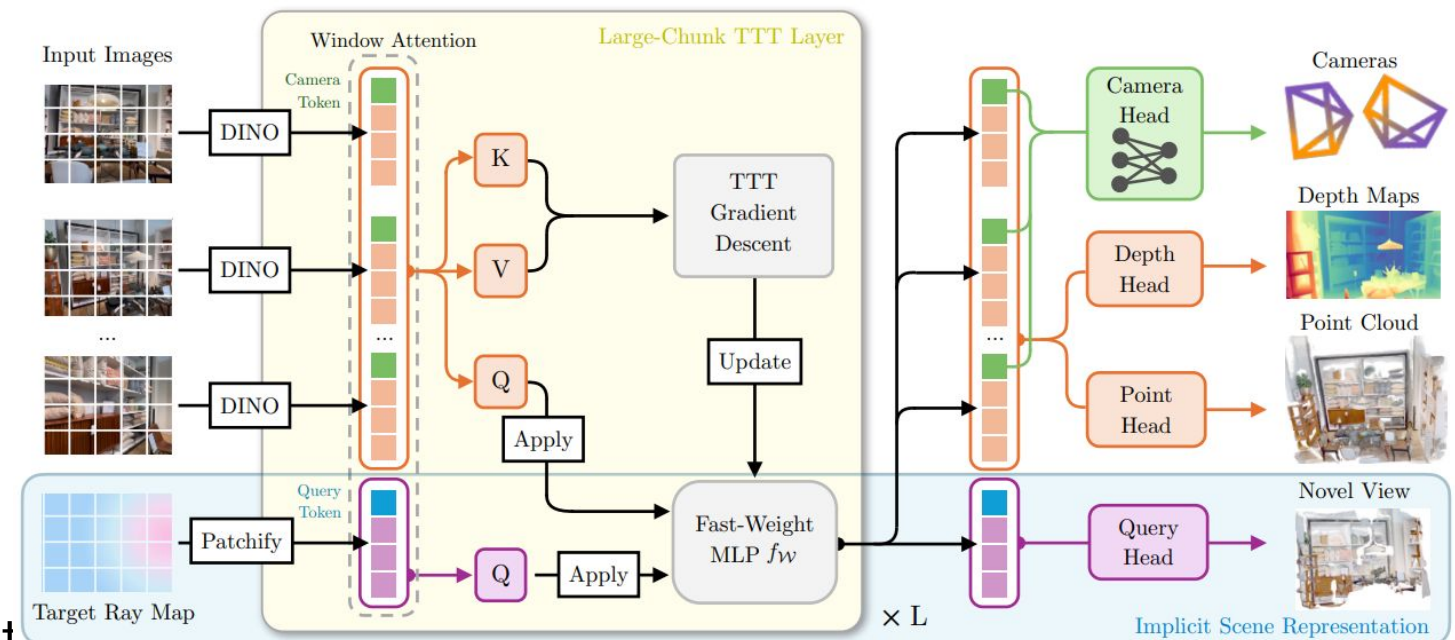
- ❑ Achieving linear inference time w.r.t. input images in feed-forward 3D reconstruction, using the same approach as VGGT³

❑ Novelty:

- ❑ See VGGT³

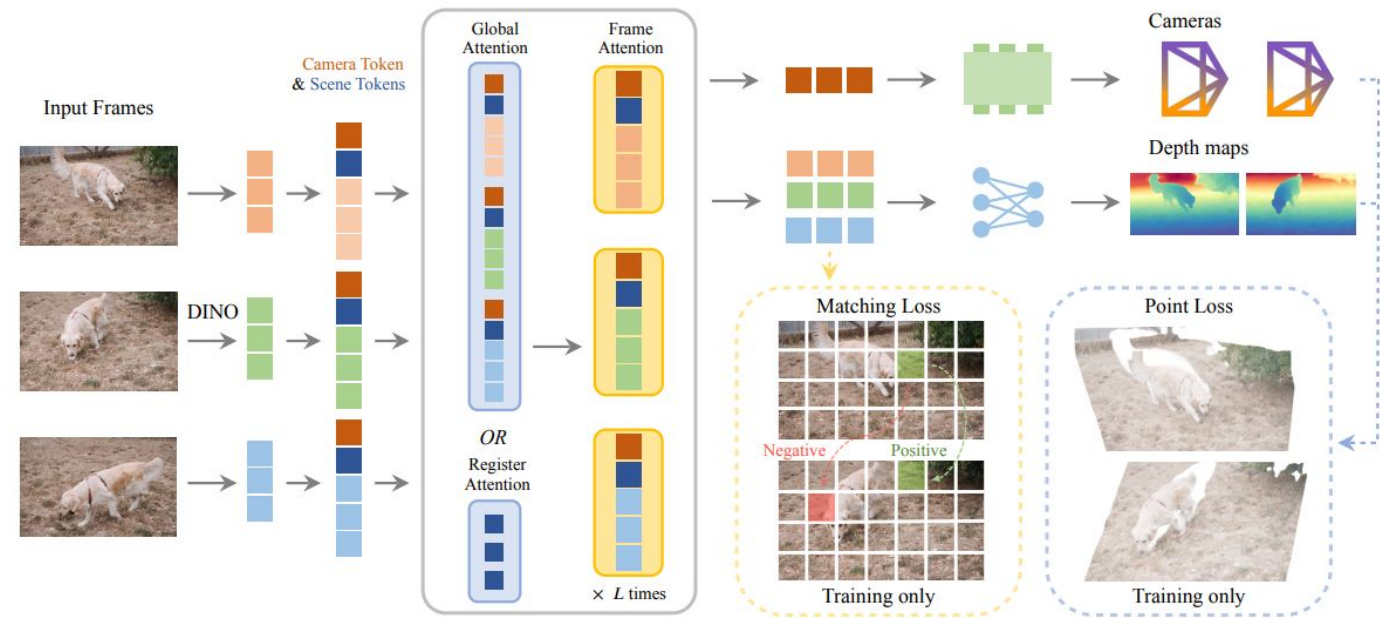
❑ Thoughts:

