ICCV 2025 Report

Hirokatsu Kataoka, Yue Qiu, Yoshihiro Fukuhara, Shumpei Takezaki,
Edgar Martinez, Ren Ohkubo, Kazuya Nishimura, Yuto Matsuo, Kohsuke Ide, Moeri Okuda,
Ryosuke Korekata, Rintaro Yanagi, Chihiro Kaneko,
Ryuichi Nakahara, Kohei Torimi, Ryo Nakamura, Daichi Otsuka, Gido Kato, Noritake Kodama, Yukinori Yamamoto, Go Ohtani, Oishi Deb

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

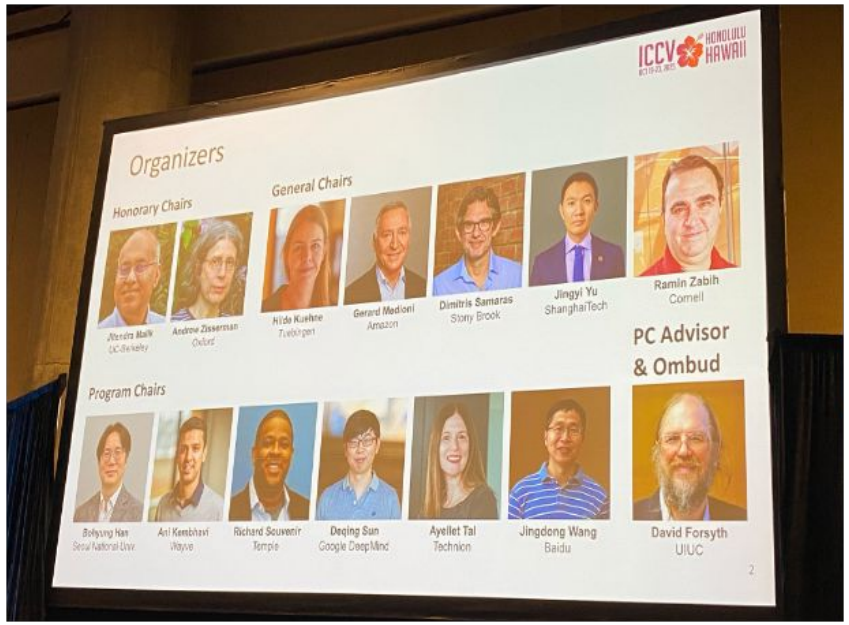
Meta Insights into Trends and Tendencies in ICCV 2025

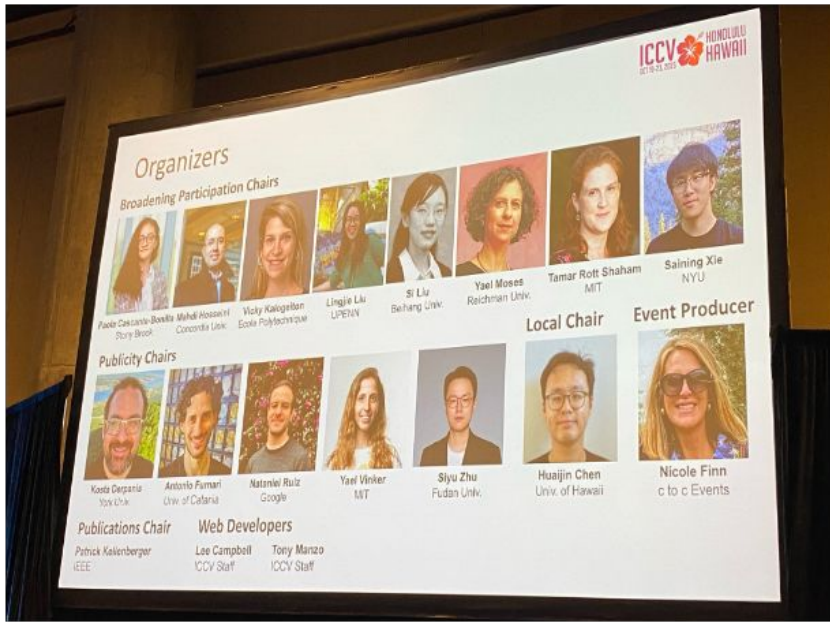
- What kind of research was trending at this moment?
- What are overseas researchers working on?
- We have compiled the “trends” and “insights.”

ICCV25: Meta Insights into Trends and Tendencies (1/153)

From opening slide

- Organizers at ICCV 2025
  - To organize the large-scale conferences over 10k submissions / 7k attendees / 5-day conference / and other things
  - 40+ CV people in this ICCV (only in core members)

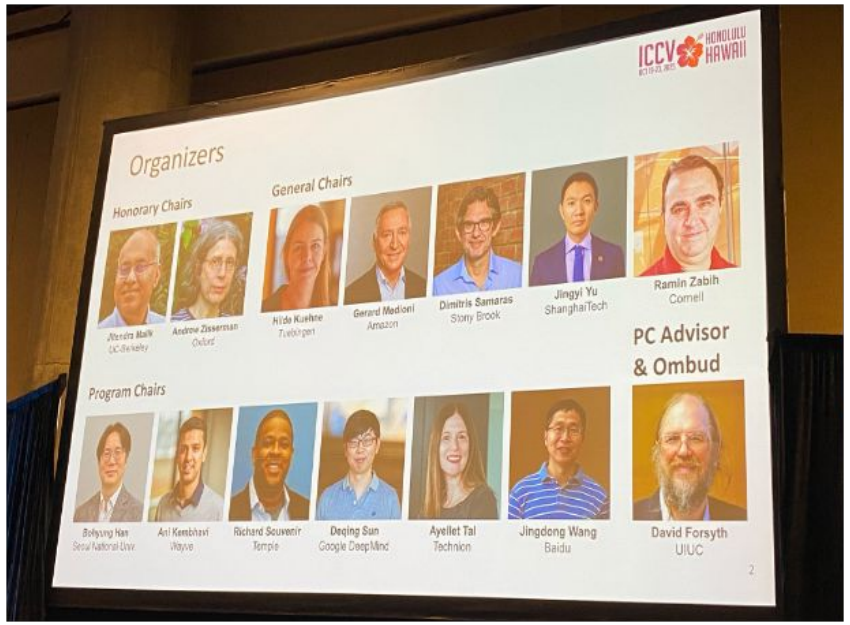


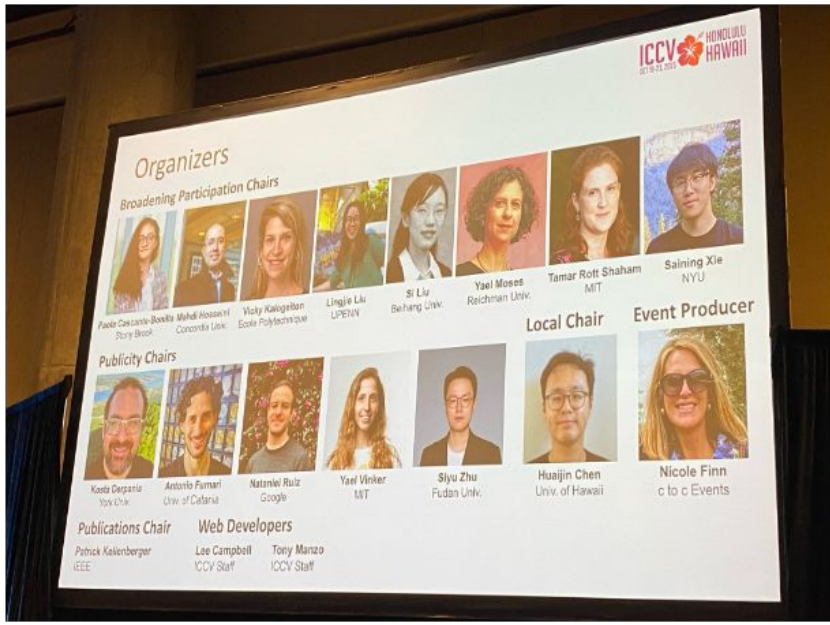


ICCV25: Meta Insights into Trends and Tendencies (2/153)

From opening slide

- Organizers at ICCV 2025
  - Honorary Chair (2), General Chair (5), Program Chair (6), PC Advisor & Ombud (1), Finance Chair (1), Technical Chair (1), Workshop Chair (5), Tutorial Chair (3), Demo Chair (1), Doctoral Consortium Chair (2), Broadening Participation Chair (8), Publicity Chair (5), Local Chair (1), Event Producer (1)

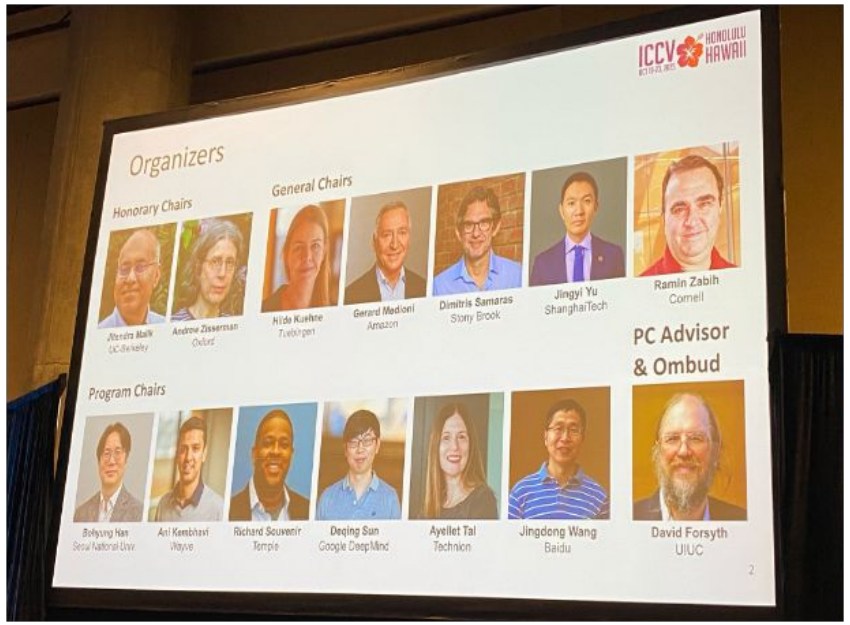


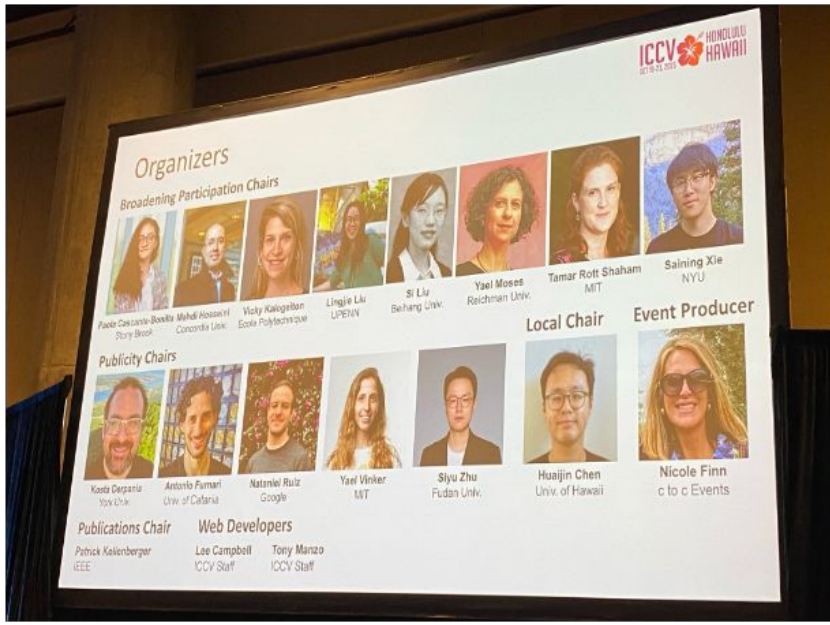


ICCV25: Meta Insights into Trends and Tendencies (3/153)

From opening slide

- What's special in ICCV 2025 (and other CV conferences)
  - More GCs (5) / PCs (6)
  - Technical Chair (1) -> OpenReview's chair
  - Broadening Participation Chairs (8) -> Distribute chances to researchers
  - Publicity chairs (5) -> Twitter (X) chair
  - Event Producer (1) -> Treating crazy number of contacts from authors/attendees

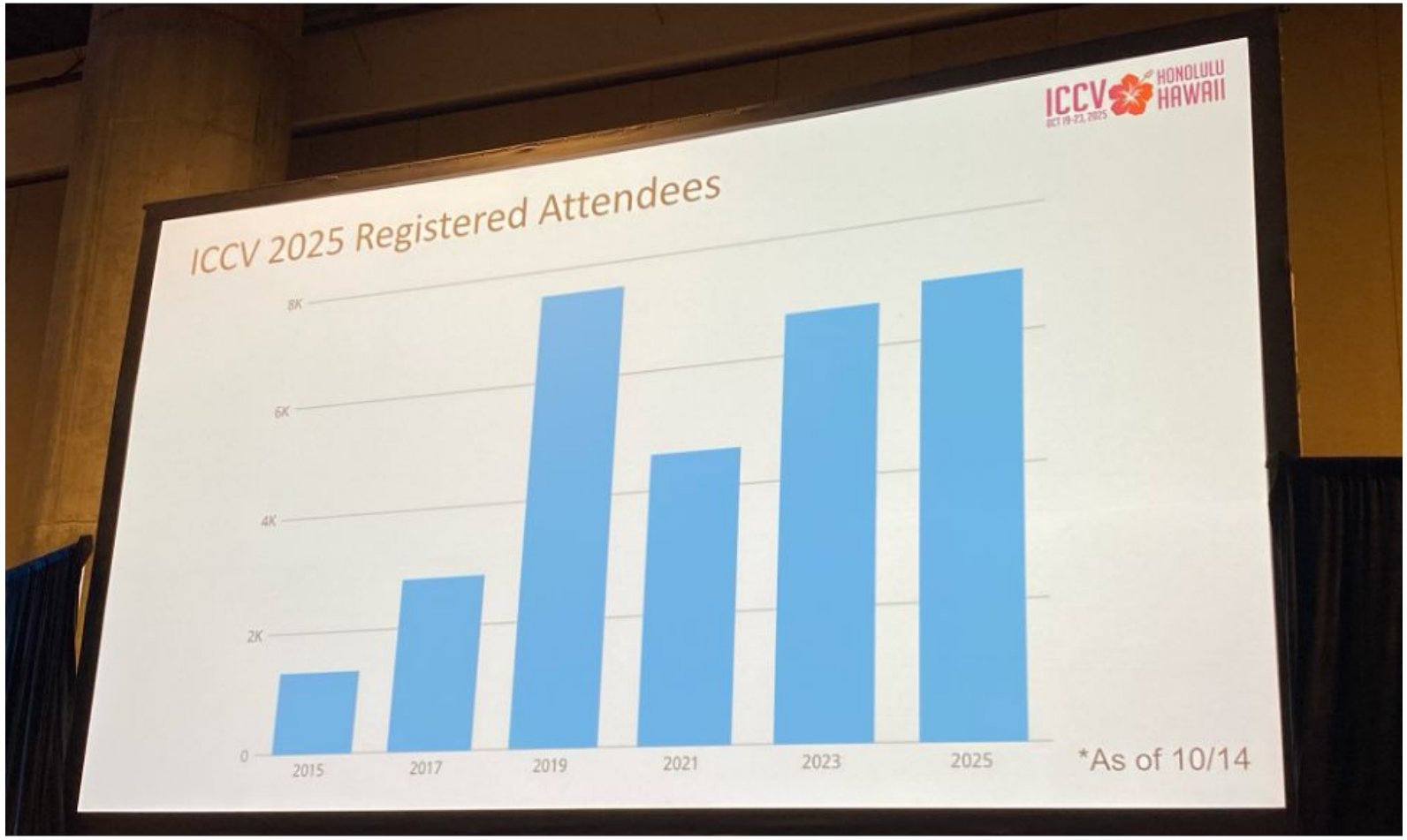




ICCV25: Meta Insights into Trends and Tendencies (4/153)

From opening slide

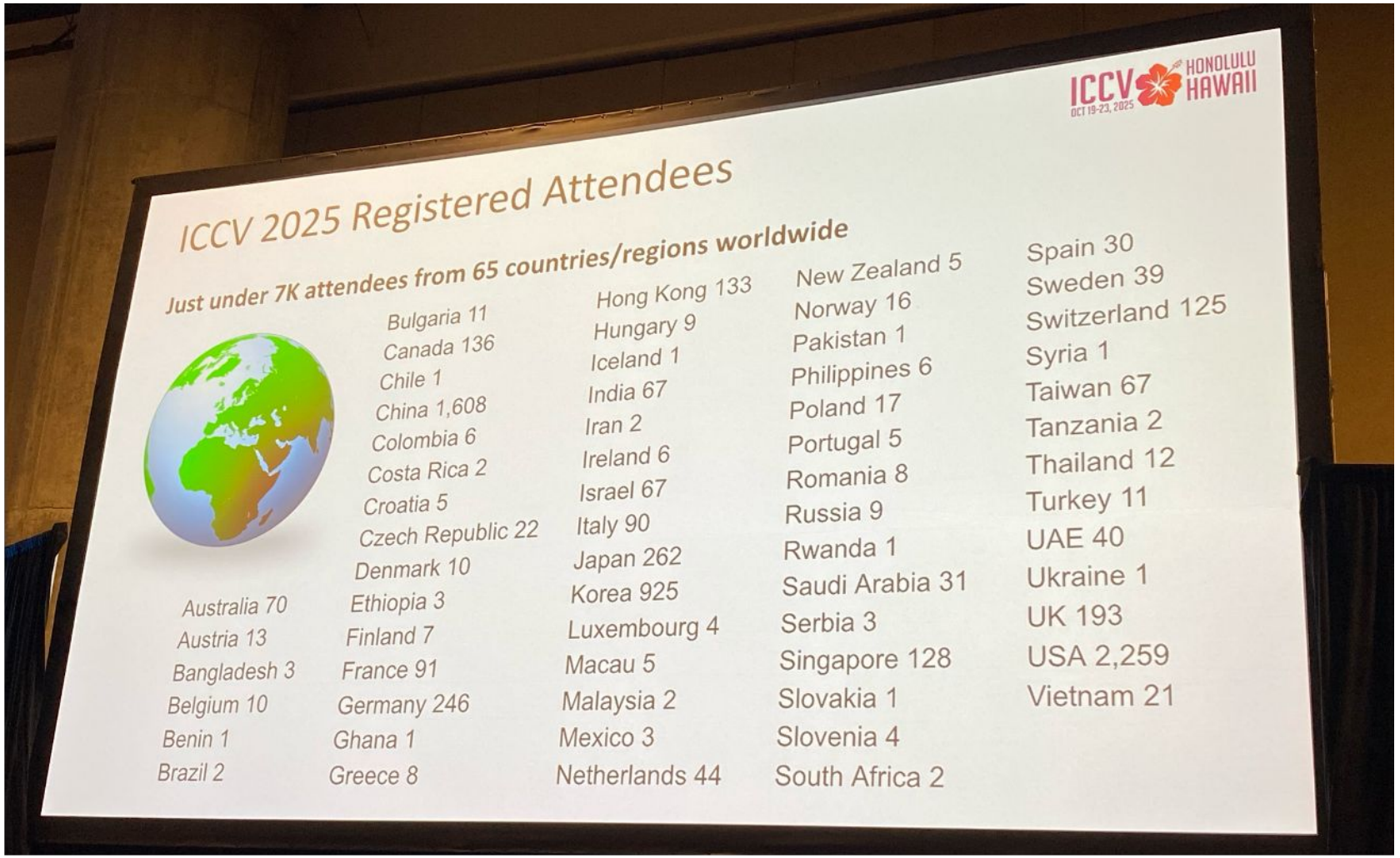
- #Attendees
  - Close to 7k attendees in-person
  - But not reached ICCV 2019 @ Korea



ICCV25: Meta Insights into Trends and Tendencies (5/153)

From opening slide

- Registered attendees



ICCV25: Meta Insights into Trends and Tendencies (6/153)

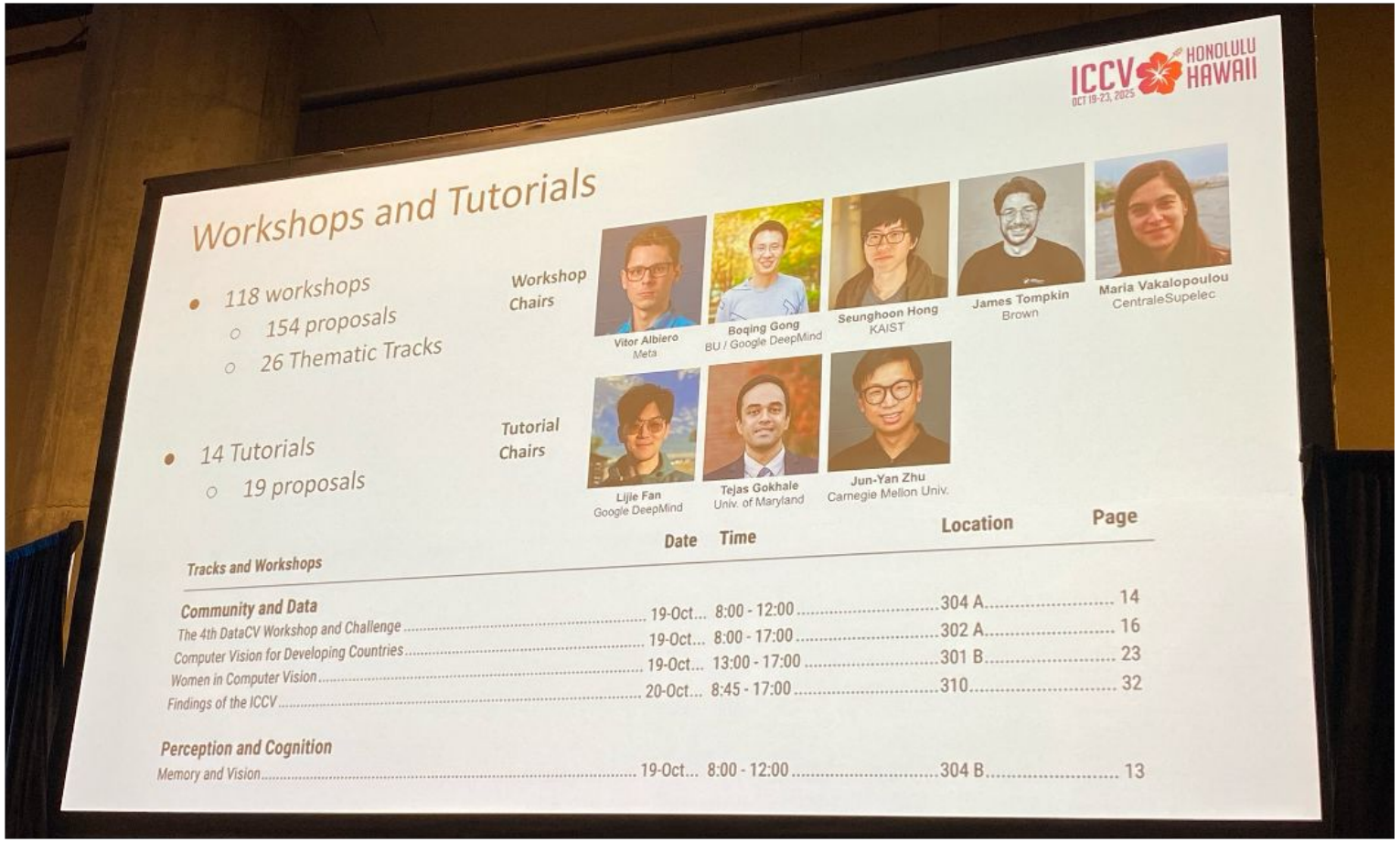
From opening slide

- Broadening participation

ICCV25: Meta Insights into Trends and Tendencies (7/153)

From opening slide

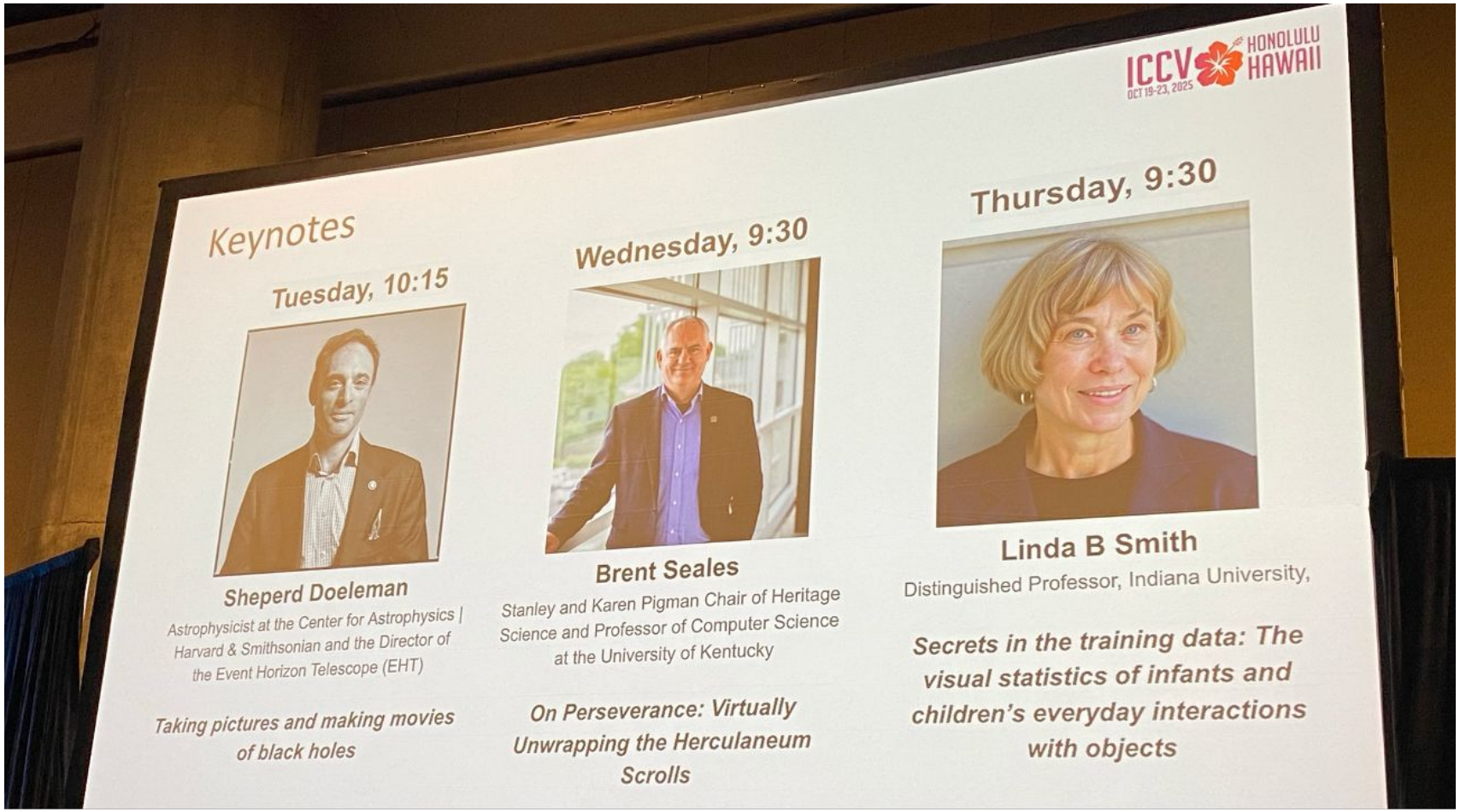
- Workshops and tutorials
  - WS chairs have adjusted & scheduled to accept many workshops
  - Tutorial proposals are less proposals in this time



ICCV25: Meta Insights into Trends and Tendencies (8/153)

From opening slide

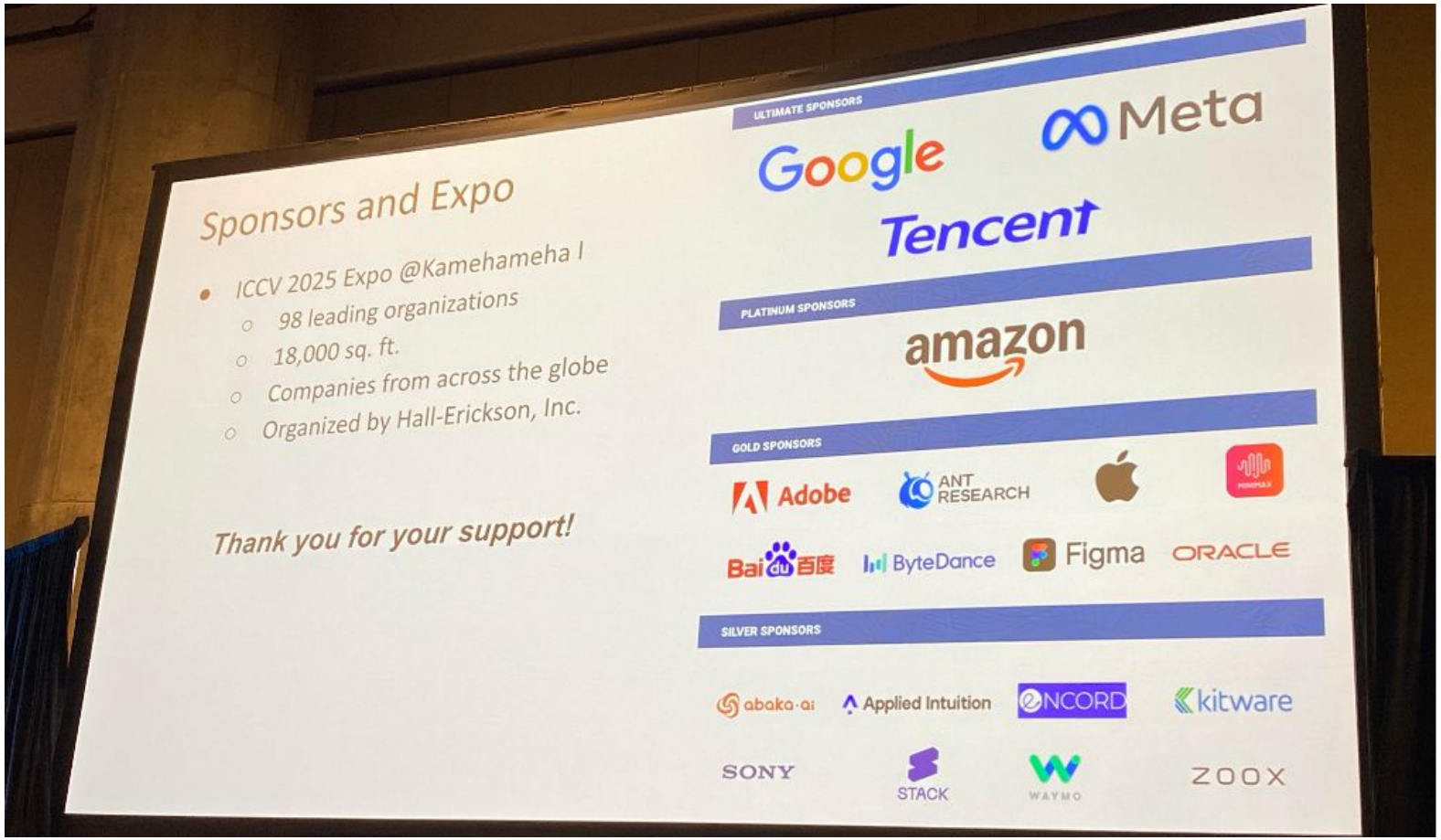
- Keynote speakers
  - Assigning three speakers for each day



ICCV25: Meta Insights into Trends and Tendencies (9/153)

From opening slide

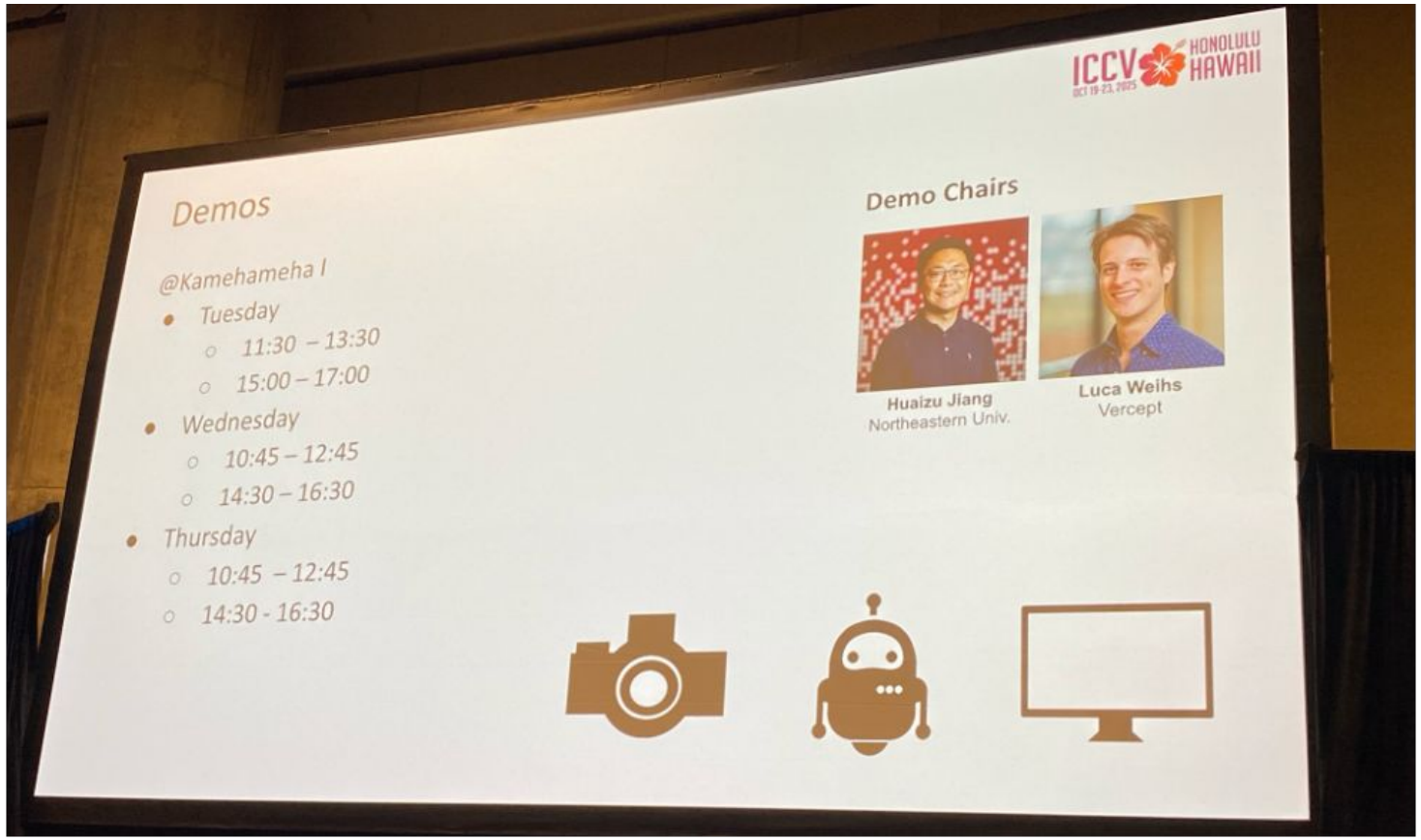
- Sponsors and expo
  - 98 sponsors
  - This conference has been run by the sponsors



ICCV25: Meta Insights into Trends and Tendencies (10/153)

From opening slide

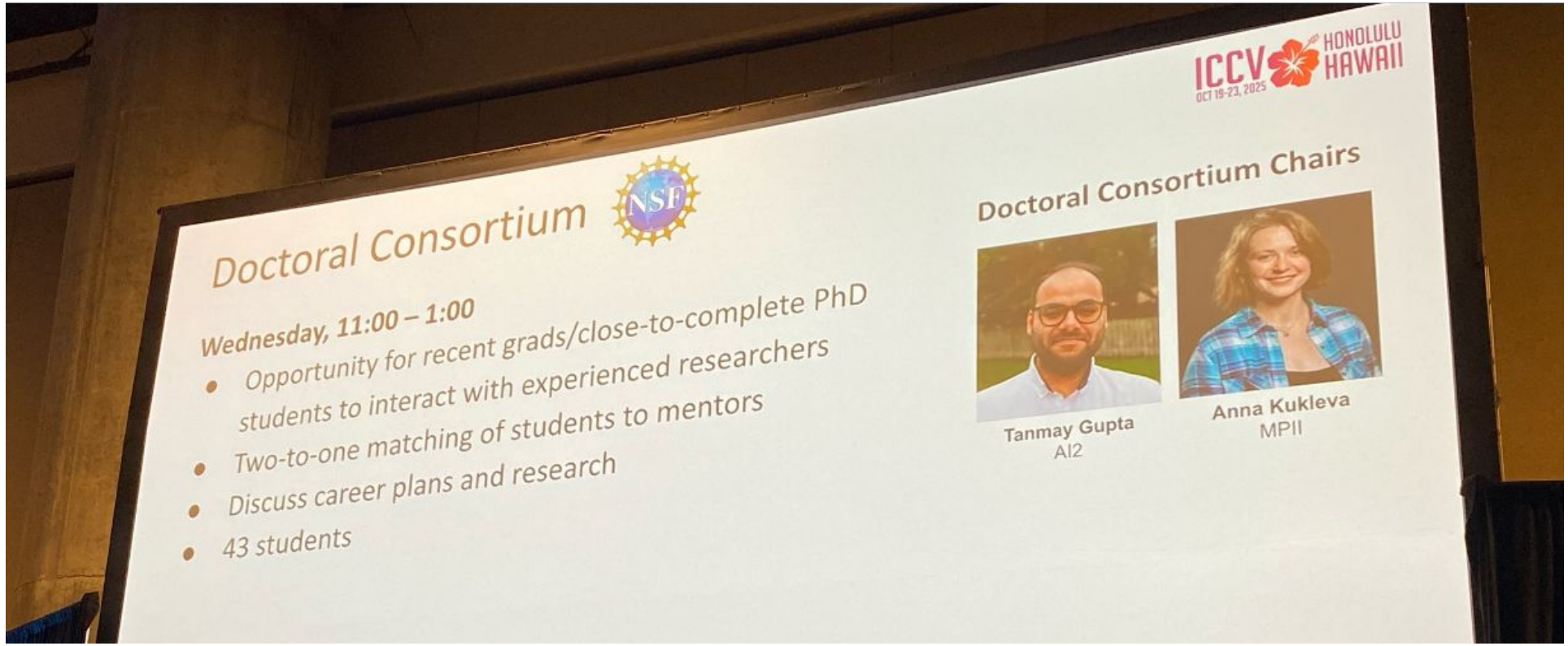
- Demonstrations
  - Demo session with CV techniques
  - Listed good algorithm, tech (close to a) product, and implementation
  - There is a demo award!



ICCV25: Meta Insights into Trends and Tendencies (11/153)

From opening slide

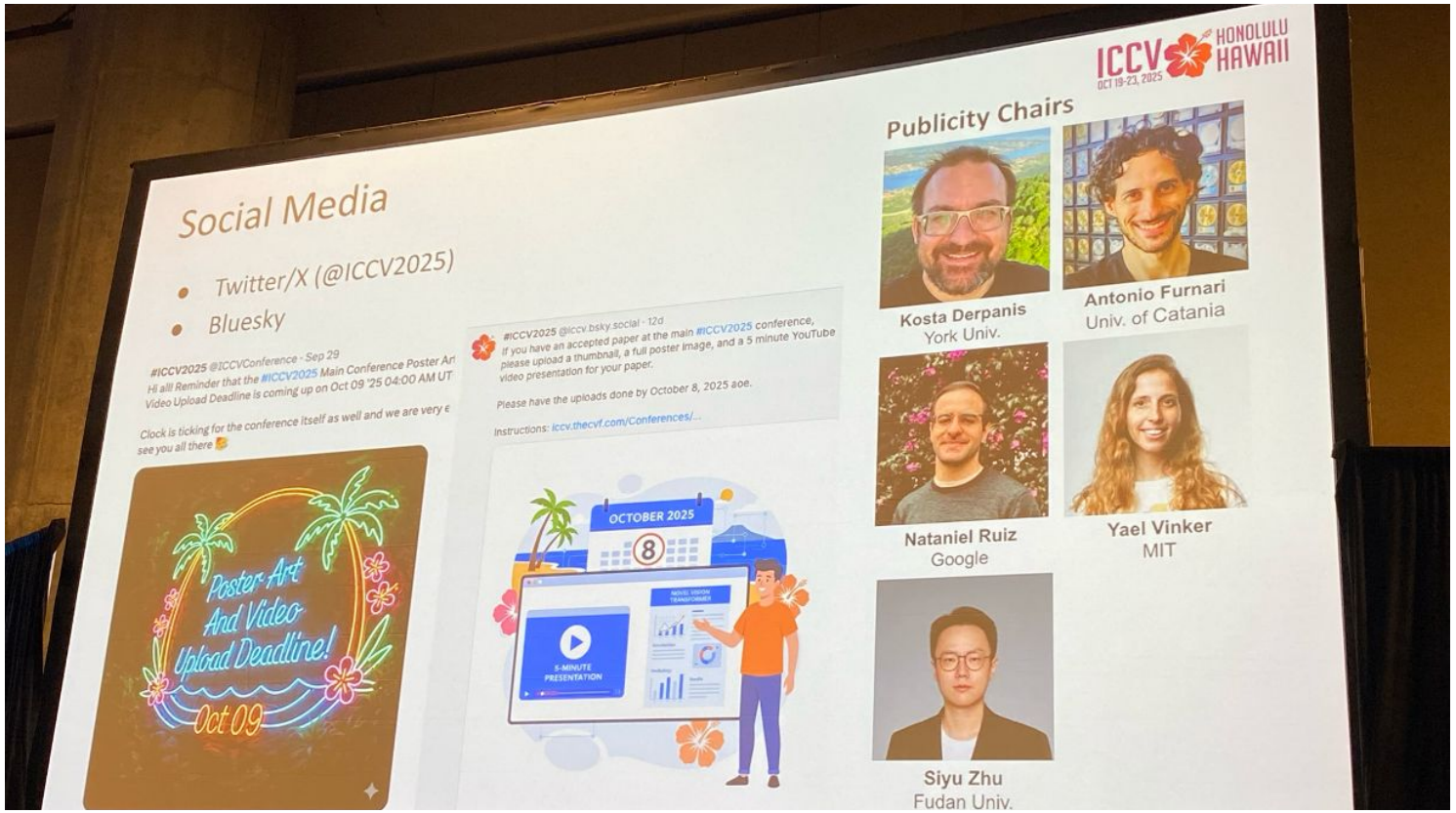
- Doctoral consortium
  - A social for doctoral students
  - Good connections, discussions, and job opportunity



ICCV25: Meta Insights into Trends and Tendencies (12/153)

From opening slide

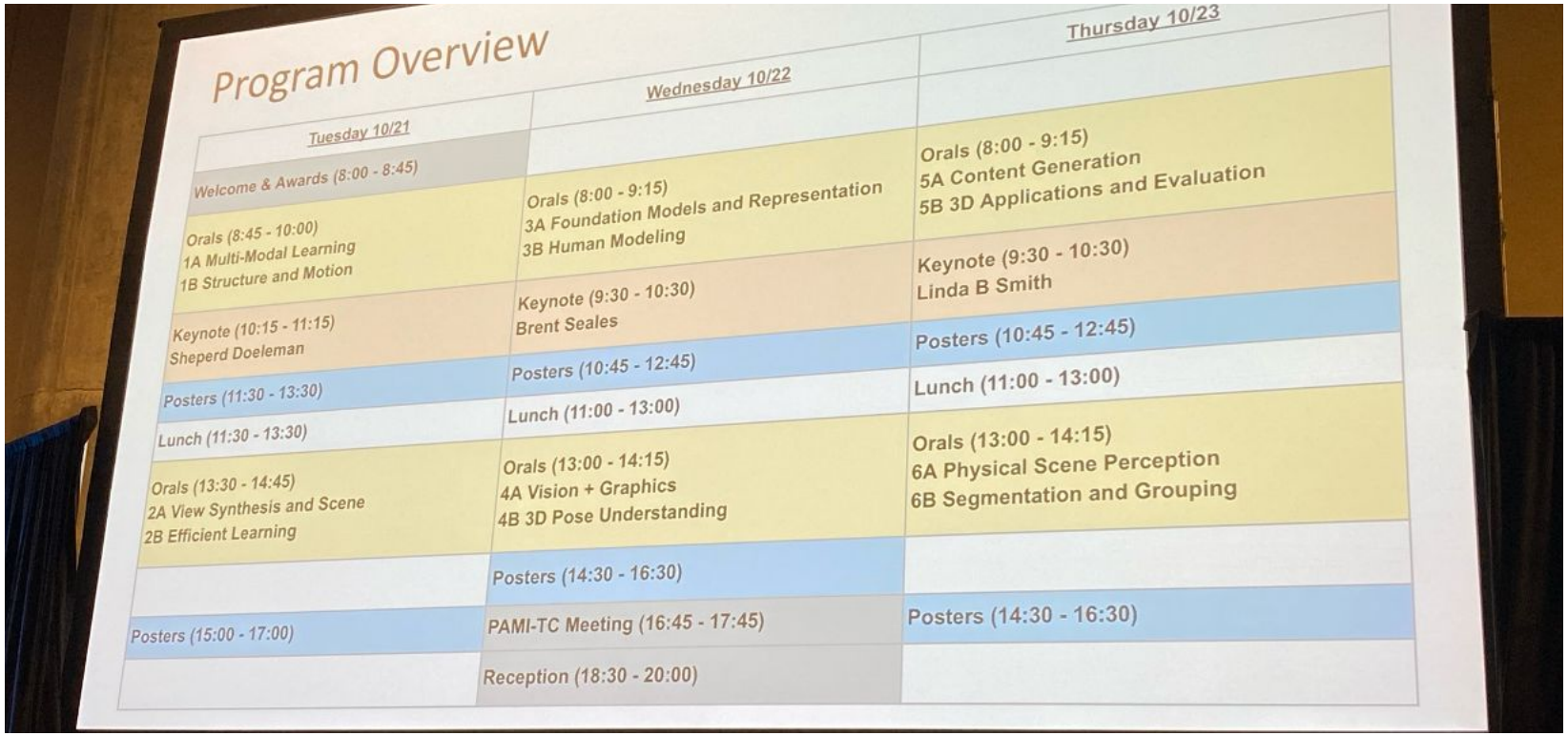
- Publicity chair = Social media chair
  - X: https://x.com/ICCVConference
  - Bluesky: https://bsky.app/profile/iccv.bsky.social



ICCV25: Meta Insights into Trends and Tendencies (13/153)