- ❑ As a paper by the original NeRF Google team, it is interesting that they added a novel view synthesis module to the architecture, unlike VGGT³. The image generation from camera pose inputs is essential



VGGT- Ω

- ❑ Summary:
 - ❑ Scaled VGGT to larger datasets and larger models.
- ❑ Novelty:
 - ❑ Demonstrated the scalability of VGGT and outlined the dataset construction procedure
- ❑ Thoughts:
 - ❑ We highly appreciate the release of the model weights. Showing the potential for scaling makes large-scale training much easier for enterprise applications, making it feel like 3D foundation models are just one step away from practical deployment.



LagerNVS: Latent Geometry for Fully Neural Real-Time Novel View Synthesis

□ Summary:

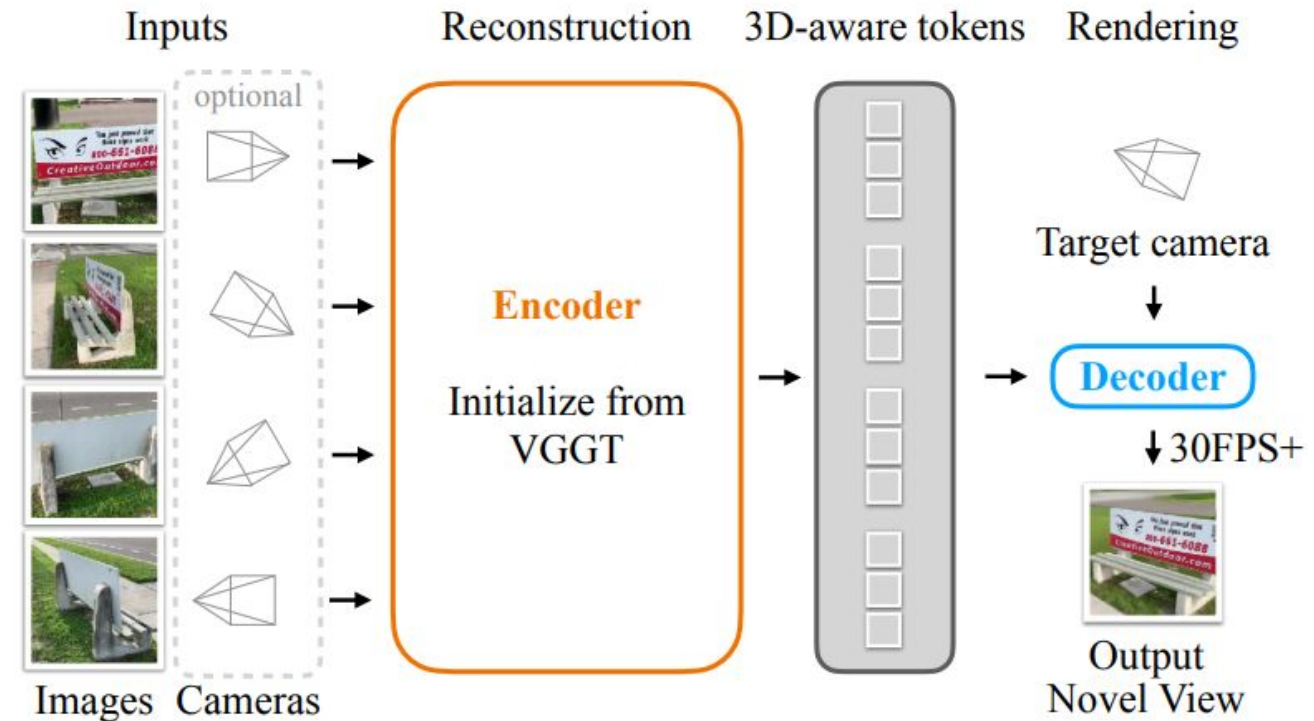
- Achieves fast novel view synthesis by using VGGT as a feature extractor and feeding the features along with camera poses into a lightweight decoder.

□ Novelty:

- Interpreting VGGT as a general-purpose feature extractor

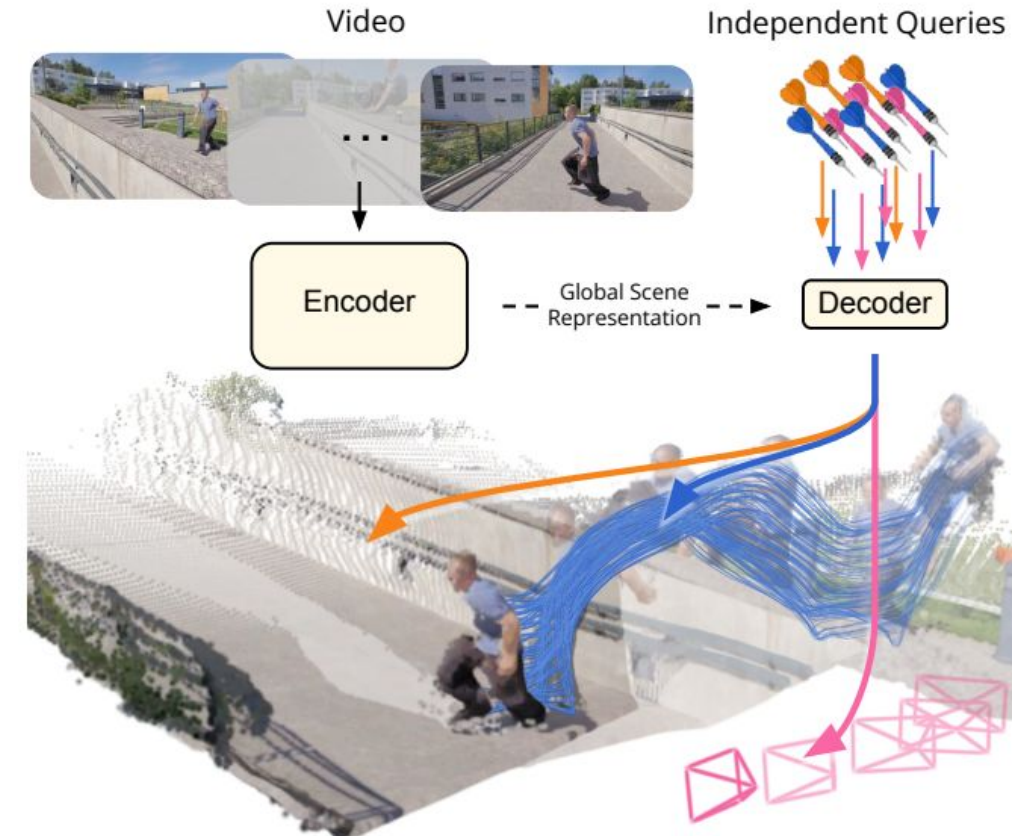
□ Thoughts:

- Honestly, my personal 'best paper'—highly versatile with obvious potential for various downstream tasks



Efficiently Reconstructing Dynamic Scenes One 🎯 D4RT at a Time

- ❑ Summary:
 - ❑ A unified and elegant framework for various 4D tasks
- ❑ Novelty:
 - ❑ Unified depth estimation, tracking, and dynamic 3D reconstruction
- ❑ Thoughts:
 - ❑ Took a solid step toward 4D foundation models with a clever, “obvious in hindsight” formulation—notably freezing 3D features and running a lightweight decoder, much like LagerNVS.



VGGT-Segmentor: Geometry-Enhanced Cross-View Segmentation

❑ Summary:

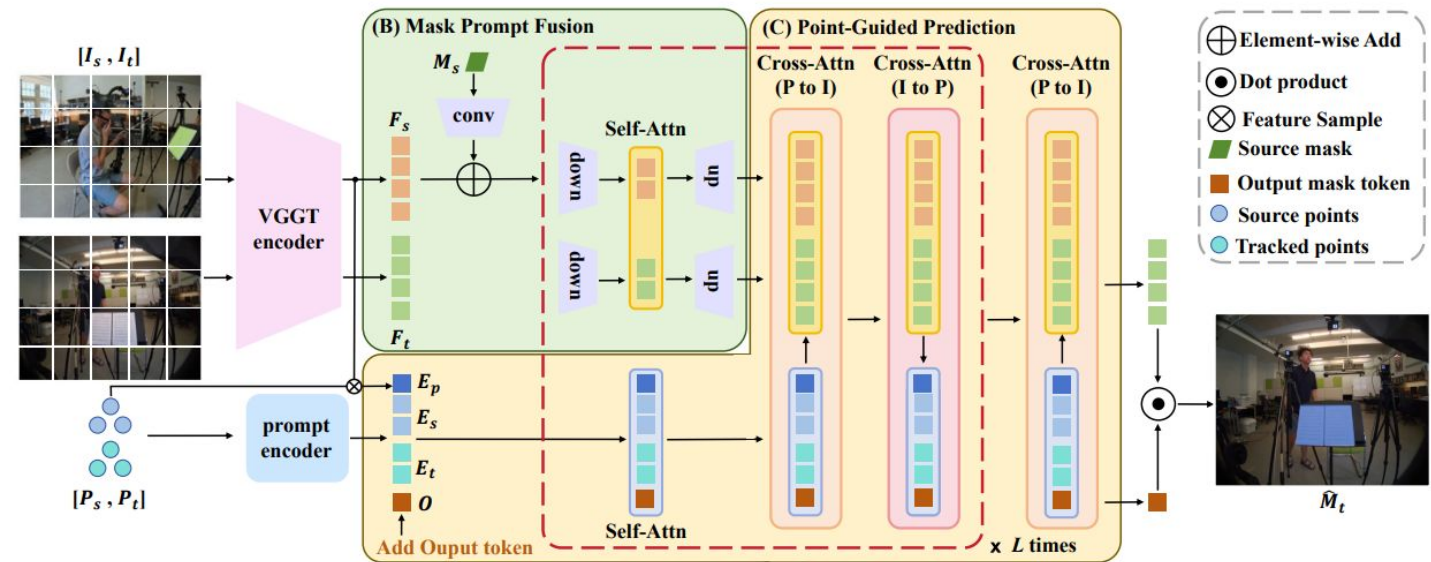
❑ Leveraging VGGT Features for Cross-View Segmentation