From opening slide

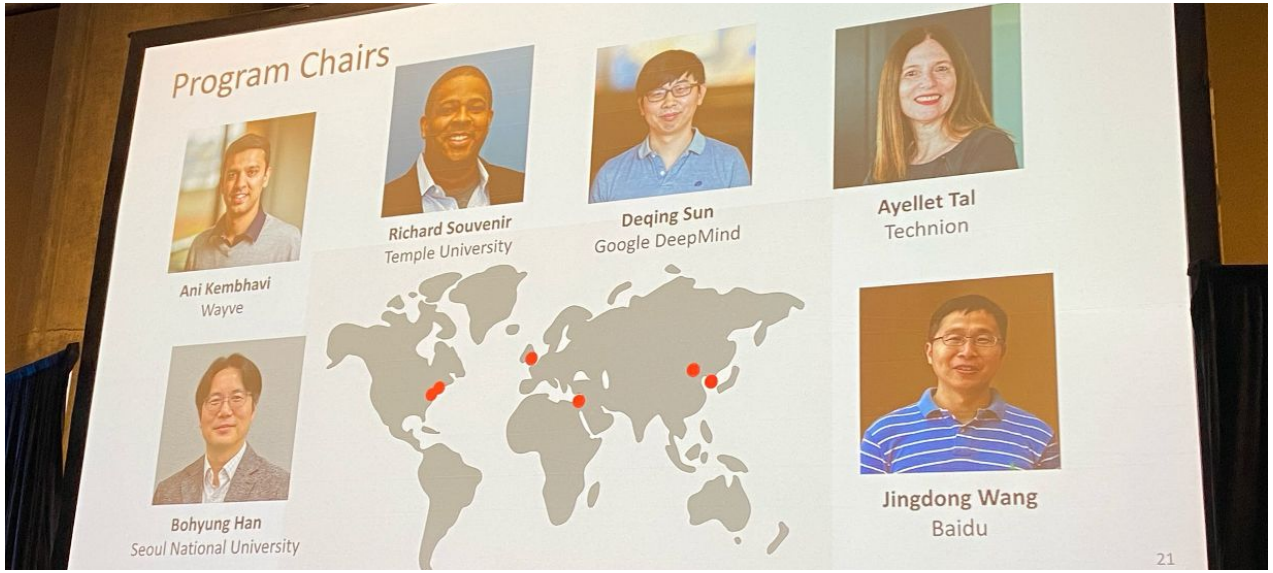
- ICCV 2025 program overview
  - No big change in this ICCV
  - Opening remarks & award ceremony at the beginning
  - 6 orals (2 parallel session) / 6 posters / 3 keynotes
  - PAMI-TC Meeting for the future conferences & community

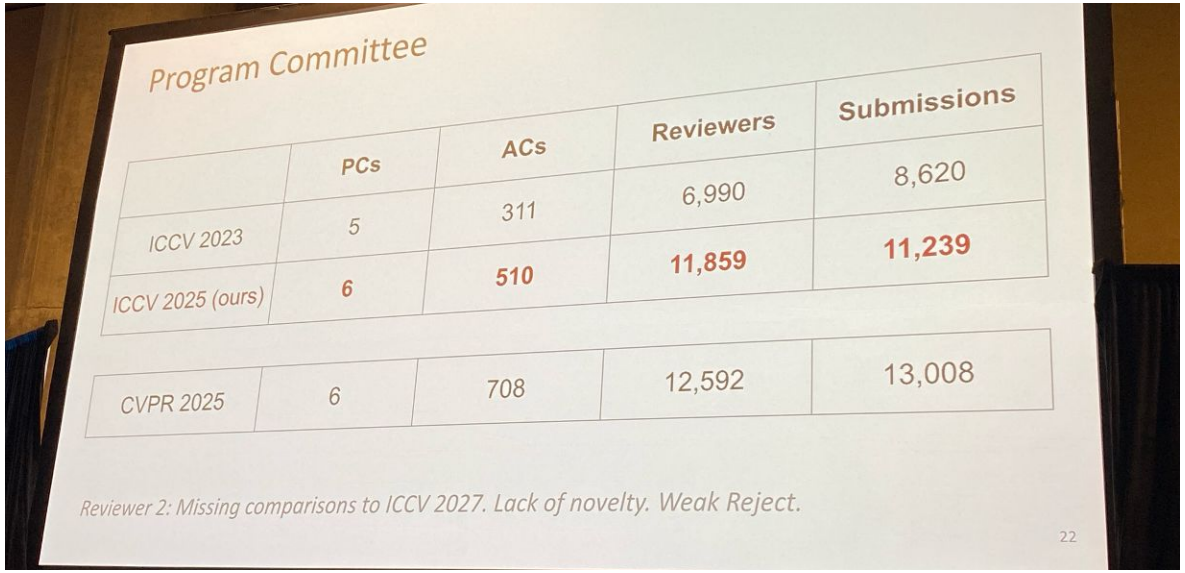


ICCV25: Meta Insights into Trends and Tendencies (14/153)

From opening slide

- 6 PCs among diverse areas (left figure)
- Comparisons with recent conferences (right figure)
  - A bit challenging for area chairs (ACs)
    - ICCV 2025: 510 ACs for 11.8k paper decisions
    - CVPR 2025: 708 ACs for 12.5k paper decisions

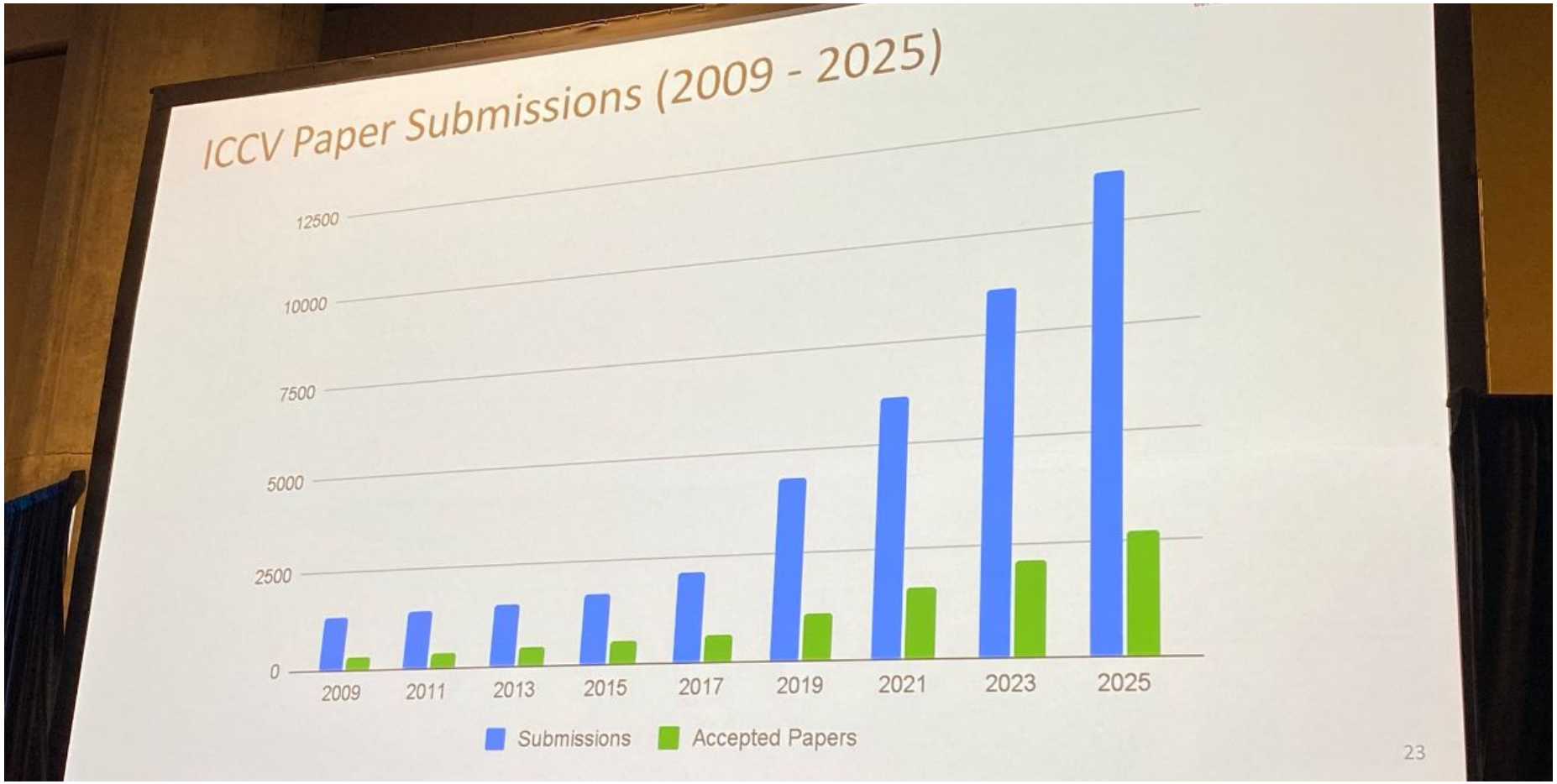




ICCV25: Meta Insights into Trends and Tendencies (15/153)

From opening slide

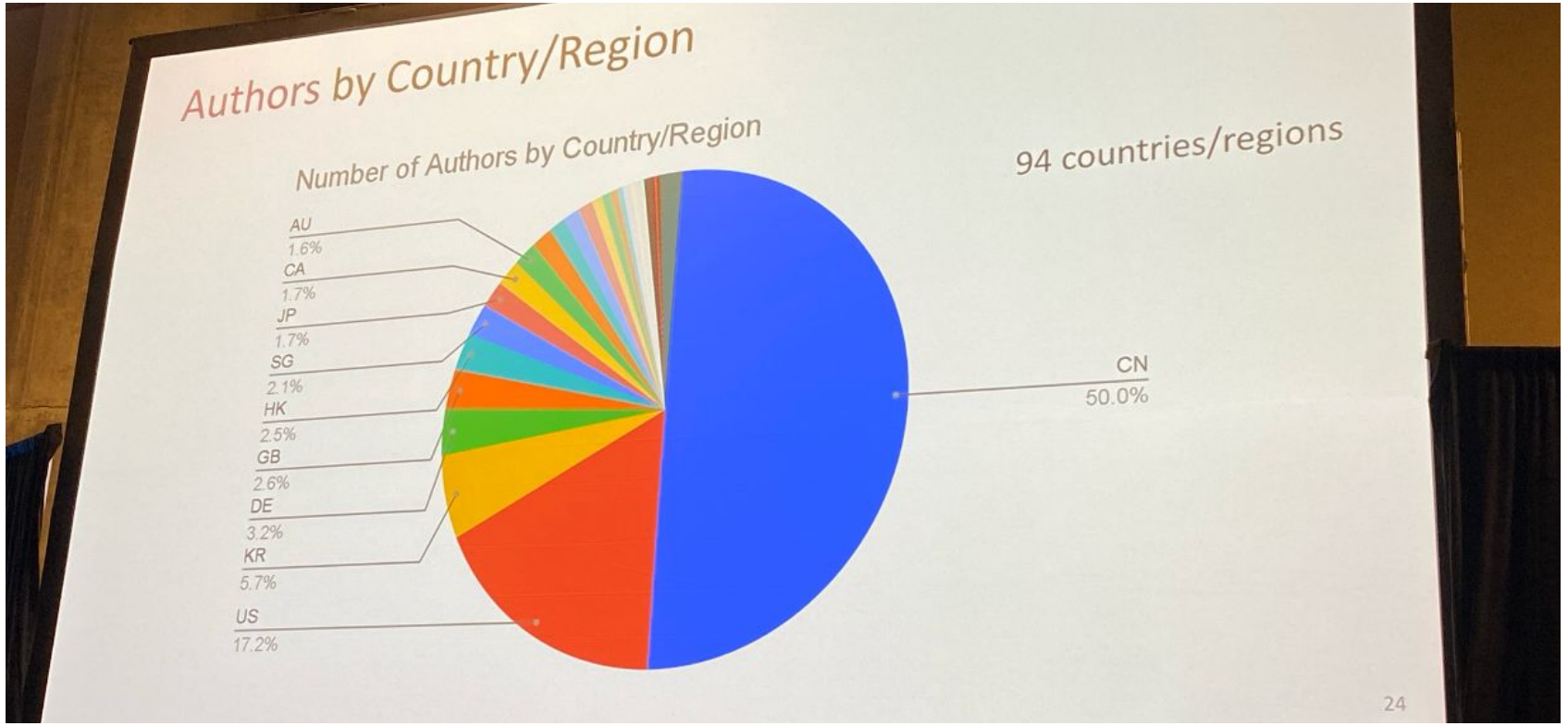
- ICCV is still growing!
  - ICCV paper submissions in the 16 years / at each 2 years



ICCV25: Meta Insights into Trends and Tendencies (16/153)

From opening slide

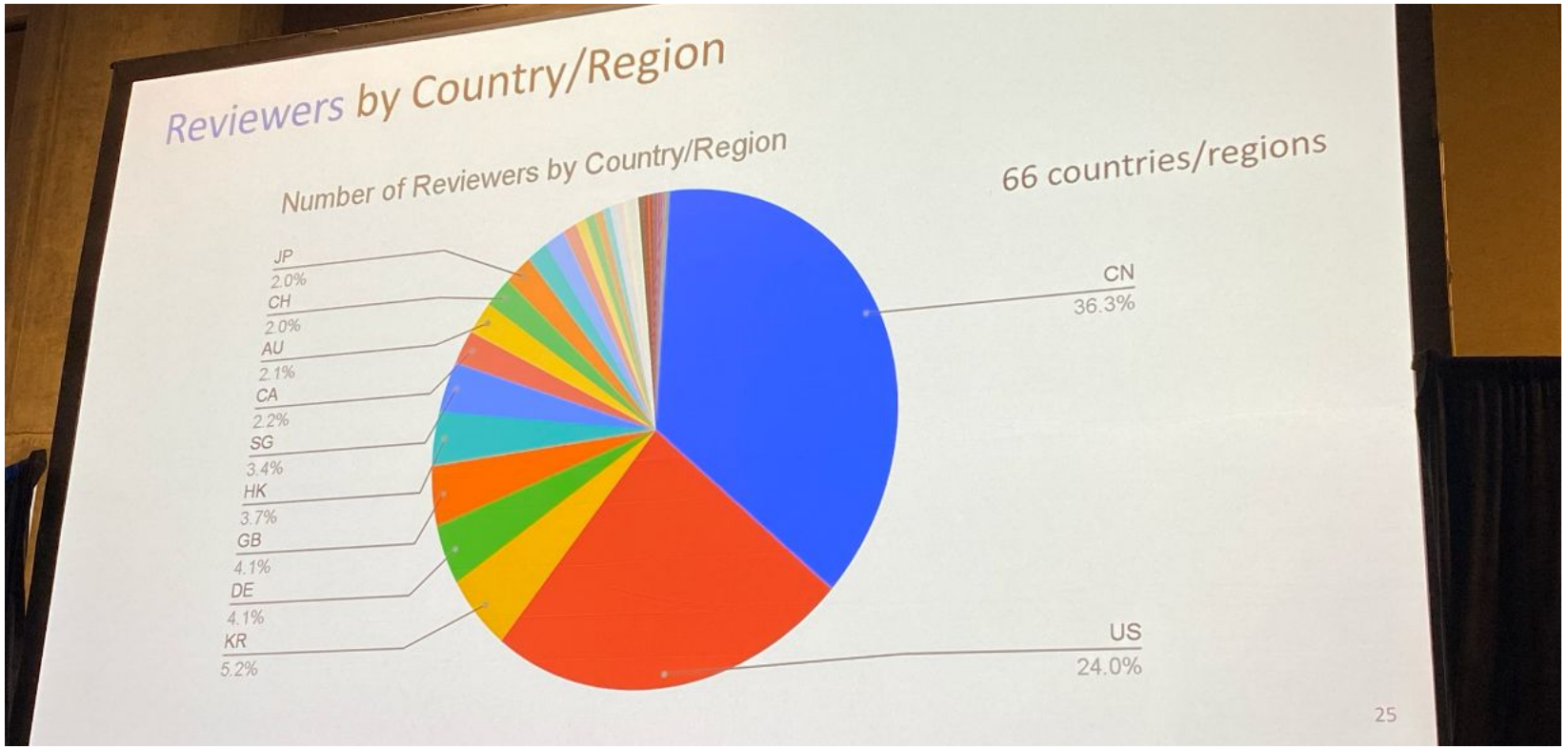
- Authors by country/region
  - Top5:
  - China 50% -> US 17.2% -> Korea 5.7% -> German 3.2% -> United Kingdom 2.6%



ICCV25: Meta Insights into Trends and Tendencies (17/153)

From opening slide

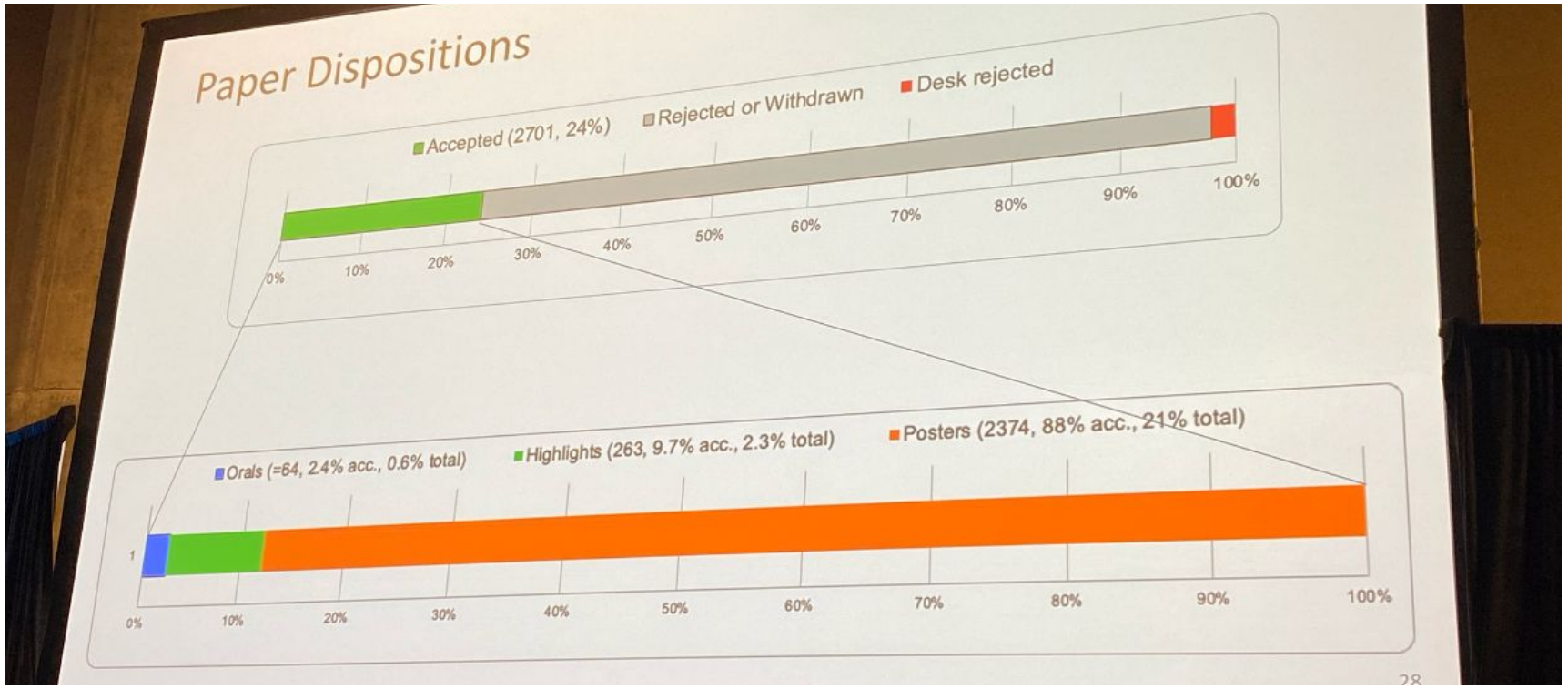
- Reviewers by country/region
  - Top5:
  - China 36.3% -> US 24.0% -> Korea 5.2% -> Germany 4.1% / United Kingdom 4.1%



ICCV25: Meta Insights into Trends and Tendencies (18/153)

From opening slide

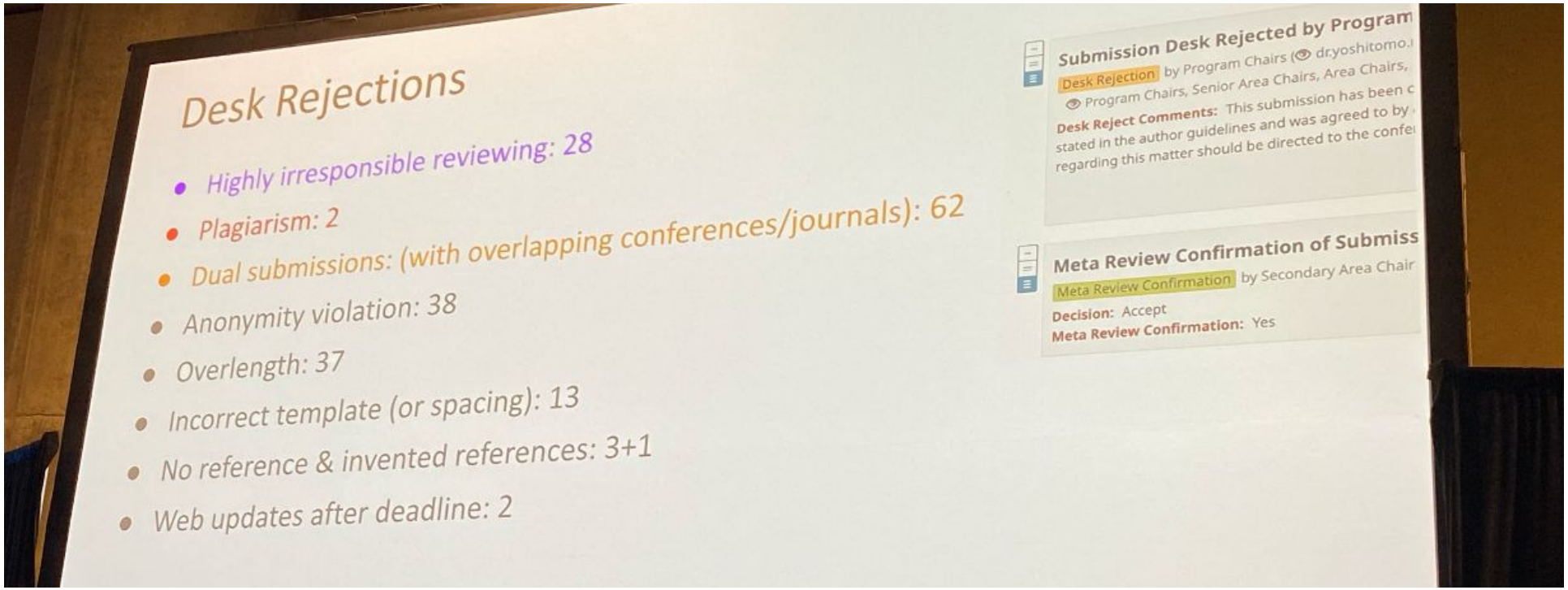
- Paper statistics
  - Acceptance rate: 24%
  - Rate of oral: 2.4% (in accepted papers) / 0.6% (in all submitted papers)
  - Rate of highlight: 9.7% (in accepted papers) / 2.3% (in all submitted papers)



ICCV25: Meta Insights into Trends and Tendencies (19/153)

From opening slide

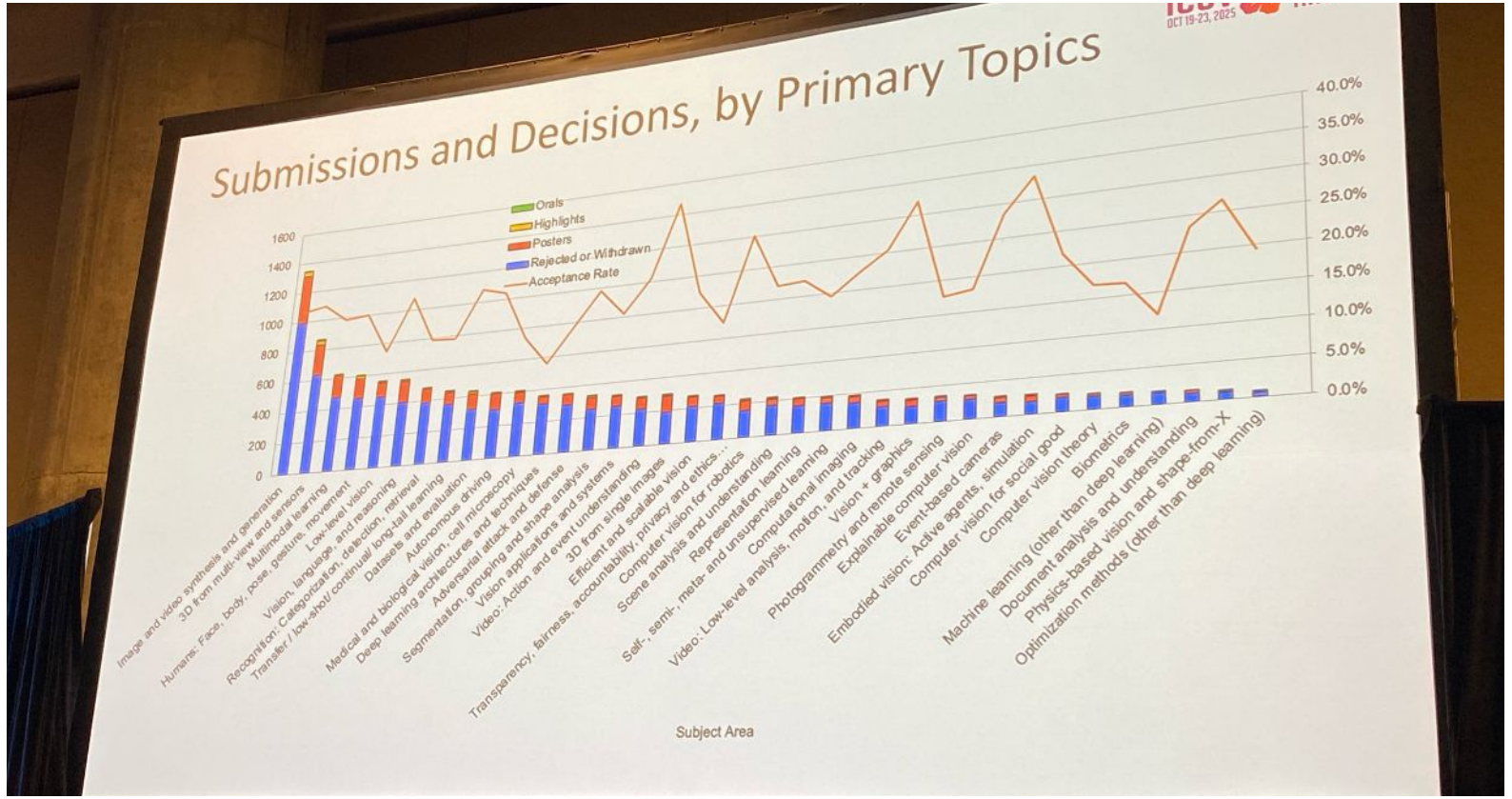
- Desk rejection
  - An irresponsible review could be a desk reject
  - Dual submissions have been checked among similar conferences
  - Most of other reasons are format violations



ICCV25: Meta Insights into Trends and Tendencies (20/153)

From opening slide

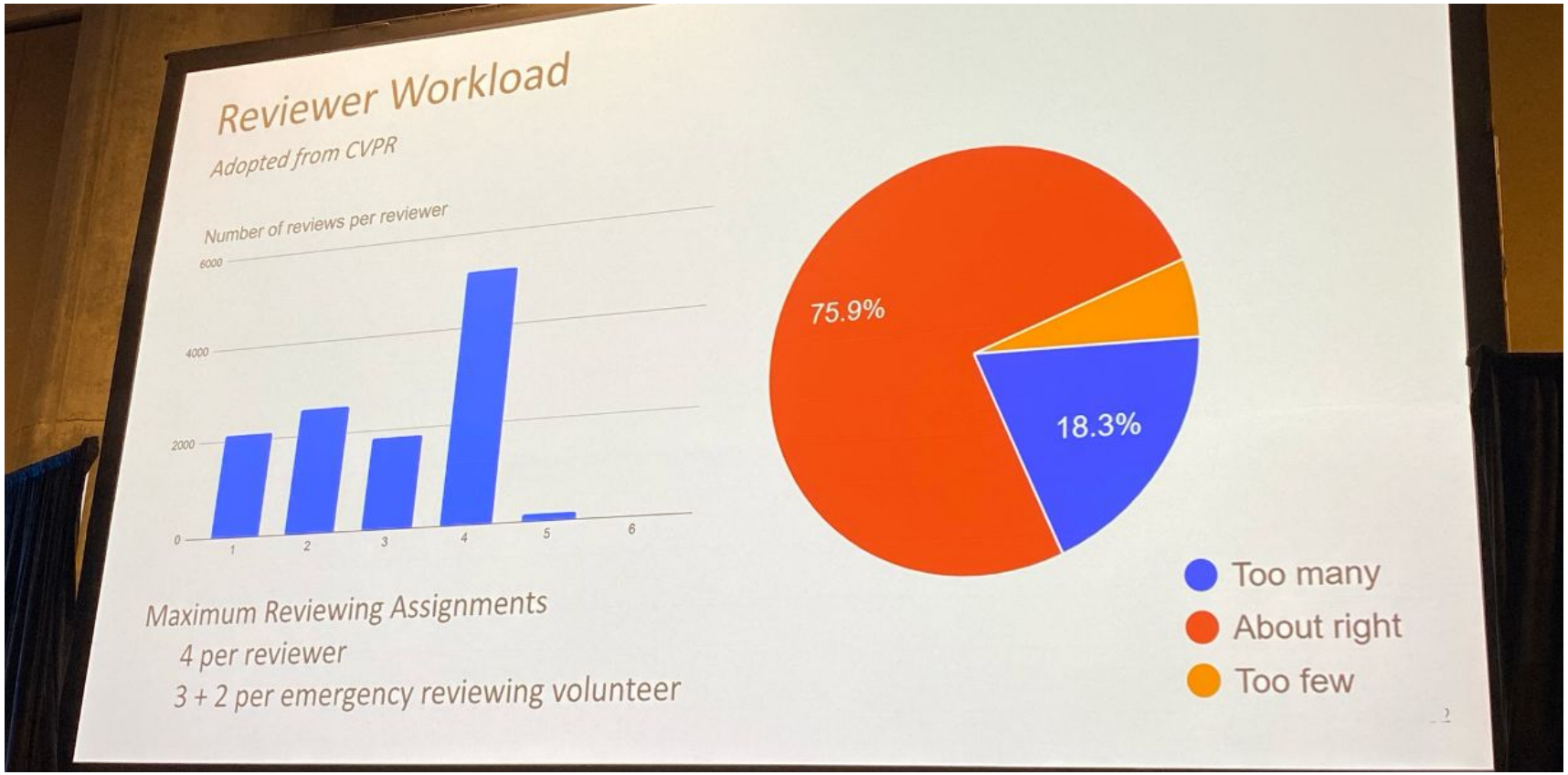
- Topics in ICCV 2025
  - GenAI (image & video)
  - 3D from multiview and sensors
  - Multimodal learning ...are TOP3 in this ICCV



ICCV25: Meta Insights into Trends and Tendencies (21/153)

From opening slide

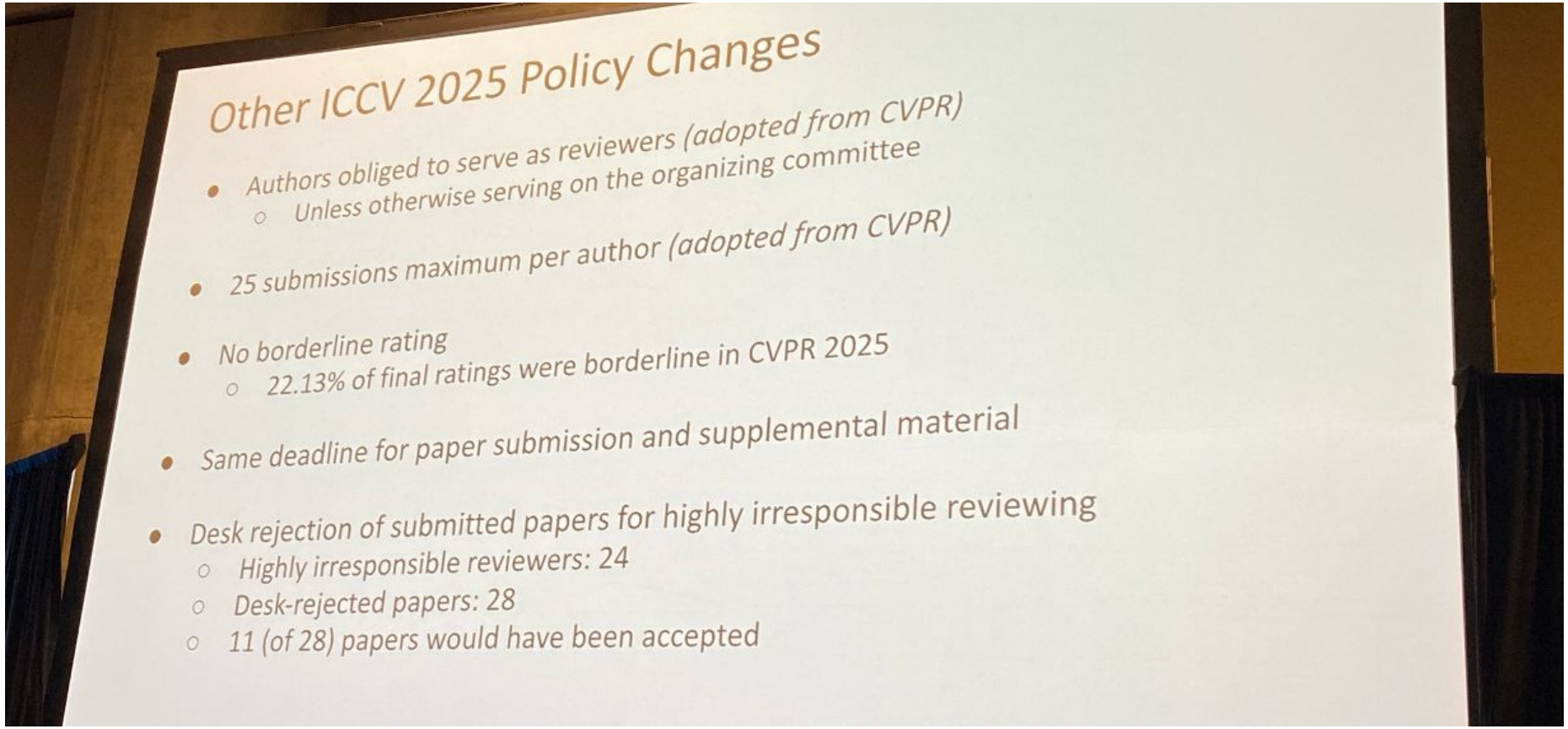
- Reviewer workload
  - 4 papers or less (1 – 3) per reviewer
  - 75% has answered “about right”



ICCV25: Meta Insights into Trends and Tendencies (22/153)

From opening slide

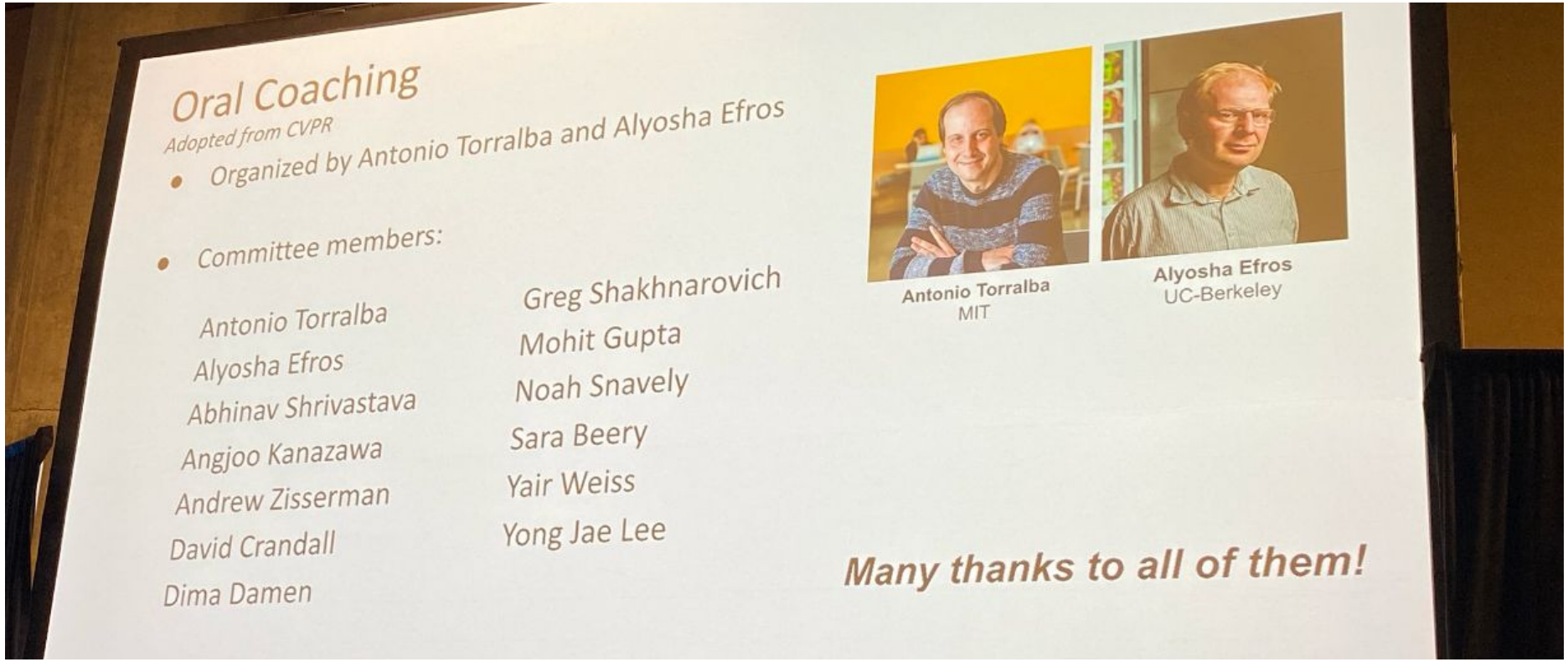
- Policy changes
  - Some policy has come from CVPR



ICCV25: Meta Insights into Trends and Tendencies (23/153)

From opening slide

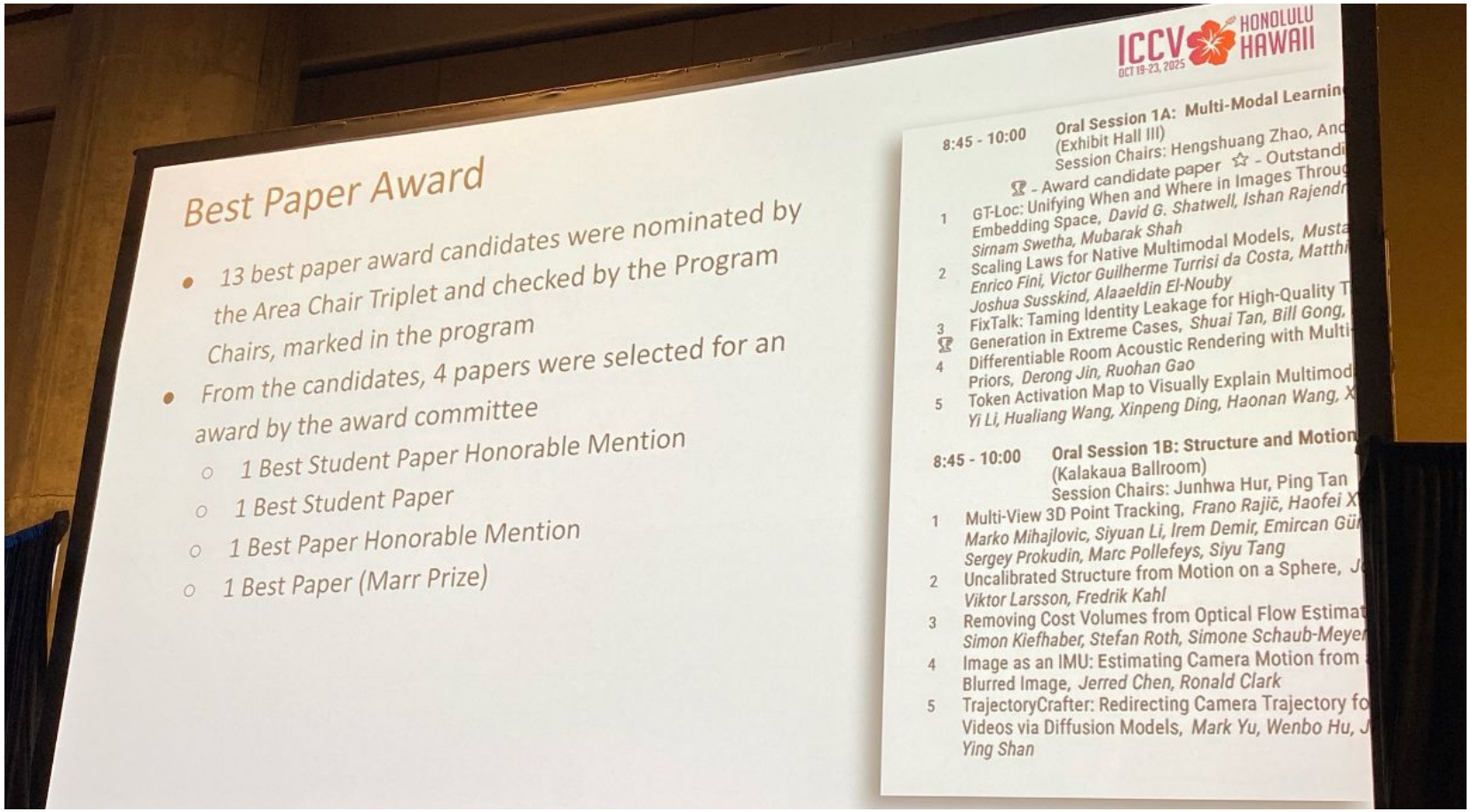
- Oral coaching
  - Oral presentation should be further improved!
  - Led by A. Torralba & A. Efros



ICCV25: Meta Insights into Trends and Tendencies (24/153)

From opening slide

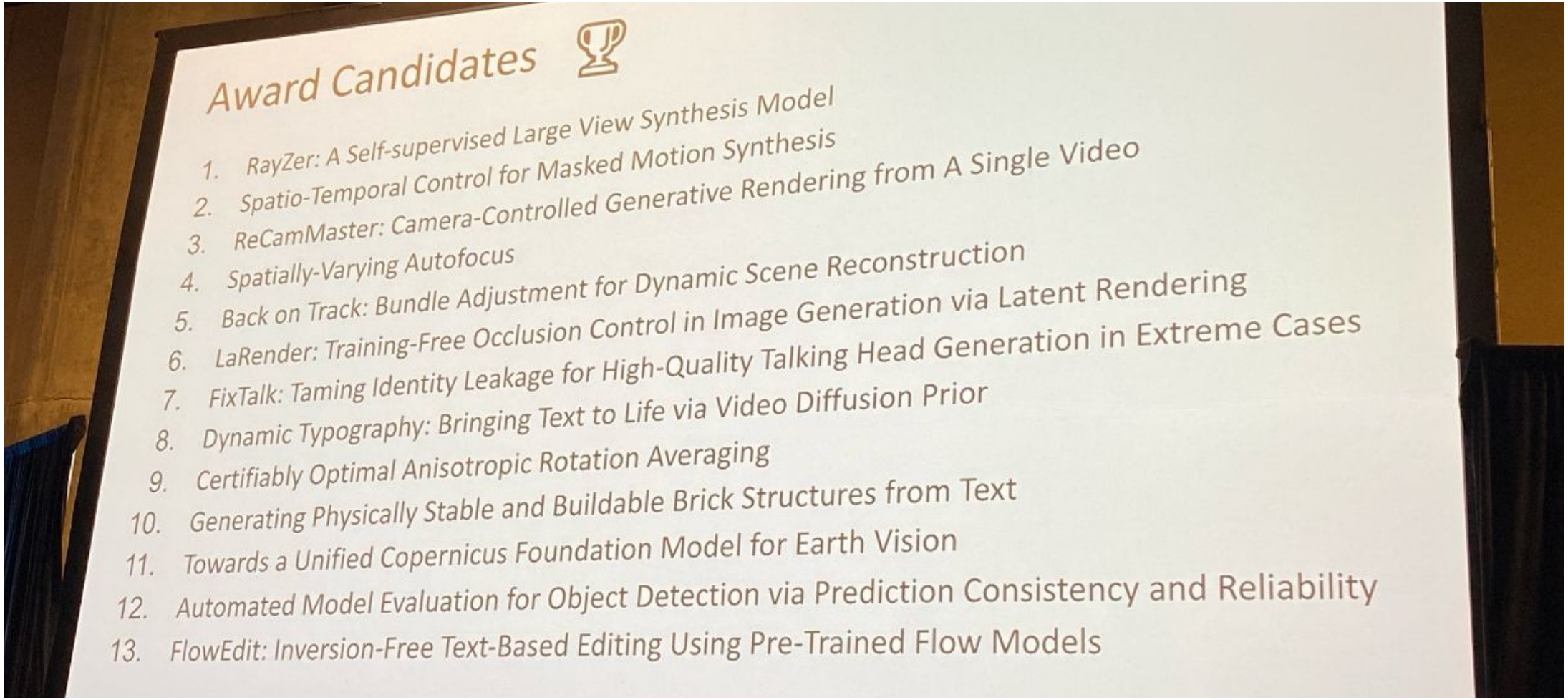
- Best Paper Award at ICCV 2025
  - 13 candidates -> 4 awards
  - Student BPHM, Student BP, BPHM, BP (Marr Prize)



ICCV25: Meta Insights into Trends and Tendencies (25/153)