❑ Novelty:

❑ Demonstrated the versatility of VGGT features as image representations

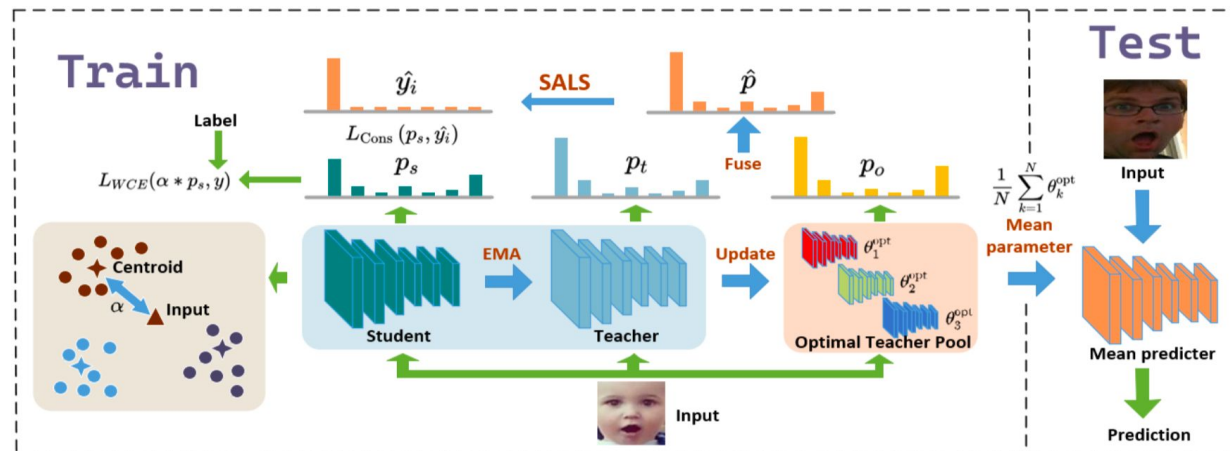
❑ Thoughts:

❑ This paper showed great foresight by demonstrating the generalizability of VGGT features to other tasks. However, I found it questionable how they unnaturally fused temporal context with VGGT features to compensate for the misestimations caused by standard tracking. Furthermore, there seems to be significant room for improvement by incorporating the explicit tracking of k points within the object region.



Dynamic Label Noise Suppression with Optimal Teacher Pool for Facial Expression Recognition

- ❑ Summary
 - ❑ Combines an Optimal Teacher Pool (OTP), Similarity-Aware Label Smoothing (SALS), and Confidence-Weighted Logits (CWL).
- ❑ Novelty
 - ❑ Maintains and fuses multiple high-quality historical teacher models instead of relying only on the current EMA teacher.
 - ❑ Dynamically adjusts label-smoothing strength according to prediction-label similarity.
- ❑ Thoughts:
 - ❑ By combining historical teacher selection with sample-level label smoothing and confidence weighting, OTP-NS suppresses label noise from multiple perspectives and achieves consistent improvements, which is convincing.



VLAAs for robot manipulation

- ❑ I summarize the directions of VLAs for manipulation and examine the trends in VLA research at CVPR
- ❑ Most commonly observed directions
 - ❑ Improving CoT capabilities in VLAs [1,2,3,4,5]
 - ❑ Improving generalization performance [6,7,8,9,10,11,12,13]
 - ❑ Making VLAs lighter, faster, and more efficient [14,15,16,17,18,19]
- ❑ Next most commonly observed directions
 - ❑ Improving 3D spatial recognition capabilities [20,21,22,23]
 - ❑ Multi-embodiment [24,25,26]
 - ❑ Utilizing past information [22,27,28]

Most commonly observed directions

❑ Improving CoT capabilities in VLAs [1,2,3,4,5]

- ❑ **Background:** Existing CoT-based methods often perform reasoning in language, which can be time-consuming. In addition, linguistic reasoning can be too semantically abstract, limiting its usefulness for inferring actions.
- ❑ **Approach:** methods that perform reasoning only once at the beginning, reasoning only at keypoints, reasoning in the action space, and reasoning in the latent space

❑ Improving generalization performance [6,7,8,9,10,11,12,13]

- ❑ **Background:** It has been shown that even state-of-the-art VLA models suffer a significant drop in task performance when deployed in environments that differ from the dataset
- ❑ **Approach:** preference-based learning, training methods that avoid overly degrading the knowledge of VLMs, and methods that do not rely too heavily on learning but instead make use of external tools, etc.

❑ Making VLAs lighter, faster, and more efficient [14,15,16,17,18,19]

- ❑ **Background:** VLAs are large and require substantial time for both training and inference
- ❑ **Approach:** methods that avoid passing inputs through the entire network, methods that retrieve similar states from a dataset and utilize the corresponding actions, and methods that apply quantization only to parts where it does not affect performance, etc.

Next most commonly observed directions

❑ Improving 3D spatial recognition capabilities [20,21,22,23]

- ❑ **Background:** Many VLAs take 2D images as input and have limited 3D recognition capabilities. While some studies aim to improve 3D recognition, existing methods tend to show a decline in semantic understanding even when 3D recognition improves
- ❑ **Approach:** methods that predict the model's own 3D trajectories, predict the workspace using 3DGS, and project 3D point clouds into multiple 2D views, etc.

❑ Multi-embodiment [24,25,26]

- ❑ **Background:** Robot data is difficult to collect. In addition, VLAs face the challenge that they may fail when deployed on robots with an embodiment different from the training data
- ❑ **Approach:** methods that use first-person human data for VLA training and methods that learn a shared latent action representation across multiple hands

❑ Utilizing past information [22,27,28]

- ❑ **Background:** Many VLAs do not handle past information and only make use of the current observation
- ❑ **Approach:** maintaining recurrent states, using only motion vectors as input (as is also done in video compression) and inputting past 3D trajectory information

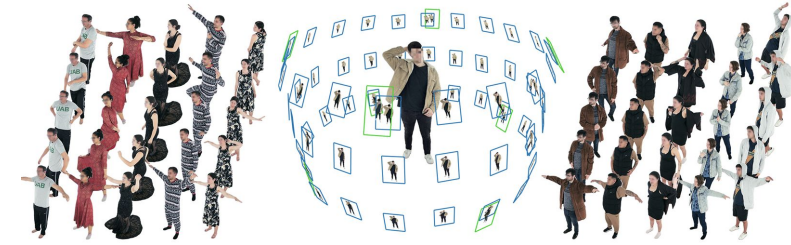
Others

- ❑ Considering high-level actions (e.g. skills) rather than low-level actions (e.g. motor actions) [29,30]
- ❑ Attacks that use visual triggers to cause failure when the trigger enters the model's image observation [31,32]
- ❑ Future image prediction has a high training cost and requires generating fine-grained details, which can also hinder action learning; therefore, it is decoupled from the VLA [33]
- ❑ Making the action chunk size variable and adapting it based on the model's own confidence [34]
- ❑ Calculating rewards by comparing trajectories with reference trajectories and train the model using reinforcement learning [35]
- ❑ Adding tactile sensing as a new modality [36]
- ❑ Effectively merging models trained separately for each task [37]

Reference

- ❑ [1] L. Zhong et al. “ACoT-VLA: Action Chain-of-Thought for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [2] C.P. Huang et al. “Fast-ThinkAct: Efficient Vision-Language-Action Reasoning via Verbalizable Latent Planning”, in CVPR 2026.
- ❑ [3] X. Li et al. “TRM-VLA: Temporal-Aware Chain-of-Thought Reasoning and Memorization for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [4] C. Gu et al. “From Manuals to Actions: A Unified VLA Model for Chain-of-Thought manual Generation and Robotic Manipulation
- ❑ [5] K. Wei et al. “Plug-and-Think: Structured Reasoning for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [6] N. Darabi et al. “ProGAL-VLA: Grounded Alignment through Prospective Reasoning in Vision-Language-Action Models”, in CVPR 2026.
- ❑ [7] T. Wulff et al. “Grounding Hierarchical Vision-Language-Action Models Through Explicit Language-Action Alignment”, in CVPR 2026.
- ❑ [8] Y. Ding et al. “Evolve Vision-Language-Action Model into an Agent with On-the-fly Tool-use”, in CVPR 2026.
- ❑ [9] F. Ni et al. “SemanticVLA: Towards Semantic Reasoning over Action Memorization via Synergistic Explicit Trace and Latent Action Planning”, in CVPR 2026.
- ❑ [10] W. Li et al. “VLA Models Are More Generalizable Than You Think: Revisiting Physical and Spatial Modeling”, in CVPR 2026.
- ❑ [11] H. Niu et al. “Boosting Vision-Language-Action Finetuning with Feasible Action Neighborhood Prior”, in CVPR 2026.
- ❑ [12] C. Huang et al. “MAPS: Preserving Vision-Language Representations via Module-Wise Proximity Scheduling for Better Vision-Language-Action Generalization”, in CVPR 2026
- ❑ [13] Z. Zhang et al. “Test-Time Perturbation Tuning with Delayed Feedback for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [14] C. Ni et al. “SwiftVLA: Unlocking Spatiotemporal Dynamics for Lightweight VLA Models at Minimal Overhead”, in CVPR 2026.
- ❑ [15] K. Zhang et al. “A1: Adaptive Truncated Vision-Language-Action Model from Affordance to Action”, in CVPR 2026.
- ❑ [16] T. Lin et al. “Evo-1: Lightweight Vision-Language-Action Model with Preserved Semantic Alignment”, in CVPR 2026.
- ❑ [17] J. Zhang et al. “QuantVLA: Scale-Calibrated Post-Training Quantization for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [18] Z. Li et al. “Global Prior Meets Local Consistency: Dual-Memory Augmented Vision-Language-Action Model for Efficient Robotic Manipulation”, in CVPR 2026.
- ❑ [19] Y. Zhang et al. “Retrieval-VLA: Training-Free In-Context Adaptation for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [20] Y. Feng et al. “Spatial-Aware VLA Pretraining through Visual-Physical Alignment from Human Videos”, in CVPR 2026.
- ❑ [21] Z. Liu et al. “ActiveVLA: Injecting Active Perception into Vision-Language-Action Models for Precise 3D Robotic Manipulation”, in CVPR 2026.
- ❑ [22] J. Qian et al. “GeoPredict: Leveraging Predictive Kinematics and 3D Gaussian Geometry for Precise VLA Manipulation”, in CVPR 2026.
- ❑ [23] Y. Zhao et al. “Localizing, Structuring, and Rendering: Bridging 3D and 2D Vision-Language-Action Models for Robotic Manipulation”, in CVPR 2026.
- ❑ [24] J. W. Kim et al. “Ego-Pi: VLA Fine-Tuning for Ego-Centric Human and Robot Data”, in CVPR 2026.
- ❑ [25] Z. Yin et al. “MiVLA: Towards Generalizable Vision-Language-Action Model with Human-Robot Mutual Imitation Pre-training”, in CVPR 2026.
- ❑ [26] G. Jiang et al. “Cross-Hand Latent Representation for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [27] L. Xiao et al. “AVA-VLA: Improving Vision-Language-Action models with Active Visual Attention”, in CVPR 2026.
- ❑ [28] M. Lin et al. “HiF-VLA: Hindsight, Insight and Foresight through Motion Representation for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [29] L. Zhang et al. “AtomicVLA: Unlocking the Potential of Atomic Skill Learning in Robots”, in CVPR 2026.
- ❑ [30] C. Xu et al. “MoEActok: A MoE-based Action Tokenizer for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [31] X. An et al. “FlowHijack: A Dynamics-Aware Backdoor Attack on Flow-Matching Vision-Language-Action Models”, in CVPR 2026.
- ❑ [32] H. Lu et al. “When Robots Obey the Patch: Universal Transferable Patch Attacks on Vision-Language-Action Model”, in CVPR 2026.
- ❑ [33] Y. Yang et al. “Mantis: A Versatile Vision-Language-Action Model with Disentangled Visual Foresight”, in CVPR 2026.
- ❑ [34] Y. Liang et al. “Adaptive Action Chunking at Inference-time for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [35] S. Fei et al. “SRPO: Self-Referential Policy Optimization for Vision-Language-Action Models”, in CVPR 2026.
- ❑ [36] X. Li et al. “AT-VLA: Adaptive Tactile Injection for Enhanced Feedback Reaction in Vision-Language-Action Models”, in CVPR 2026.
- ❑ [37] Y. Fu et al. “MergeVLA: Cross-Skill Model Merging Toward a Generalist Vision-Language-Action Agent”, in CVPR 2026.