From opening slide

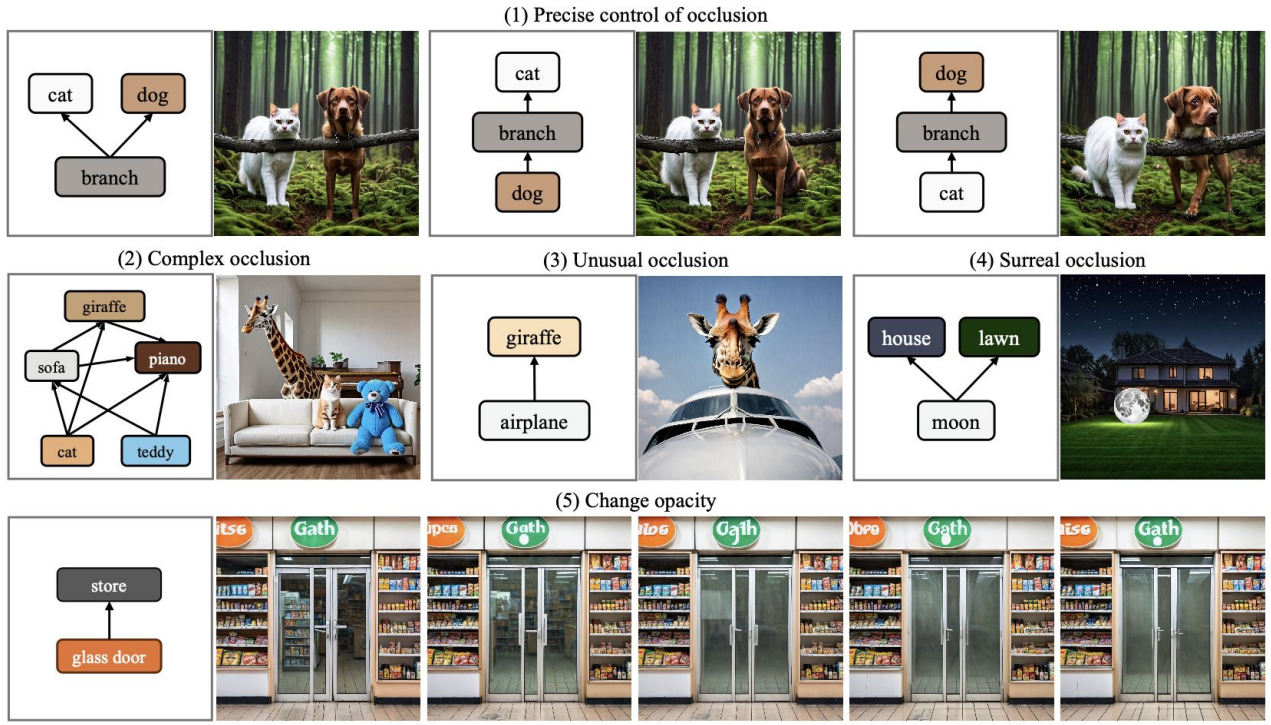
- Award candidates

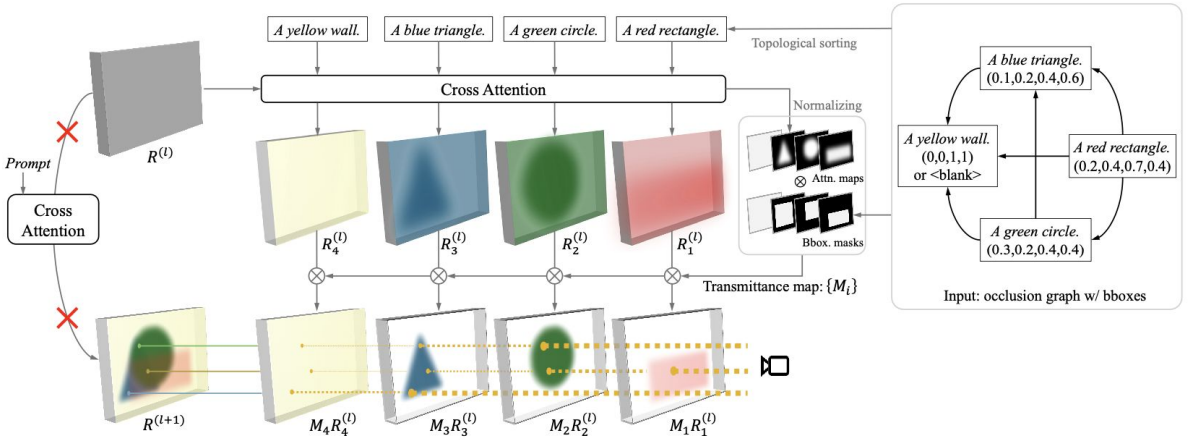


ICCV25: Meta Insights into Trends and Tendencies (26/153)

Summary of best paper candidates 1: LaRender: Training-Free Occlusion Control in Image Generation via Latent Rendering

- Training-free text-to-image rendering to control fine-grained object occlusions with latent space & pre-trained diffusion models
- This method leverages ‘volume rendering’ which has no additional training and accurate occlusion control

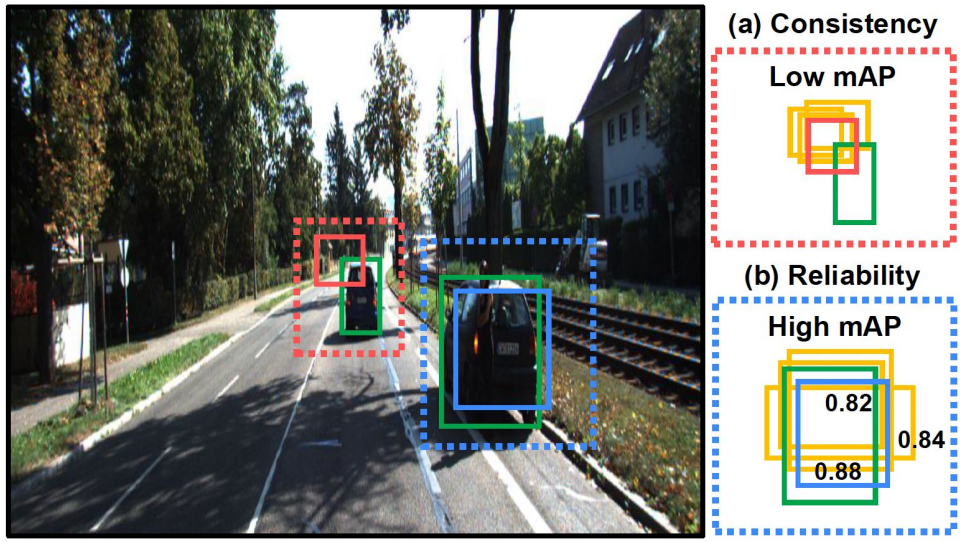


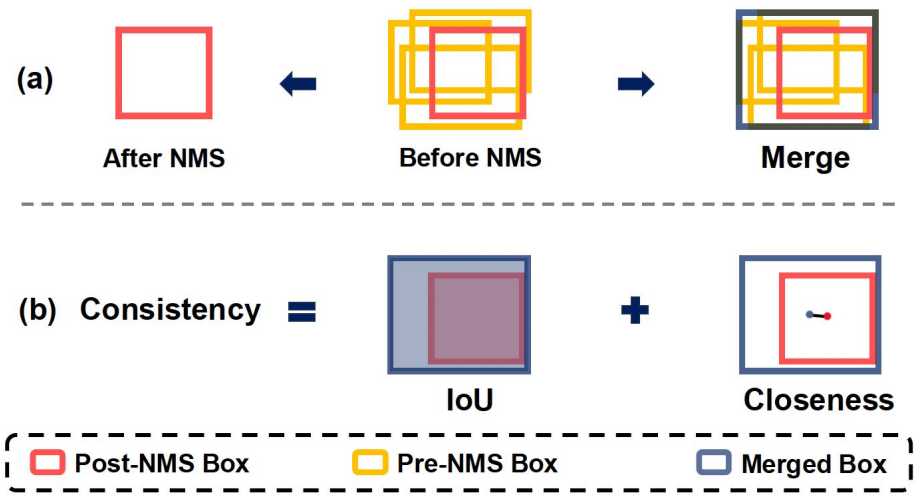


https://xiaohangzhan.github.io/projects/larender/

Summary of best paper candidates 2: Automated Model Evaluation for Object Detection via Prediction Consistency and Reliability

- Extending AutoEval approaches into object detection topic (prior to this work, this approach is quite limited in image classification)
- This evaluation considers two metrics in consistency (how to fix low mAP bboxes) and reliability (how to find & assign high mAP bboxes) for non-maximum suppression



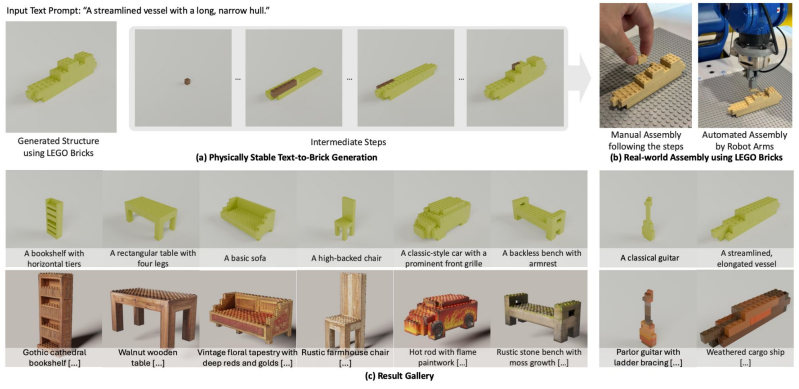


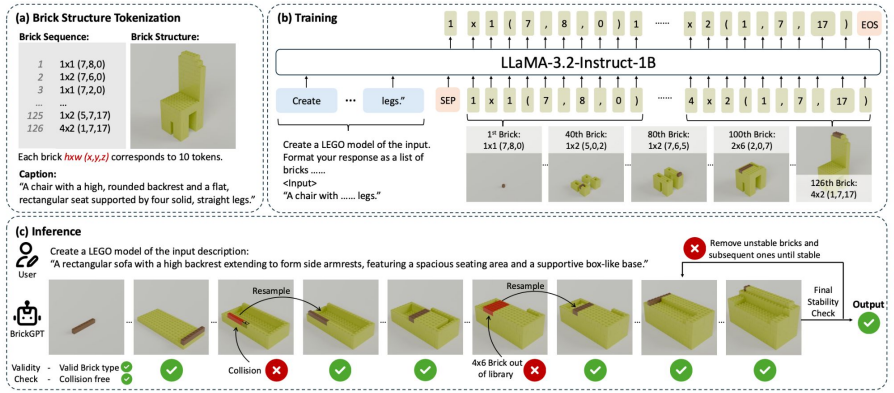
ICCV25: Meta Insights into Trends and Tendencies (28/153)

Summary of best paper candidates 3: Generating Physically Stable and Buildable Brick Structures from Text

https://avalovelace1.github.io/BrickGPT/

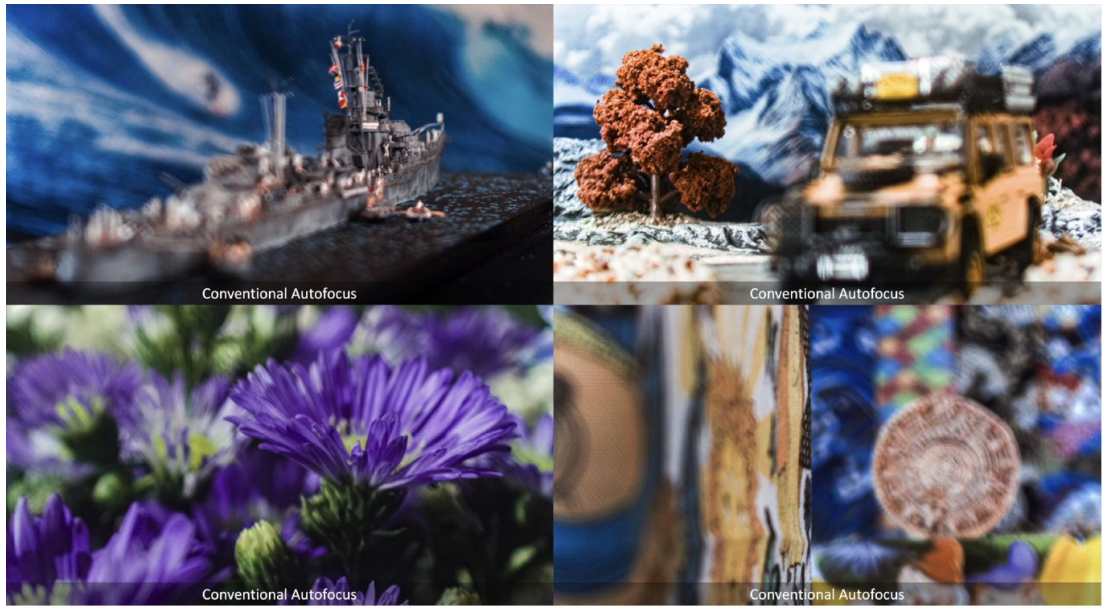
- BrickGPT the first autoregressive large language model to generate physically stable interconnecting bricks assembly models from text.
  - Trained on large-scale brick structures along with their captions. Core taken from ShapeNet.
- Pre-train – base model LLaMA-3.2. Fine-tune using LEGO model of captions.
  - Included physical structural stability on autoregressive inference.
- Showed better stability and validity metrics against LLaMA, LLaMA-Mesh, XCube, Hunyuan.





Summary of best paper candidates 4: Spatially-Varying Autofocus

- Beyond the conventional cameras, this paper introduces & develops multi-focus & depth camera system
- Programmable/computational Split-Lohmann lens and spatially-varying autofocus algorithm enables to capture all-in-focus at 21 fps without post-processing

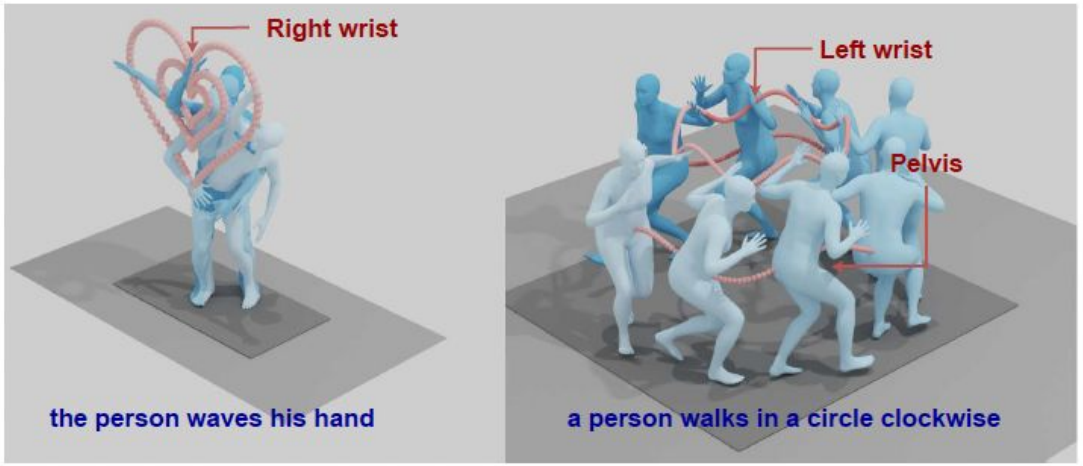


https://imaging.cs.cmu.edu/svaf/

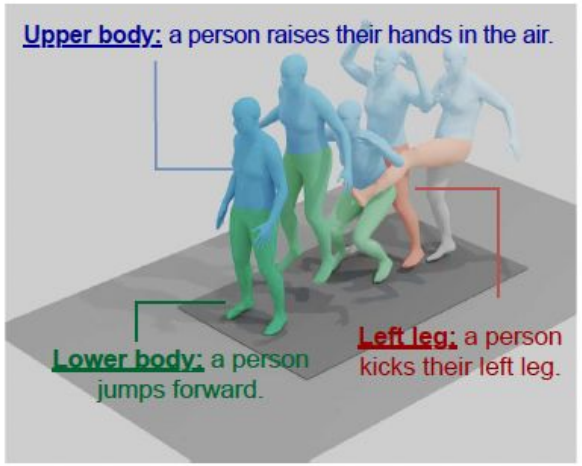
ICCV25: Meta Insights into Trends and Tendencies (30/153)

Summary of best paper candidates 5: MaskControl: Spatio-Temporal Control for Masked Motion Synthesis

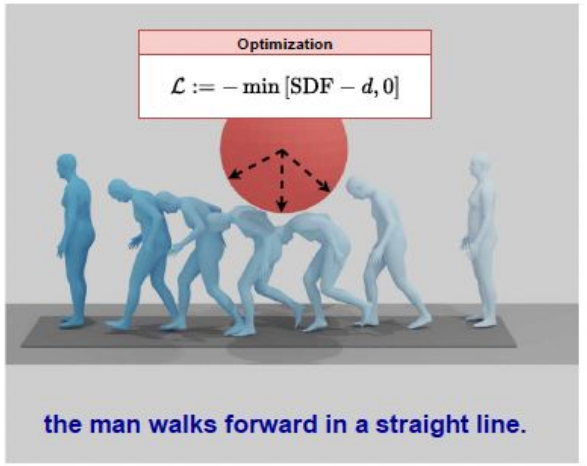
- MaskControl, a controllable 3D human motion, with masked motion models in text-to-motion model
- Logits regularizer extends the motion distribution in training phase
- Logit optimization optimizes the predicted logits in inference phase
- MaskControl enables to conduct any-to-any/zero-shot control



(a) Any-Joint-Any-Frame Control



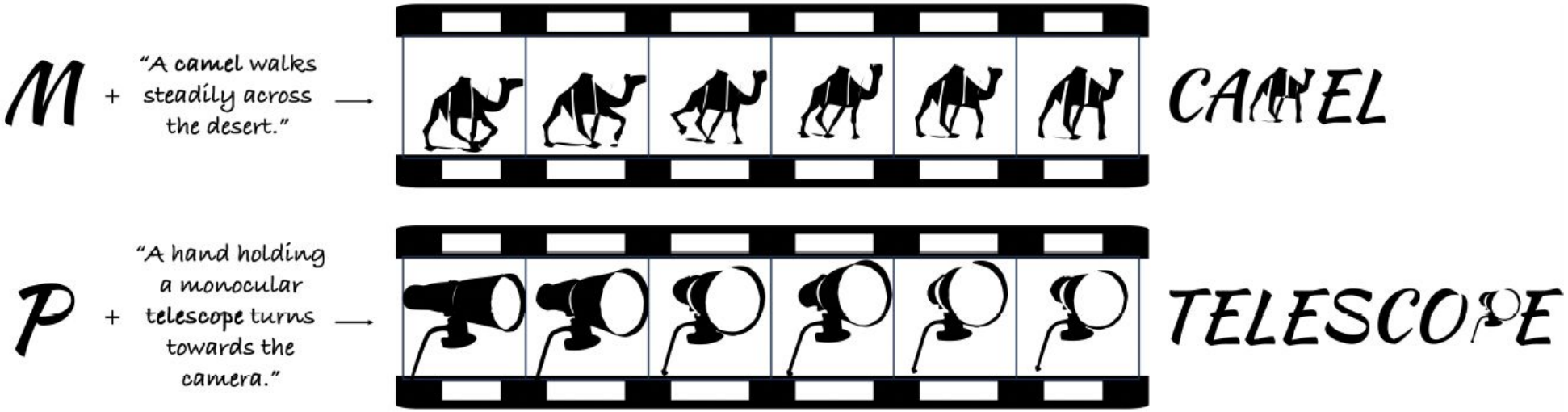
(b) Body-Part Timeline Control



(c) Zero-shot Objective Control

Summary of best paper candidates 6: Dynamic Typography: Bringing Text to Life via Video Diffusion Prior

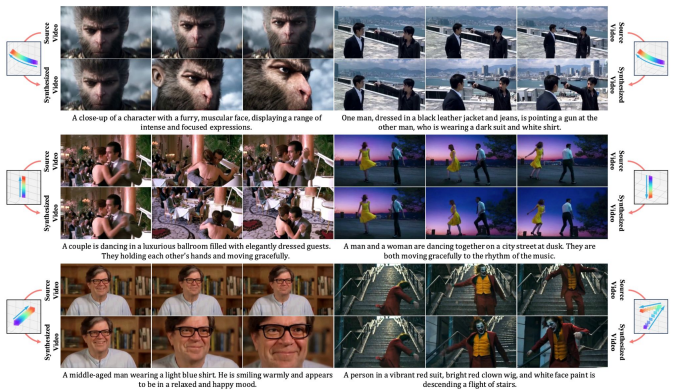
- From given letter & text prompt, Dynamic Typography automatically generates an animation to the letter by aligning the text semantics
- The text–to–video framework is literally trained in end–to–end, treated as a vector format, and preserves base letter shapes

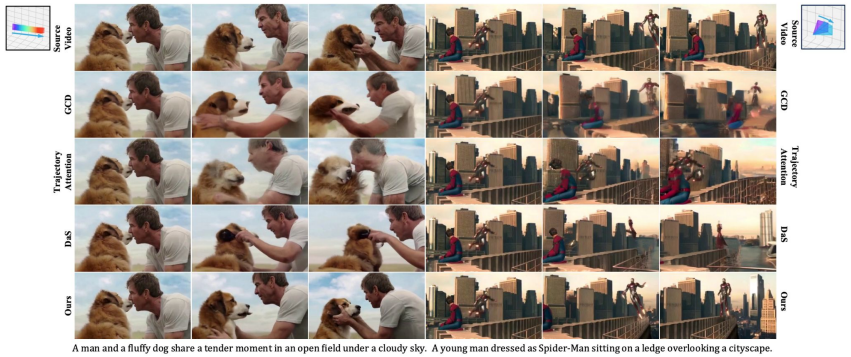


ICCV25: Meta Insights into Trends and Tendencies (32/153)

Summary of best paper candidates 7: ReCamMaster: Camera-Controlled Generative Render from a Single Video

- ReCamMaster a camera controlled generative video re-rendering framework that reproduces the dynamic scene of an input video at novel camera trajectories.
  - New trajectories are performed using generative capabilities of pre-trained text-to-video models.
  - Dataset created on Unreal Engine 5 curated to follow real-world filming (136k realistic videos).
- Better scores on metrics on camera acc., source-target synchronization and visual quality compared to GCD, Trajectory-Attention and DaS.
  - Method applicable to video stabilization, super-resolution and video outpainting.



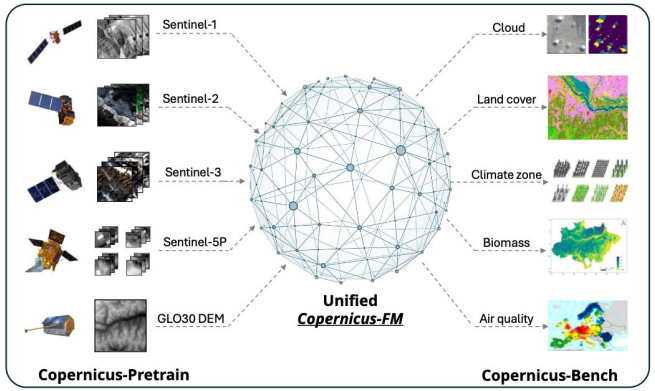


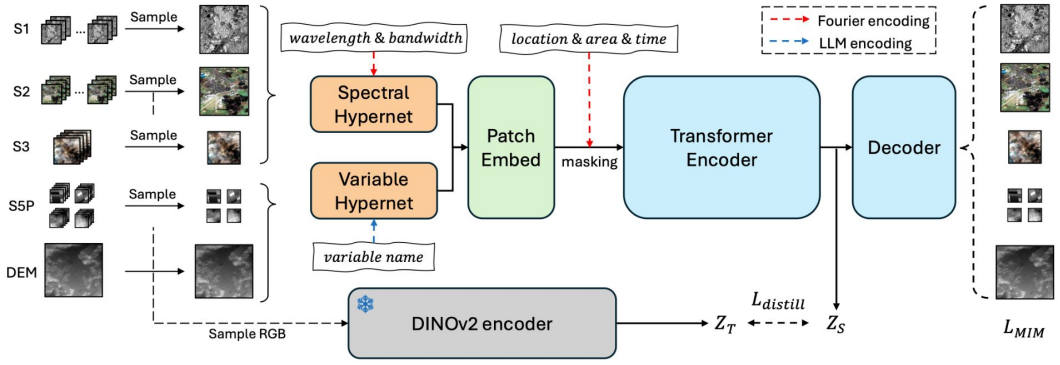
https://github.com/KwaiVGI/ReCamMaster

ICCV25: Meta Insights into Trends and Tendencies (33/153)

Summary of best paper candidates 8: Towards a Unified Copernicus Foundation Model of Earth Vision

- Copernicus-Pretrain massive dataset 18.7M surface and atmosphere aligned images.
- CopernicusFM unified foundation model that process any spectral or non-spectral sensor using hyper networks and flexible metadata providing better accuracy on k-NN classification and segmentation Earth Observation tasks.
  - It utilizes a dynamic patch embedding layer that patchifies the input image into a sequence of patch tokens.
- CopernicusBench a comprehensive evaluation benchmark with 15 downstream tasks.

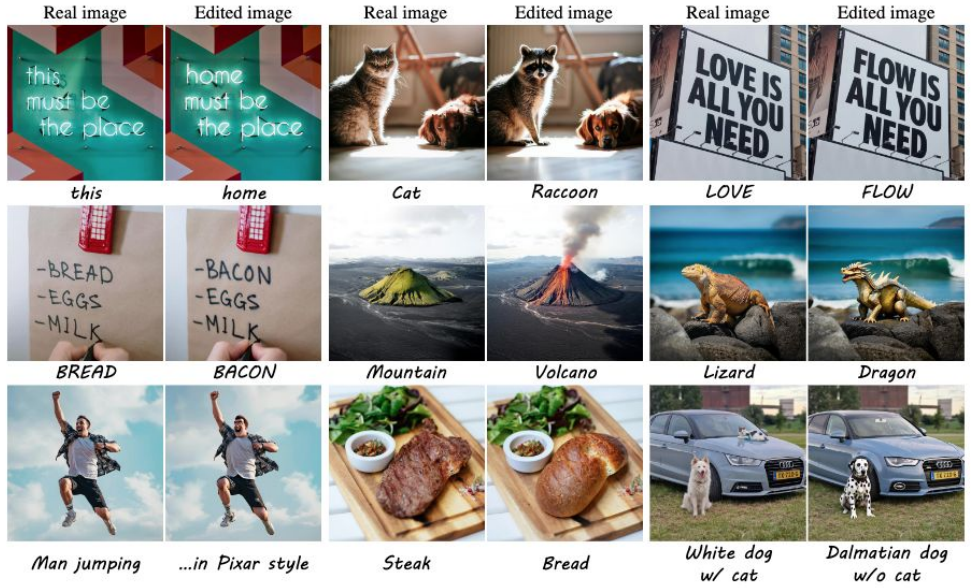


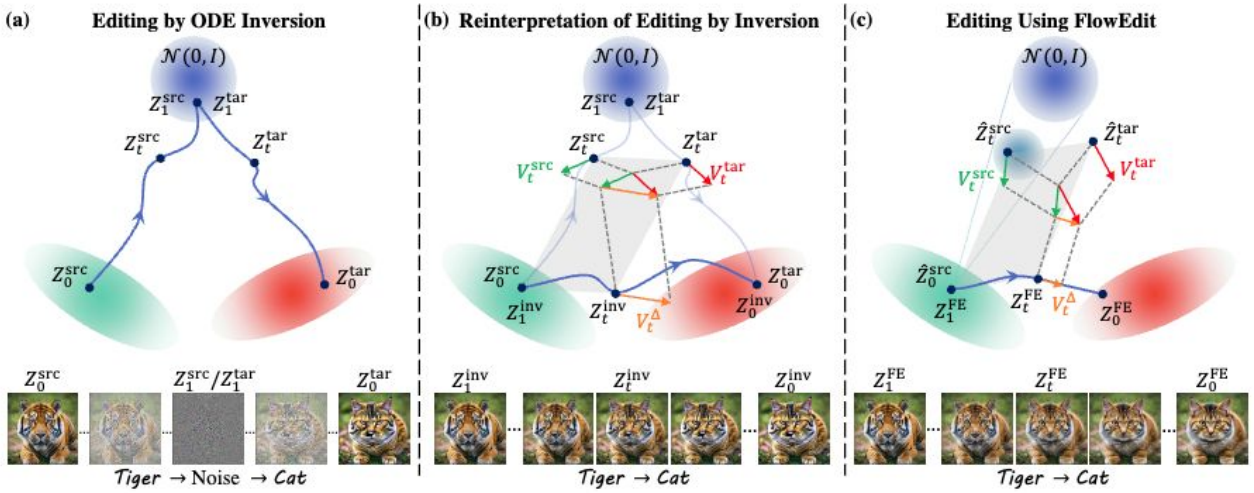


ICCV25: Meta Insights into Trends and Tendencies (34/153)

Summary of best paper candidates 9: FlowEdit: Inversion-Free Text-Based Editing Using Pre-Trained Flow Models

- FlowEdit, text-based editing method for pre-trained T2I flow models (e.g., Stable Diffusion 3, Flux)
- Constructs ODE that directly maps between the source and target distributions (without inversion from image to initial noise)
  - Archives lower transport costs than the inversion approach

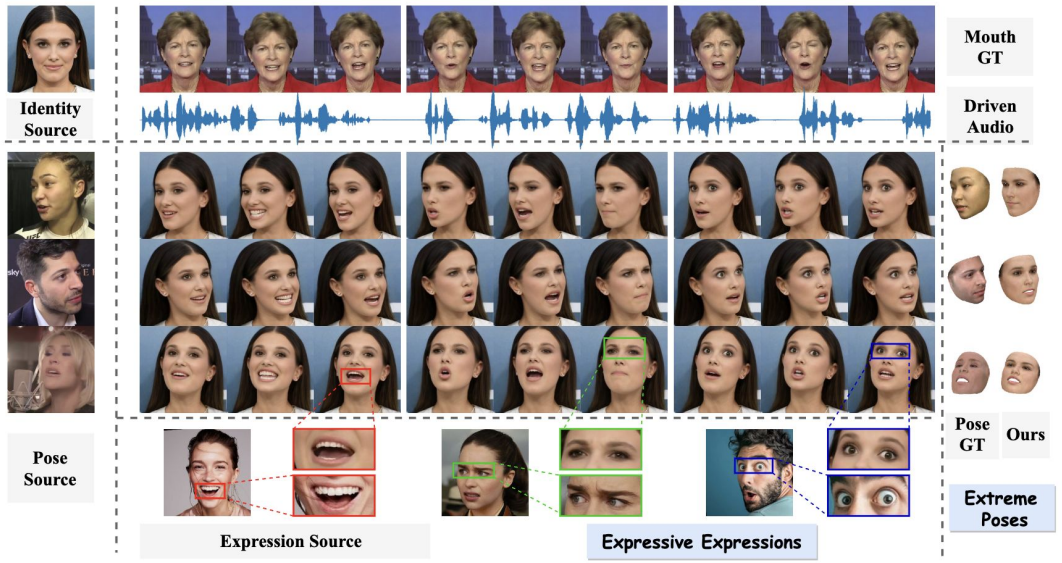


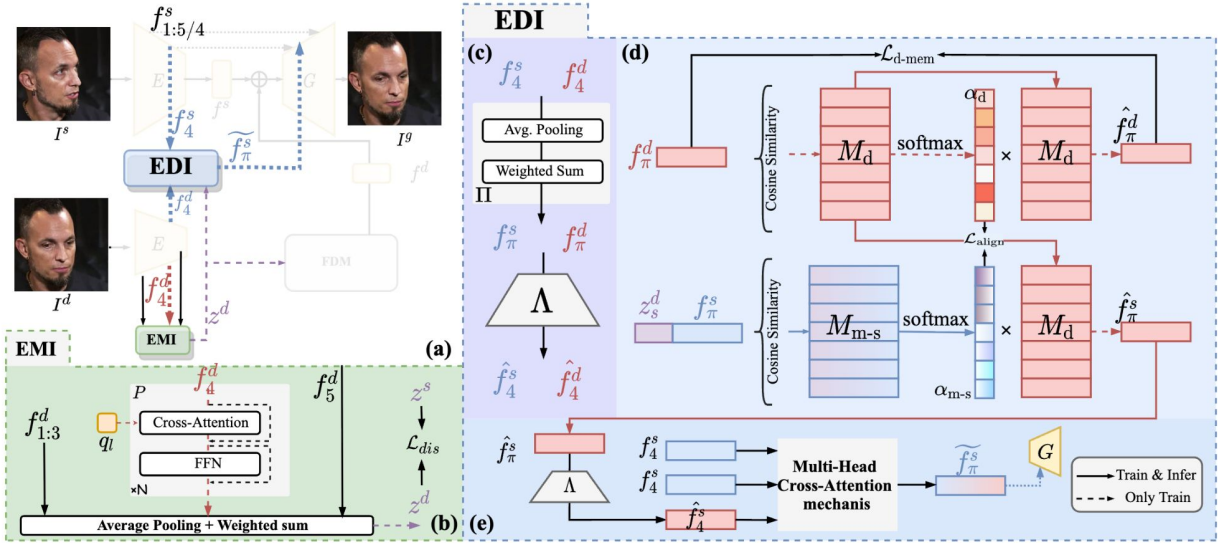


ICCV25: Meta Insights into Trends and Tendencies (35/153)

Summary of best paper candidates 10: FixTalk: Taming Identity Leakage for High-Quality Talking Head Generation in Extreme Cases

- FixTalk, which facilitates talking head generation to achieve three critical goals: System Efficiency, Decoupled Control, and High-Quality Rendering
- EMI and EDI, which tame the identity leakage for high-quality talking head generation

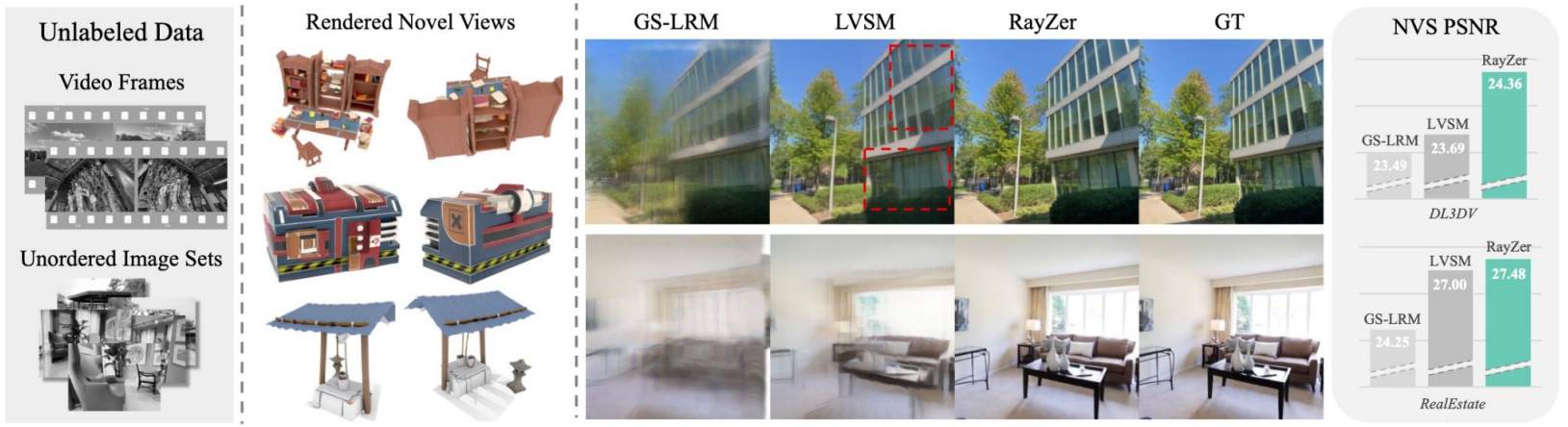


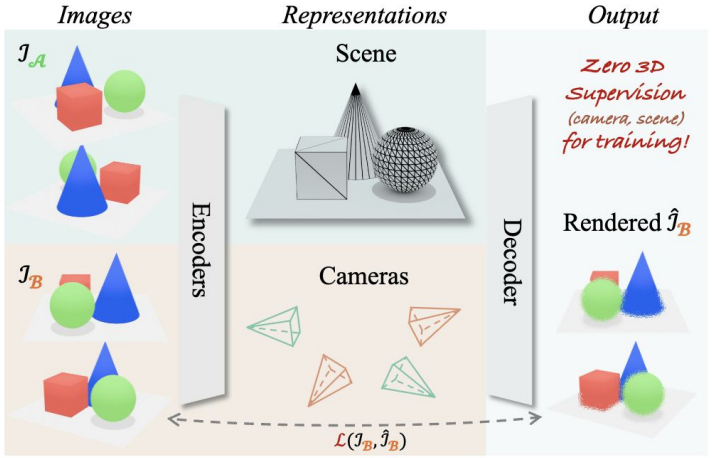


ICCV25: Meta Insights into Trends and Tendencies (36/153)

Summary of best paper candidates 11:
RayZer: A Self-supervised Large View Synthesis Model

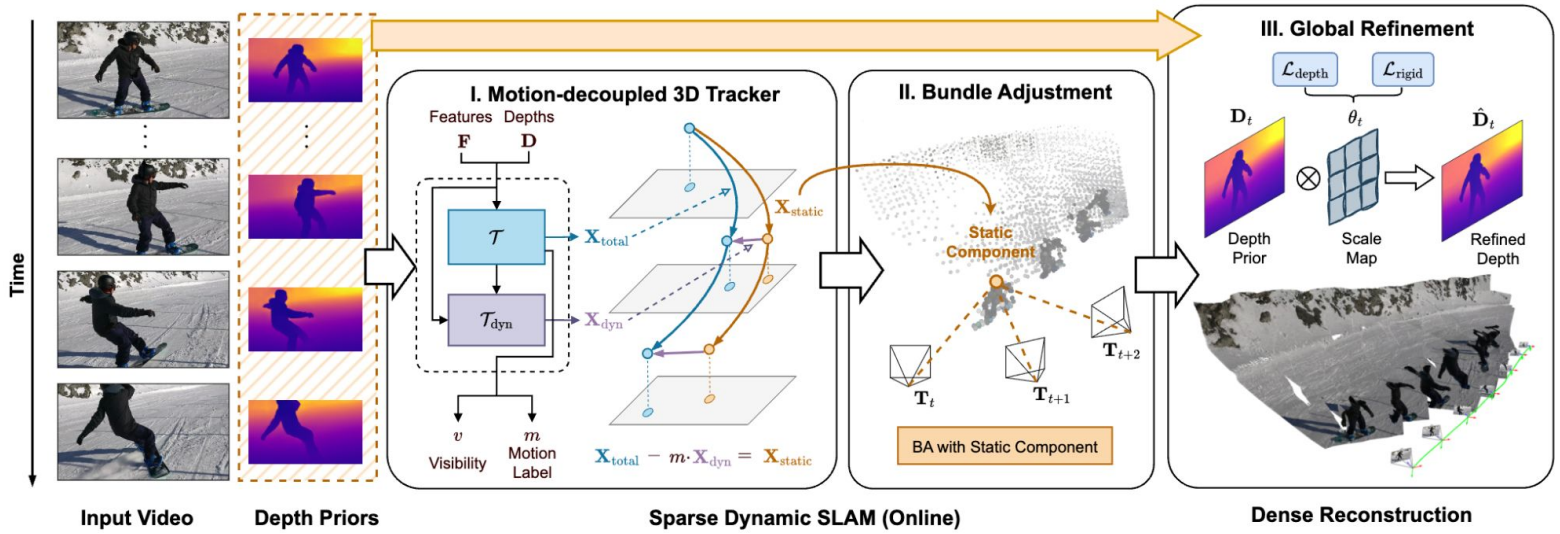
- RayZer, a self-supervised (SSL) multi-view 3D Vision model trained without any 3D supervision (i.e., camera poses and scene geometry, etc.)
- This method is attributed to two key factors:
  - SSL framework by disentangling camera and scene representations
  - Transformer based model in which the only the ray structure





Summary of best paper candidates 12: Back on Track: Bundle Adjustment for Dynamic Scene Reconstruction

- BA-Track, a framework that jointly reconstructs static and dynamic scene geometry from casual video sequences
  - Decoupling the camera-induced motion from the observed (total) motion
  - Bundle adjustment back-end for accurate pose and depth estimation
  - Global refinement to achieve dense, scale-consistent depth



ICCV25: Meta Insights into Trends and Tendencies (38/153)

Summary of best paper candidates 13: Certifiably Optimal Anisotropic Rotation Averaging

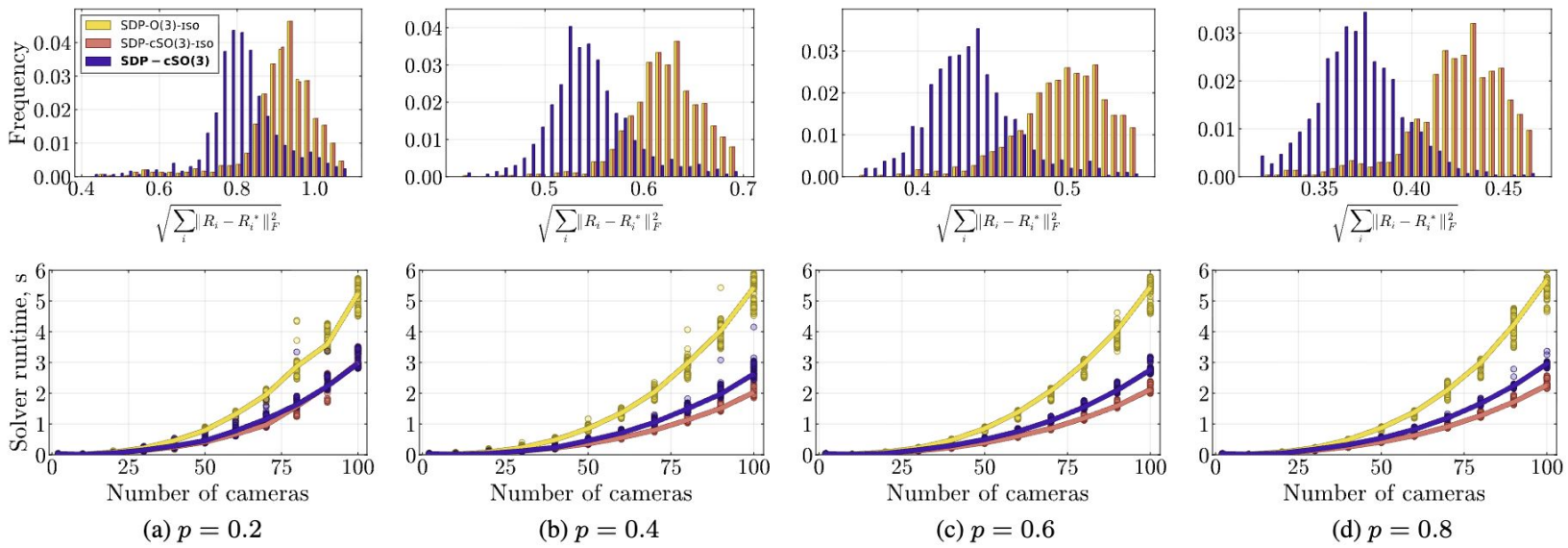
- This study show how anisotropic costs can be incorporated in certifiably optimal rotation averaging.
- It also demonstrates how existing solvers, designed for isotropic situations, fail in the anisotropic setting

If $X_{ii} \neq \mathtt{I}$ it is clear that the maximum over Υ_{ii} will be unbounded. The second term is $\mathcal{I}^{**}_{SO(3)}(X_{ij})$, which is the convex envelope of the indicator function $\mathcal{I}_{SO(3)}$. This is also the indicator function of $\mathrm{conv}(SO(3))$ [37]. Therefore the bidual program—and consequently our proposed relaxation—is given by

$$\begin{aligned} \min_{\mathbf{X}\succeq 0} \; & -\mathrm{tr}(\mathbf{N}\mathbf{X}) \\ \text{s.t. } & X_{ii} = \mathtt{I}, \; X_{ij} \in \mathrm{conv}(SO(3)). \end{aligned} \qquad \text{(SDP-cSO(3))}$$

The constraint $X_{ij} \in \mathrm{conv}(SO(3))$ has been shown to be equivalent to a semidefinite constraint [40, 41]. A 3×3 matrix Y is in $\mathrm{conv}(SO(3))$ if and only if $\mathcal{A}(Y) + \mathtt{I} \succeq 0$, where $\mathcal{A}(Y) =$

$$\begin{pmatrix} -Y_{11}-Y_{22}+Y_{33} & Y_{13}+Y_{31} & Y_{12}-Y_{21} & Y_{23}+Y_{32} \\ Y_{13}+Y_{31} & Y_{11}-Y_{22}-Y_{33} & Y_{23}-Y_{32} & Y_{12}+Y_{21} \\ Y_{12}-Y_{21} & Y_{23}-Y_{32} & Y_{11}+Y_{22}+Y_{33} & Y_{31}-Y_{13} \\ Y_{23}+Y_{32} & Y_{12}+Y_{21} & Y_{31}-Y_{13} & -Y_{11}+Y_{22}-Y_{33} \end{pmatrix}.$$



ICCV25: Meta Insights into Trends and Tendencies (39/153)

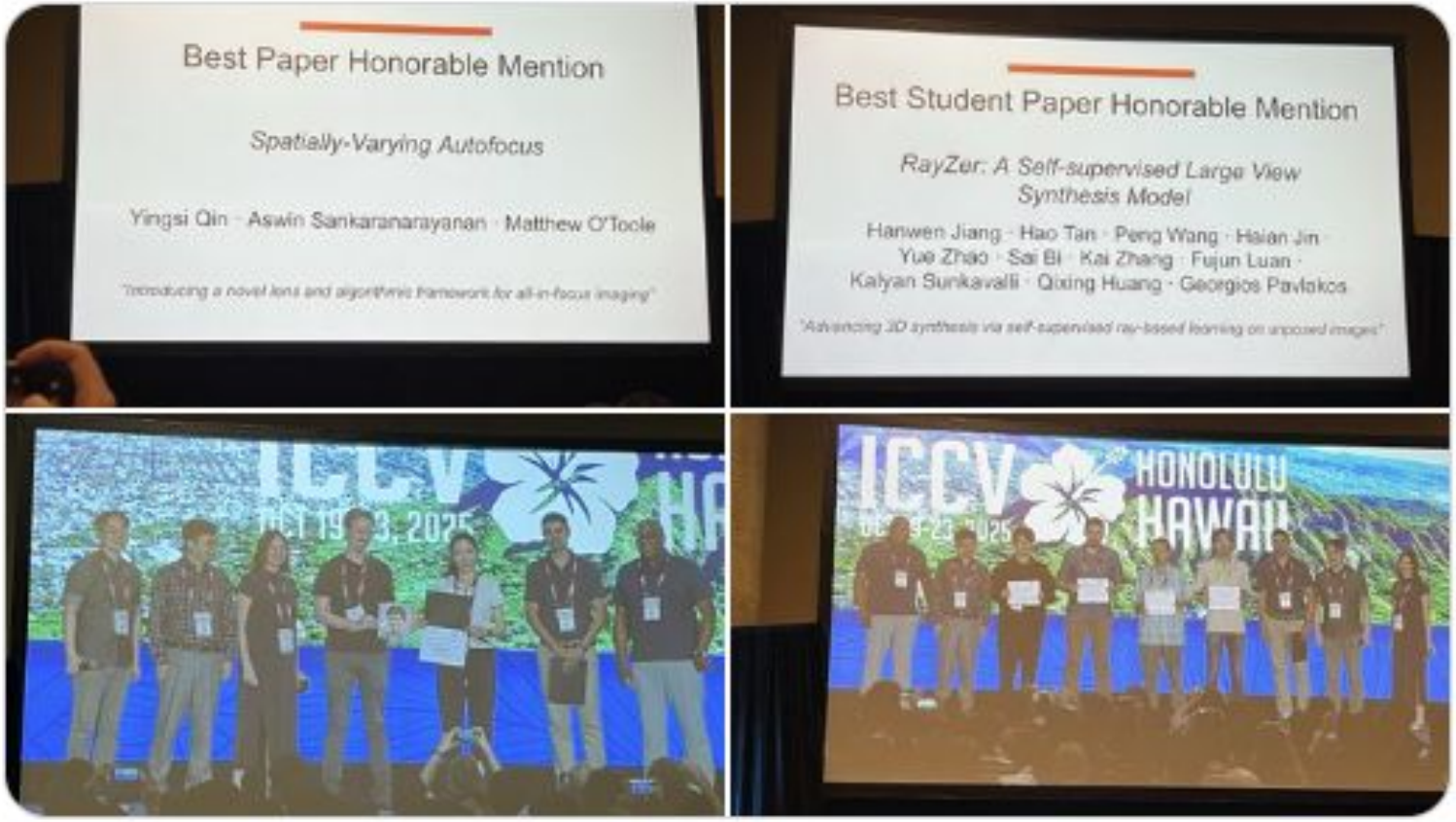
From opening slide

- Best Paper Awards

ICCV25: Meta Insights into Trends and Tendencies (40/153)