MV-Fashion: Towards Enabling Virtual Try-On and Size Estimation with Multi-View Paired Data



❑ Summary

- ❑ This study introduces 'MV-Fashion,' a large-scale dataset pairing flat catalogue images with multi-view videos and rich annotations, establishing a foundation for advanced virtual try-on, size estimation, and novel view synthesis

❑ Novelty

- ❑ Paired Data: First to pair multi-view videos with flat catalogue images.
- ❑ Real Sizing: Includes actual tape measurements (cm) and material properties.
- ❑ Complex Styling: Captures multi-layered outfits and real-world styles (e.g., tucked shirts).
- ❑ Advanced AI Control: Enables fine-grained, text-guided virtual try-on.

❑ Thoughts

- ❑ I feel that the emergence of multifaceted evaluation criteria—which balance visual realism from every innovative perspective with a precise fit for the body—will drive the development of next-generation AI.

❑ Reference

- ❑ Hunor Laczkó et al. "MV-Fashion: Towards Enabling Virtual Try-On and Size Estimation with Multi-View Paired Data" ,in CVPR 2026

Sign Language AI: from language benchmarks to motion-native communication models

❑ Summary

- ❑ Sign language AI is moving beyond multilingual VLM benchmarks.
- ❑ Recent works focus on motion-native multimodal modeling.
- ❑ Key topics include: sign-language motion generation, latent pose representation, multimodal sign recognition, radar-based privacy-preserving recognition

❑ Novelty

- ❑ Language is treated as structured human motion, not only text or gloss labels.
- ❑ Handshape, location, movement, and timing carry linguistic meaning.
- ❑ Pose-latent design strongly affects diffusion-based sign production.
- ❑ Recognition is expanding beyond RGB video to radar and sensor-based inputs.

❑ Thoughts

- ❑ This topic is different from multilingual/cultural LMM evaluation. The main question is whether the model can understand and generate embodied language over time.

❑ Reference

- ❑ [1] R. Hong and J. Kosecka. “Toward Phonology-Guided Sign Language Motion Generation: A Diffusion Baseline and Conditioning Analysis”, in CVPR Workshops 2026, GenSign.
- ❑ [2] G. Fauré et al. “The Impact of VAE Design on Latent Pose Representations for Diffusion-based Sign Language Production”, in CVPR Workshops 2026, GenSign.
- ❑ [3] A. Hasanaath et al. “SignEval 2026 Challenge Results”, in CVPR Workshops 2026, MSLR.
- ❑ [4] D.-T. Tran et al. “CAPY at SignEval 2026: Range-Aware Transformer with Kinematic Feature Augmentation for Robust Radar-Based Sign Language Recognition”, in CVPR Workshops 2026, MSLR.