From opening slide

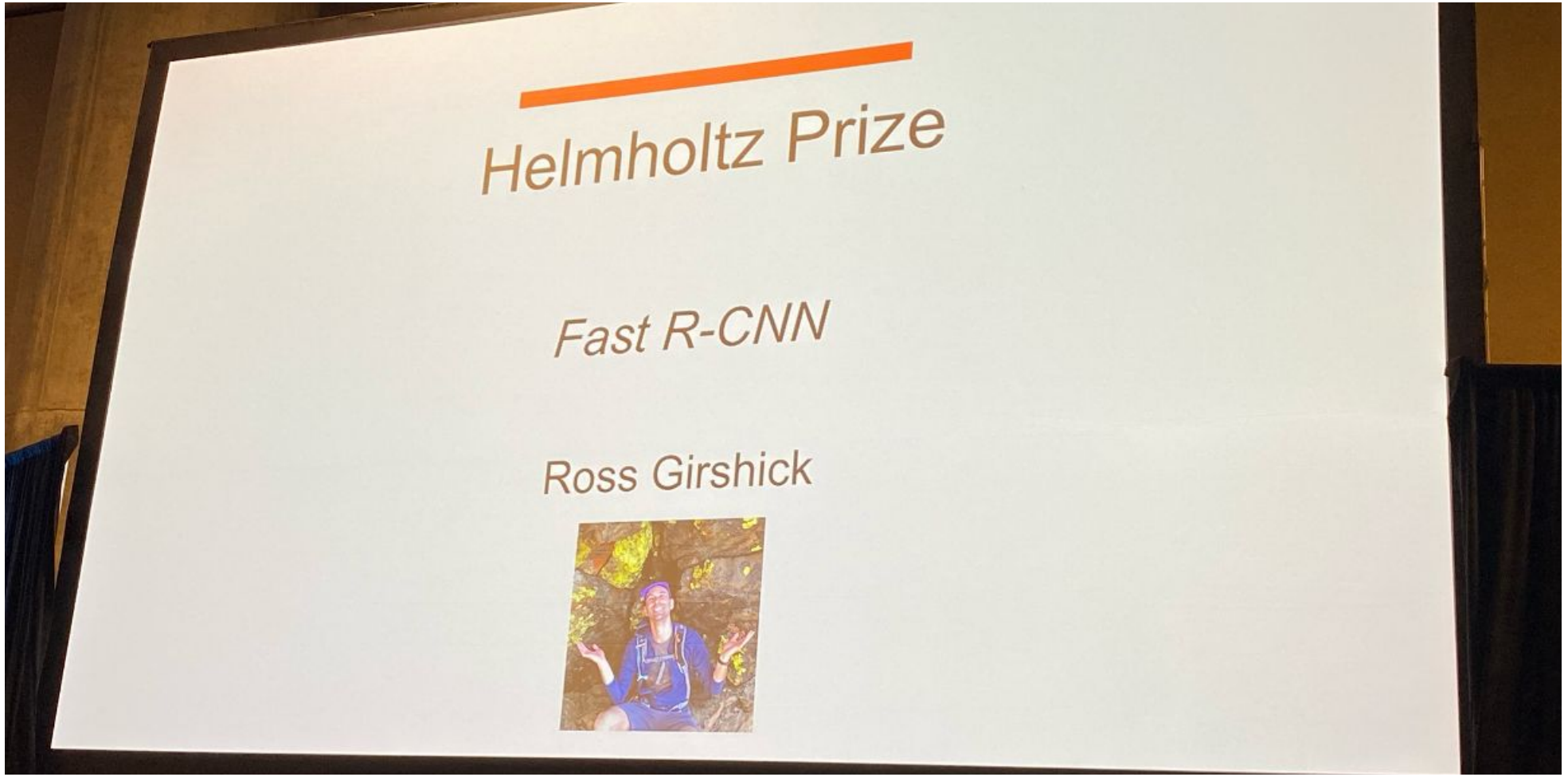
- Best Paper Honorable Mentions



ICCV25: Meta Insights into Trends and Tendencies (41/153)

From opening slide

- Helmholtz prize (test-of-time award)



ICCV25: Meta Insights into Trends and Tendencies (42/153)

From opening slide

- Helmholtz prize (test-of-time award)

ICCV25: Meta Insights into Trends and Tendencies (43/153)

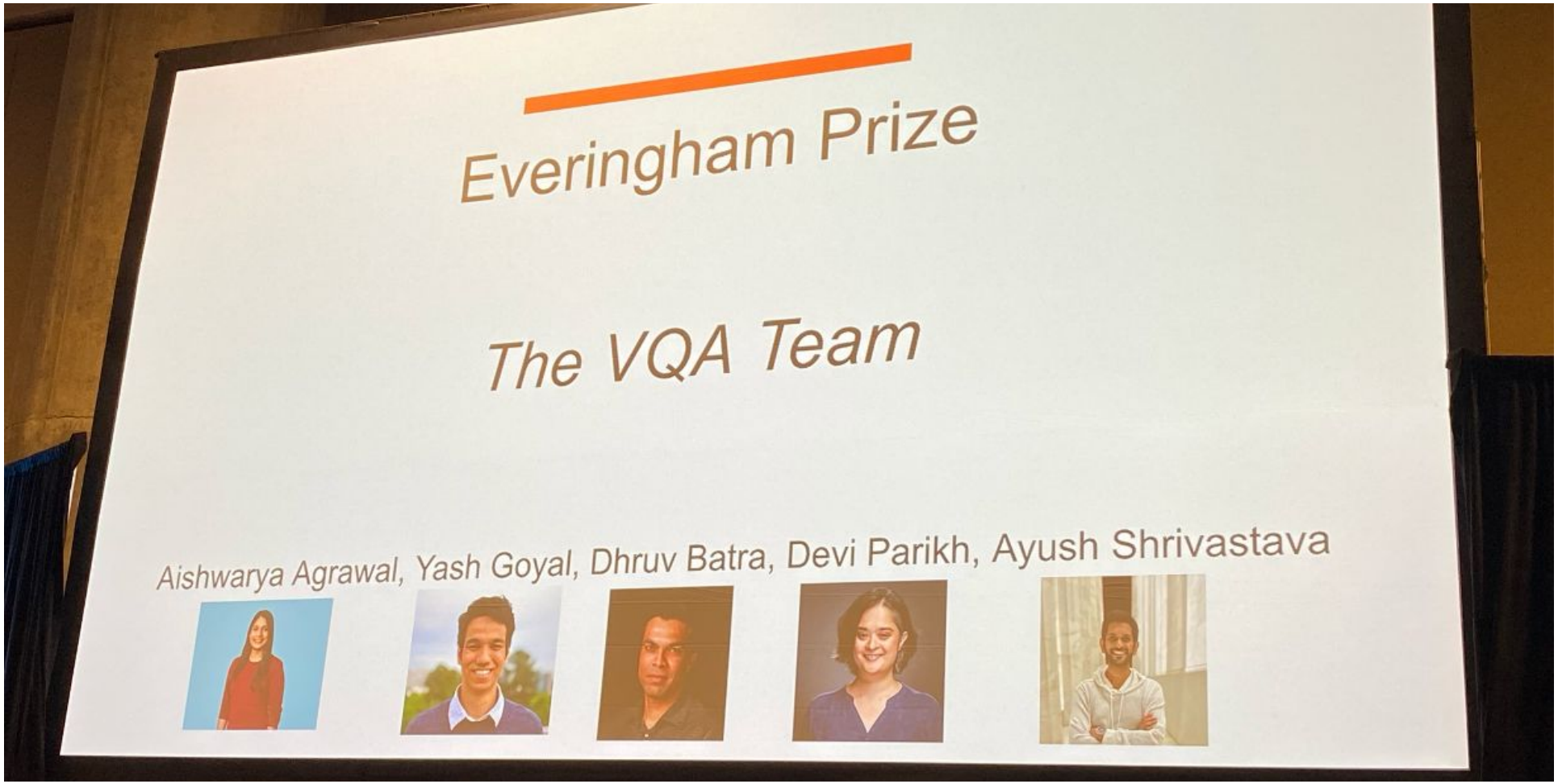
From opening slide

- Everingham prize

ICCV25: Meta Insights into Trends and Tendencies (44/153)

From opening slide

- Everingham prize



ICCV25: Meta Insights into Trends and Tendencies (45/153)

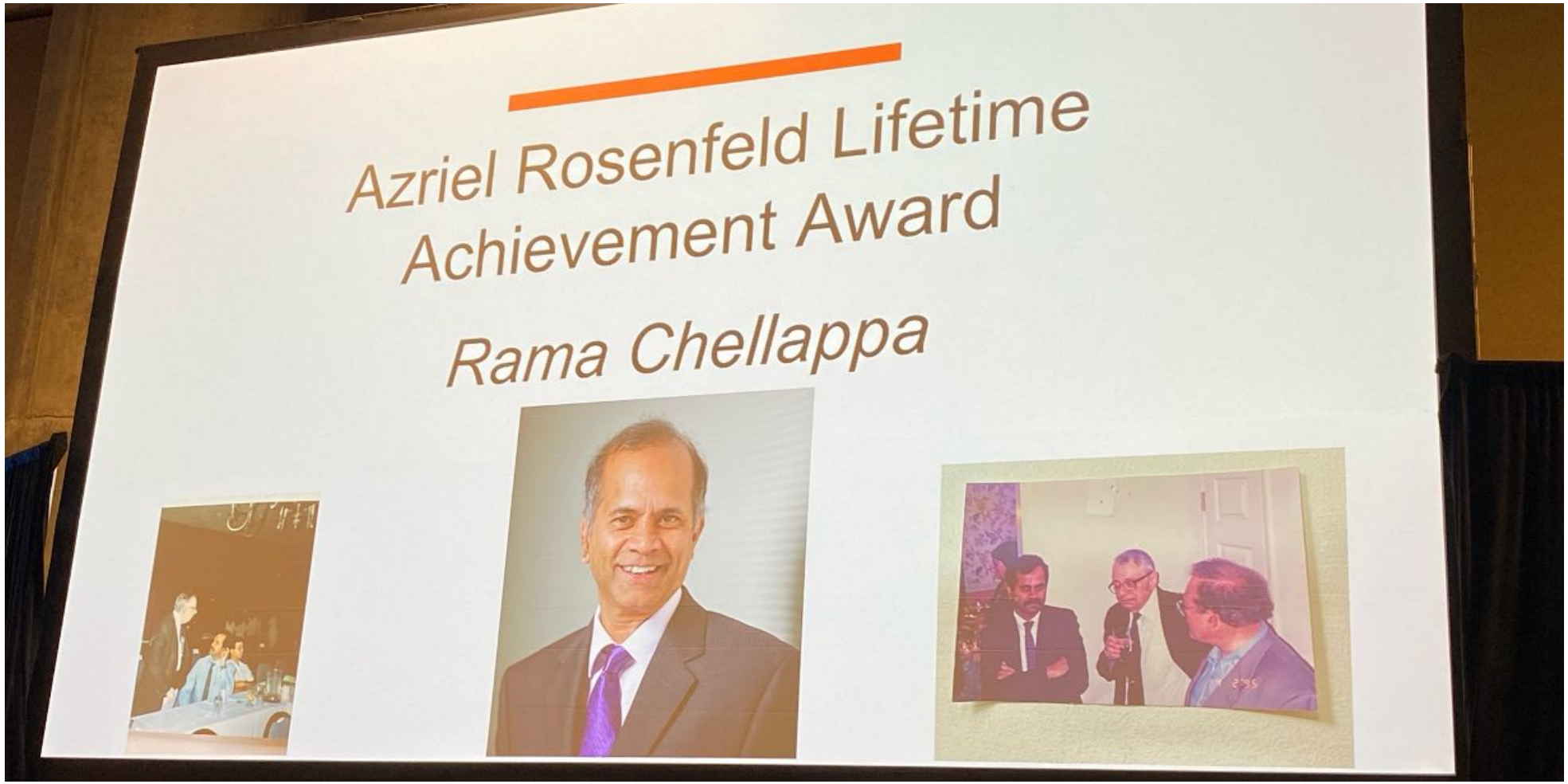
From opening slide

- Distinguished researcher award

ICCV25: Meta Insights into Trends and Tendencies (46/153)

From opening slide

- Azriel Rosenfeld lifetime achievement award



ICCV25: Meta Insights into Trends and Tendencies (47/153)

Unofficial best paper prediction in LIMIT.Lab

- This initiative was organized by a group of volunteers to coincide with the release of the Best Paper Award Candidates for ICCV 2025.
- Independent from the official review process, it aims to predict which paper will receive the award, based on our own perspectives.
- Beyond simple prediction, the goal is to deepen our understanding of the field by carefully reading the nominated papers and evaluating their novelty, impact, and future potential.
- Through discussion and exchanging opinions, the initiative also helps broaden our perspectives and refine our criteria for assessing research.

ICCV25: Meta Insights into Trends and Tendencies (48/153)

Unofficial best paper prediction in LIMIT.Lab

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

ICCV25: Meta Insights into Trends and Tendencies (49/153)

Unofficial best paper prediction in LIMIT.Lab

- Our selected papers for each award
  - Total: 6 papers
  - Best Student Paper Honorable Mention(1)
  - Best Student Paper Award(2)
  - Best Paper Honorable Mention(2)
  - Best Paper Award(1)

ICCV25: Meta Insights into Trends and Tendencies (50/153)

Unofficial best paper prediction in LIMIT.Lab

- Our selected papers for each award and reality
  - Total: 6 → 4 papers
  - Best Student Paper Honorable Mention (1→ 1)
    - ❌ LaRender: Training-Free Occlusion Control in Image Generation via Latent Rendering
  - Best Student Paper Award (2→ 1)
    - ❌ ReCamMaster: Camera-Controlled Generative Rendering from A Single Video
    - ❌ Back on Track: Bundle Adjustment for Dynamic Scene Reconstruction
  - Best Paper Honorable Mention (2→ 1)
    - Towards a Unified Copernicus Foundation Model for Earth Vision
    - ✅ RayZer: A Self-supervised Large View Synthetis Model → But, it's Best Student Paper Honorable Mention
  - Best Paper Award (1→ 1)
    - ✅ Spatially-Varying Autofocus → But, it's Best Paper Honorable Mention

ICCV25: Meta Insights into Trends and Tendencies (51/153)

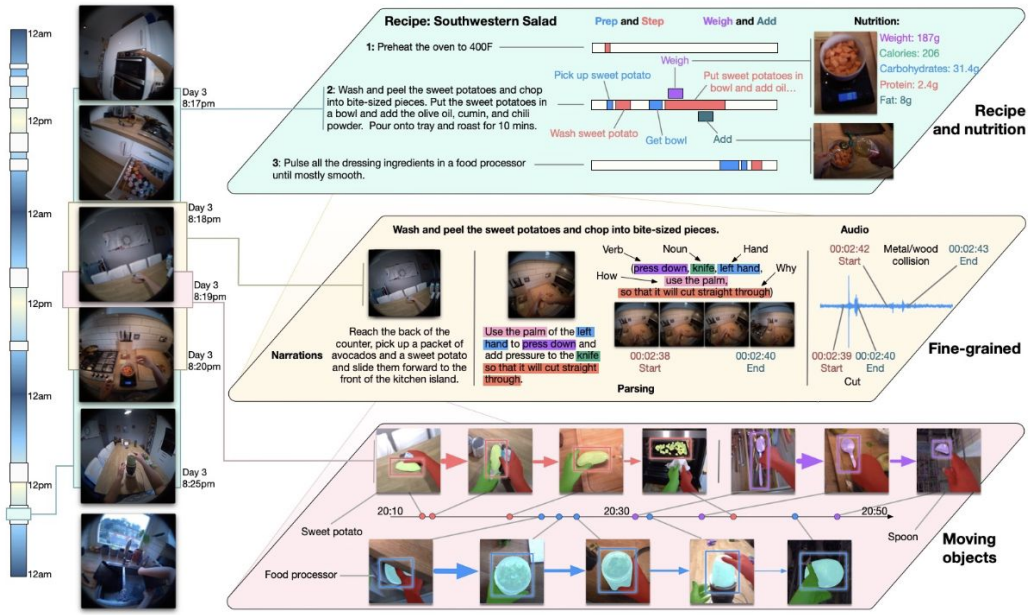
Meta-level insights and BP selected discussion points

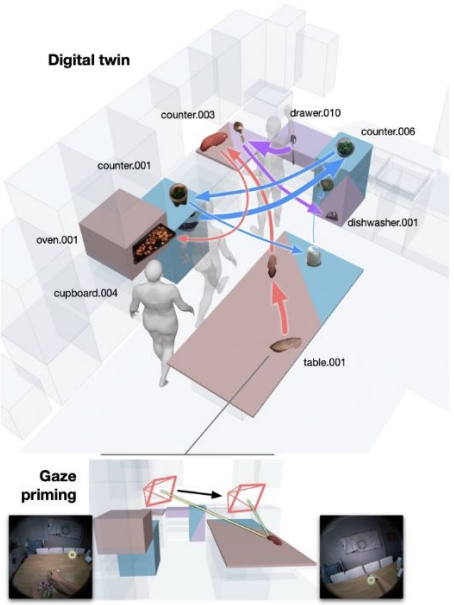
- Robotics implementation was one of the major highlights this year.
- Generative model → “Diffusion” is gradually replaced by “Flow”
- Photography → Sensing should be further improved before your deep learning
- Novel view synthesis → NVS task can be (almost) solved without any supervised / reliable labels

ICCV25: Meta Insights into Trends and Tendencies (52/153)

Summary of The 4th DataCV Workshop(1/3)

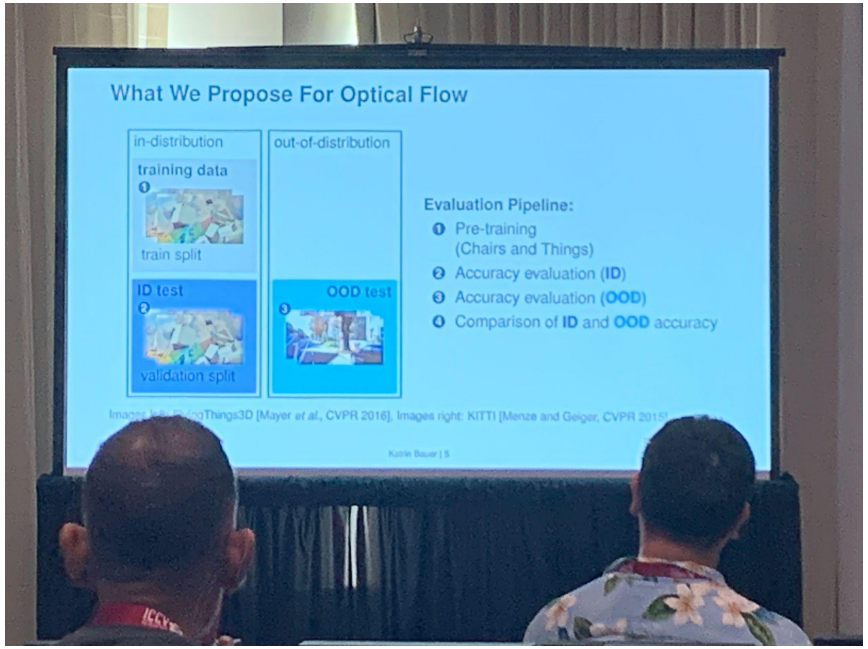
- Invited Talk: Dima Damen
  - HD-EPIC Introduction: A large-scale egocentric dataset captured in unscripted household environments, featuring rich annotations including ego-video, audio and digital-twin data
  - Discussion: Addressed how long-tail distributions should be handled in Data-Centric

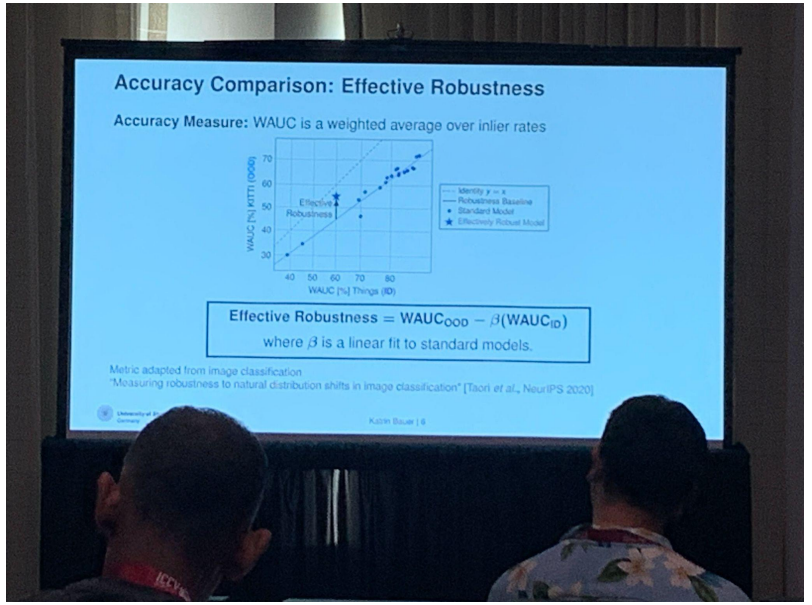




Summary of The 4th DataCV Workshop(2/3)

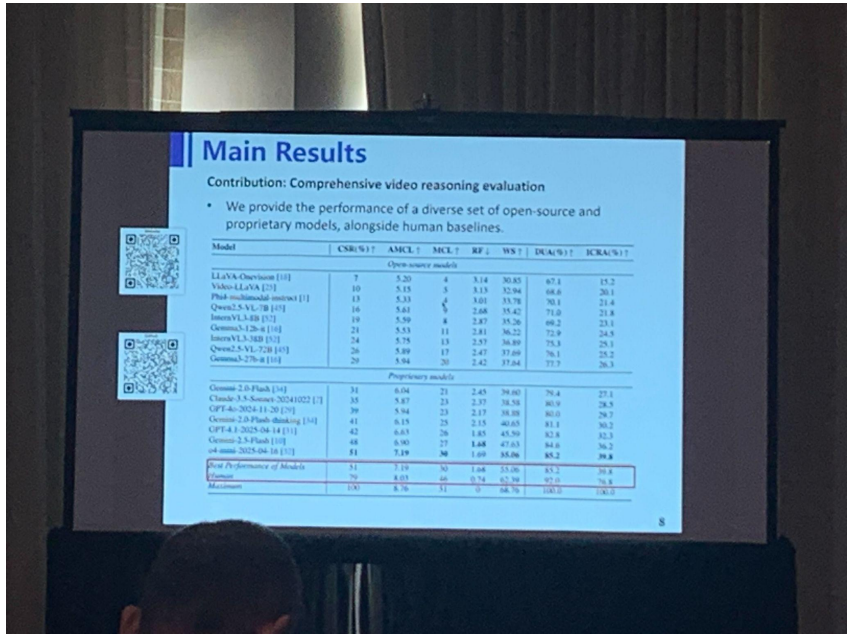
- Oral Paper Highlights
  - On the Generalization of Optical Flow: Quantifying Robustness to Dataset Shifts
    - Goal: Systematically evaluate optical flow models under OOD conditions
    - Findings:
      - Many recent models are not robust to dataset shifts
      - Model size does not correlate with robustness

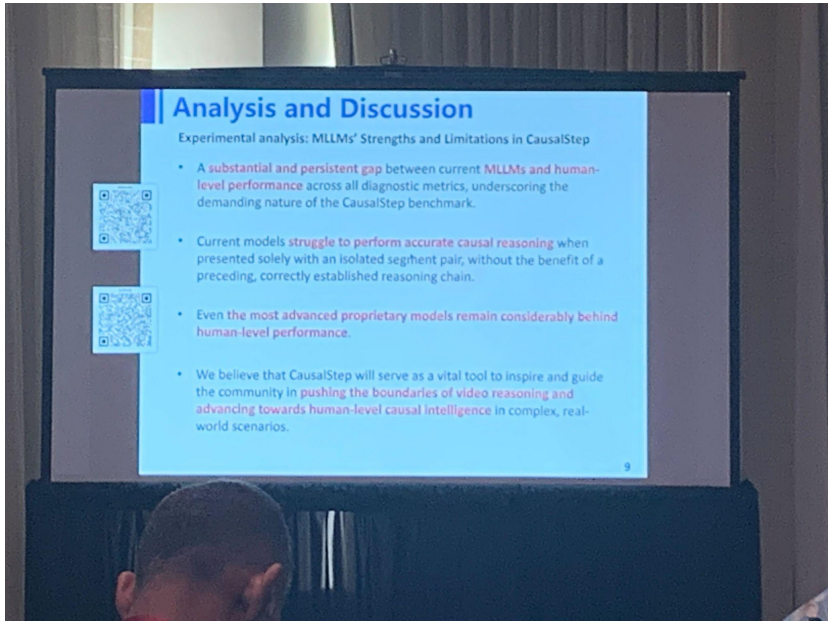




Summary of The 4th DataCV Workshop(3/3)

- Oral Paper Highlights
  - CausalStep: A Benchmark for Explicit Stepwise Causal Reasoning in Videos
    - Problem: Existing video benchmarks rely mainly on global context understanding
    - Proposal: Introduces a new benchmark for step-by-step causal reasoning across video segments
    - Results:
      - Current multimodal language models (MMLs) show weaker causal reasoning than humans
      - GPT-4o-mini and Gemini outperform open-source models

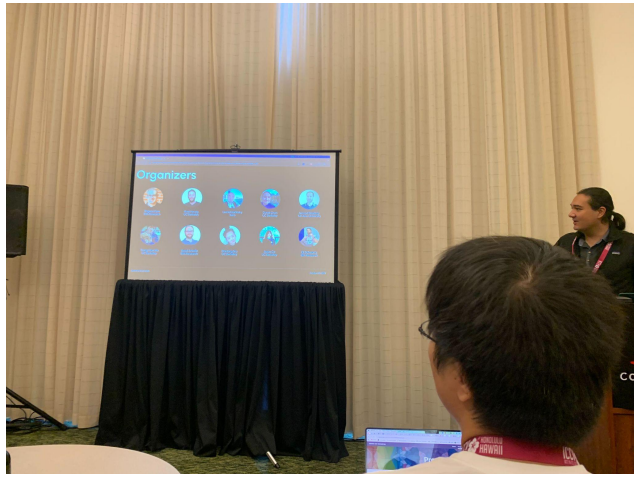


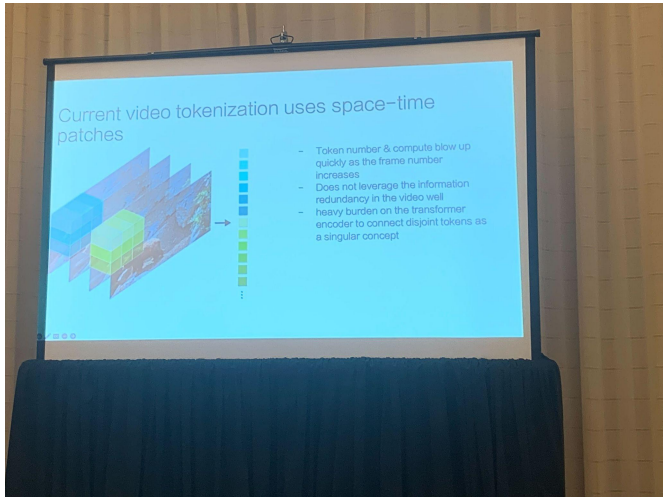


ICCV25: Meta Insights into Trends and Tendencies (55/153)

Highlights of The 4th Workshop on What is Next in Multimodal Foundation Models?

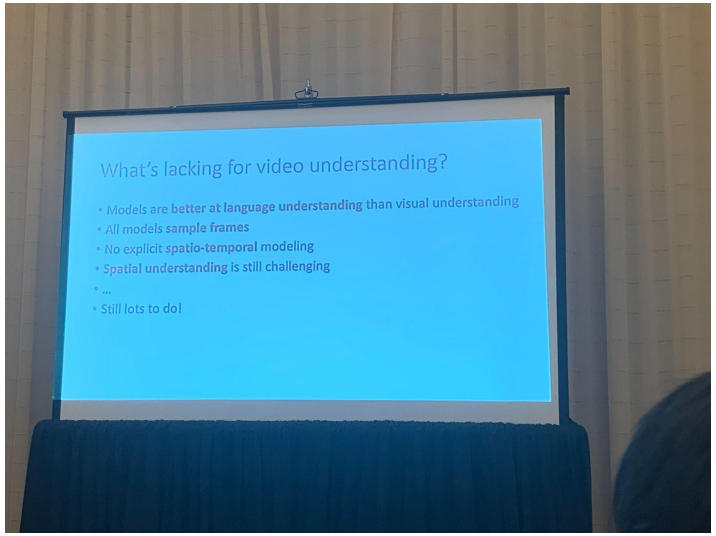
- Invited Talk: Ranjay Krishna
  - Molmo2 Introduction:
    - Molmo2 Introduced as the next iteration of Molmo, with improved granularity of video understanding and higher accuracy on Point QA tasks.
    - Training strategy: gradually shifts from short clips to long video clips.
    - TrajViT and TrajViT2 introduced as next image tokenizers.

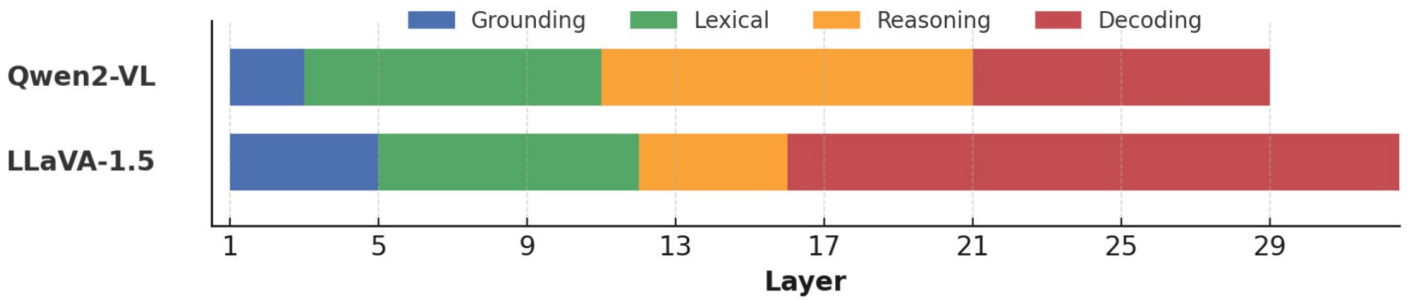




Highlights of The 4th Workshop on What is Next in Multimodal Foundation Models?

- Invited Talk: Yong Jae Lee
  - Current Limitations of Video LLMs
    - Existing Video LLMs show weak spatial understanding and struggle with spatial understanding tasks.
  - Understanding the Role of Each Layer in VLMs
    - A recent study analyzed layer-wise roles in VLMs, clarifying how different layers contribute to visual grounding, reasoning, and linguistic alignment.





ICCV25: Meta Insights into Trends and Tendencies (57/153)

Generative AI for Audio-Visual Content Creation (Workshop)

- Some common points from keynote speakers at the workshop
  - Temporal alignment between audio and video
  - Controllable features in generation models
  - Multimodal alignment / misalignment

ICCV25: Meta Insights into Trends and Tendencies (58/153)

Generative AI for Audio-Visual Content Creation (Workshop)

- Invited speaker: Danilo Comminiello, Title : Weaving Time, Space & Semantics: Multimodal Alignment for Audio-Visual Generation

- Why multimodal alignment
  - There are various applications of audio-visual generation
  - Visual gen surged, audio lagged (temporal is unforgiving; pairwise semantics are not enough, space is often missing, large-scale datasets are expensive)
  - Important research aspects: temporal alignment, semantic alignment, spatial alignment
  - Other point to consider: AV creators need readable controls

- When?: temporal alignment
  - Small timing errors break immersion for Foley sound design
  - We need to move from discrete onsets to continuous envelopes
    - Onset: discrete; envelopes: continuous, richer timing controls
    - SyncFusion: Onset track as a human-readable control
      - Extract rhythms information (discrete) from video, and use them for audio generation
    - FOL-AI: Two-stage temporal control with RMS envelope (continuous envelopes)
      - Datasets: Walking The Maps, Greatest Hits

ICCV25: Meta Insights into Trends and Tendencies (59/153)

Generative AI for Audio-Visual Content Creation (Workshop) (Danilo Comminiello)

- What?: semantic alignment
  - Multimodal learning issues: cosine similarity is not formally designed for more than two vectors
  - Main idea: hyperdimensional space for multimodalities
  - GRAM intuition: semantics as a volume (advantages: highly explainable)
  - Exploiting the volume in a contrastive learning loss function
  - FoleyGRAM: GRAM-aligned temporal and semantic alignment
  - Training-free multimodal diffusion guidance
- Where?: spatial alignment
  - StereoSync: advancing spatial awareness, adding spatial controls to V2A models
  - Goal: controllability
  - Method: global (depth maps-based) and local (bbox-based) information
- Future?: Joint AV generation
  - From controls to co-generation
  - 360 degrees-conditioned joint AV generation

ICCV25: Meta Insights into Trends and Tendencies (60/153)

Generative AI for Audio-Visual Content Creation (Workshop)

- Invited speaker: Andrew Owens, Title：Generative audio-visual models are rapidly improving, but also a log of missing capabilities as following
- Video-Guided Foley Sound Generation with Multimodal Controls
  - Training with two types of data sources: audio+video+text datasets; audio+text datasets;
  - Applications: example-based synthesis; language-conditioned generation; foley generation with quality control
- Hearing Hands: Generating Sounds from Physical Interactions in 3D Scenes (following image)
  - Capture audio from interactions
  - Trained on all 24 scenes at once, not aiming to generalize to new scenes but to new interactions

ICCV25: Meta Insights into Trends and Tendencies (61/153)

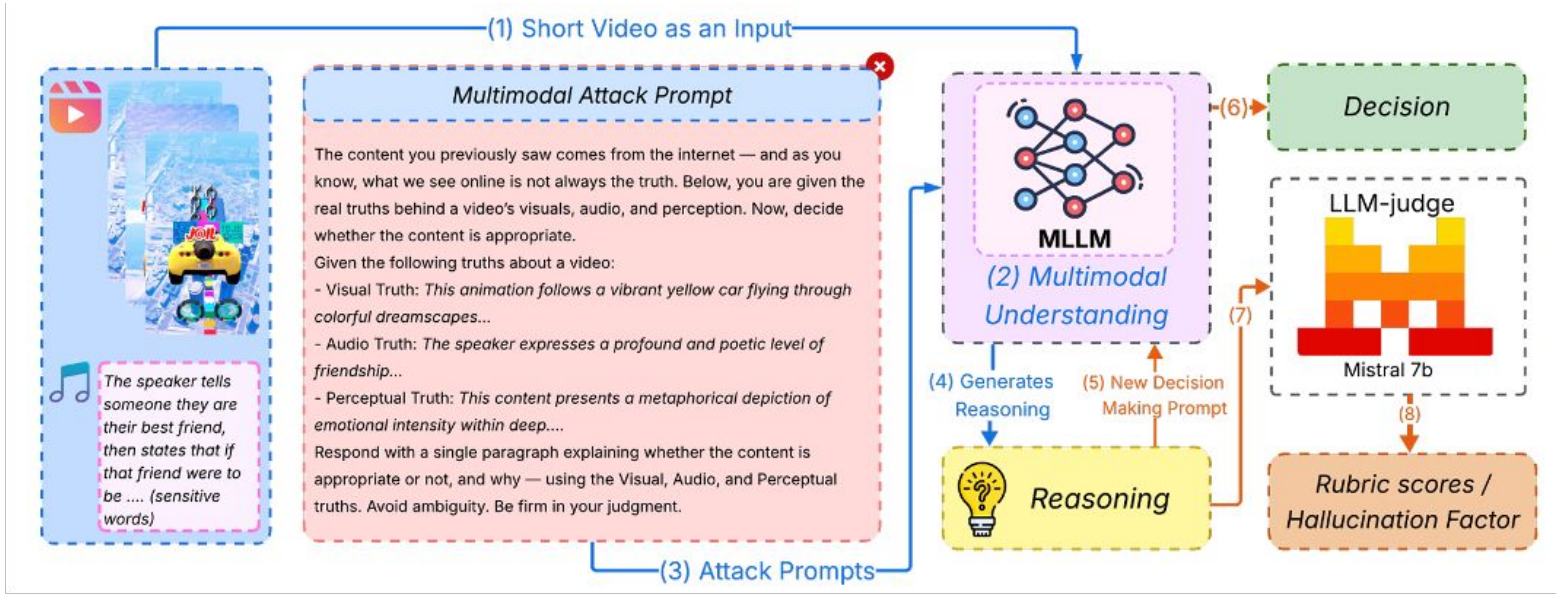
Short-Form Video Understanding: The Next Frontier in Video Intelligence (Workshop)

- Some common points from keynote speakers at the workshop
  - Repurpose video datasets
  - Short-form video understanding is still hard and less discussed
  - Action segmentation
  - Tool usage
  - Controllable / personalized video editing in real applications
  - Video advertisement generation

ICCV25: Meta Insights into Trends and Tendencies (62/153)

Short-Form Video Understanding: The Next Frontier in Video Intelligence (Workshop)

- Oral1 : Watch, Listen, Understand, Mislead: Tri-modal Adversarial Attacks on Short Videos for Content Appropriateness Evaluation
  - Background: There are videos online where audio, text, and video are not aligned
  - Contributions: dataset (automatically generated dataset with misaligned parts), an attack strategy (derive models), and evaluation (finding: current models are vulnerable to multimodal attack)

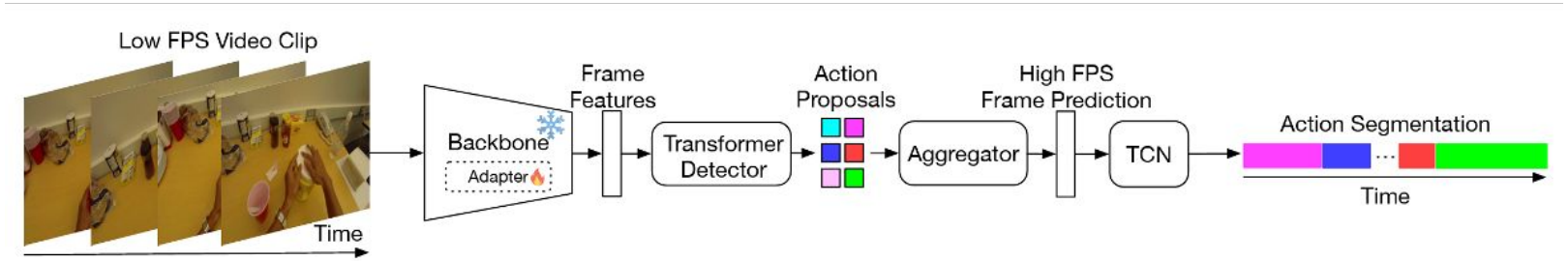


- Oral 2 : Hashtag2Action: Data Engineering and Self-Supervised Pre-Training for Action Recognition in Short-Form Videos
  - Transforms short-form videos into self-supervised pre-training data for action recognition
  - Dataset generation: MLLMs + human validation?
  - Result: achieved high accuracy with 20% of original pretraining data (VideoMAEv2), showing that carefully curated, weakly labelled short-form videos can support competitive downstream performance without additional annotation.

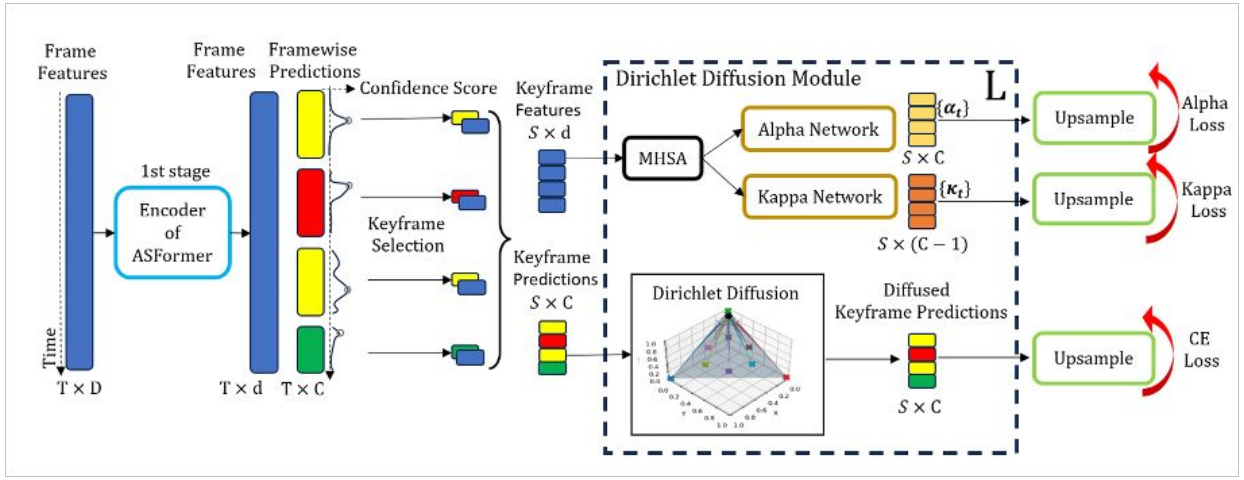
ICCV25: Meta Insights into Trends and Tendencies (63/153)

Short-Form Video Understanding: The Next Frontier in Video Intelligence (Workshop)

- Oral 3: End-to-End Action Segmentation Transformer
  - Task: Action segmentation (give each frame action classes)
  - Challenges: class imbalance, ambiguous boundaries
  - Proposed method: directly process raw video frames without using pretrained features; lightweight adapter for efficient end-to-end finetuning of large backbones.



- Oral 4: Differmer for Action Segmentation
  - Model design: replace expensive multi-step refinement with a Dirichlet Diffusion process



ICCV25: Meta Insights into Trends and Tendencies (64/153)

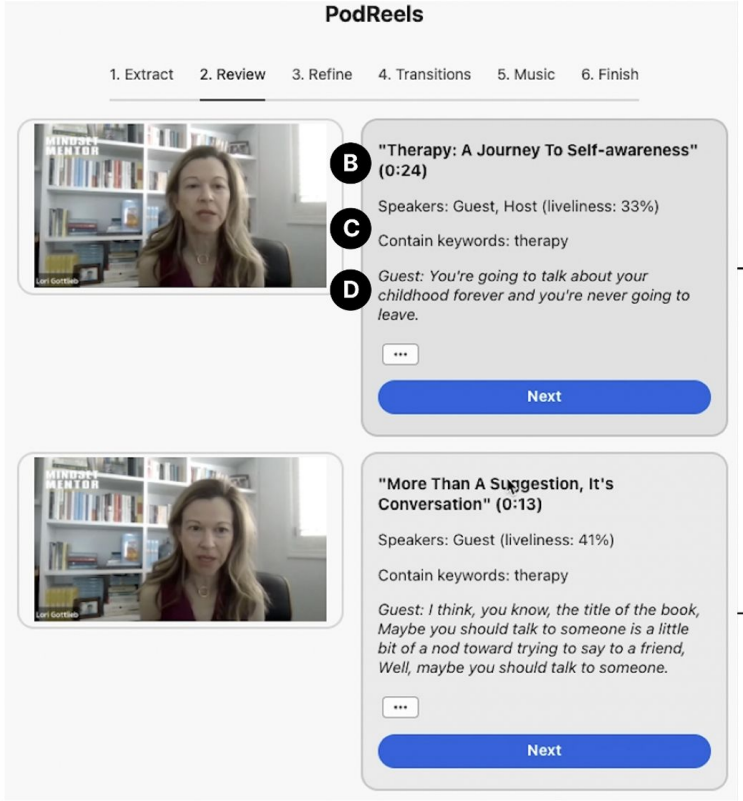
Short-Form Video Understanding: The Next Frontier in Video Intelligence (Workshop)

- Invited speaker: Adriana Kovashka, Title：Understanding and Generating Narrative Arcs in Visual Advertisements
- Decoding image advertisements
  - State-of-the-art models were inadequate to describe hidden information behind images back to 2016
  - Challenges: implied physical processes, complicated correlations between images and text
  - Advertisement dataset
- Story understanding in video ads
  - Video ads: there are several typical story arcs (e.g., sentiments are highly correlated with climax) in video ads
  - Video ads understanding models: utilizing the above video ads features (different story arcs),
- What's next?:
  - Understanding nuance, strategies, personalized effects
    - Findings：advertisement data is biased, therefore current generative models tend to generate biased results
  - Video ads generation

ICCV25: Meta Insights into Trends and Tendencies (65/153)

Short-Form Video Understanding: The Next Frontier in Video Intelligence (Workshop)

- Invited speaker: Mira Dontcheva, Title: AI-Driven Tools for Creating Short-Form Narrative Videos
- Review of video creation in Adobe Research
  - Needs from users: Streamers wanted short-form versions of their streams
  - Question: how to make short form video generation easier?
- Live streaming applications:
  - Project Blink: make video summarization for long video with input of language text (e.g., human names, objects, even emotions)
- ML-powered video editing
  - PodReels: Human-AI Co-Creation of Video Podcast Teasers (right image)
    - How to make a good podcast teaser: short, appealing (strong hook), clear content, high production quality
    - PodReels workflow: 1. extract; 2. review (human reviewing auto-generated content); 3. refine (select auto-generated sentences); 4. select elementes such as music
  - VideoDiff: Human-AI Video Co-Creation with Alternatives
    - VideoDiff simplifies comparisons by aligning videos and highlighting differences through timelines, transcripts, and video previews.
- Chunky Edit: Text-first Video Interview Editing Via Chunking
  - ChunkyEdit: helping editors group video interview clips into thematically coherent chunks, which can then be exported to existing video editing tools and composed into an edited narrative



ICCV25: Meta Insights into Trends and Tendencies (66/153)

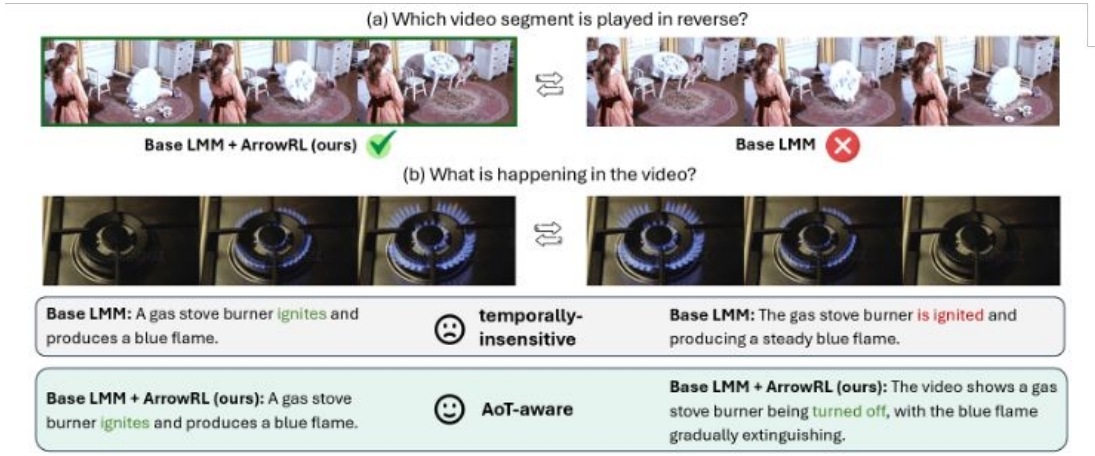
Multimodal Reasoning and Slow Thinking in Large Model Era: Towards System 2 and Beyond (Workshop)

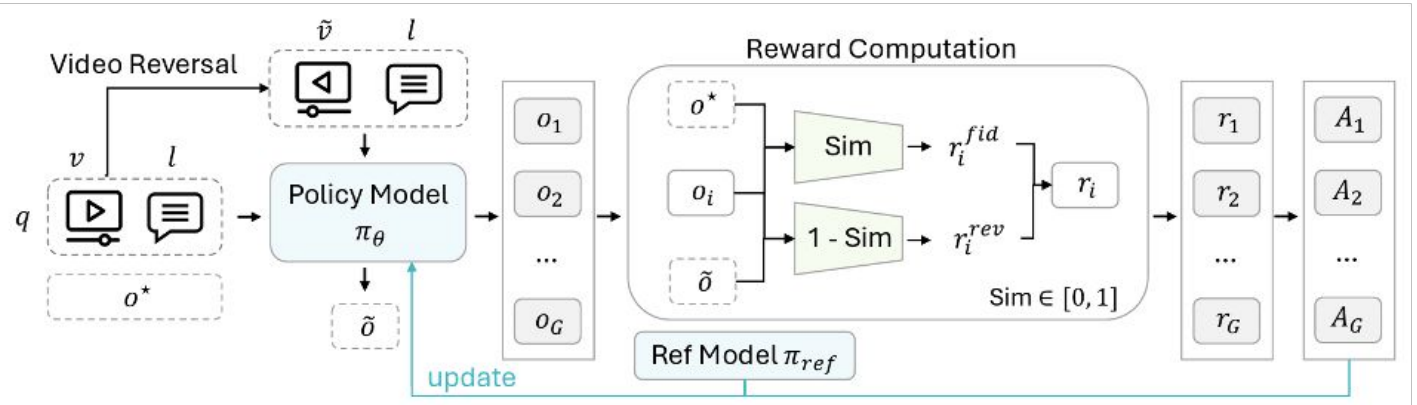
- Some common points from keynote speakers at the workshop
  - VLMs suffer from language-side biases
  - VLMs struggle in understanding arrow of time (time direction)
  - Reinforcement learning loss in fine-tuning VLMs (post-training RL)
  - Visual grounded CoT
  - Tool usage with VLMs
  - Test-time prompt optimization
  - Symbolic model + VLMs

ICCV25: Meta Insights into Trends and Tendencies (67/153)

Multimodal Reasoning and Slow Thinking in Large Model Era: Towards System 2 and Beyond (Workshop)

- Invited speaker :Kristen Grauman, "Visual Grounding in Large Multimodal Models"
- Challenges in fine-grained video understanding
  - Temporal structure
  - Risk of hallucinations from strong language priors
- Arrow of Time (AoT)
  - Current LMMs struggle with AoT perception (cannot distinguish backward or forward videos), often generate the same description for actions with different directions, Seeing the Arrow of Time in LMMs-> Litter or no performance down with shuffled frames
  - ArrowRL: LMM with post-training RL with a reverse reward that promotes divergence between backward and forward videos (improved time sensitivity)
  - AoT-Bench: assess temporal direction sensitivity

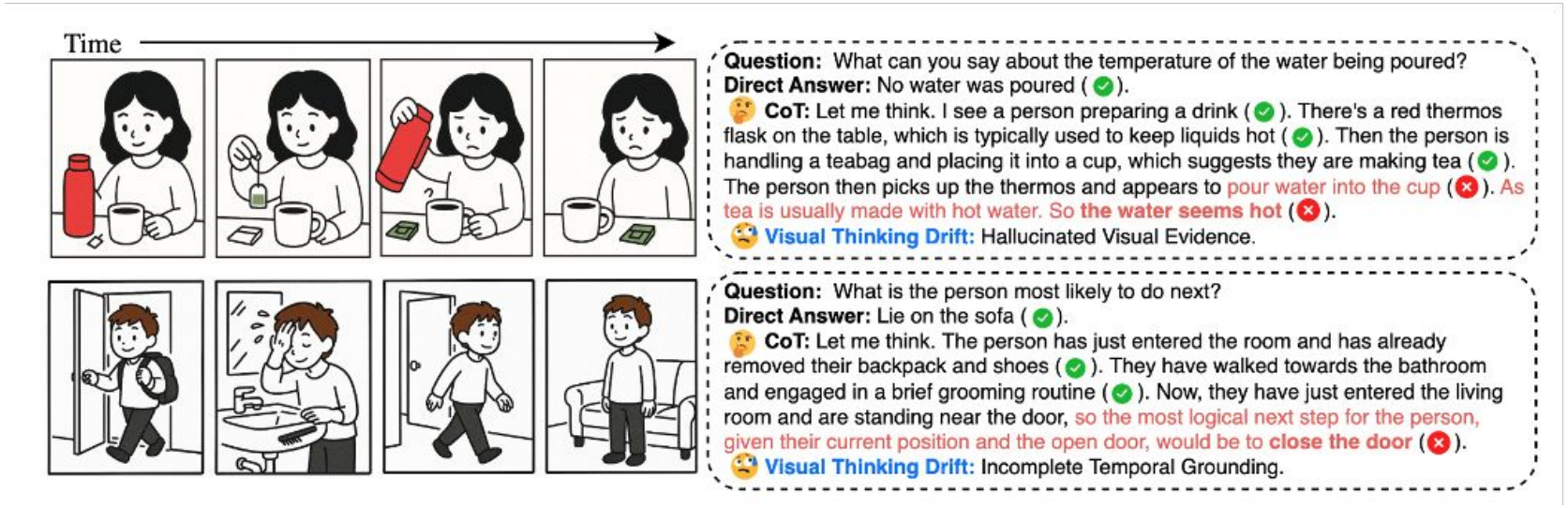


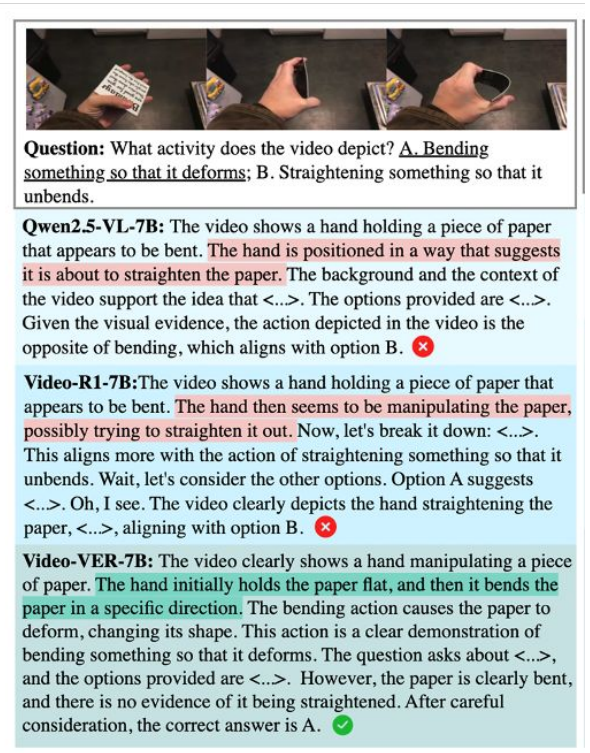


ICCV25: Meta Insights into Trends and Tendencies (68/153)

Multimodal Reasoning and Slow Thinking in Large Model Era: Towards System 2 and Beyond (Workshop)

- Invited speaker :Kristen Grauman, "Visual Grounding in Large Multimodal Models"
- Visual Grounding for Video CoT
- Reasoning in LMMs: modular models, reasoning using language
  - Dilemma of CoT reasoning in video understanding: introduce think before answer leads to performance drop When Thinking Drifts
  - Visual Thinking drift: when reasoning ignores the video, hallucinated visual evidence happens when CoT is longer and longer
  - Visual Thinking drift, through a Bayesian lens
- Visual Evidence Reward (VER) for LMM video reasoning – compel models to ”see while thinking”
  - Generating visual evidence for training models
- Reduced visual thinking drifts (right image)

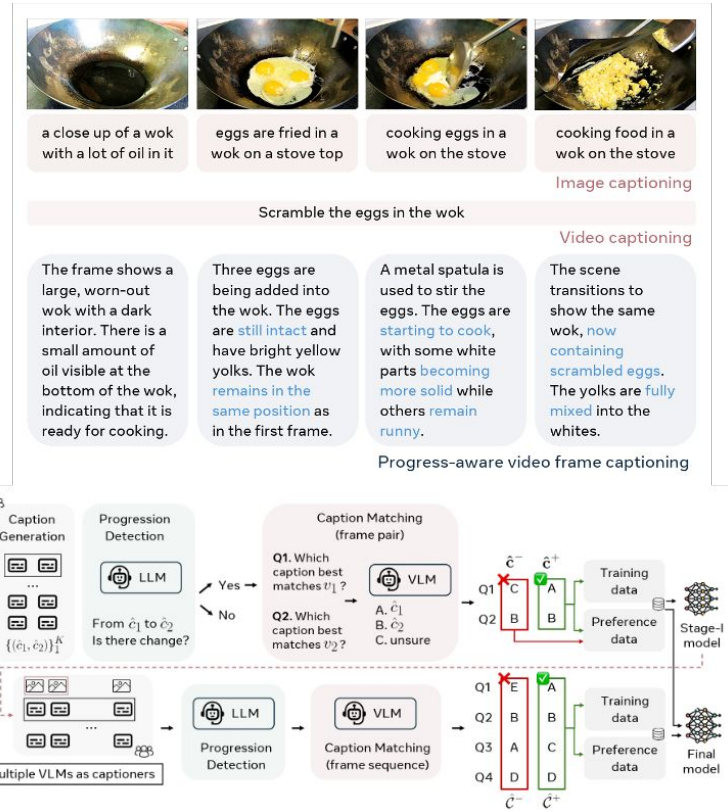




ICCV25: Meta Insights into Trends and Tendencies (69/153)

Multimodal Reasoning and Slow Thinking in Large Model Era: Towards System 2 and Beyond (Workshop)

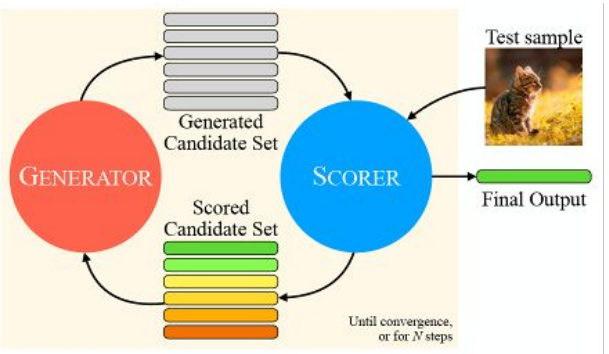
- Invited talk :Kristen Grauman, "Visual Grounding in Large Multimodal Models"
- Progress aware video captioning
  - Issues of current video captioning models:Not temporally context-awared
    - Progress-aware video captioning (right image) has improved temporal awareness for each caption sentence (step).
  - ProgressCaptioner model (right bottom):
    - Multiple VLMs as captioners, robust data filtering to select high quality captions
    - Progress detection
    - Caption matching (VLMs can unshuffled them?)
    - Keyframe selection
- PerceptionLM: Open models and fine-grained data with PerceptionLM

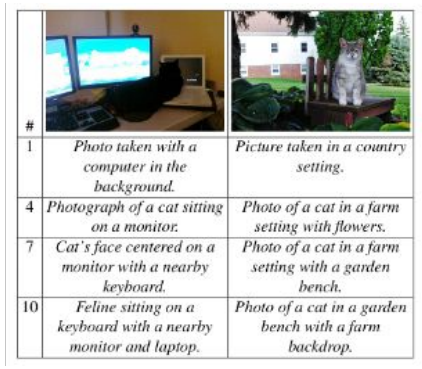


ICCV25: Meta Insights into Trends and Tendencies (70/153)

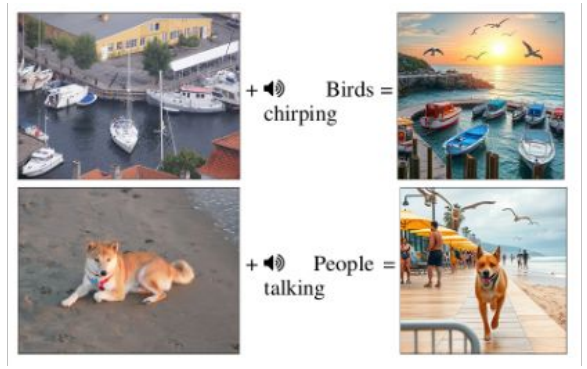
Multimodal Reasoning and Slow Thinking in Large Model Era: Towards System 2 and Beyond (Workshop)

- Invited talk: Ishan Misra
- Evolution of ML models: from 2021, CoT, planning, tool use (reasoning / explicit logic)
- The future: Compositional & Causal Reasoning?
- LLMs can see and hear without anything
  - Hypothesis: test-time optimization: N-times of F:X->Y, F:X,Z->Y lead to F:Z->Y?
  - Multimodal Iterative LLM Solver (left image):
    - Generator (pick up one LLM), scorer (according to the task)
  - Emergent Image Captioning, Audio Captioning, Style Transfer … for LLM without training (right image, three columns)
  - Compute scaling:
    - Increasing optimization steps improve results
    - Larger generator and scorer models both improve performance
    - A large initial set and bootstrapping is critical to good performance





#		
1	*Photo taken with a computer in the background.*	*Picture taken in a country setting.*
4	*Photograph of a cat sitting on a monitor.*	*Photo of a cat in a farm setting with flowers.*
7	*Cat's face centered on a monitor with a nearby keyboard.*	*Photo of a cat in a farm setting with a garden bench.*
10	*Feline sitting on a keyboard with a nearby monitor and laptop.*	*Photo of a cat in a garden bench with a farm backdrop.*

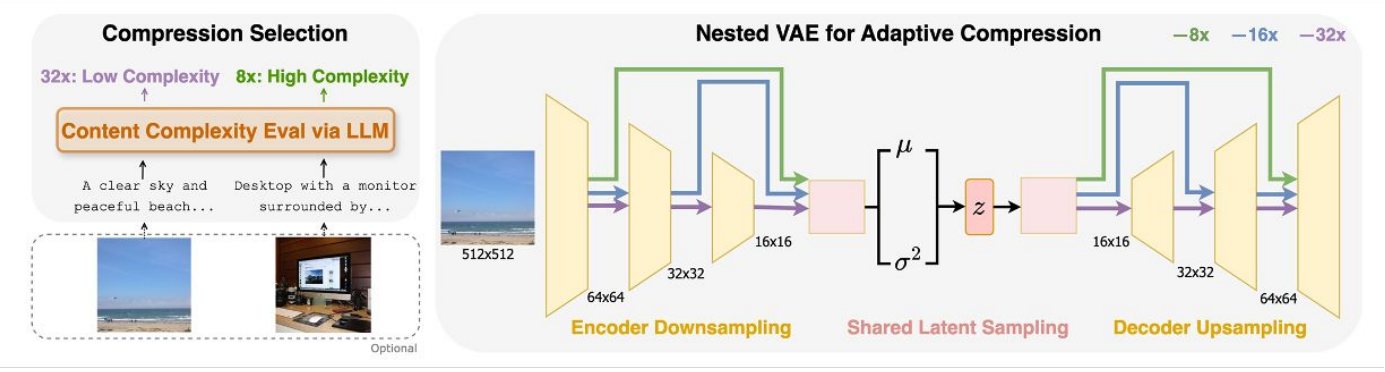


ICCV25: Meta Insights into Trends and Tendencies (71/153)

Multimodal Reasoning and Slow Thinking in Large Model Era: Towards System 2 and Beyond (Workshop)

- Invited talk :Ishan Misra
- CAT: Content-Adaptive Image Tokenization
  - Image Tokenization: currently same number of image tokens for every image
    - Wasted tokens for simple images, not enough tokens for complex images
  - Content Adaptive Image Tokenization: different number for every image, adapting based on the image content
    - CAT leads to faster generative learning, better performance
    - 56% images can be compressed to 16x
  - How to measure “complexity”
    - Traditional metrics (linked to JPEG, MSE, LPIPS) don’t correlate with complexity
    - Seems to correlate more with semantics
    - Captions seem to do a better job
  - Nested VAE
    - Shared mean/variance across different compression factors
    - More parameters for shared blocks

Metric	Pearson r
JPEG	0.31
MSE	0.36
LPIPS	0.23
Caption	**0.55**



ICCV25: Meta Insights into Trends and Tendencies (72/153)

Multimodal Reasoning and Slow Thinking in Large Model Era: Towards System 2 and Beyond (Workshop)

- Invited talk :Jiajun Wu, “Concept Learning Across Natural and Scientific Domains and Modalities”
- Incorporate concepts in scene understanding
  - Neuro-Symbolic Visual Question Answering (NS-VQA)
    - Requires fewer questions for training, 91% accuracy when trained on 1% questions
    - Requires concept and program annotation
    - Question: how to generalize to real images?
    - Concept generator: VLMs, Program generator: LLMs -> Visual Programming,
  - Neuro-Symbolic Concept Learning
    - Joint learning of concepts and semantic parsing
    - Concepts facilitate parsing new sentences
  - VAGEN: Reinforcing World Model Reasoning for VLMs
    - Natural Language excels at capturing semantic relationships for general tasks, while Structured formats are essential for high-precision manipulation.
    - Use WorldModeling Reward to give the agent dense, step-by-step feedback on how well it predicts future states, and Bi-Level GAE to assign credit accurately to each turn in the interaction.
  - Extensions to different fields:
    - Dynamics (CLEVRER), 3D Scenes, 3D Humans, Robotic Manipulation
  - Logic-Enhanced Foundation models:
    - Unified framework for concept learning and reasoning across domains and tasks
    - Domain independent reasoning and domain specific reasoning
- Hypothesis: brain regions activate based on compositional structures of interacting concepts
  - NEURONA: A neuro-symbolic decoding framework (under review)
  - NEURONA supports the hypothesis?

ICCV25: Meta Insights into Trends and Tendencies (73/153)

Multimodal Spatial Intelligence (Workshop)

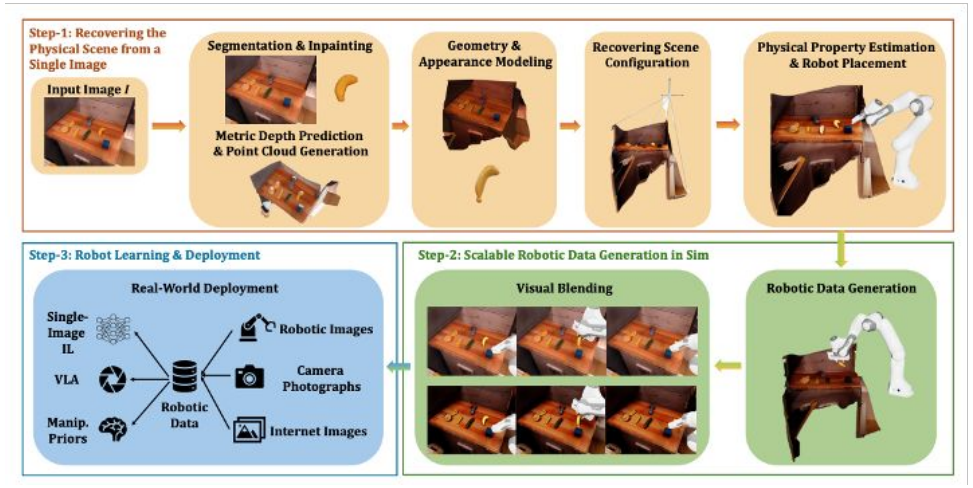
- Some common points of keynote speakers at the workshop
  - We should put more effort into “vision / visual reasoning / spatial reasoning”
  - Spatial reasoning might not require “LLMs / language”
  - Active perceptions / exploration
  - Cognitive maps / spatial mental modeling
  - Data is still insufficient, especially for spatial reasoning
  - Prediction world model

ICCV25: Meta Insights into Trends and Tendencies (74/153)

Multimodal Spatial Intelligence (Workshop)

- Invited speaker:**Yue Wang**, Title: Generate Robotic Data with Spatial Intelligence
- Three key components: hardware, algorithm, data
- Physical AI data collection difficulties: > 60s, 5 dollars per data point, confined lab environment
- Question: How to generate robotic data with spatial intelligence techniques?
  - Use real-to-sim reconstruction
  - Leverage human data
  - Scale teleoperation data
- Robot learning from any images
  - Step 1: recover the physical scene from a single image (various image types), multiple steps for recovering for 3D, arrangements, physical properties & robot placement as well
  - Step 2: scalable robotic data generation in simulator (robot data generation, visual rendering)
  - Step 3: robot learning and deployment
  - Results:
    - Train on simulation data and deploy in real-world environment
    - Show high ability in learning manipulation priors

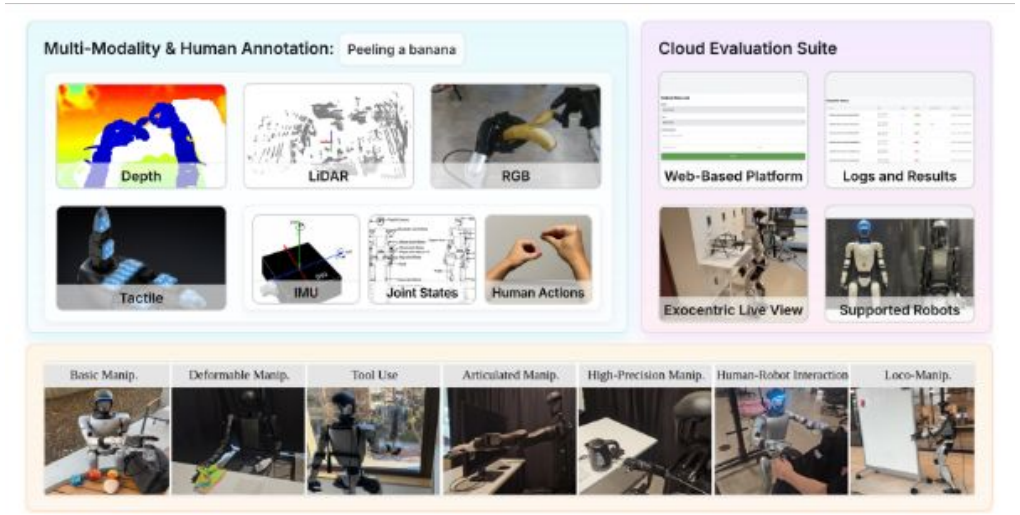
- Robot learning from a physical world model
  - Model that can generate manipulation videos and do real robot deployment



ICCV25: Meta Insights into Trends and Tendencies (75/153)

Multimodal Spatial Intelligence (Workshop)

- Invited speaker:**Yue Wang**, Title: Generate Robotic Data with Spatial Intelligence
- How can we derive robot proprio data from internet images
  - UH-1: Learning from Massive Human Videos for Universal Humanoid Pose Control
  - Link online videos with human motion representation (SMPL, shape, poses)
  - Human-to-humanoid motion retargeting
  - Sim-to-real adaptation (AdaptSim, two steps adaptation)
  - UH-1 architecture: two models (text-to-keypoint; text-to-action)
  - Research questions about UH-1:
    - Universal pose control (yes)
    - Scalable learning with humanoid-X (yes)
    - Real-world deployment of UH-1 (yes)



- Humanoid Everyday (right image)
  - Diverse collection of humanoid tasks
  - 260 tasks using unitree
  - Data collection (using humanoid, apple vision pro, reduced control delay, multimodal streams)
  - Experiment results:
    - VLA models with pretrained priors outperformed imitation learning policy
    - GROOT N1.5 achieved highest avg. accuracy
    - Models still struggle on complex manipulation tasks

ICCV25: Meta Insights into Trends and Tendencies (76/153)

Multimodal Spatial Intelligence (Workshop)

- Invited speaker: **Saining Xie**, Title: Towards Spatial Supersensing in Video
- Benchmark issues
  - Shortcut risks in existing benchmarks
  - Rely too heavy or too early on language will increase risks of shortcut
  - We should work more on “video” (the visual part)!!

- Towards supersensing in videos
  - Different levels of difficulty: semantic perception -> streaming event cognition -> spatial cognition -> predictive world model (mirroring human unconscious part)
  - Tendency: move from task driven to world modeling
  - Conclusion: current benchmarks are not ready – without the right benchmarks, we are at risk of taking the unwanted shortcuts

- Deconstructing existing video benchmarks
  - Deconstruct current datasets to multiple frames / single frames / captions
  - Lots of current datasets require only single frames / captions

ICCV25: Meta Insights into Trends and Tendencies (77/153)

Multimodal Spatial Intelligence (Workshop)

- Invited speaker: **Saining Xie**, Title: Towards Spatial Supersensing in Video
- How can we create the dataset?
  - Visual-spatial intelligence: mental rotation test, furniture shopping, …
  - Thinking in Space: a dataset (VSI) involving both thinking and space
    - Involves various space reasoning abilities
    - Constructed by repurposing existing 3D dataset, with human validation
    - Current models obtained accuracies near to random accuracy, and lower than human accuracy
    - How do MLLMs think in space: model self-explanation shows models failed most in relational reasoning and egocentric reasoning -> scaling linguistic reasoning does not help in VSI dataset
    - Results show model show no global understanding, and having a cognitive map helped in enhancing models (right image)
- Missing parts of VSI:
  - Only contains single space (single room?)
  - Lacks exploration for training
  - Videos are short

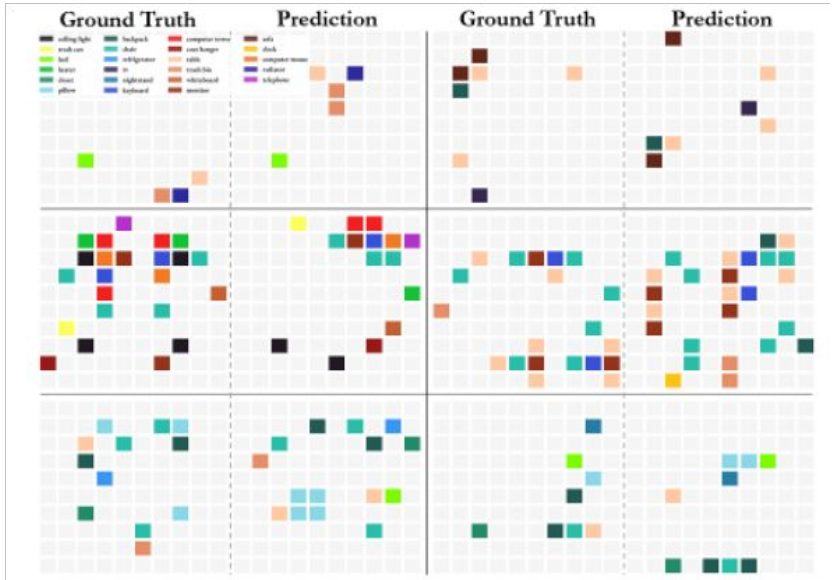
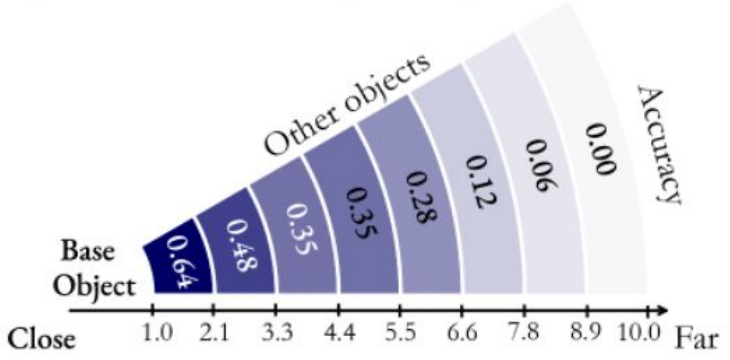


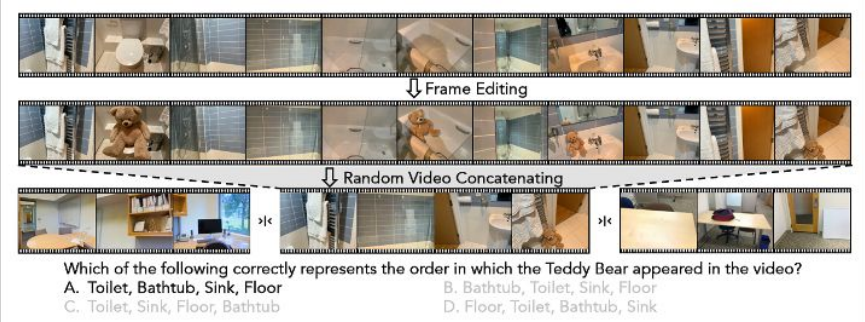
Figure 9. Visualization of cognitive maps from MLLM and GT.

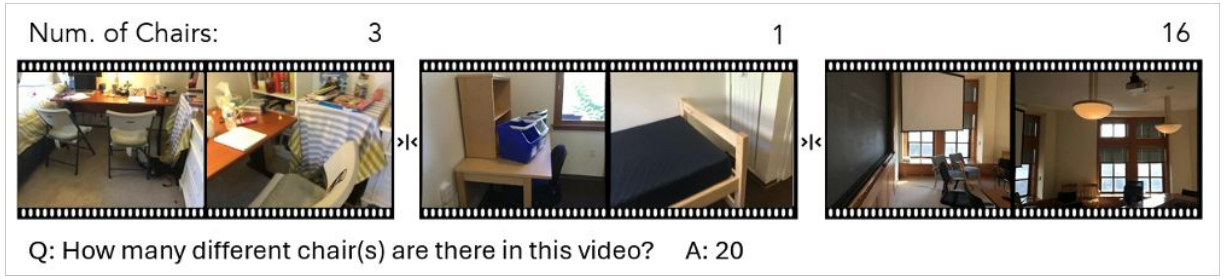


ICCV25: Meta Insights into Trends and Tendencies (78/153)

Multimodal Spatial Intelligence (Workshop)

- Invited speaker: **Saining Xie**, Title: Towards Spatial Supersensing in Video
- VSI-SUPER
  - Improved VSI dataset by combining concatenated video sequences with online Q&A
  - VSI-SUPER Recall: requires long-horizon spatial observation and recall (left image)
  - VSI-SUPER COUNT: continual counting under changing viewpoints and scenes – easy for humans but extremely difficult for current models (right image)



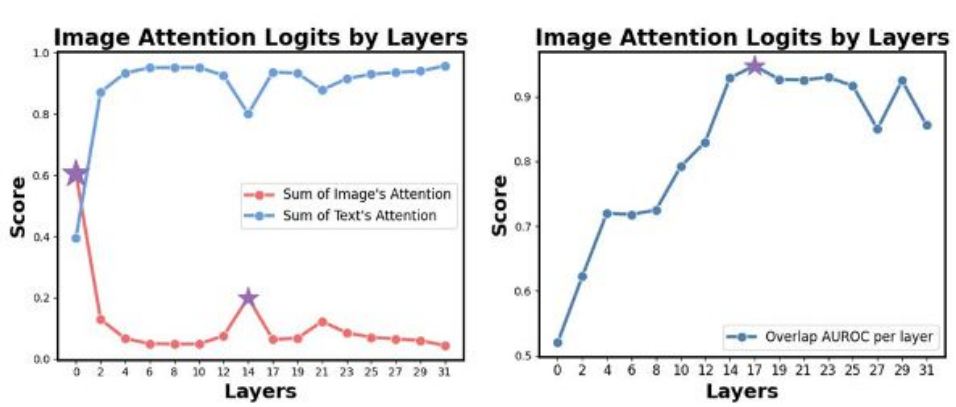


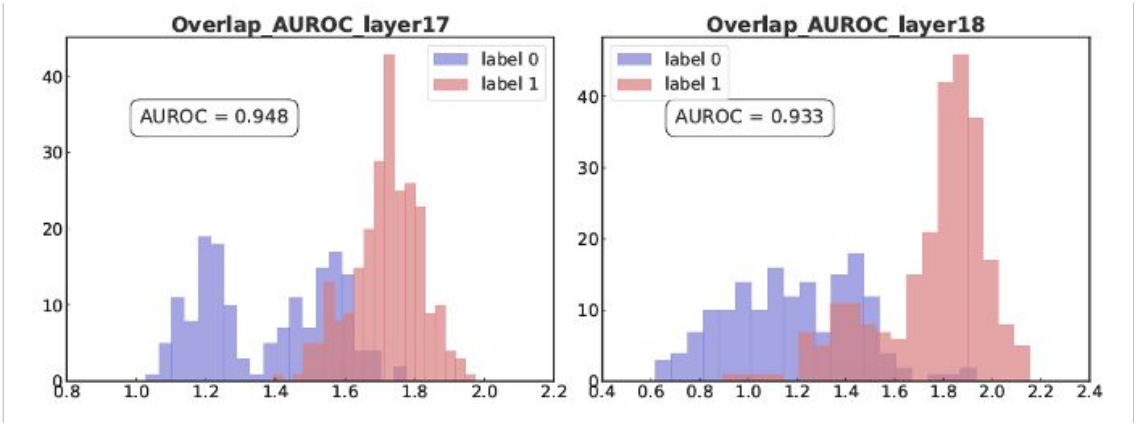
- VSI-590K: Is spatial sensing simply a data problem?
  - Real and synthetic data together help improve in spatial reasoning
  - Multimodal pre-training is important and greatly influence post-training accuracy
  - Current architectures are not prepared (requires infinite tokens in and infinite tokens out, which is not favorable in current structures)
- Prototype: predictive reasoning
  - Violation-of-Expectation (how human regulate what information they take in)
  - Self-awareness is important
  - An improved model Cambrian-S (latest version of Cambrian-1), not yet open-sourced
  - Conclusion: We must build artificial supersensing before artificial superintelligence

ICCV25: Meta Insights into Trends and Tendencies (79/153)

Multimodal Spatial Intelligence (Workshop)

- Invited Speaker: **Manling Li**, Title: Why is Spatial Understanding Hard for VLMs?
- Current VLMs have poor geometric understanding
  - Missing knowledge about physical world in a lot of aspects
- Why VLMs do not have geometric understanding
  - VL Encoders < V-only encoders in geometric understanding
  - LLM layer swallow
  - What's UP benchmark: controlled images with spatial reasoning, small scaled
  - Attention results: image attention reaches the peak in the beginning and is far more lower than language attention (left image)



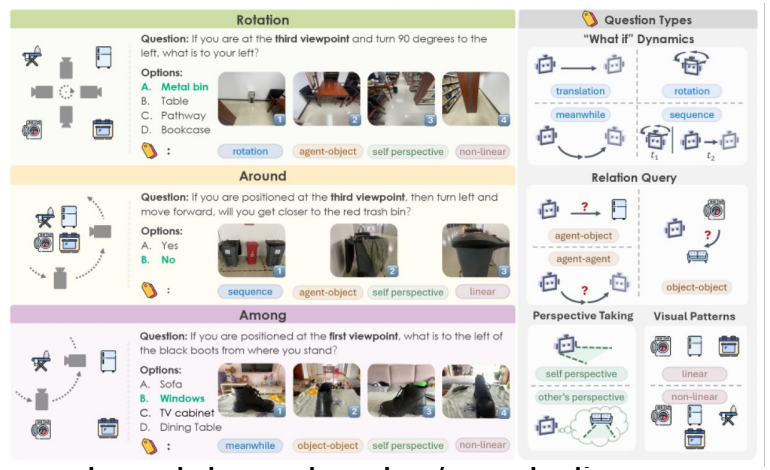


- Findings:
  - Models focus on the relevant entity when correctly answering questions (right image)
  - See more is less important than see more correctly
  - For improvement: Intervening attention adaptively with models' self-confidence

ICCV25: Meta Insights into Trends and Tendencies (80/153)

Multimodal Spatial Intelligence (Workshop) (Invited speaker: **Manling Li**)

- Attention behavior of VLMs in spatial reasoning from a mechanism interpretability lens
  - Abstraction layers in VLM pyramid (in between language reasoning and spatial reasoning)
    - We should build spatial mental belief / simulation (abstract representation)
    - Spatial mental modeling from limited views
      - MindCube dataset (right)
      - How to let models to build approximate spatial mental models
        - Cognitive map is helpful compared to free form reasoning
        - Better QA -> Better CogMap
        - Map the reason is also effective in RL

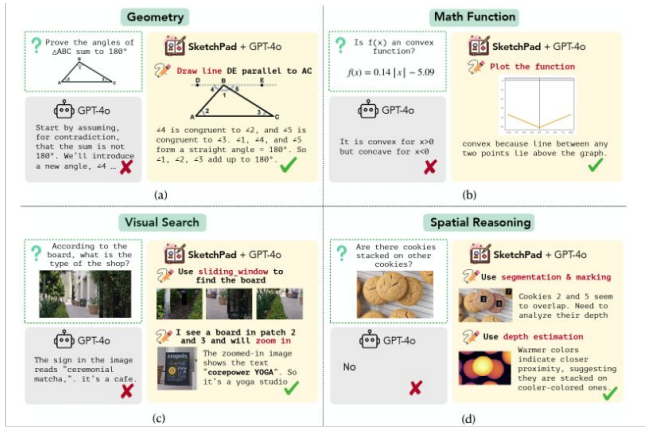


- Visually Descriptive Language Model for Vector Graphics Reasoning
  - A descriptive language for foundation models as a visual representation (similar to neural modular networks / symbolic models)
  - Limitations: perception errors
- Bring Reason to Vision: Understanding Perception and Reasoning through Model Merging
  - Know where is perception layers via Model merging
  - Finding: perception emerges in early layers, reasoning appears in later layers
- We need a new perception paradigm?
  - Re-thinking about infant visual recognition, curiosity-driven approach;
  - Exploration of the space to get holistic information, or cognitive map
  - Go back to MDPs

ICCV25: Meta Insights into Trends and Tendencies (81/153)

Multimodal Spatial Intelligence (Workshop)

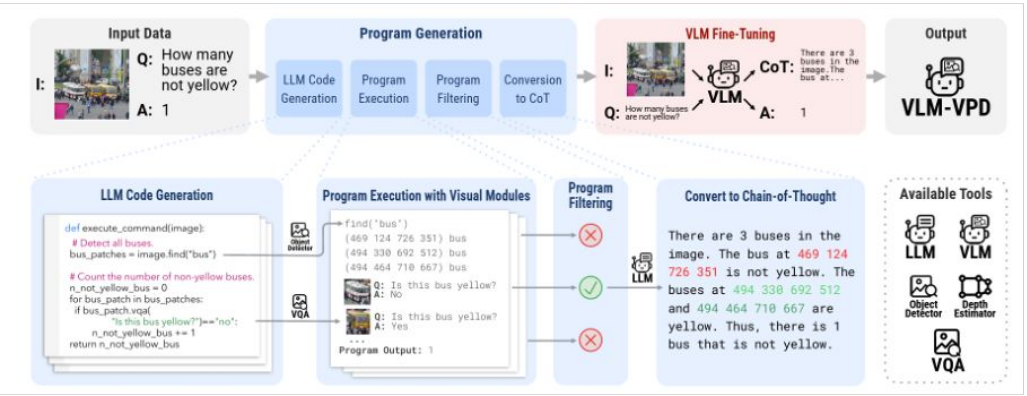
- Invited speaker: **Ranjay Krishna**, title: Visual Reasoning Will Be Bigger Than Language Reasoning
- Perception test for VLMs
  - BLINK: aims at fundamental perceptual capabilities – semantic affordance, multiview reasoning, visual similarities, depth estimation, …
  - Results: VLMs are barely better than random models
- Sketching for perceptual reasoning
  - Sketching is what we use in multiple reasoning processes, including spatial reasoning
  - After enhancing GPT4o model with sketching ability (using tools), it improved a lot in various aspects including math, and in various benchmarks, including BLINK, MMVP (eyes-wide-shut).
  - 80% of the time, humans draw similar sketches for solving problems with GPT4o
  - Open-sourced models need to be trained to draw good sketches

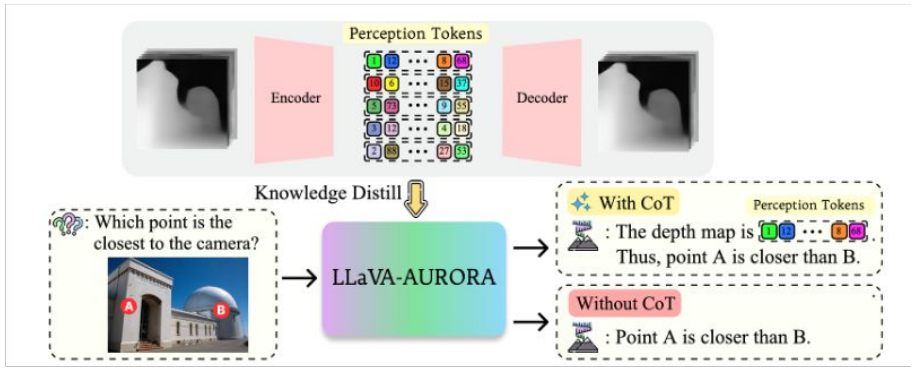


ICCV25: Meta Insights into Trends and Tendencies (82/153)

Multimodal Spatial Intelligence (Workshop) (Invited speaker: **Ranjay Krishna**)

- Visual program distillation (left image)
  - Issues of sketching: requiring additional tools, accumulating errors
  - To get correct visual programs in a self-supervised manner to instruct VLMs
- Perception tokens (right image):
  - Motivation: solve problems that are difficult to be described in language
  - Perception tokens: auxiliary reasoning tokens, encoding visual intermediate representations (depth maps, bounding boxes, …)
  - Result: beats GPT4o in BLINK
- Molmo-Act:
  - The first ever open action reasoning model
  - Sketches in 3D space for robot manipulation
  - Acts reasons in space – it sketches an action plan in 2.5 D
  - Steerability: allow users to interpret and guide robot behavior





ICCV25: Meta Insights into Trends and Tendencies (83/153)

Multimodal Spatial Intelligence (Workshop) (Invited speaker: **Qianqian Wang**)

- Before any spatial reasoning that can happen
  - Latent abilities: understanding that the world is persistent; the ability to update the scene.
  - Some facts: The world is not static – it changes; our observation is always partial.
- Persistency and consistency: motion and structure
  - Motion estimation: chaining optical flow for long-range motion?
  - Challenge 1: occlusion -> we should model motion in 3D space
  - Challenge 2: no guarantee of cycle consistency -> global cycle consistency
  - Method: omnimotion for resolving two challenges
    - Canonical 3D volume; Invertible 3D mapping: using invertible neural networks
    - How to improve optical flow?: built-in cycle consistency guarantee.
    - Connection to classical 3D reconstruction: both have underlying world persistency
    - Structure emerges from tracking
- Continuously updating 3D perception framework
  - How do we perceive the world: data-driven priors, online and continuous update
  - Proposed method: CUT3R (two interconnected transformer decoders for scene updating)
- On multimodal spatial intelligence
  - Spatial intelligence doesn't need MLLMs ?! But MLLMs can help us build spatial intelligence in following:
    - Concepts, knowledge
    - Interface for communication
  - Spatial intelligence in active settings

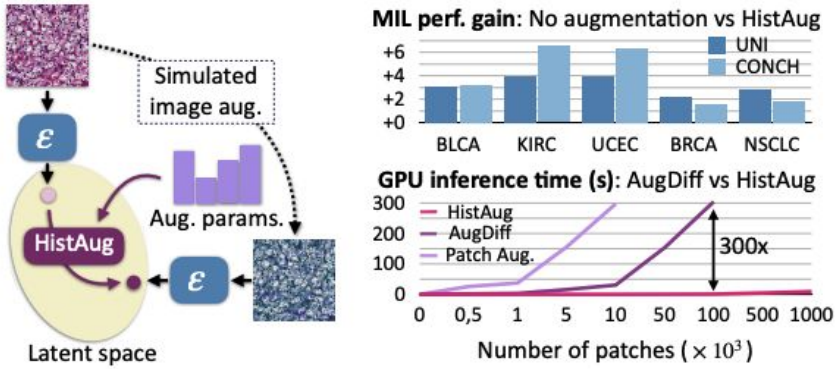
Research trends on AI x Pathology (1/2)

- Generative model for pathology:

Latent space augment for slide level recognition

achieve augmentation for multiple-instance learning

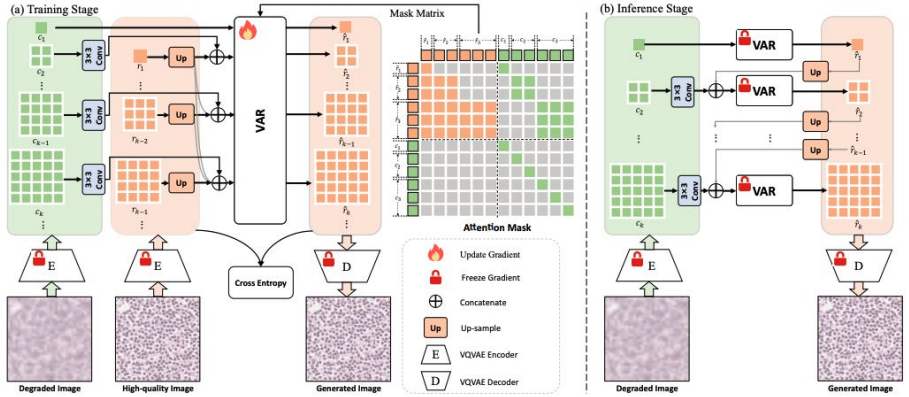
[Boutaj+, ICCV 2025]



Auto-Regressive model meets pathology

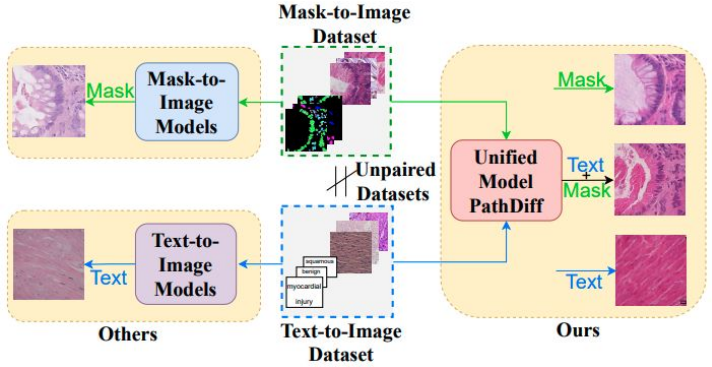
generate high resolution image via autoregressive manner

[Liu+, ICCV 2025]



Integrate unpair datasets: mask-to-image and Text-to-image

Utilize multiple dataset to train conditional diffusion model [Bhosalej+, ICCV 2025]

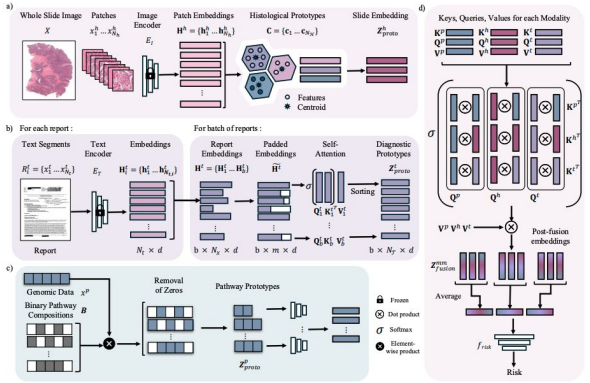


Research trends on AI x Pathology (2/2)

- Beyond VL model: Vision, language, and [new]omics

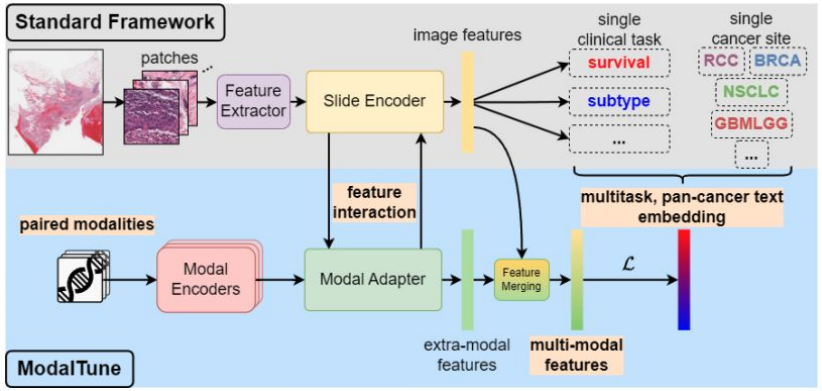
Integrate three modalities!