Computer Vision in Education: from visual understanding to learning intervention

❑ Summary

- ❑ Educational CV is moving from passive classroom observation to active learning support.
- ❑ Recent works use multimodal models to analyze: student engagement, confusion, classroom behavior, speech and whiteboard content, and learning material comprehensibility

❑ Novelty

- ❑ Treating language as structured human motion, not only as text or gloss labels
- ❑ Handshape, hand location, movement, pose-latent design, and sensing modality strongly affect both generation and recognition performance.

❑ Thoughts

- ❑ The topic is different from multilingual/cultural LMM evaluation.
- ❑ Evaluation should include naturalness, linguistic correctness, temporal consistency, and privacy-aware sensing.

❑ Reference

- ❑ [1] R. Hong and J. Kosecka. “Toward Phonology-Guided Sign Language Motion Generation: A Diffusion Baseline and Conditioning Analysis”, in CVPR Workshops 2026, GenSign.
- ❑ [2] G. Fauré et al. “The Impact of VAE Design on Latent Pose Representations for Diffusion-based Sign Language Production”, in CVPR Workshops 2026, GenSign.
- ❑ [3] A. Hasanaath et al. “SignEval 2026 Challenge Results”, in CVPR Workshops 2026, MSLR.
- ❑ [4] D.-T. Tran et al. “CAPY at SignEval 2026: Range-Aware Transformer with Kinematic Feature Augmentation for Robust Radar-Based Sign Language Recognition”, in CVPR Workshops 2026, MSLR.

Event-based Vision: sensing is becoming part of the model

❑ Summary

- ❑ Event-based vision is gaining attention as a hardware-aware vision paradigm.
- ❑ It targets high-speed, low-latency, and low-power perception.
- ❑ key topics include: in-sensor denoising, event-based visual odometry, event-guided deblurring, neuromorphic onboard pose estimation

❑ Novelty

- ❑ The camera is no longer treated as a passive image provider.
- ❑ Event cameras capture brightness changes asynchronously.

❑ Thoughts

- ❑ This topic shifts the discussion away from larger foundation models.
- ❑ Future CV progress may come from better sensing, not only bigger models.

❑ Reference

- ❑ [1] J. So et al. “EventGuard: Sparsity-Aware In-Sensor Denoising for Frame-Based Event Vision Sensors”, in CVPR Workshops 2026, The 22nd Embedded Vision Workshop, EVW
- ❑ [2] A. Marchei et al. “TinyDEVO: Deep Event-based Visual Odometry on Ultra-low-power Multi-core Microcontrollers”, in CVPR Workshops 2026, The 22nd Embedded Vision Workshop, EVW
- ❑ [3] Y. Qiu et al. “EVFormer: A Dual-Model Event-Guided Framework for Motion Deblurring”, in CVPR Workshops 2026, 11th New Trends in Image Restoration and Enhancement Workshop and Challenges, NTIRE
- ❑ [4] L. Sun et al. “The Second Challenge on Event-Based Image Deblurring at NTIRE 2026: Methods and Results”, in CVPR Workshops 2026, 11th New Trends in Image Restoration and Enhancement Workshop and Challenges, NTIRE
- ❑ [5] A. Rathinam et al. “Efficient Onboard Spacecraft Pose Estimation with Event Cameras and Neuromorphic Hardware”, in CVPR Workshops 2026, Artificial Intelligence for Space, AI4Space

Thermal-Det: Language-Guided Cross-Modal Distillation for Open-Vocabulary Thermal Object Detection

Summary

- Thermal-Det enables zero-shot open-vocabulary object detection for thermal images by transferring knowledge from RGB models via cross-modal distillation and a synthetic dataset derived from GroundingCap-1M without manual annotations.

Novelty

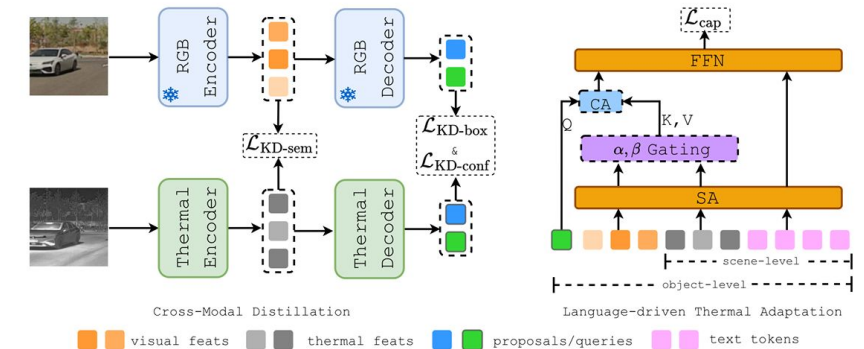
- Annotation-free thermal OVD: Enables zero-shot open-vocabulary detection in thermal imagery without thermal labels.
- Semantic alignment and distillation: Uses the Thermal-Text Alignment Head (TTAH) and RGB-to-thermal distillation to transfer open-vocabulary knowledge.

Thoughts

- Knowledge distillation is particularly interesting because it transfers semantic knowledge learned by RGB models to thermal images. This suggests that object concepts learned from RGB data can remain useful even when the visual appearance differs significantly across modalities.

Reference

- Y.Ranasinghe et al. “Thermal-Det: Language-Guided Cross-Modal Distillation for Open-Vocabulary Thermal Object Detection”, in CVPR 2026.



Discussion after the Report

1-hour discussion in researchers & students

CVPR26: Discussion after the Report (1/7)

Discussion

- ❑ Should awards be required to release?
 - ❑ Codes / models / datasets / evaluation protocols
- ❑ Should CVPR tracks be separated?
 - ❑ Scientific best paper
 - ❑ Open science best paper
 - ❑ Industrial best paper

What do you think?

Discussion

- ❑ Importance of visibility at poster presentation
 - ❑ Poster design
 - ❑ Panels: award candidate, oral, highlight
- ❑ Many attendees decide whether to stop at a poster within seconds
- ❑ How can conferences surface important work beyond popularity?

CVPR26: Discussion after the Report (3/7)

Discussion

- ❑ Is it a right time for a CVPR position paper track?
 - ❑ Inspired by discussions, e.g., “The Bitter Lesson” workshop
 - ❑ & emerging AI paradigms
- ❑ Is it enough to discuss on the workshop?
 - ❑ Debate future directions
 - ❑ Challenge assumptions
 - ❑ Introduce research visions
 - ❑ (But extreme numbers of workshops!)

Discussion

- ❑ Toward computer-aware awards
 - ❑ Movement: Compute Reporting Initiative (Good tendency?)
 - ❑ Could a CPU-only or laptop-scale challenge reveal new innovation?
 - ❑ If compute becomes abundant, should efficiency become the new frontier?

Discussion

- ❑ Industry vs Academia (Part 1)
 - ❑ Are current review criteria equally fair to both communities?
 - ❑ Should industry & academia compete in the same award pool?
 - ❑ Can academia still generate paradigm shifts using limited resources?
 - ❑ How should we evaluate contributions when compute budgets differ by $\times 100 - \times 1000$?

Discussion

- ❑ Industry vs Academia (Part 2)
 - ❑ Different Resources, Different Contributions?
 - ❑ Industry
 - ❑ Large-scale compute and data
 - ❑ Ability to run massive experiments
 - ❑ Faster deployment to real products
 - ❑ Strong engineering and infrastructure
 - ❑ Academia
 - ❑ New questions and hypotheses
 - ❑ Small-scale exploratory experiments
 - ❑ Alternative viewpoints
 - ❑ Long-term scientific curiosity
- ❑ Discussion
 - ❑ How many attendees are doing ML as their primary profession?
 - ❑ Is the gap in compute becoming too large?
 - ❑ Are we seeing enough negative results and failed trials?
 - ❑ Should conferences encourage sharing trial-and-error insights?

Discussion

- ❑ Where are we heading?
 - ❑ Recent award candidates show increasing emphasis on:
 - ❑ 3D understanding
 - ❑ 4D dynamic scene modeling
 - ❑ World simulation
 - ❑ Physical reasoning
 - ❑ Discussion
 - ❑ Is the field becoming biased toward 3D/4D?
 - ❑ Does 4D naturally subsume 3D and 2D?
 - ❑ Are we moving from representation learning toward simulation learning?
 - ❑ How much progress still comes from pure representation learning?



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

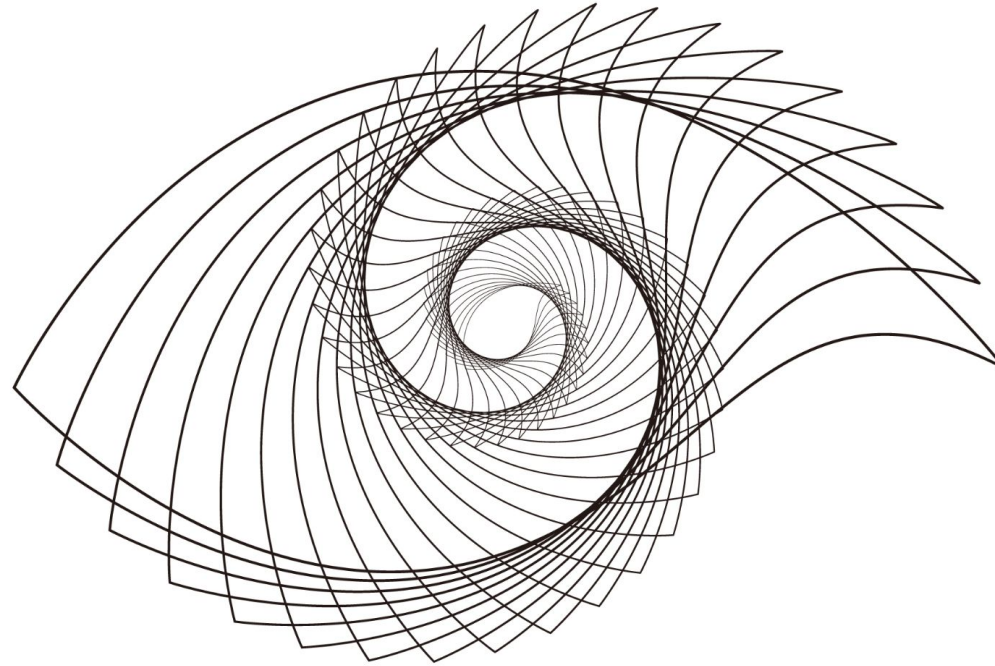
Oxford VGG



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



LIMIT.LAB

Join us! → Slack invitation [[Link](#)]