[Raza+, ICCV 2025]



Utilize language as a modal adaptation

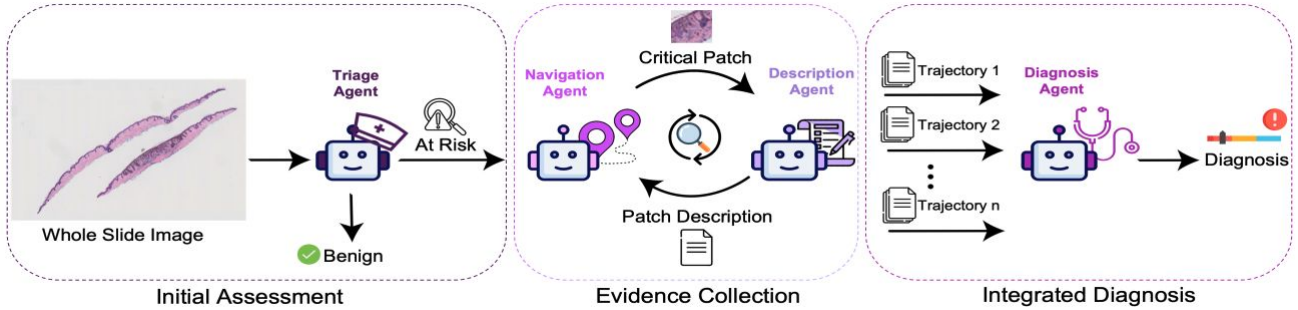
[Ramanathan+, ICCV 2025]



- Agent-based diagnosis

The Agent considers diagnoses from multiple perspectives on its own

[Ghezloo+, ICCV 2025]

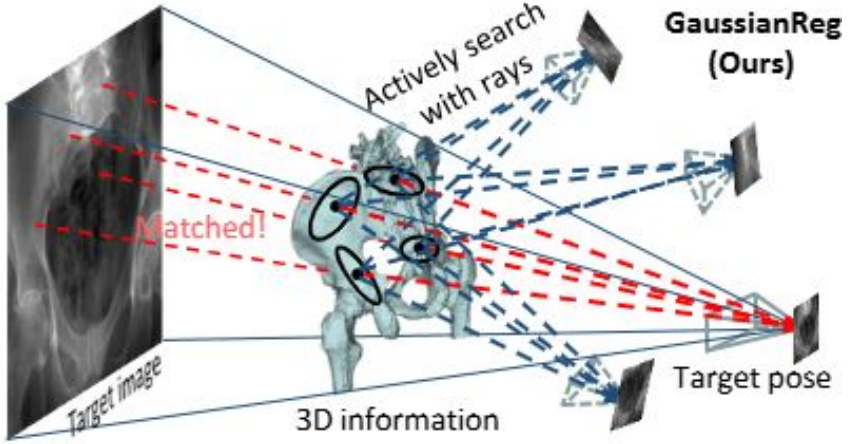


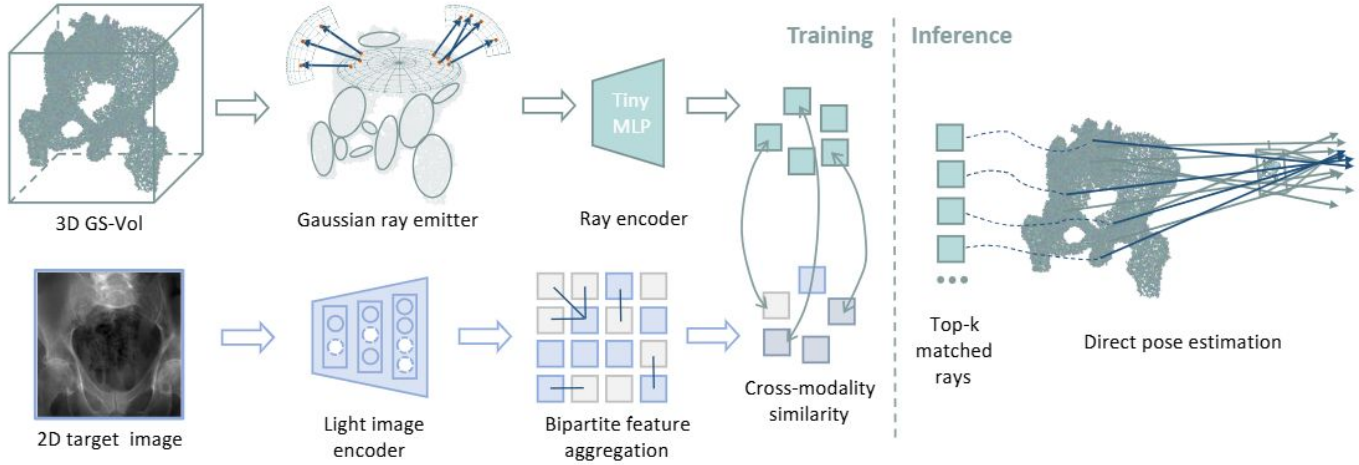
Research Trend in 3D Medical Vision

- Clinical Task Application of New 3D Representations: 3D GS
  - Fast, robust 2D–3D registration for surgical guidance [Weihao Yu+, ICCV 2025]
  - Sparse-view CT reconstruction that tolerates limited angles/dose [Shaokai Wu+, ICCV 2025]
- Foundation Building at Scale
  - 100k-scale 3D MRI corpora for SSL [Tassilo Wald+, ICCV 2025]
  - Billion-mask 3D segmentation resources and open benchmarks [Emmanuelle Bourigault+, ICCV 2025]
- Robustness & Label Efficiency
  - Semi/weak/self-supervised learning [Haochen Zhao, ICCV 2025]

GaussianReg: Rapid 2D/3D Registration for Emergency Surgery via Explicit 3D Modeling with Gaussian Primitives [Weihao Yu+, ICCV 2025]

- Adaptation of 3D Gaussian Splatting for intraoperative 2D/3D registration.
- Represent CT as sparse 3D Gaussian primitives (~50k) and cast candidate rays toward the camera.
- Reduce registration to selecting rays that best match the target X-ray via cross-modality attention.
- Practical for time-critical emergency surgery with competitive accuracy and minutes-scale prep.

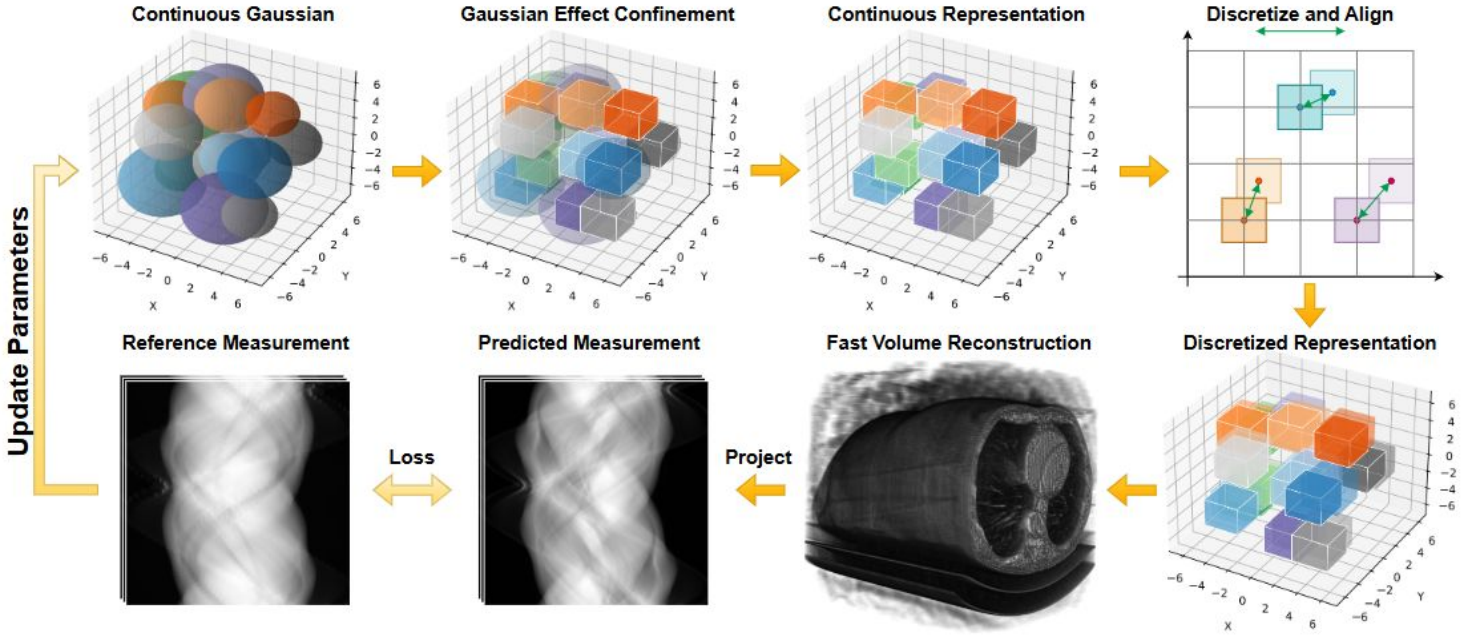




Discretized Gaussian Representation for Tomographic Reconstruction

[Shaokai Wu+, ICCV 2025]

- Fast, high-quality CT from sparse/low-dose views; existing methods need heavy training and don't generalize.
- Represent the volume as voxel-aligned isotropic Gaussians and optimize directly from projections (no pre-training).
- Reaches state-of-the-art reconstruction quality with fast convergence, robust across datasets and scan setups.



ICCV25: Meta Insights into Trends and Tendencies (89/153)

An OpenMind for 3D medical vision self-supervised learning

[Tassilo Wald+, ICCV 2025]

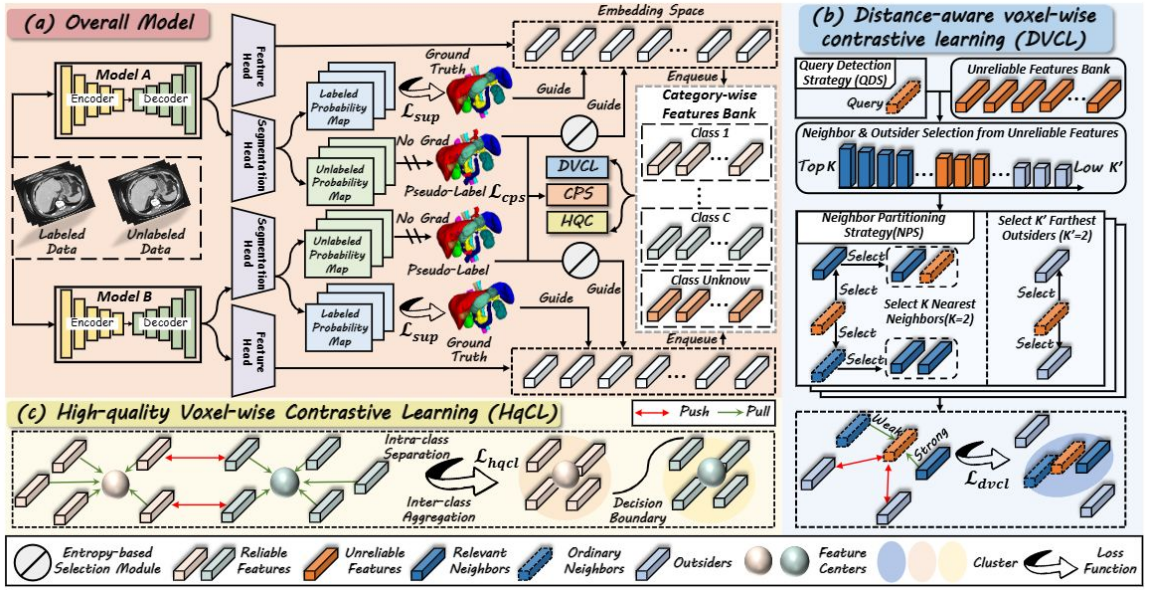
- OpenMind dataset: 114k 3D head-and-neck MRI volumes across 23-24 modalities from ~800 studies.
- Trains multiple 3D SSL methods on OpenMind and evaluates on 15 downstream datasets with two architectures: ResEnc-L (CNN) and Primus-M (Transformer).
- Provides a standard, large-scale, open foundation for 3D medical SSL - enabling fair comparisons, faster adoption, and data-/method-centric research.

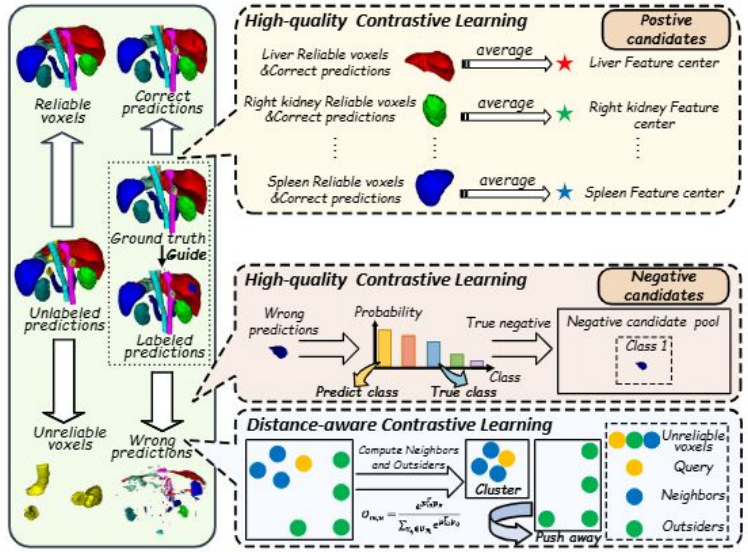
	Dice Similarity Coefficient (DSC) [%] on ...															
	Dataset of same anatomical region (ID)									Dataset of OOD region			Average across ...			
PT Method	ATL	SBM	ISL	HNT	HAN	MSF	TPC	YBM	COS	ACD	AMO	KIT	ID	OOD	All	
nnU-Net def. 1k	58.70	59.98	78.40	62.98	53.37	52.19	79.50	58.43	46.19	91.10	88.00	87.21	61.08	88.77	68.00	
nnU-Net def.	56.08	60.41	78.22	59.37	32.27	55.84	76.78	56.73	49.31	90.72	83.88	77.61	58.33	84.07	64.77	
ResEnc-L (CNN)																
Scratch 1k	58.21	53.43	79.14	65.75	58.24	54.90	79.94	56.12	71.57	92.09	88.73	87.48	64.15	89.43	70.47	
Scratch	57.02	54.29	78.09	63.30	56.11	55.47	76.18	54.42	65.20	91.97	85.24	84.03	62.23	87.08	68.44	
VoCo	57.14	59.62	77.50	63.48	51.12	54.90	75.12	56.92	63.49	91.44	85.60	85.71	62.14	87.58	68.50	
SwinUNETR	56.07	57.25	77.45	61.64	49.42	54.82	74.69	57.05	65.68	90.53	84.95	85.54	61.56	87.01	67.92	
SimCLR	57.15	59.72	78.01	63.32	51.56	55.68	77.77	59.14	68.20	91.76	86.06	84.85	63.40	87.56	69.44	
VF	57.42	59.88	78.18	64.32	51.67	57.42	76.11	59.31	63.98	91.57	85.38	86.21	63.14	87.72	69.29	
MG	58.03	61.57	77.58	65.11	54.69	55.25	77.14	58.67	71.27	91.74	86.35	86.17	64.37	88.09	70.30	
MAE	58.25	62.41	77.89	66.58	55.14	56.84	77.96	60.07	70.85	91.98	86.78	86.12	65.11	88.30	70.91	
S3D	58.76	64.09	78.05	65.74	52.81	56.08	78.81	59.18	66.66	92.01	86.16	86.01	64.46	88.06	70.36	
Primus-M (Transformer)																
Scratch 1k	56.77	48.50	76.59	58.40	53.40	53.27	76.32	52.53	64.68	90.89	87.24	85.57	60.05	87.90	67.01	
Scratch	51.51	43.26	75.23	55.30	50.60	54.00	73.31	50.30	62.11	90.93	80.17	76.73	57.29	82.61	63.62	
VoCo	46.80	34.15	73.29	51.06	47.64	52.64	65.52	44.75	52.16	87.26	65.81	70.21	52.00	74.43	57.61	
SwinUNETR	47.23	36.31	73.84	50.15	46.49	52.80	66.23	44.49	54.82	87.92	66.17	70.39	52.49	74.82	58.07	
SimCLR	54.61	42.62	75.43	56.75	50.80	53.59	70.08	48.36	58.20	89.97	75.75	81.27	56.72	82.33	63.12	
VF	58.62	47.37	77.56	62.37	56.18	55.00	74.98	53.96	69.74	91.41	84.95	86.17	61.75	87.51	68.19	
MG	56.50	47.34	76.76	58.42	54.02	54.67	73.65	49.77	60.74	90.89	82.15	84.45	59.10	85.83	65.78	
MAE	61.16	56.67	77.12	66.12	57.24	56.02	78.31	54.35	72.02	92.16	87.16	86.74	64.34	88.69	70.42	
SimMIM	60.28	51.68	77.53	62.76	56.74	55.91	77.00	52.90	70.87	91.98	86.57	85.92	62.85	88.16	69.18	

ICCV25: Meta Insights into Trends and Tendencies (90/153)

Keep Your Friends Close, and Your Enemies Farther: Distance-aware Voxel-wise Contrastive Learning for Semi-supervised Multi-organ Segmentation [Haochen Zhao, ICCV 2025]

- Problem: Pseudo-labels are noisy; standard voxel contrastive learning amplifies errors and wastes uncertain voxels.
- Even with noise, voxel features cluster locally. Pull close neighbors, push far outsiders for uncertain voxels; use standard contrastive learning for reliable ones.
- SOTA across FLARE'22, AMOS, MMWHS, BTCV with ~+2–5.5 Dice gains; largest boosts on hard organs; faster learning.





ICCV25: Meta Insights into Trends and Tendencies (91/153)

Trends in End-to-End 3D Vision

- **Learning over optimization:**
  - clear shift from classical optimization (SfM, MVS) to deep learning-driven approaches.
    - e.g., Dream-to-Recon, POMATO
- **Transformers & big priors:**
  - Transformer architectures are increasingly used to solve correspondence and integration problems.
  - Large pre-trained models (e.g., diffusion models, foundation models) are being repurposed to inject world knowledge / priors into 3D reconstruction (or 3D Synthesis?)
- **Dynamic scenes & real-time:**
  - Growing focus on handling realistic scenarios→longer sequences, moving objects, real-time operation.
  - The field is converging on solutions that brin 3D reconstruction out of static settings into the wild.
    - e.g., LONG3R
- **Open challenges:**
  - Despite progress, fully robust E2E3D remains challenging. Many approaches still rely on external inputs (e.g., known camera poses or pre-trained models) and face issues in extreme conditions (e.g. heavy occlusion, lighting changes).
  - Benchmarking and consensus on evaluation metrics lags behind the rapid innovations

Research Trends Remote Sensing

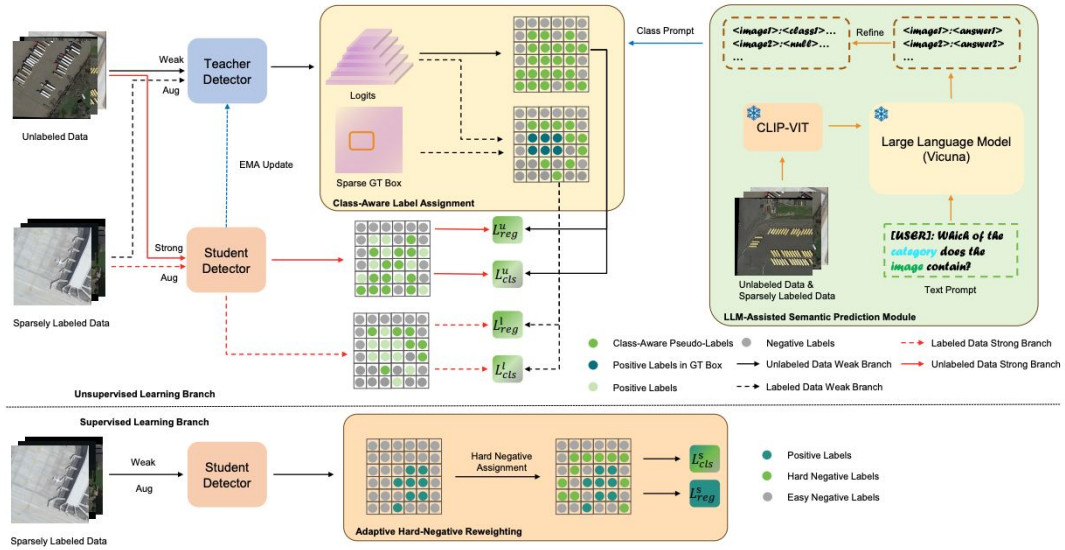
- Rise of multimodal models
  - Many papers address multi-sensor settings, open-vocabulary vision-language models, and links to behavioral/action data (e.g., navigation).
- Deep Learning-based Pan-Sharpening
  - converging on alignment-aware, interpretable, and deployment-oriented pan-sharpening
  - marrying in-network alignment and auxiliary supervision (PAN-Crafter) with unrolled, prior-driven optimization (Deep Adaptive Unfolded Network).
- Fusion of time series and physics constraints
  - physics-guided spatio-temporal reconstruction (e.g., air-temperature fields)
  - Koopman × ViT for EO time-series forecasting
  - S1/S2-based flood monitoring with gap filling.
  - introducing physical constraints and linear-dynamics priors yields time-series methods that are more robust to missing data, noise, and distribution shift.

ICCV25: Meta Insights into Trends and Tendencies (93/153)

Object detection for remote sensing

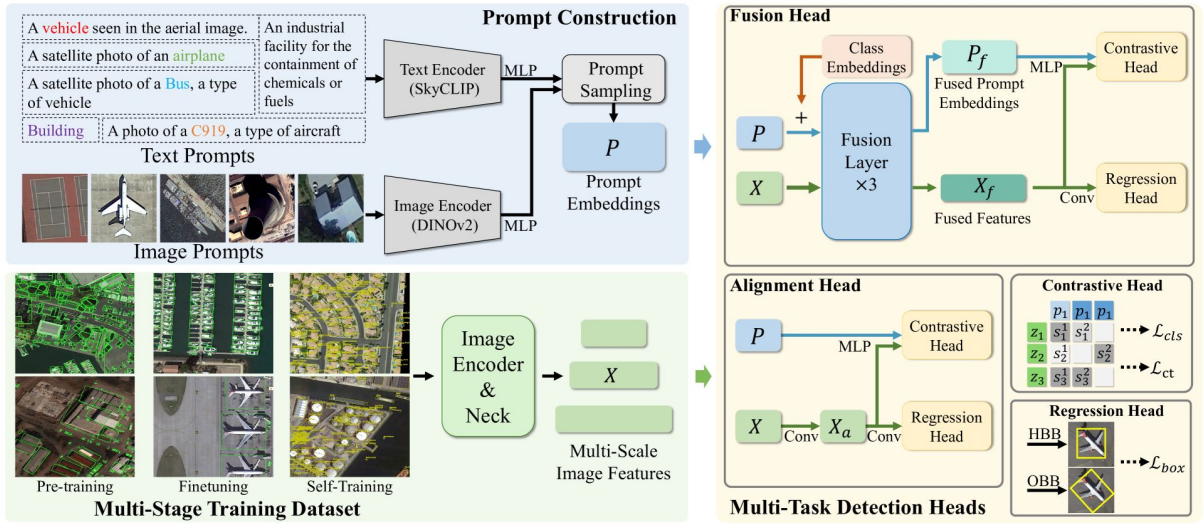
- Trend toward performance improvement with limited datasets by using language model semantics
  - Using language model semantics to support dense pseudo-labels
  - Augment pairs by using multi-prompt

Semantic Support for LLM on Sparsely Annotated Data



Wei Liao et al. (2025), "LLM-Assisted Semantic Guidance for Sparsely Annotated Remote Sensing Object Detection", ICCV.

OpenRSD : Multimodal Prompt-Based Object Detector



Ziyue Huang et al. (2025), "OpenRSD: Towards Open-prompts for Object Detection in Remote Sensing Images", ICCV.

Research Trends City Scale 3D Model

- Satellite to 3D City Generation Goes Mainstream
  - From single-image multi-view synthesis to geometry-consistent reconstruction. Synthetic city datasets improve generalization
  - key metrics: multi-view consistency, height/normal accuracy, visual fidelity.
- 3D Gaussian Splatting at City Scale & with DynamicsWide-area
  - high-speed reconstruction and online updates (4D/7D/SLAM integration) enable continuously updating Digital Twin(DT)s.
  - GS as a core representation accelerates rendering, editing, and map integration.
- Workshops of the Digital Twins
  - Generating Digital Twins from Images and Videos: focus on dynamic reconstruction and view consistency.
  - From street to space: EO-oriented NeRF scaling (Tile-and-Slide) and 3DGS to point-cloud conversion, bridging ground to satellite.
  - Neural-SLAM: direct GS × SLAM coupling (e.g., DROID-Splat) boosts practicality.

Accelerating of Diffusion or Flow based models (1/7)

- Contrastive Flow Matching
  - Enforce uniqueness across all conditional flows, enhancing condition separation
  - Add a contrastive objective that maximizes dissimilarities between predicted flows from arbitrary sample pairs
  - Improve training speed 9× faster, FID and require up to 5× fewer de-noising steps

Algorithm 1 Contrastive Flow-Matching Batch Step

Input: A model v_θ, batch of N flow examples $F = \{(x_1, y_1, \epsilon_1), \ldots, (x_N, y_N, \epsilon_N)\}$ where $(x_i, y_i) \sim p(x, y)$ and $\epsilon_i \sim \mathcal{N}(0, \mathrm{I})$, β learning rate, $\lambda = 0.05$.
Output: Updated model parameters θ
$L(\theta) = 0$
for i in range(N) **do**
  $t \sim U(0, 1), x_t = \alpha_t x_i + \sigma_t \epsilon_i$
  sample $(\tilde{x}, \tilde{y}, \tilde{\epsilon}) \sim F$, s.t. $(\tilde{x}, \tilde{y}, \tilde{\epsilon}) \neq (x_i, y_i, \epsilon_i)$
  $\hat{v} = v(x_t, t, y_i), v = \dot{\alpha}_t x_i + \dot{\sigma}_t \epsilon, \tilde{v} = \dot{\alpha}_t \tilde{x} + \dot{\sigma}_t \tilde{\epsilon}$
  $L(\theta) += ||\hat{v} - v||^2 - \lambda ||\hat{v} - \tilde{v}||^2$
end for
$\theta \leftarrow \theta - \frac{\beta}{N} \nabla_\theta L(\theta)$

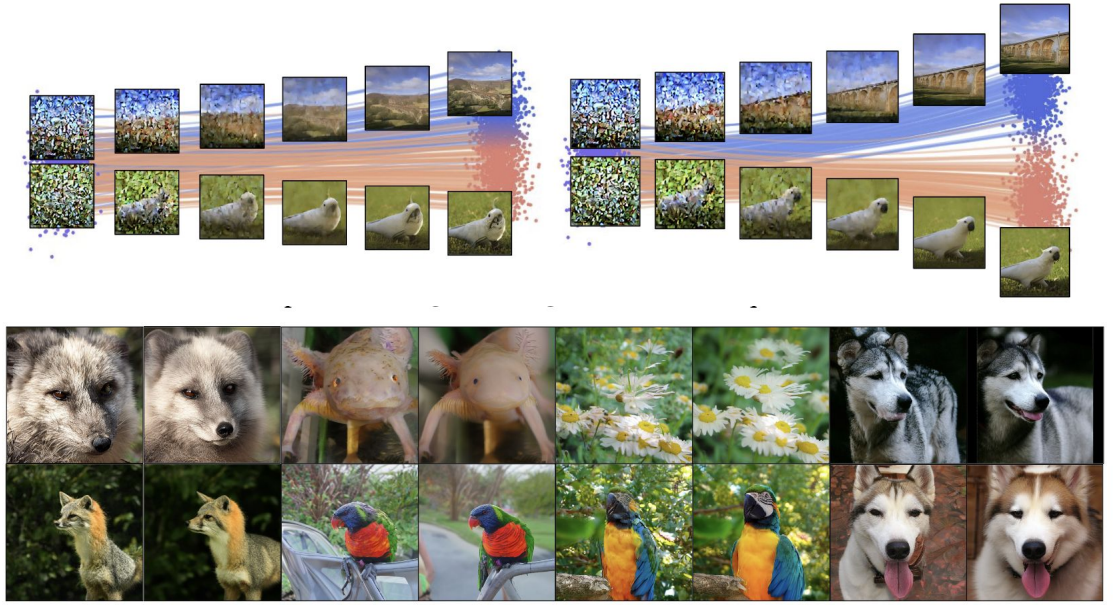
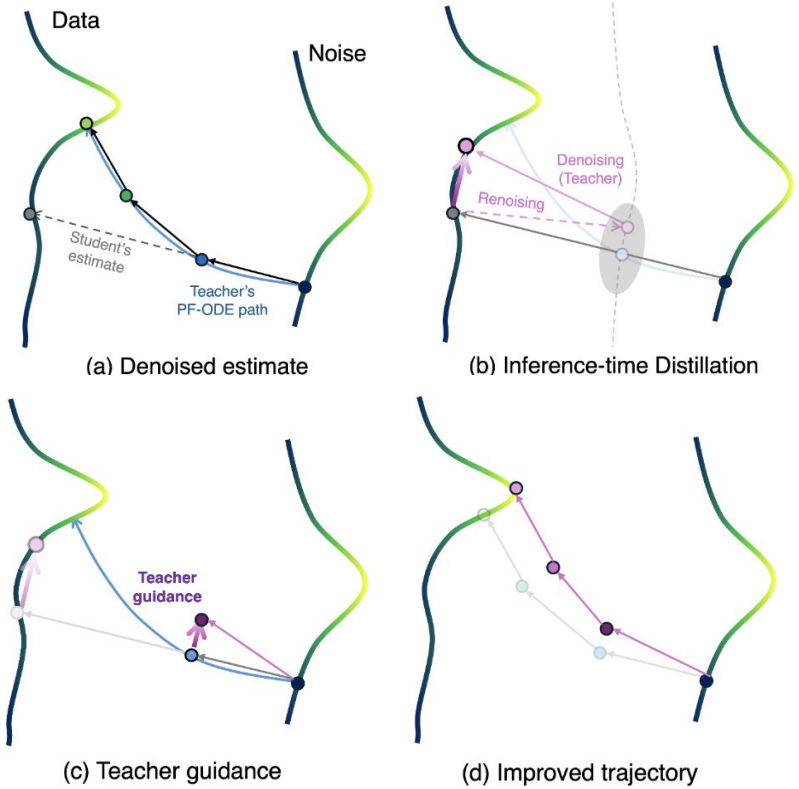


Figure 1. **Training with Contrastive Flow-Matching (ΔFM) improves natural image generation. (left** is baseline, **right** is with ΔFM) Here we show comparisons between images generated by diffusion models trained on ImageNet-1k (512 × 512). Each pair of images is generated with the same class and initial noise to ensure similar image structure for comparability. We see that our ΔFM objective encourages significantly more coherent images and improves the consistency of global structure.

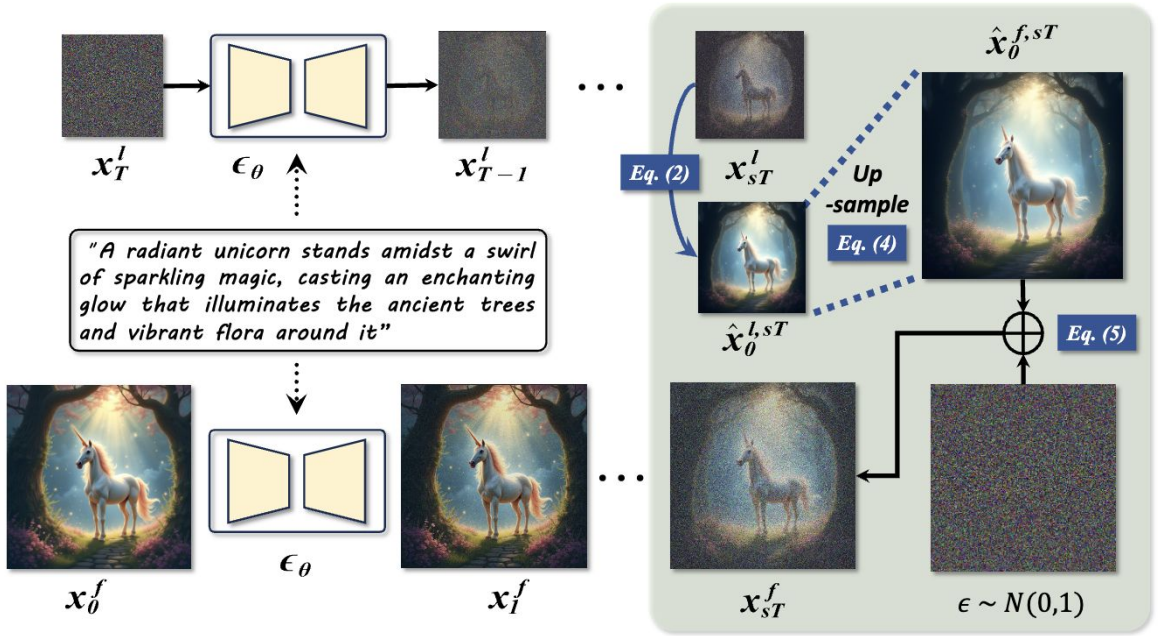
Accelerating of Diffusion or Flow based models (2/7)

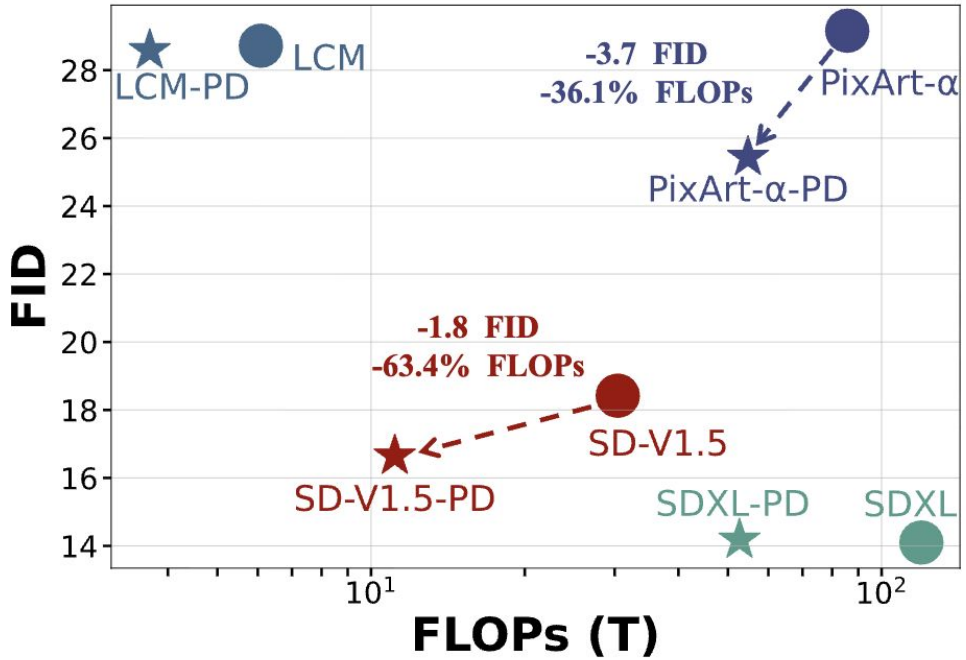
- Inference Time Diffusion Models Distillation
  - Distillation++: Inference-time distillation framework that reduces this gap by incorporating teacher-guided refinement during sampling
  - Improve particularly in early sampling stage over SOTA distillation baselines



Accelerating of Diffusion or Flow based models (3/7)

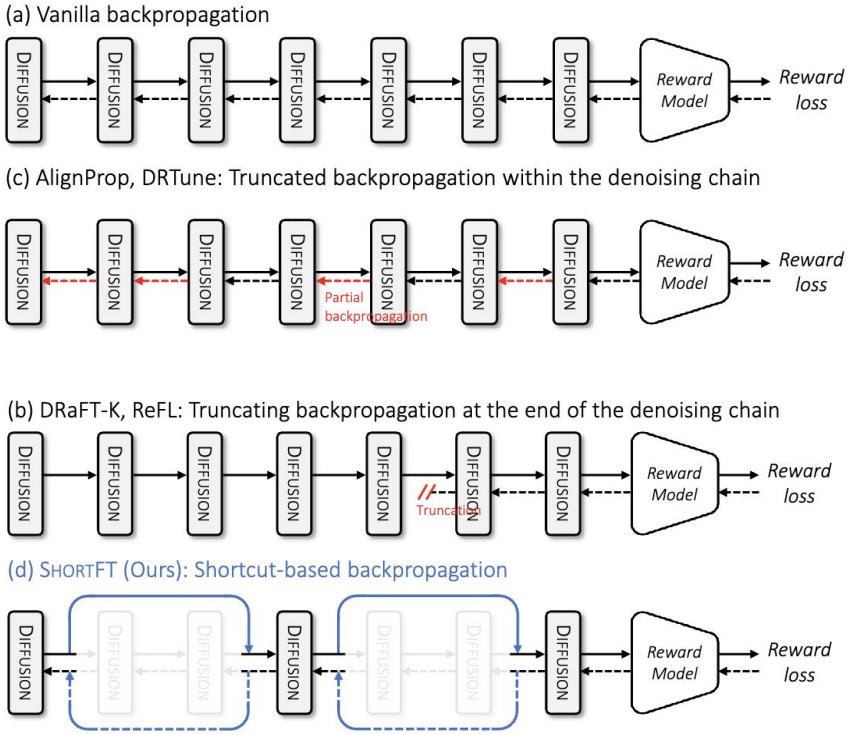
- Fewer Denoising Steps or Cheaper Per-Step Inference: Towards Compute-Optimal Diffusion Model Deployment
  - Mixed-resolution denoising scheme, which leverages the resolution in early denoising steps, achieves a win-win in both accuracy and efficiency
  - Hybrid module caching strategy to reuse computations across denoising steps

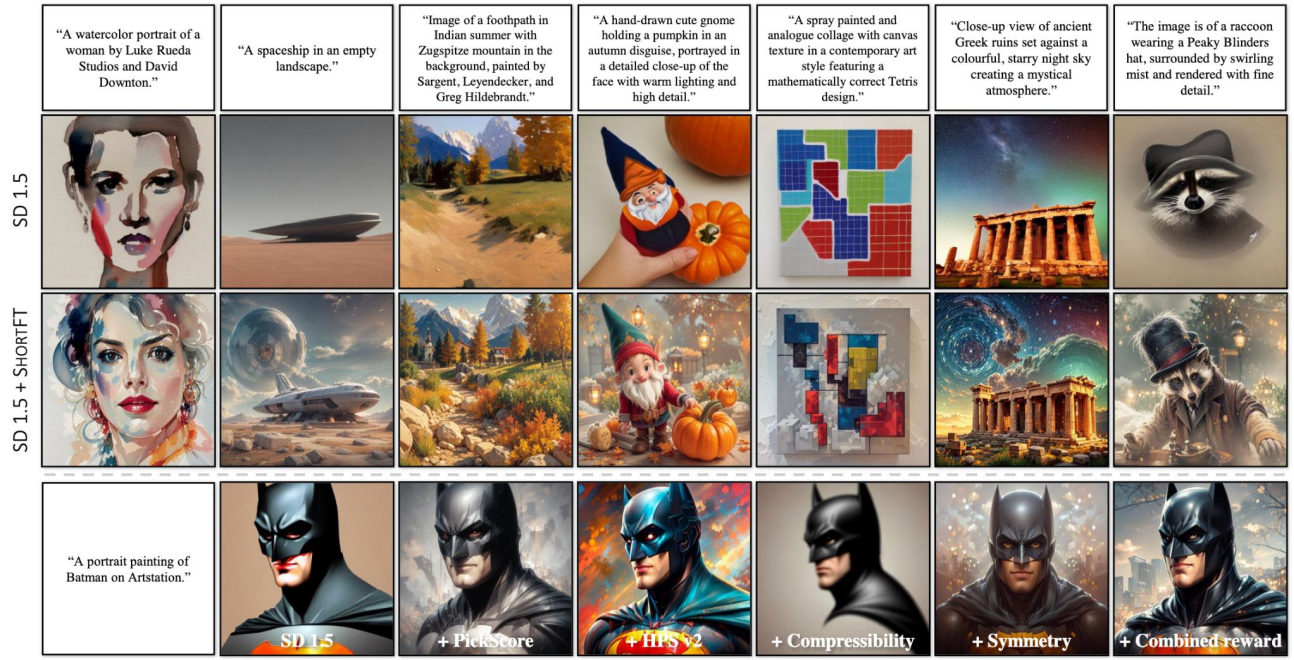




Accelerating of Diffusion or Flow based models (4/7)

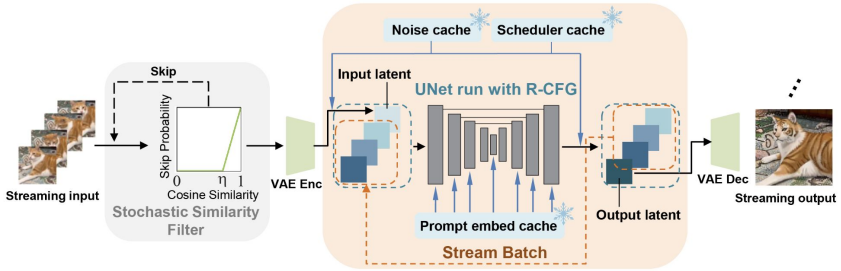
- SHORTFT: Diffusion Model Alignment via Shortcut-based Fine-Tuning
  - Shortcut-based Fine-Tuning (SHORTFT): an efficient fine-tuning strategy that utilizes the shorter denoising chain
  - Enhance the efficiency and effectiveness of fine-tuning the foundational mode.

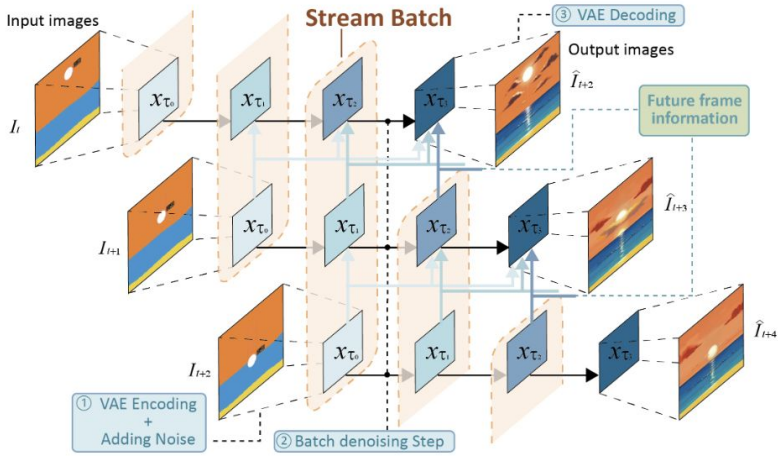


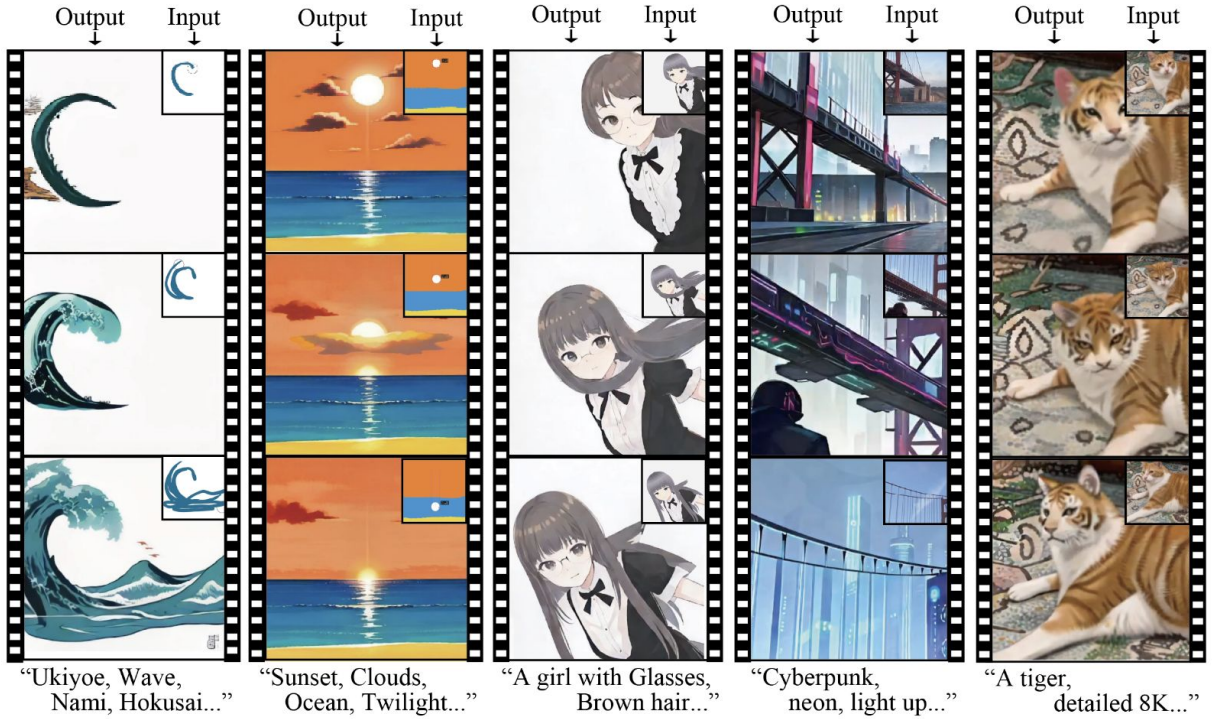


Accelerating of Diffusion or Flow based models (5/7)

- StreamDiffusion: A Pipeline-level Solution for Real-time Interactive Generation
  - Real-time diffusion pipeline designed for streaming image generation.
  - The authors also propose StreramDiT for Text-to-Video!!

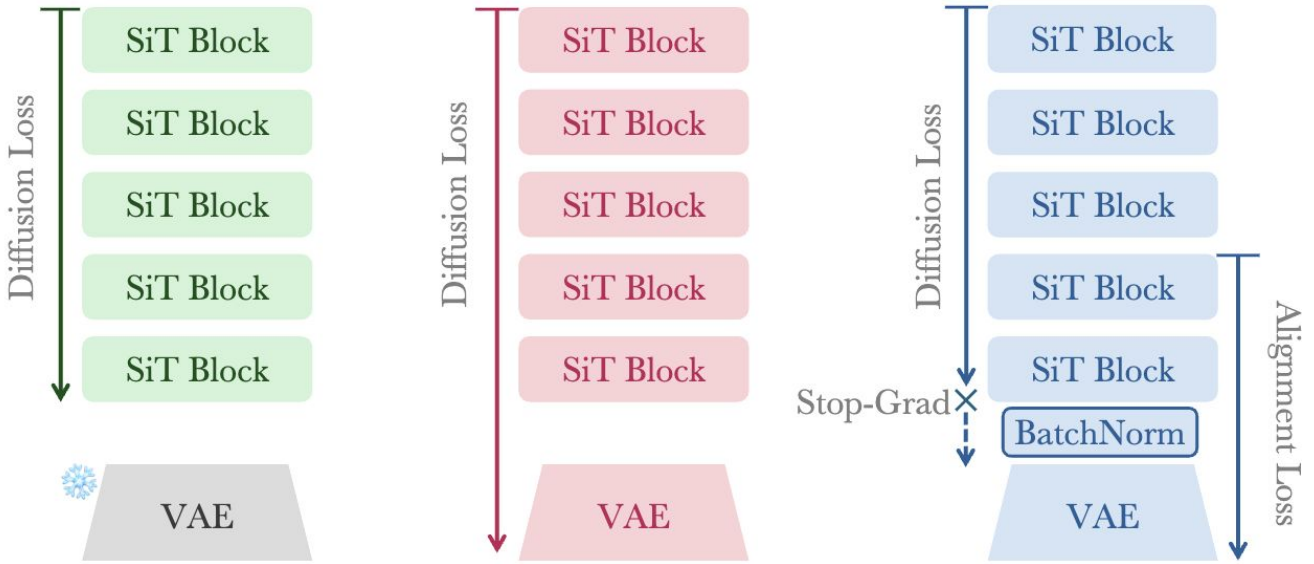






Accelerating of Diffusion or Flow based models (6/7)

- REPA-E: Unlocking VAE for End-to-End Tuning with Latent Diffusion Transformers
  - RQ: Can we train LDMs together with the VAE in an end-to-end manner?
  - Diffusion loss is ineffective, end-to-end training can be unlocked through the representation-alignment (REPA) loss



a) Traditional LDM Training

b) Naïve End-to-End LDM Training

c) REPA-E (Ours)

d) Training Steps vs. FID-50K Improved Generation Performance

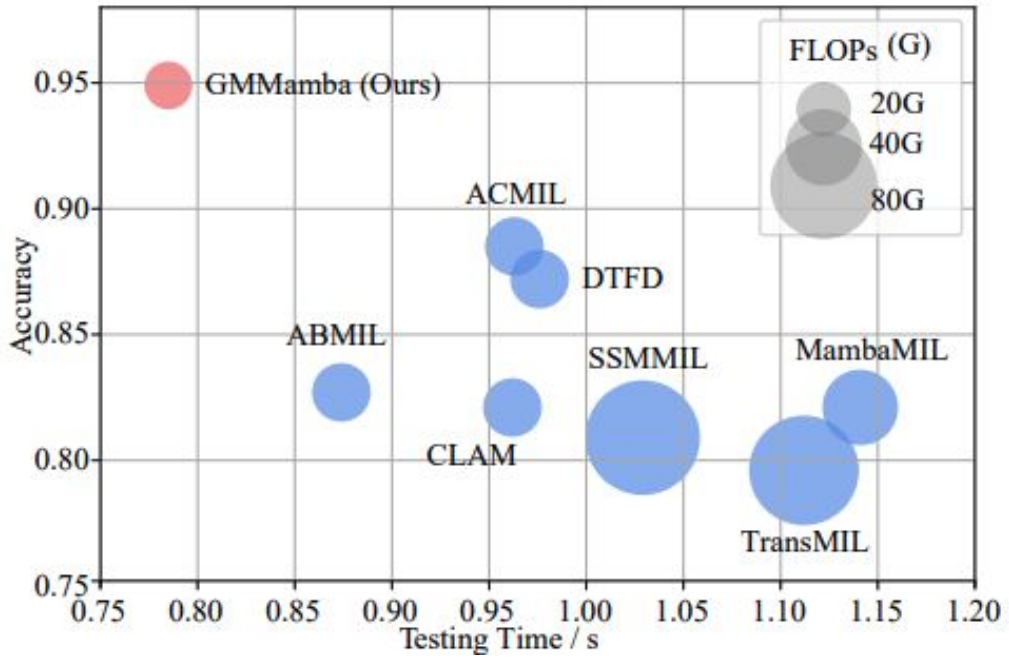
Accelerating of Diffusion or Flow based models (7/7)

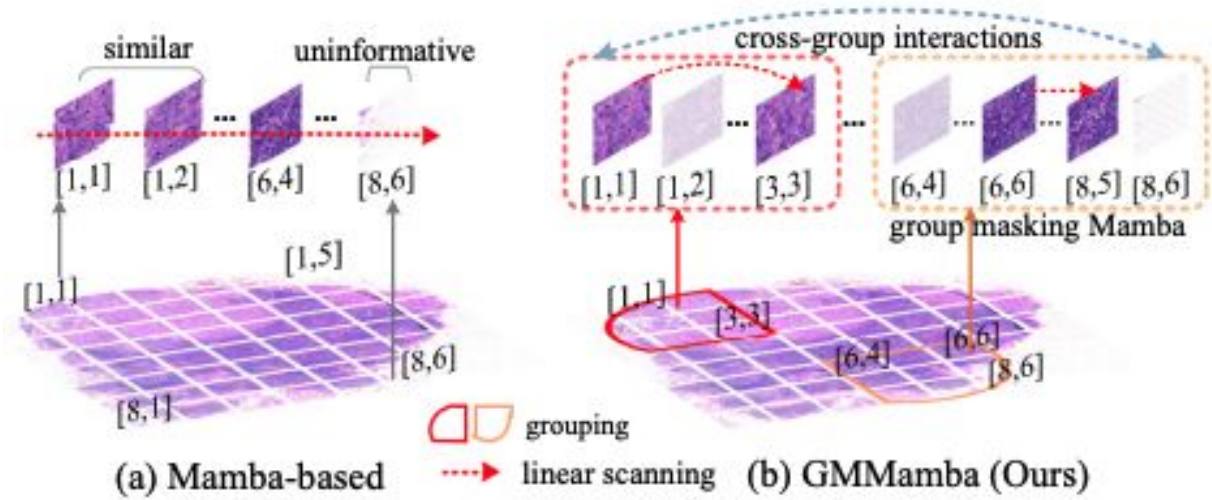
- Many studies related to "acceleration" across all domains of training (or fine-tuning) and inference were accepted
- The emergence of DiT is likely to further increase their importance
- In particular, integration with representation learning methods like REPA-E should play a crucial role in accelerating learning speed

ICCV25: Meta Insights into Trends and Tendencies (102/153)

GMMamba: Group Masking Mamba for Whole Slide Image Classification

- Proposing group masking Mamba (GMMamba)
  - Faster testing time, and higher Acc. on TCGA-ESCA datasets.
- GMMamba enhances the efficiency of Mamba-based MIL methods through 2 modules, IMM (Intra-group masking Mamba) and CSS (Cross-group Super-feature Sampling)
  -





SIC: Similarity-Based Interpretable Image Classification with Neural Networks

- SIC provides faithful local, global, and pixel-level explanations.
- Strength: SIC is readily applicable to datasets with hundreds of classes—more so than k-means and SVM.

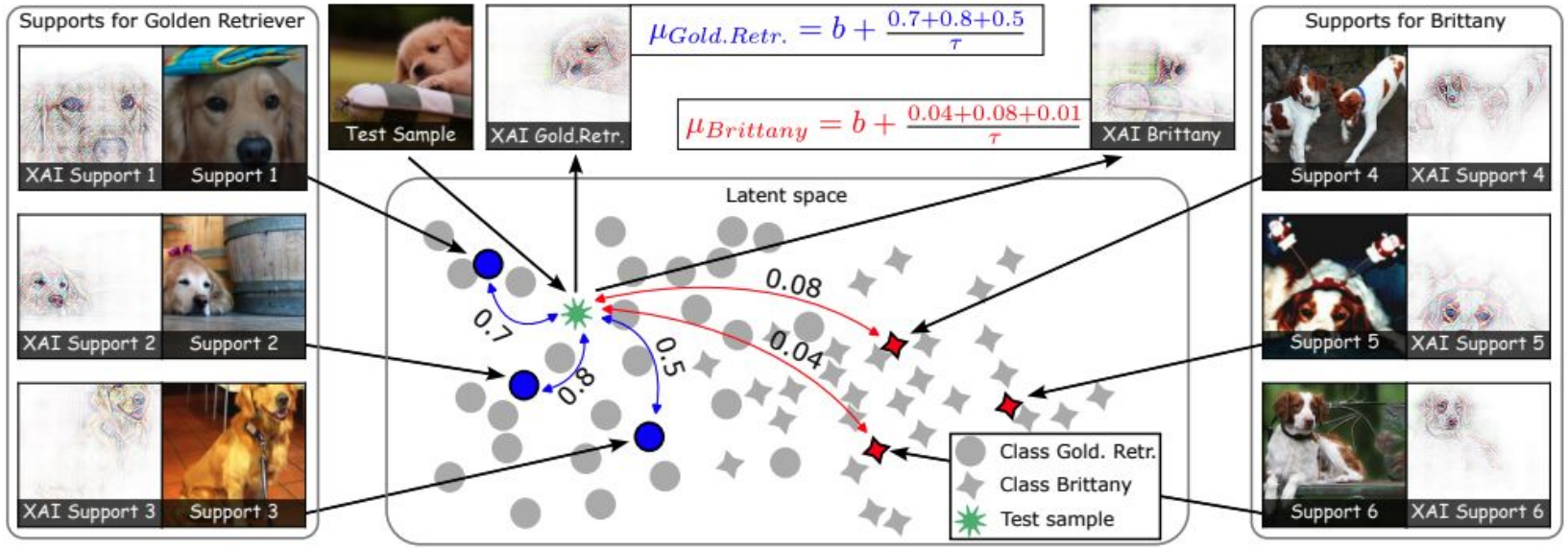
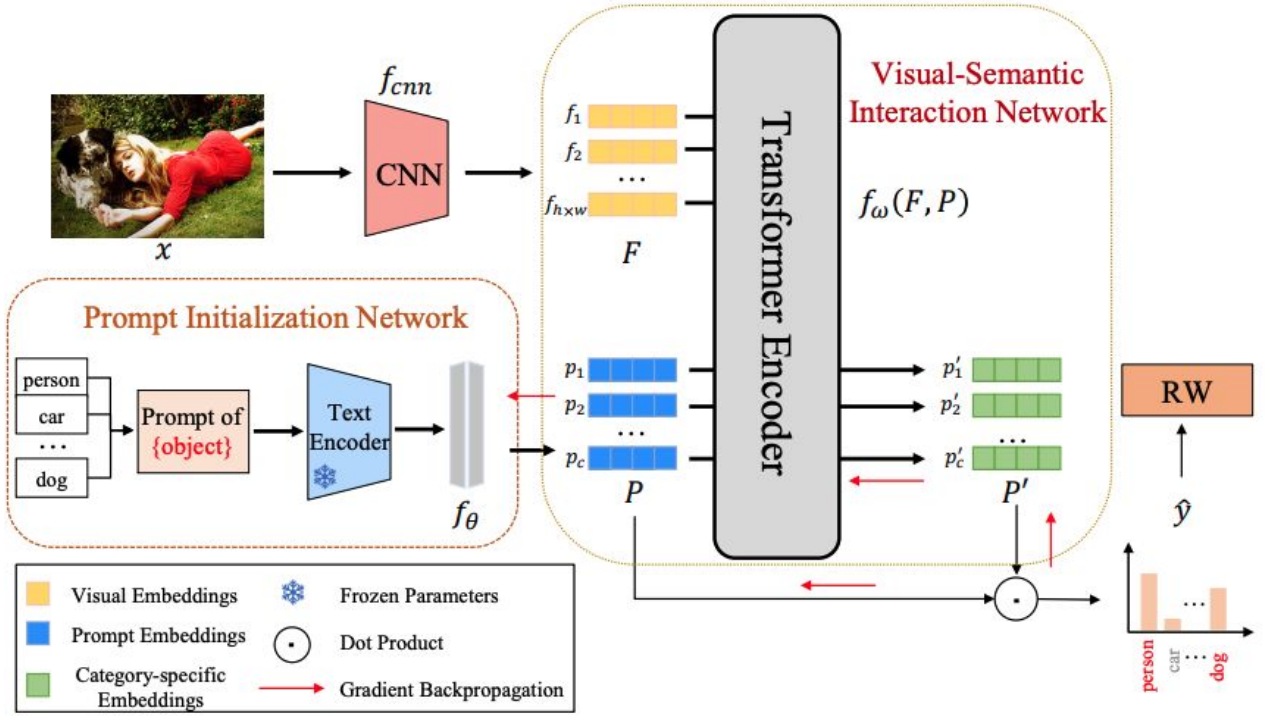


Figure 1. SIC trains a neural network to extract class-representative support feature vectors (red and blue) from training images. It computes class logits μ as the sum of temperature-normalized similarity scores between a class's support feature vectors and a test sample's feature vector (green), as shown for the predicted *Golden Retriever* class and non-predicted *Brittany* class. Its B-cos backbone permits the computation of faithful local and global explanations. Stars and circles denote samples of different classes.

ICCV25: Meta Insights into Trends and Tendencies (104/153)

Category-Prompt Refined Feature Learning for Long-Tailed Multi-Label Image Classification

- For multi-object recognition (in a single image), this work provides CPRFL.
- CPRFL establish the semantics correlations between the head and tail classes by applyi CLIP
  encoder.
- This work's knowledge offers an innovative solution tailored to the unique characteristics of the data.



ICCV25: Meta Insights into Trends and Tendencies (105/153)

SUB: Benchmarking CBM Generalization via Synthetic Attribute Substitutions

- CBM: Concept Bottleneck Models
  - CBMs make AI applications more transparent.
- This work demonstrates the weakness of CBMs under distribution shifts and introduces SUB, a benchmark of photorealistic single-attribute substitutions for fine-grained evaluation.
- Meta insights: Explainability for long-tailed image recognition is now a recognized research trend in the computer-vision academia.

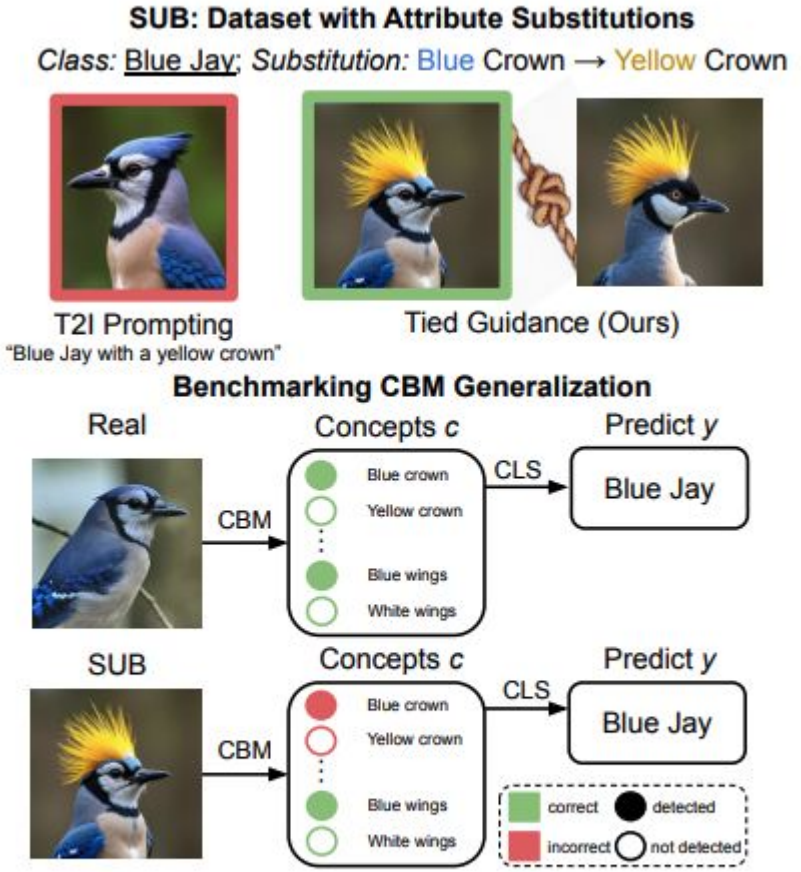
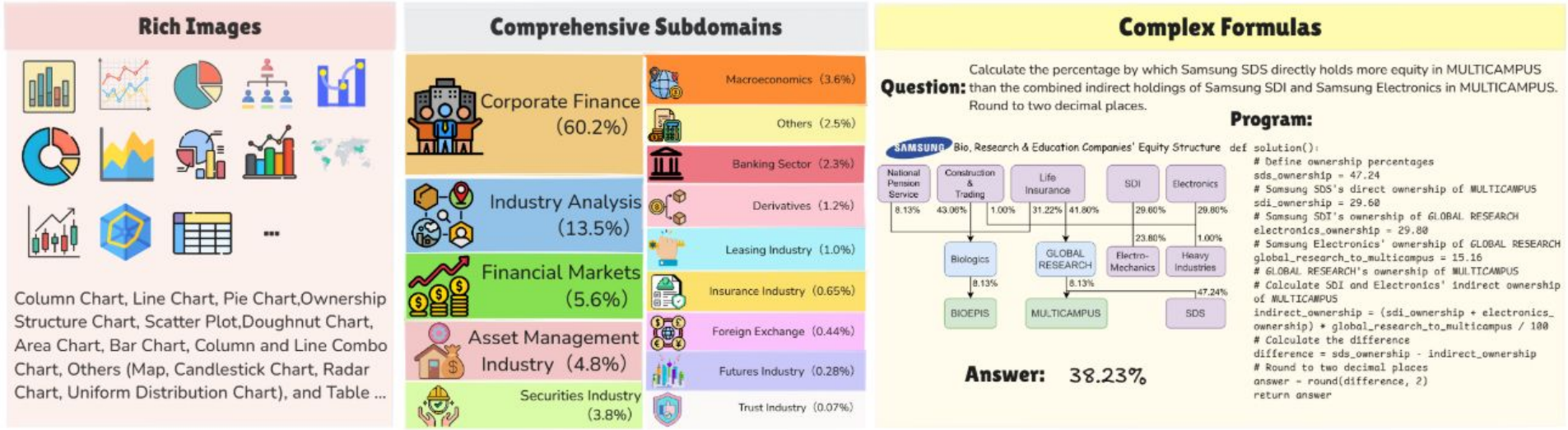


Figure 1. (Top) TGD modifies attributes where prompting fails. (Bottom) The CBM generalizes poorly, memorizing the "Blue Jay" concept vector and mis-classifying the modified concept.

ICCV25: Meta Insights into Trends and Tendencies (106/153)

FinMMR: Make Financial Numerical Reasoning More Multimodal, Comprehensive, and Challenging



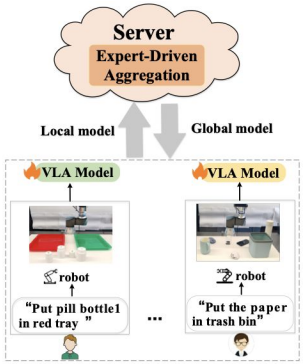
Overview of the FinMMR dataset. FinMMR presents three challenges: (1) visual perception: 8.7K financial images of 14 categories; (2) knowledge reasoning: 4.3K financial questions of 14 subdomains; (3) numerical computation: multi-step precise calculation.

- bilingual multimodal benchmark
  - Multimodality: Existing financial reasoning datasets was transformed.
  - Comprehensiveness: FinMMR encompasses 14 financial subdomains.
  - Challenge: Models are required to perform multi-step precise numerical reasoning.

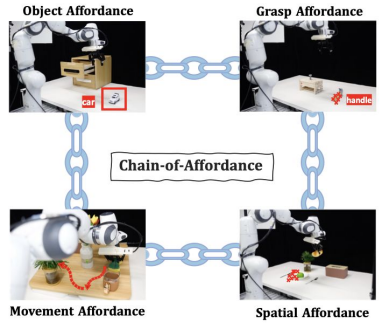
ICCV25: Meta Insights into Trends and Tendencies (107/153)

Exploration for constructing Vision-Language-Action model(VLA)

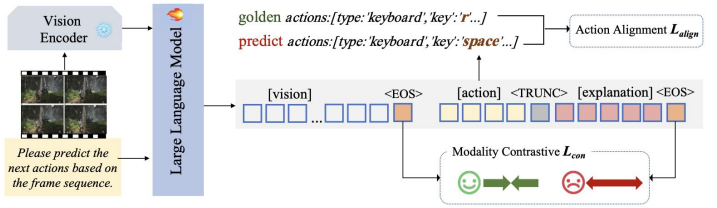
- Efficient training methods
  - C. Miao+: Mixture of Experts
  - P. Chen+: Chain of Thought with Action for RPG
  - J. Li+: Chain of Thought with Multiple Affordance
  - Y. Wang+: Action tokenization with Vector Quantization
  - Z. Hou+: Diffusion Transformer based VLA without VLM
- Long-horizon task
  - D. Li+: Reasoning and Memorization for Navigation
- Robustness
  - T. Wang+: Adversarial Attack to VLA



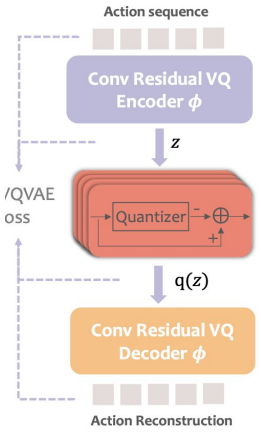
C. Miao+



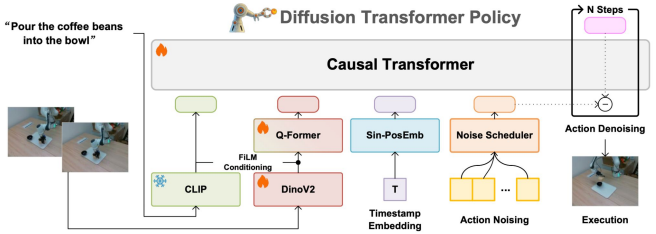
J. Li+



P. Chen+



Y. Wang+

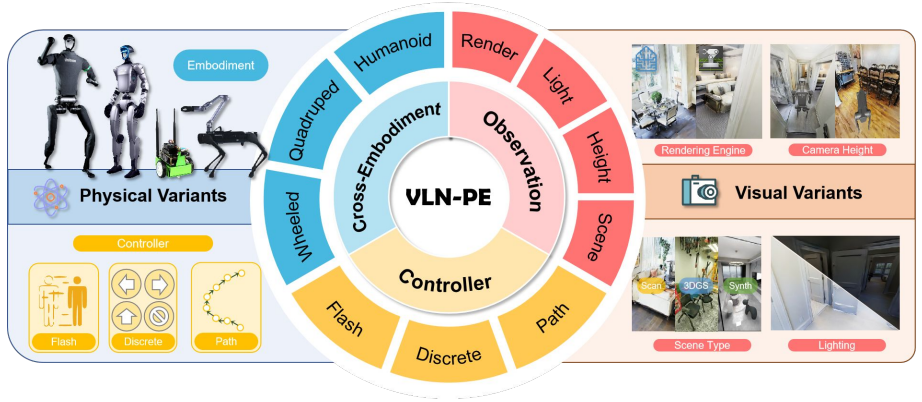


Z. Hou+

ICCV25: Meta Insights into Trends and Tendencies (108/153)

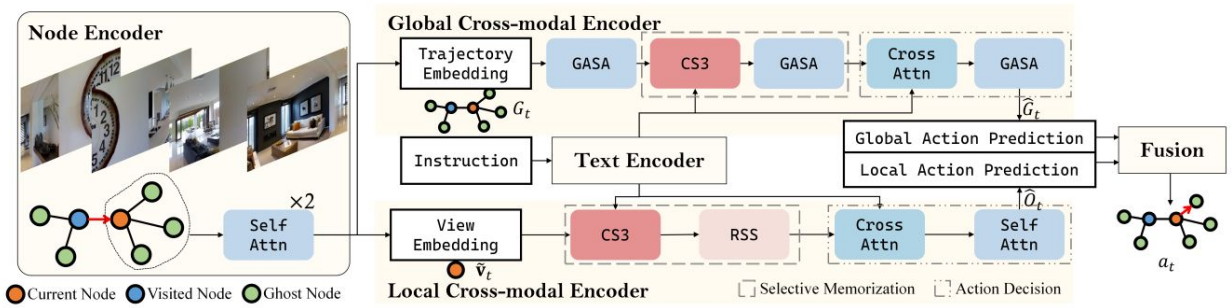
Robotics: Vision-and-Language Navigation (VLN)

❑ Cross-embodiment platform



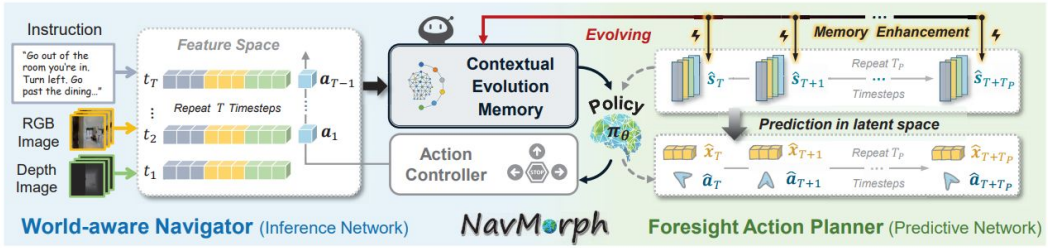
L. Wang, et al. "Rethinking the Embodied Gap in Vision-and-Language Navigation: A Holistic Study of Physical and Visual Disparities", in ICCV 2025.

❑ Selective memorization with SSMs



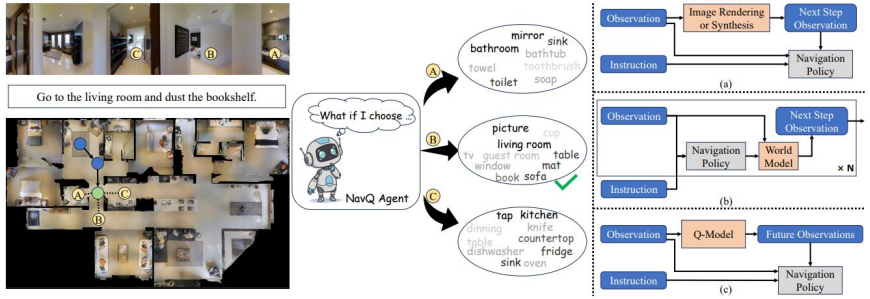
S. Zhang, et al. "COSMO: Combination of Selective Memorization for Low-cost Vision-and-Language Navigation", in ICCV 2025.

❑ Self-evolving world models



X. Yao, et al. "NavMorph: A Self-Evolving World Model for Vision-and-Language Navigation in Continuous Environments", in ICCV 2025.

❑ Q-learning based future modeling

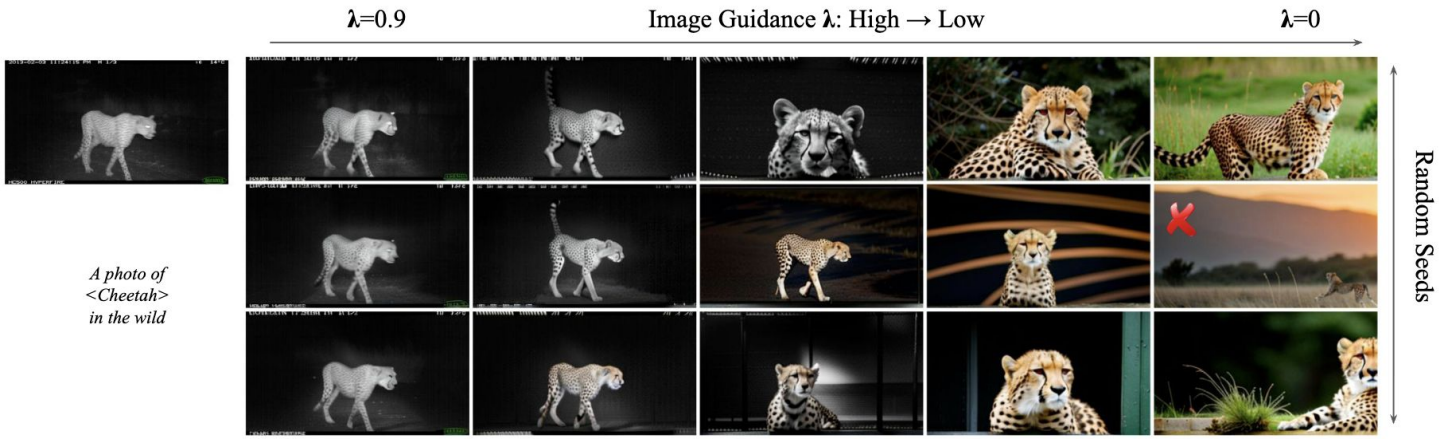


P. Xu, et al. "NavQ: Learning a Q-Model for Foresighted Vision-and-Language Navigation", in ICCV 2025.

ICCV25: Meta Insights into Trends and Tendencies (109/153)

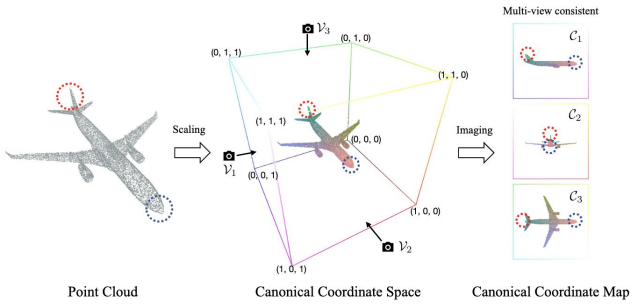
Diffusion Curriculum: Synthetic-to-Real Data Curriculum via Image-Guided Diffusion

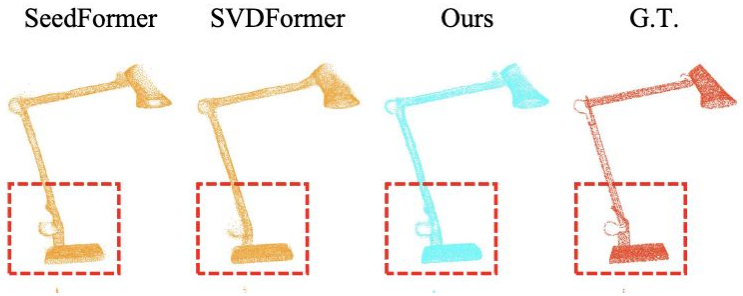
- Challenge:
  - Hard samples (long-tailed, low-quality) degrade model performance
  - Text-only synthetic training data causes distribution gap to real data
- Key idea:
  - Image guided diffusion with strength weight: synthetic-to-real spectrum
  - Curriculum: progressively shift from low to high guidance images
- Meta insights:
  - For data with limited or biased quantities compared to images (such as point clouds), wouldn't this method be even more effective?



GeoFormer: Learning Point Cloud Completion with Tri-Plane Integrated Transformer

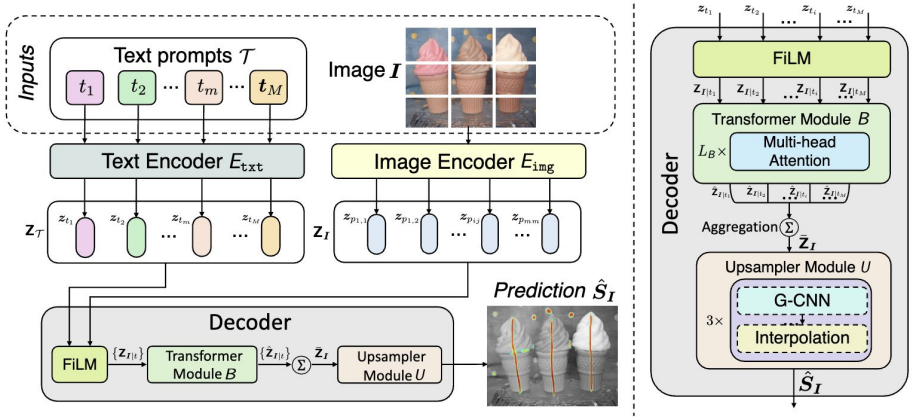
- Challenge:
  - Research on Point Cloud Completion
    Given a partial 3D object point cloud, the goal is to recover the missing geometry.
- Meta insights:
  - Conclusion：Achieves higher-accuracy point cloud completion than prior methods.
  - Limitations of Prior Work：When leveraging multi-view depth maps, inconsistencies arise due to non-unified viewpoints, hurting geometric coherence.
  - Method：
    - Enforce viewpoint consistency via a canonical coordinate map.
    - Fuse the canonical view information with point-cloud feature representations.
    - This combination yields high-precision reconstruction of missing regions.





CLIPSym: Delving into Symmetry Detection with CLIP

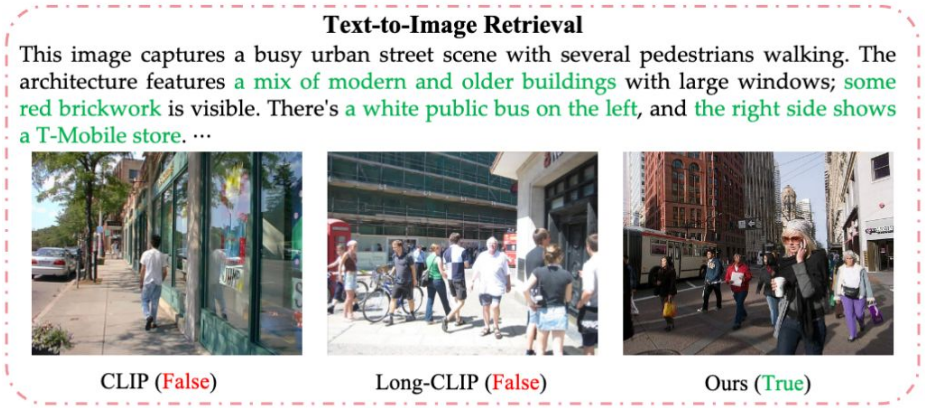
- Challenge:
  - This paper propose CLIPSym, a new method for detecting symmetry in images.
- Meta insights:
  - Datasets: DENDI, SDRW, LDRS (symmetry detection benchmarks)
  - Tasks: Reflection and rotation symmetry
  - Result: New state of the art on both tasks across all datasets
  - Contributions:
    - End-to-end framework for symmetry detection with CLIP
    - SAPG to strengthen language-side understanding of symmetry
    - Rotation-equivariant decoder for improved robustness

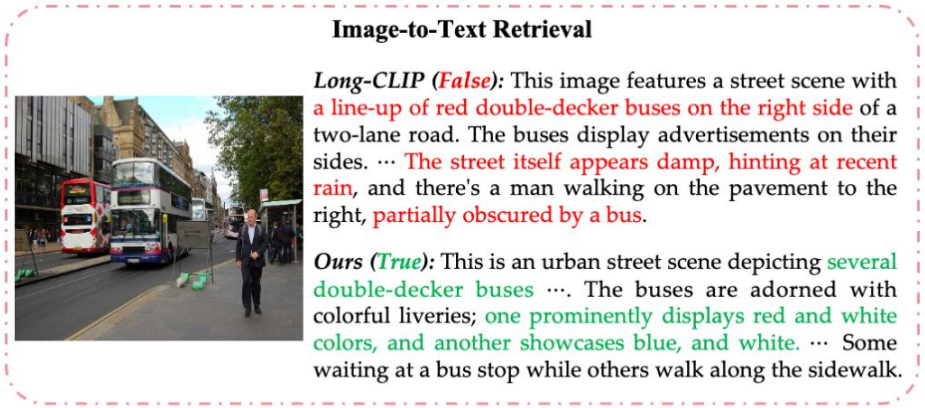


ICCV25: Meta Insights into Trends and Tendencies (112/153)

FIX-CLIP: Dual-Branch Hierarchical Contrastive Learning via Synthetic Captions for Better Understanding of Long Text

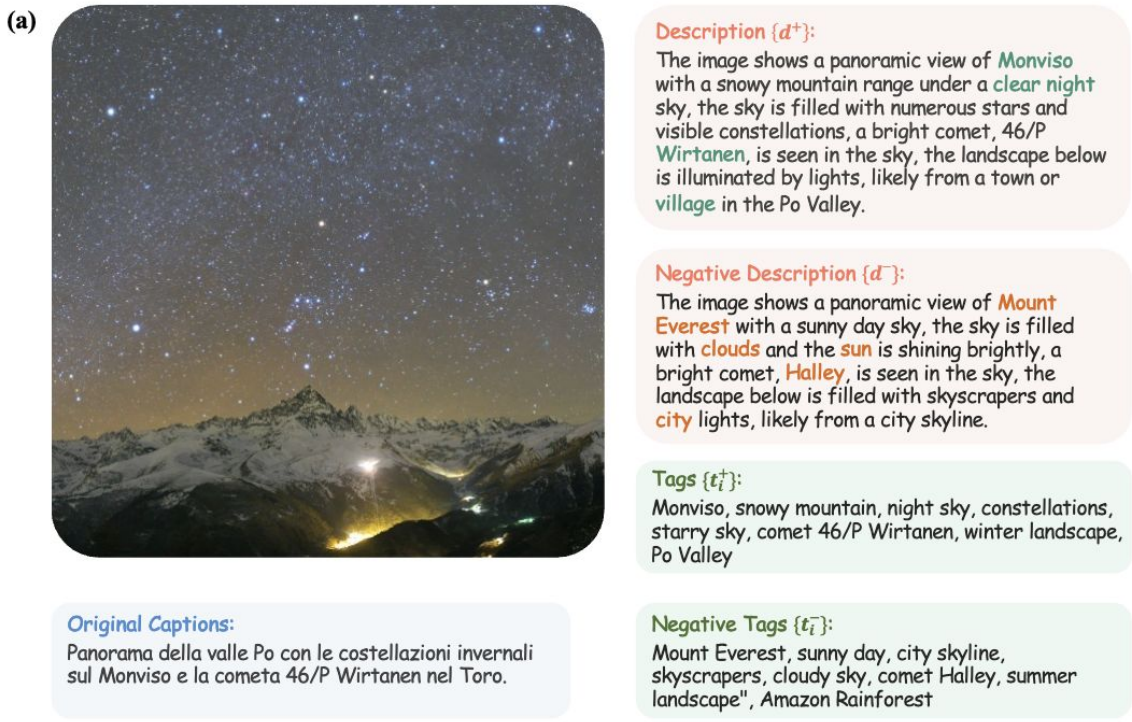
- Challenge:
  - Limitation of CLIP (Text Side):
    77-token cap in the text encoder, Weak with long inputs (> 77 tokens)
    For long descriptions (retrieval / generation), performance plateaus
- Meta insights:
  - Datasets: CC3M/12M, YFCC15M, VG, SBU, ShareGPT4V
  - Method: Use Llama3-LLaVA-NeXT-8B to generate long-form captions
  - Result: Re-caption existing images at large scale with richer descriptions
  - With fast caption generation, this framework becomes highly useful.





HQ-CLIP: Leveraging Large Vision-Language Models to Create High-Quality Image-Text Datasets and CLIP Models

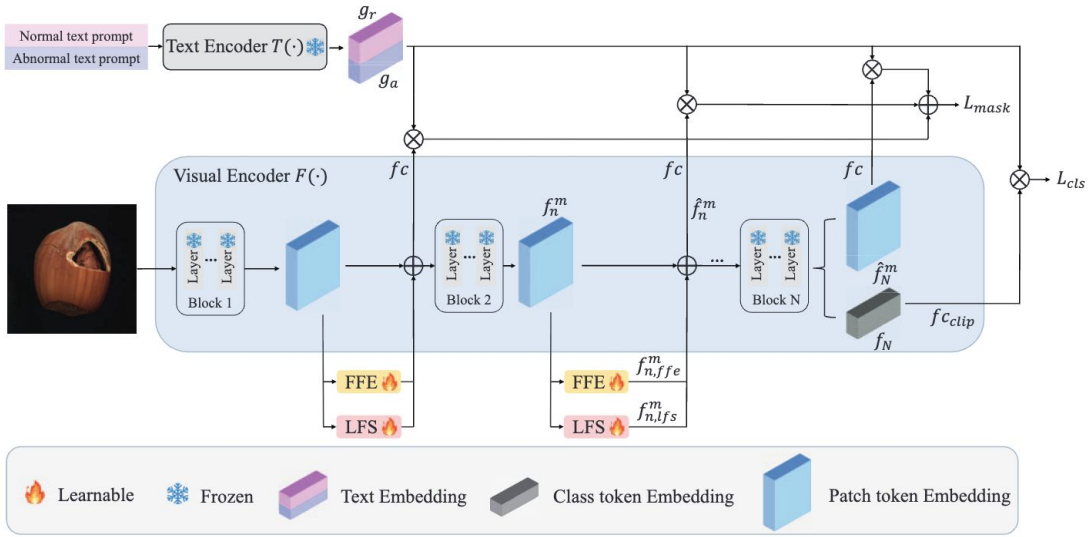
- Challenge:
  - Can we use LVLMs in reverse to improve data quality and self-improve CLIP
- Meta insights:
  - Method:
  - Input: image + existing alt-text
  - Outputs (4 complementary texts):
    - Long positive/negative description
    - Short positive/negative tags
    - use all four as multi-granular positive and negative training signals.



ICCV25: Meta Insights into Trends and Tendencies (114/153)

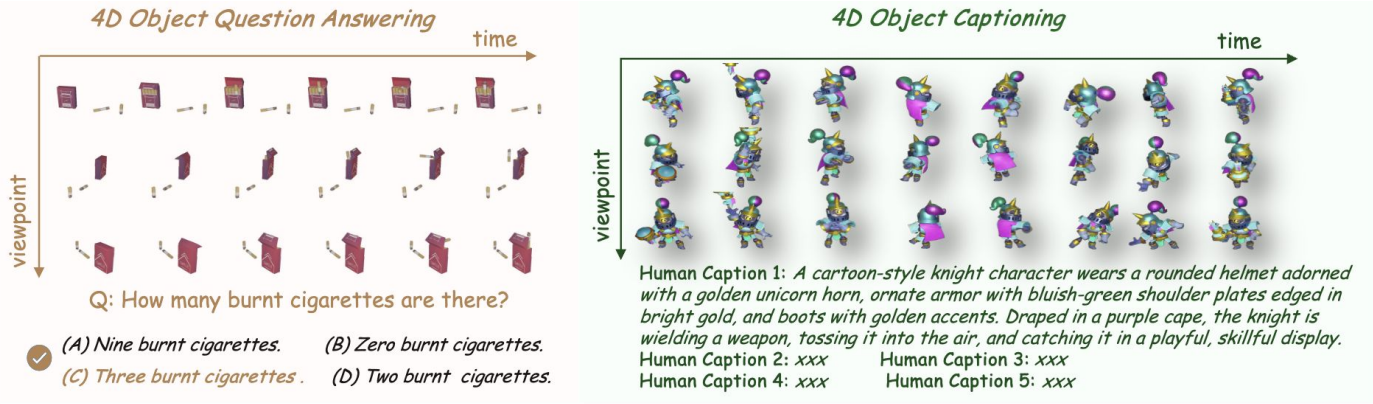
FE-CLIP: Frequency Enhanced CLIP Model for Zero-Shot Anomaly Detection and Segmentation

- Challenge:
  - This paper improves ZSAD/ZSAS generalization by leveraging frequency information often overlooked by CLIP-based methods.
- Meta insights:
  - Result:
    Zero-shot detection & segmentation: SOTA-level across all 10 datasets
    Beats prior CLIP-based methods in many cases (e.g., WinCLIP, AnomalyCLIP, AdaCLIP)



4D-Bench: Benchmarking Multi-modal Large Language Models for 4D Object Understanding

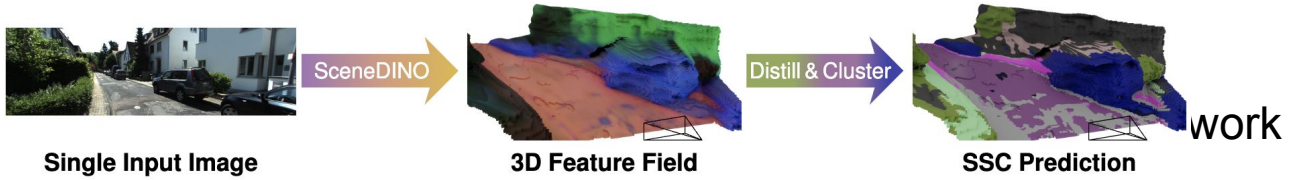
- Purpose and Method
  - Investigate whether existing MLLMs can directly perform 4D object recognition (spatio-temporal understanding)
  - Propose two human-curated benchmarks: Question Answering and Captioning
    - QA: Tests counting, temporal reasoning, and action understanding
    - Captioning: Requires describing actions and events within 4D scenes
- Results
  - Overall, MLLMs perform far below human level (GPT-4o: 62.98% vs Human: 91.08%)
  - Action and temporal understanding remain the weakest aspects
  - Closed-source models outperform open-source models



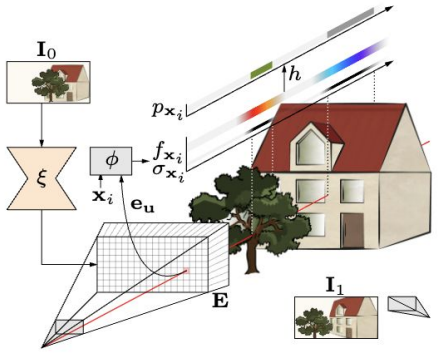
ICCV25: Meta Insights into Trends and Tendencies (116/153)

Feed-Forward SceneDINO for Unsupervised Semantic Scene Completion

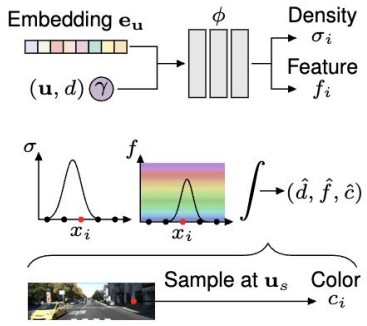
- Overview
  - Previous SSC methods relied on supervi work that enables single-image Semantic Scene Completion from multi-view consistency



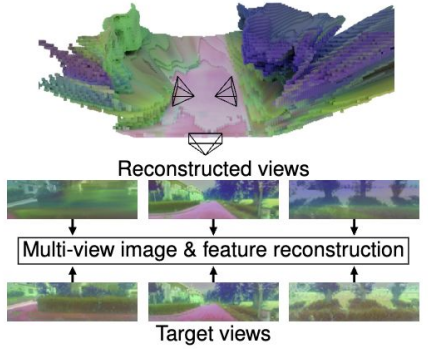
  - The model predicts 3D geometry and 3D features in a single feed-forward pass using a 2D encoder–decoder and an MLP decoder
- Method
  - Using 2D-DINO features as supervision targets, the network reconstructs consistent features across sampled views, poses, and transformations
  - To achieve unsupervised SSC, a projection head maps features to a low-dimensional semantic space, where corresponding pairs are sampled via surface points and depths estimated from the model's predicted densities



(a) 3D feature field & semantic inference



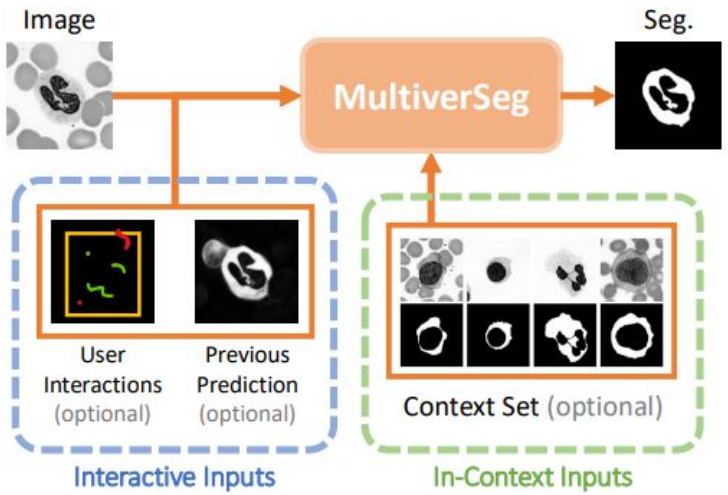
(b) Volumetric feature & image rendering

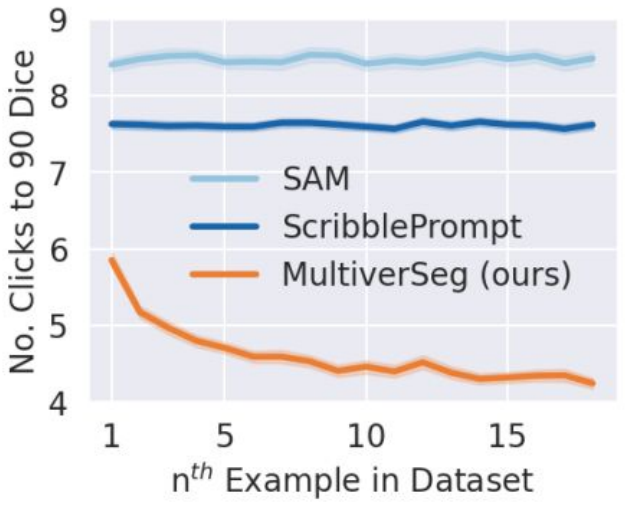


(c) Self-supervised multi-view training

MultiverSeg: Scalable Interactive Segmentation of Biomedical Imaging Datasets with In-Context Guidance

- MultiverSeg is a revolutionary framework designed to reduce the "annotation fatigue" typically associated with labeling new medical image datasets.
- It is the first to merge interactive segmentation with in-context learning, enabling immediate, zero-shot labeling of new tasks without requiring pre-existing domain labels.
- The system progressively learns from the user: Every time a segmentation is finalized, that high-quality label is added to the Context Set, ensuring the model gets smarter and demands fewer manual corrections on every subsequent image.

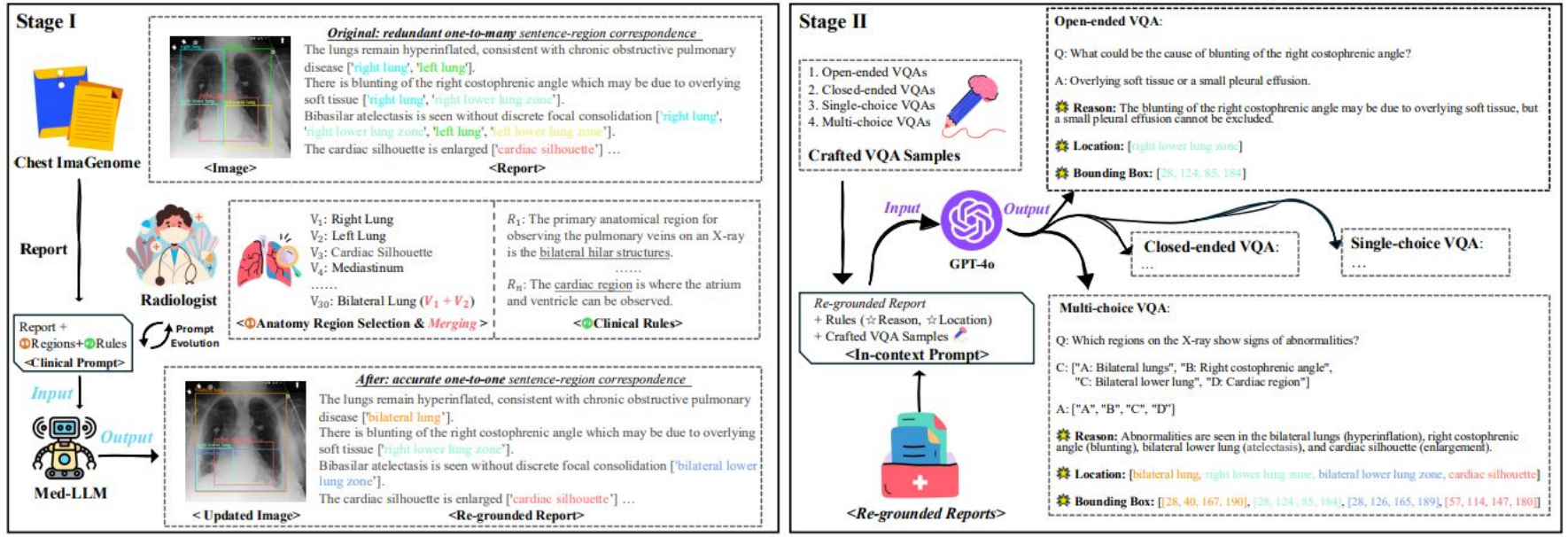




ICCV25: Meta Insights into Trends and Tendencies (118/153)

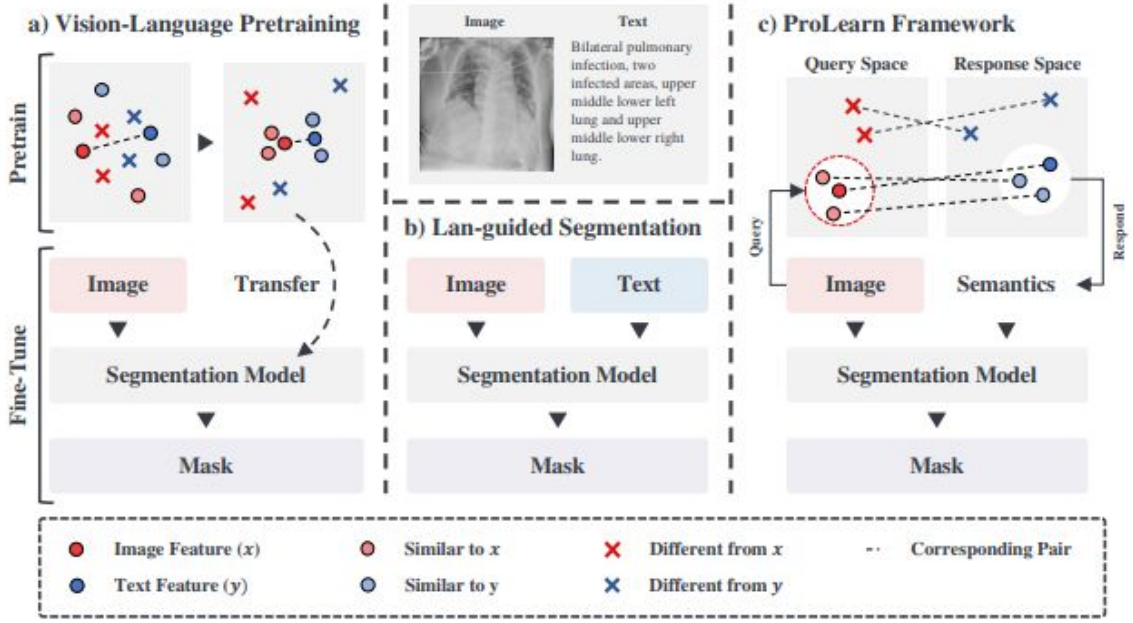
GEMeX: A Large-Scale, Groundable, and Explainable Medical VQA Benchmark for Chest X-ray Diagnosis

- GEMeX is a massive, AI-generated, and explainable chest X-ray VQA dataset.
- It includes over 151,000 images and approximately 1.6 million QA pairs.
- The benchmark features four diverse question types: Open-ended, Closed-ended, Single-choice, and Multi-choice.
- The data generation pipeline leveraged a medical LLM, OpenBioLLM-70B, for re-grounding reports



Alleviating Textual Reliance in Medical Language-guided Segmentation via Prototype-driven Semantic Approximation

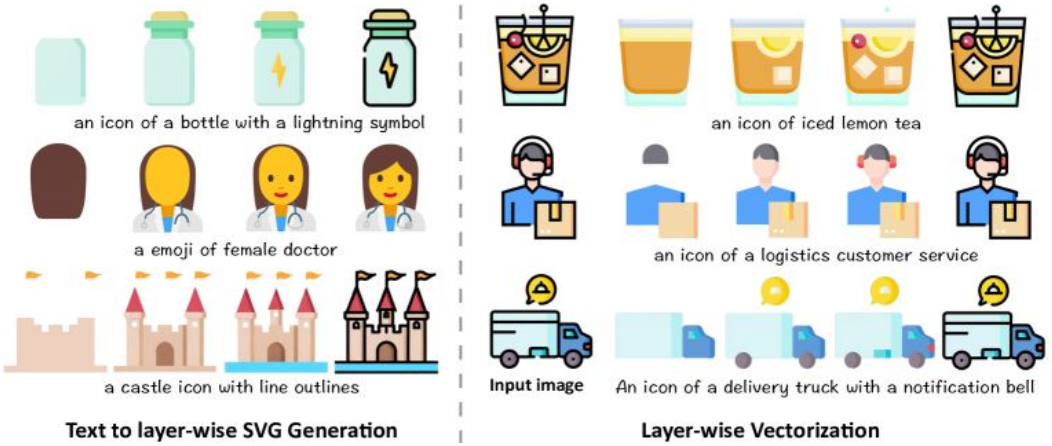
- Medical language-guided segmentation inherently relies on paired image-text input.
- ProLearn addresses this by distilling segmentation-relevant semantics from clinical reports into a discrete, compact prototype space.
- This approach approximates semantic guidance without text, enabling a compact model for text-free inference that alleviates textual reliance.



ICCV25: Meta Insights into Trends and Tendencies (120/153)

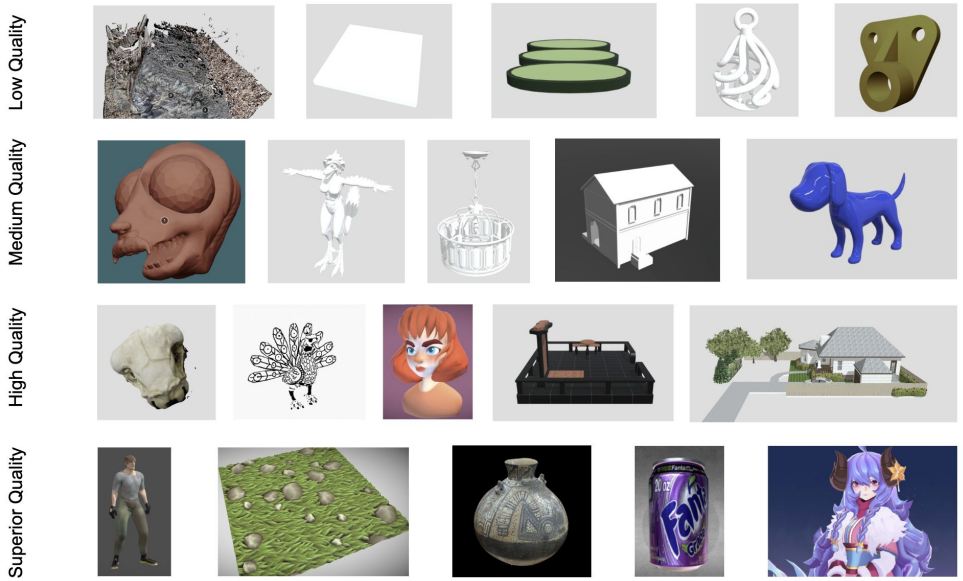
LayerTracer: Cognitive-Aligned Layered SVG Synthesis via Diffusion Transformer

- Challenge:
  - Existing SVG generation often produces single-layer or redundant vector shape
  - No model captures designers' cognitive workflow (layer logic, spatial grouping)
  - Lack of large-scale layered SVG datasets limits training diversity
- Key idea:
  - Learn human-like design sequences using a Diffusion Transformer (DiT)
    - Text-conditioned DiT generates sequential raster "construction blueprints"
    - Layer-wise vectorization removes redundancy → clean editable SVGs
  - For image input: conditional diffusion predicts likely layer-building steps
- Meta insights:
  - This could also be applied to other data with layered structures, such as 2D illustrations or CAD models, that humans construct procedurally.



Objaverse++: Curated 3D Object Dataset with Quality Annotations

- Objaverse++ is a paper of a curated dataset of 3D objects with quality annotations, addressing the prevalence of low-quality models in existing large-scale datasets like Objaverse.
- The authors manually annotate 10,000 3D objects with detailed attributes, including aesthetic quality scores, and then train a neural network to automatically annotate the remaining dataset.

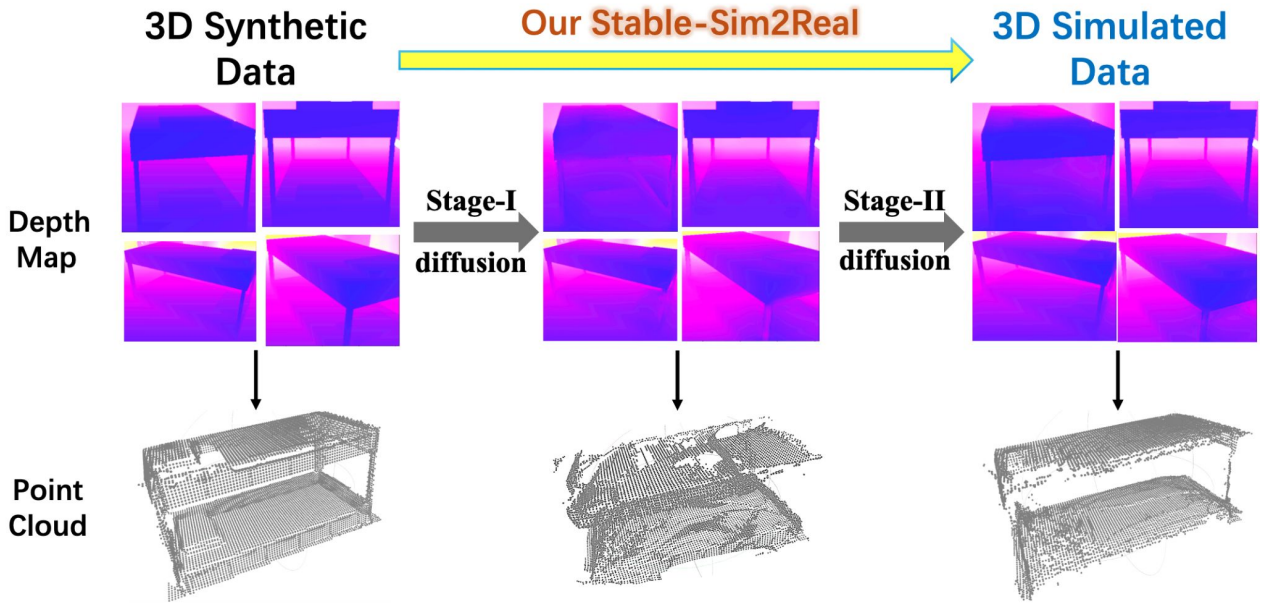


Real3D: Scaling Up Large Reconstruction Models with Real-World Images

- Real3D is the first Large Reconstruction Model (LRM) designed to overcome the limitations of training on synthetic or multi-view data by instead learning from abundant single-view, real-world images.
- The proposed method uses a novel self-training framework that combines supervised training on synthetic data with unsupervised losses on real images, employing pixel-level cycle consistency and semantic guidance to improve reconstruction quality.

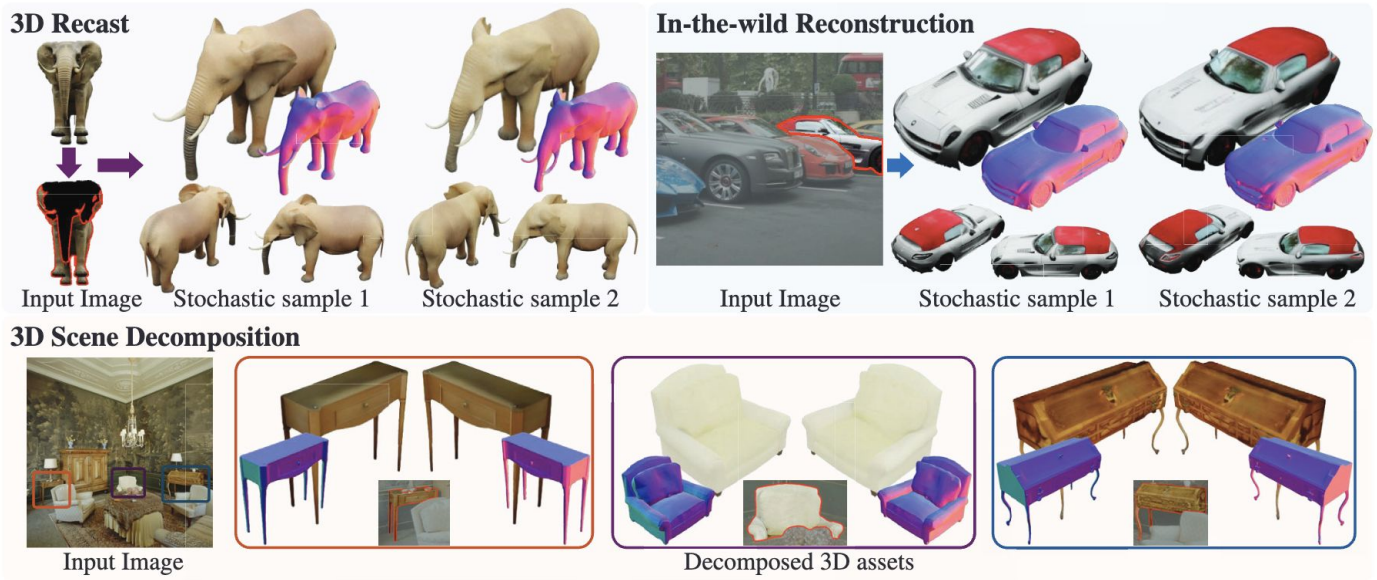
Stable-Sim2Real: Exploring Simulation of Real-Captured 3D Data with Two-Stage Depth Diffusion

- Stable-Sim2Real is a novel method for simulating realistic 3D data from synthetic inputs to bridge the sim-to-real gap.
- The approach uses a two-stage depth diffusion model that first generates a coarse depth map by learning the difference between real and synthetic data, and then refines specific regions using a second diffusion stage guided by a 3D discriminator.



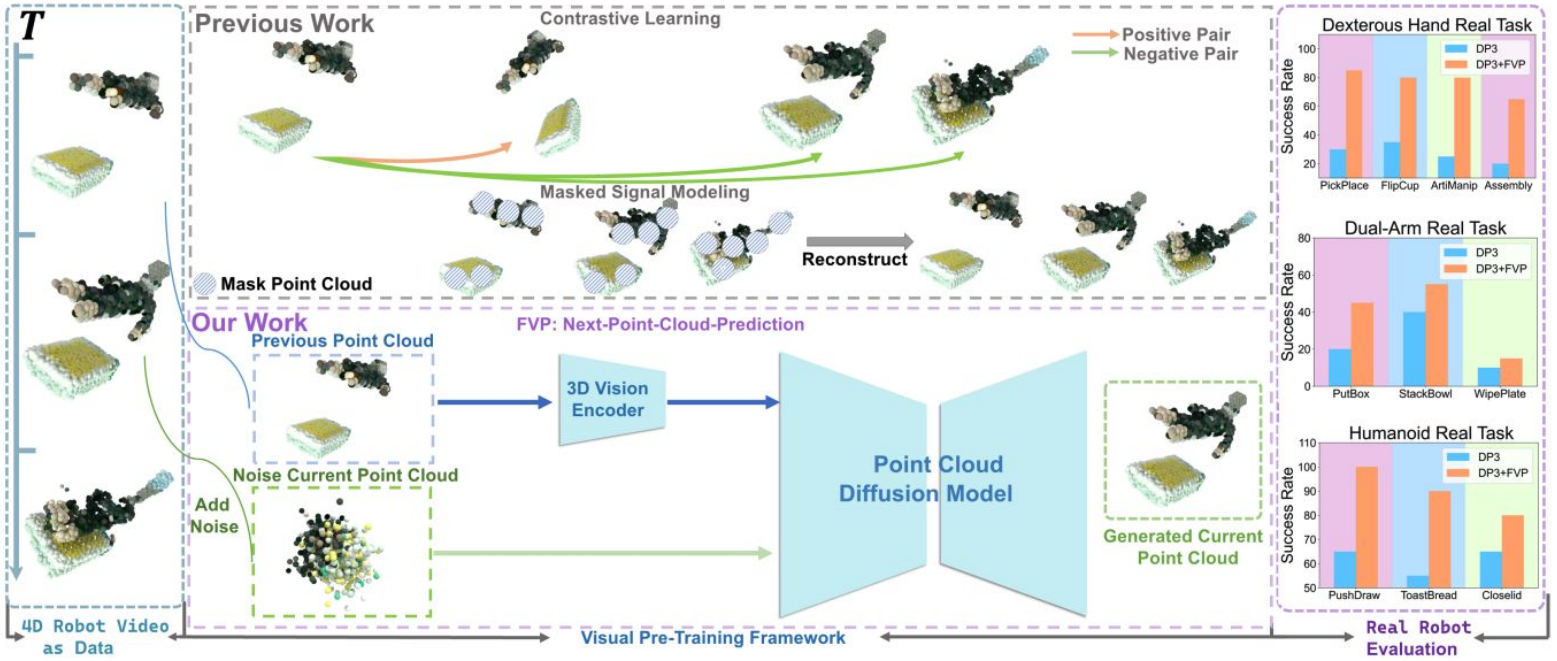
Amodal3R: Amodal 3D Reconstruction from Occluded 2D Images

- Amodal3R is a model that reconstructs complete 3D objects from single 2D images where the object is partially hidden.
- By using novel attention mechanisms that focus on visible parts while being aware of occlusions, the model reasons about the object's full geometry and appearance directly in 3D space.



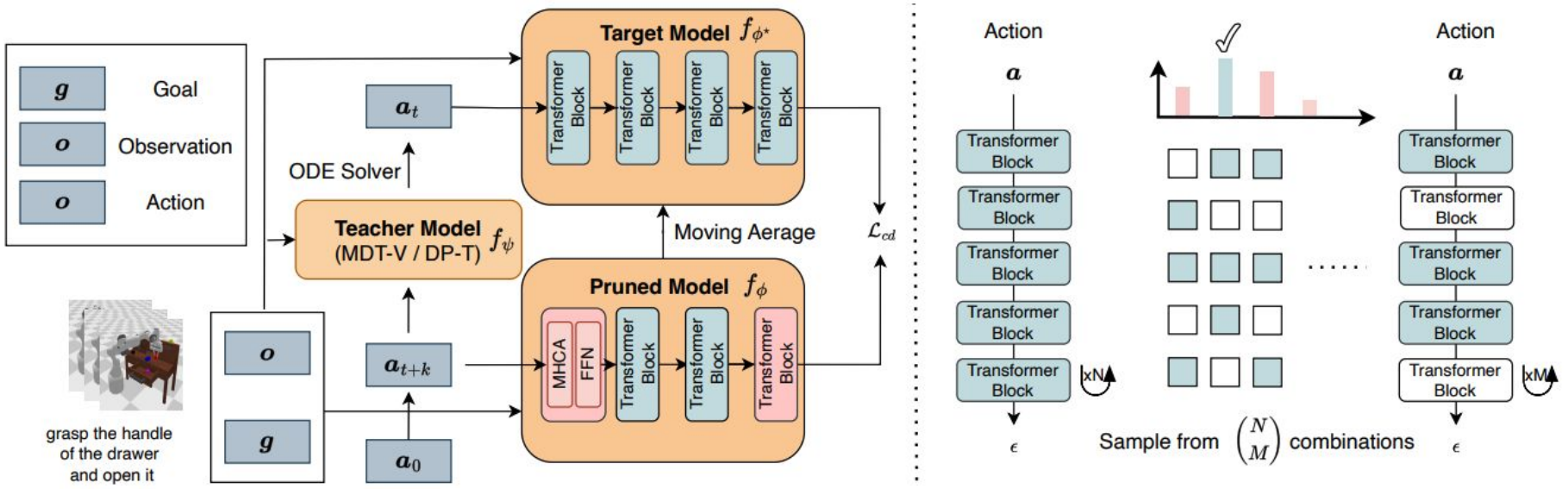
4D Visual Pre-training for Robot Learning

- A novel framework called FVP performs pre-training of 3D visual representations by predicting future point-cloud frames using a conditional diffusion model on robot manipulation data.
- It greatly improves performance on imitation and vision-language-action tasks in both simulation and real robots.



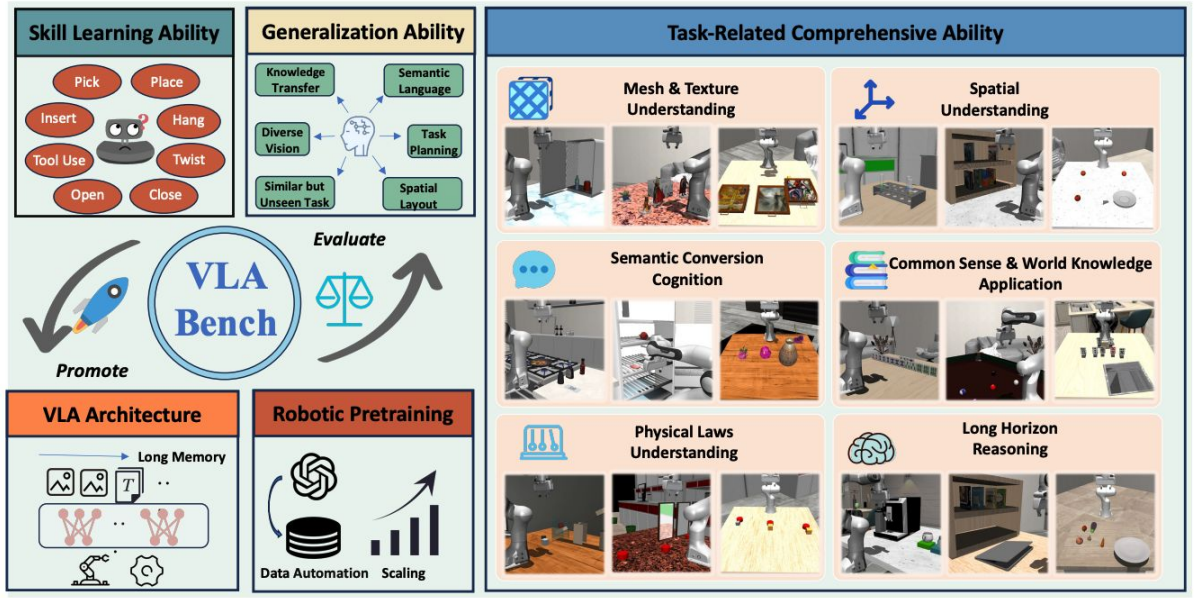
On-Device Diffusion Transformer Policy for Efficient Robot Manipulation

- A lightweight diffusion-transformer policy is proposed to enable efficient on-device robot manipulation under resource-limited hardware.
- It integrates step distillation and model pruning within a diffusion-based transformer architecture to reduce computation while preserving expressive action generation.



On-Device Diffusion Transformer Policy for Efficient Robot Manipulation

- VLABench evaluates vision-language-action models on 100 categories of language-conditioned, long-horizon manipulation tasks involving over 2000 objects.
- It emphasizes comprehension of human intentions, world knowledge transfer and multi-step reasoning, providing training data and highlighting significant gaps in current VLA model capabilities.



ICCV25: Meta Insights into Trends and Tendencies (128/153)

Meta insights in Image Restoration

❑ Module and Architecture Design

Attention Mechanism Enhancement

[Enhancing Image Restoration Transformer via Adaptive Translation Equivariance]

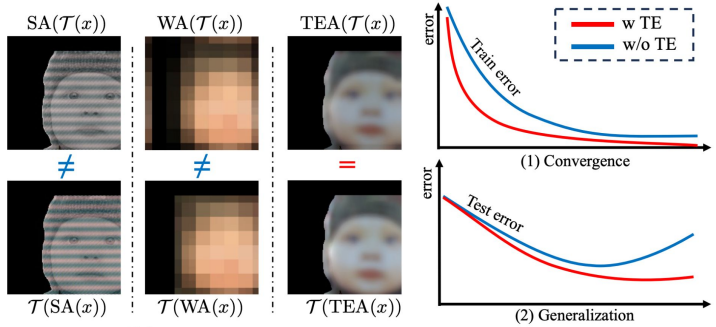


Figure 1. $\mathcal{T}(\cdot)$ means the translation function. "SA" and "WA" are commonly used for self-attention and window attention, respectively, but disrupt translation equivariance (TE) due to position encoding and feature shifting. "TEA" is our proposed translation equivariance adaptive attention, which satisfies TE. TE promotes faster convergence and better generalization.

Novel Module Design

[Reverse Convolution and Its Applications to Image Restoration]

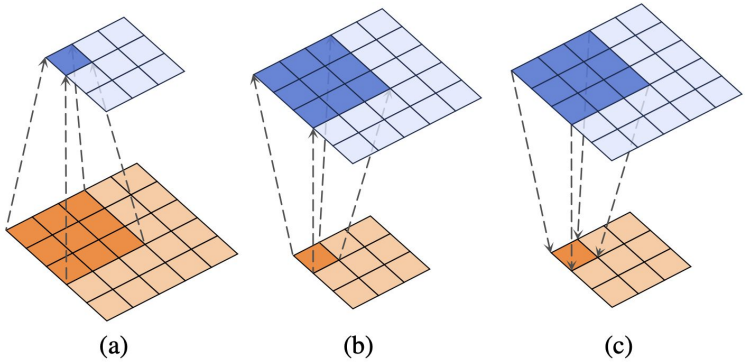


Figure 1. Illustration of the structural differences among (a) standard convolution, (b) transposed convolution, and (c) reverse convolution. Each subfigure shows a single input feature map (in orange) and its corresponding output feature map (in blue).

Mamba Architecture

[EAMamba: Efficient All-Around Vision State Space Model for Image Restoration]

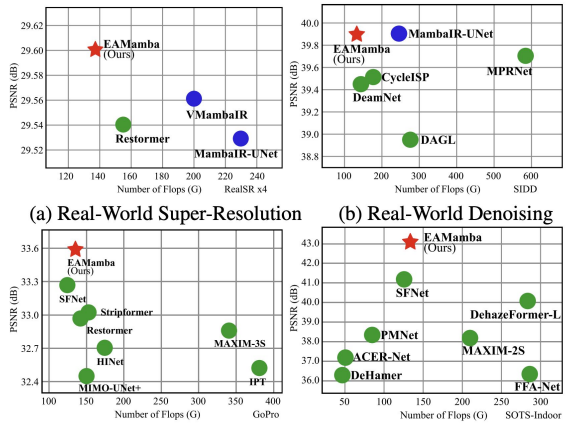


Figure 1. Computational efficiency versus image quality across model architectures. Our method (denoted by ★) demonstrates superior efficiency compared to other Vision Mamba-based methods (●) and existing approaches (●). EAMamba establishes a new efficiency frontier for Vision Mamba-based image restoration.

ICCV25: Meta Insights into Trends and Tendencies (129/153)

Meta insights in Image Restoration

- Innovative IR Approaches

New Task Proposal

[MoFRR: Mixture of Diffusion Models for Face Retouching Restoration]

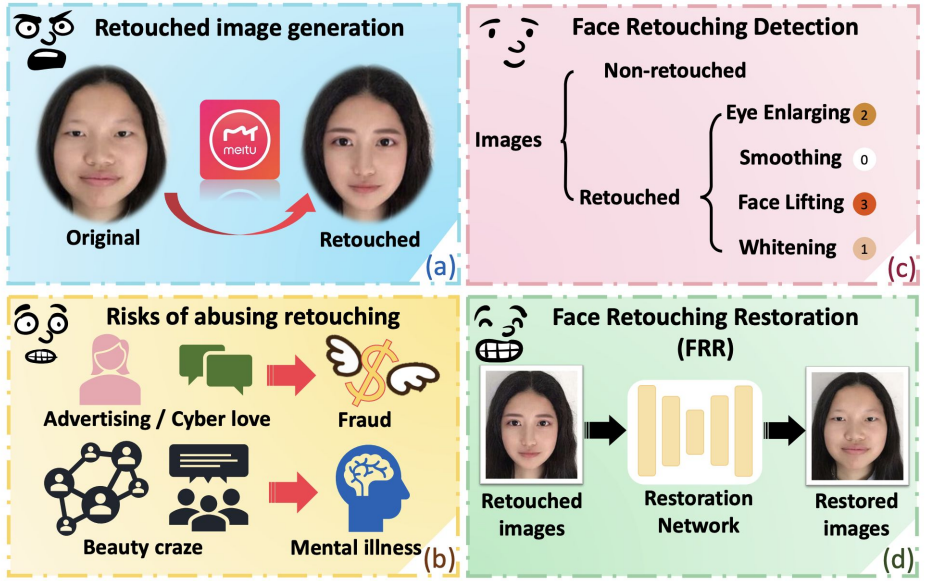
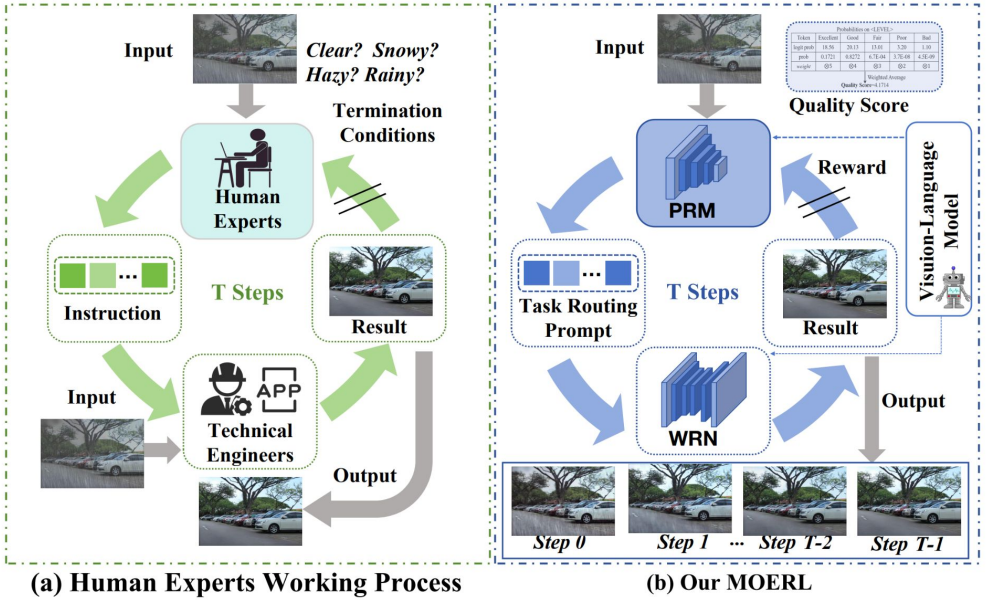


Figure 1. Application scenario of the proposed scheme (MoFRR). (a) People use face retouching in various applications, (b) face retouching poses risks such as fraud, societal security, and cultural psychological issues, (c) existing work for face retouching detection, (d) our proposed MoFRR method recovers the original image from the retouched version, thereby offering an additional layer of protection against face retouching fraud.

Introduction of Reinforcement Learning

[MOERL: When Mixture-of-Experts Meet Reinforcement Learning for Adverse Weather Image Restoration]



(a) Human Experts Working Process (b) Our MOERL

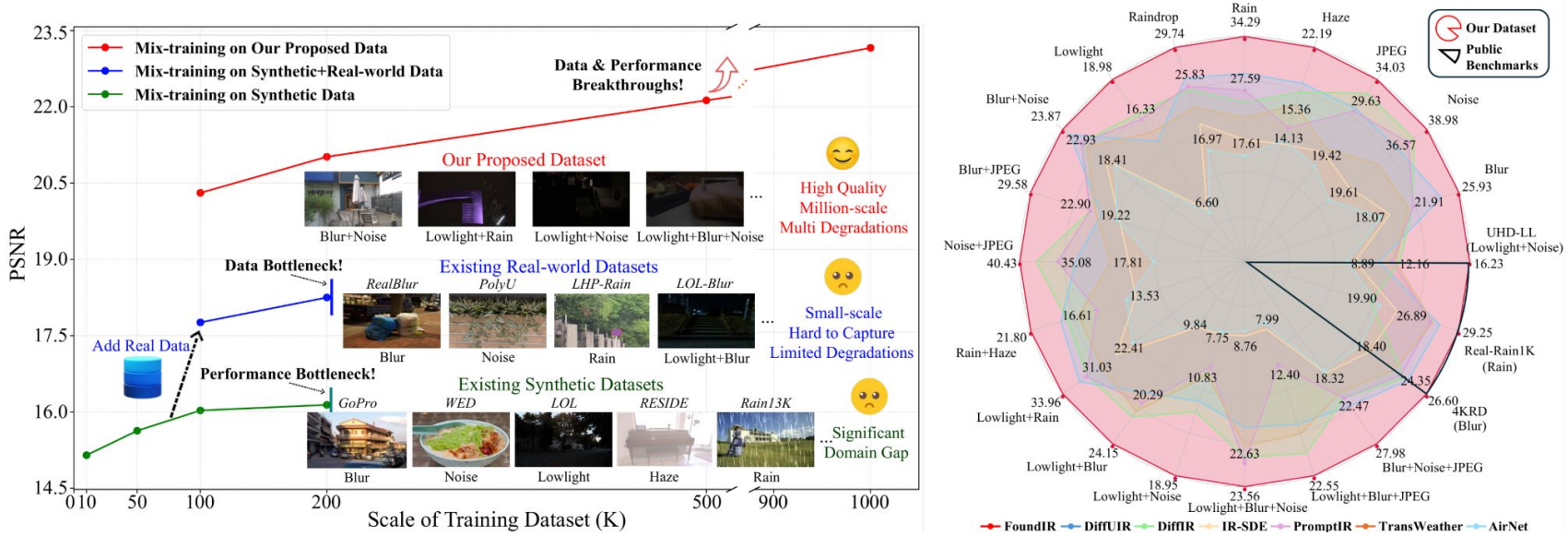
Figure 1. The Motivation of our MOERL. (a) Human Expert Process: The experts iteratively refine images based on perceptual feedback. (b) Our MOERL: Inspired by this, MOERL uses deep reinforcement learning to progressively optimize restoration with guidance from the pre-trained vision-language model, enabling adaptive enhancement for diverse weather degradations.

ICCV25: Meta Insights into Trends and Tendencies (130/153)

Meta insights in Image Restoration

- Toward Large-Scale Data
  - Construct a million-scale high-quality paired dataset for image restoration
  - Propose FoundIR, a robust and versatile restoration model
  - Demonstrate the effectiveness of the dataset and achieve SOTA performance

FoundIR: Unleashing Million-scale Training Data to Advance Foundation Models for Image Restoration



(a) Scaling up real-world training **data** (b) Scaling up **model** generalization

Figure 1. The potential of large-scale training data for universal image restoration. (a) Analysis of universal image restoration performance in real-world scenarios as training data vary. As the size of real-world training data increases, the image restoration model (the generalist model of FoundIR) can achieve significant performance improvement. (b) Our proposed FoundIR, trained on our million-scale dataset, achieves state-of-the-art performance across a broad range of restoration tasks compared to existing universal image restoration methods.

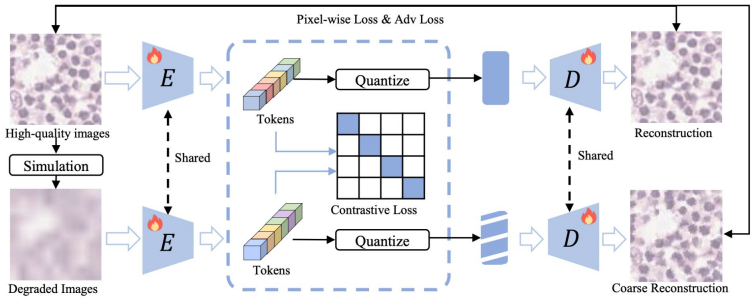
ICCV25: Meta Insights into Trends and Tendencies (131/153)

Meta insights in Image Restoration

- ❏ Expanding Super-Resolution Across Other Domains

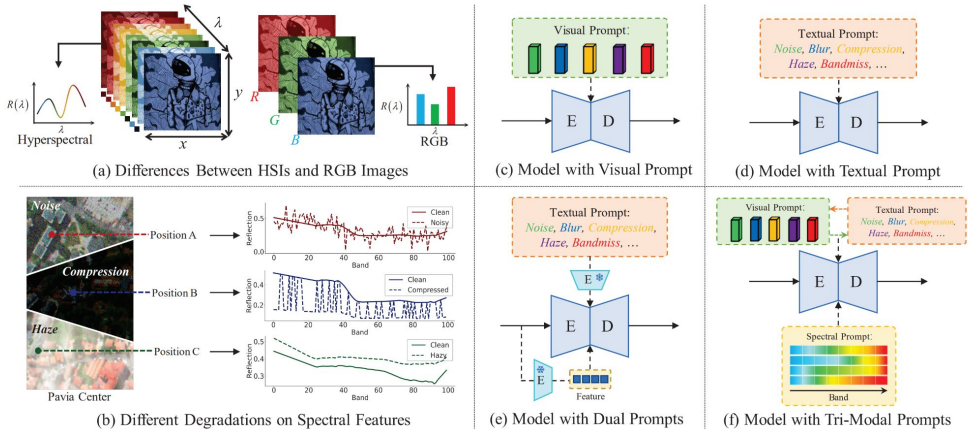
Medical

[Conditional Visual Autoregressive Modeling for Pathological Image Restoration]



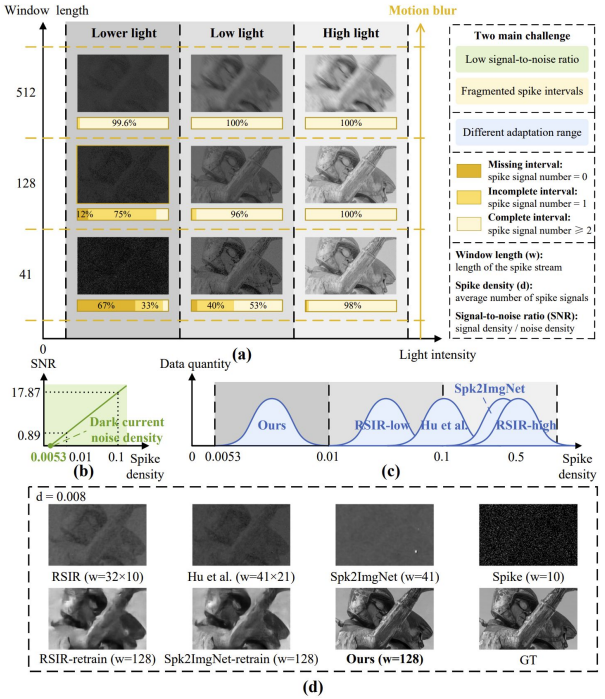
HSI

[MP-HSIR: A Multi-Prompt Framework for Universal Hyperspectral Image Restoration]



Spike

[Noise-Modeled Diffusion Models for Low-Light Spike Image Restoration]



ICCV25: Meta Insights into Trends and Tendencies (132/153)

Meta insights in Super-Resolution

- Generative Models Lead the Pursuit of Photo-Realism
  - Real-World Super-Resolution (SR) is the primary focus of current research, while Classic SR (defined by Bicubic downsampling only) continues to be studied in the context of improving computational efficiency.
  - DMs, including Latent Diffusion Models (LDM) and Diffusion Transformers (DiT/MM-DiT), are widely adopted to generate realistic high-frequency textures for Real-World SR

- The Race for Computational Efficiency and Alternative Architectures
  - Diffusion Acceleration:
    - Consistency Trajectory Matching for One-Step Generative Super-Resolution
  - Mamba/State-Space Models (SSMs) Introduction:
    - VSRM: A Robust Mamba-Based Framework for Video Super-Resolution
    - MedVSR: Medical Video Super-Resolution with Cross State-Space Propagation

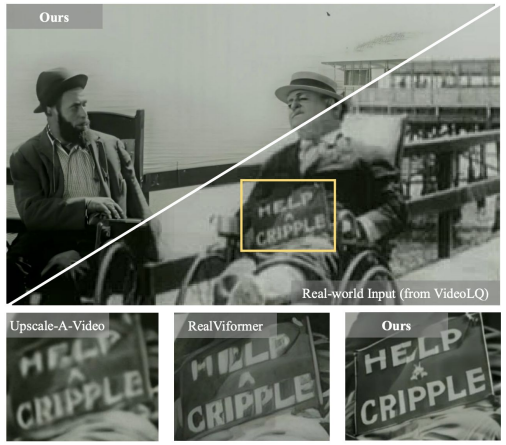
ICCV25: Meta Insights into Trends and Tendencies (133/153)

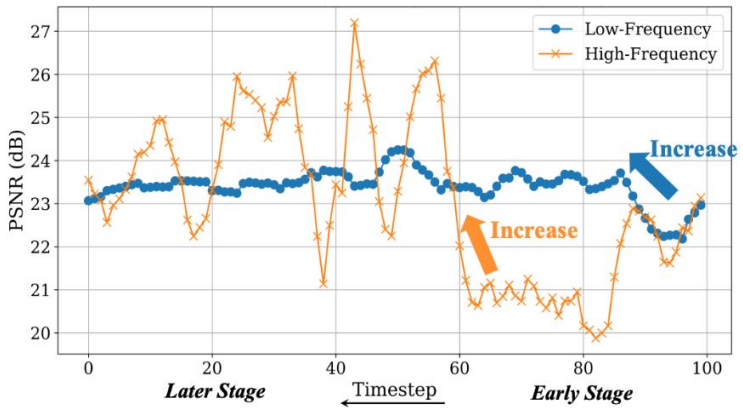
Meta insights in Super-Resolution

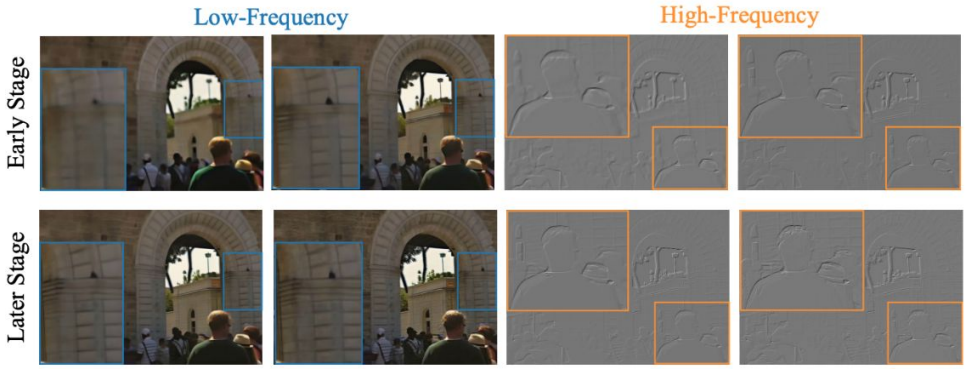
- Expanding Super-Resolution Across Other Domains
  - Polarization Imaging
    - Benchmarking Burst Super-Resolution for Polarization Images: Noise Dataset and Analysis
    - introduce the first dedicated dataset and model for BurstSR in polarization imaging, addressing low light efficiency and resolution limitations through multi-frame reconstruction
  - Hyperspectral Imaging
    - Hipandas: Hyperspectral Image Joint Denoising and Super-Resolution by Image Fusion with the Panchromatic Image
    - integrate denoising and super-resolution via fusion with panchromatic images, enhancing both spectral fidelity and spatial sharpness for hyperspectral image restoration
  - Remote Sensing
    - NeurOp-Diff: Continuous Remote Sensing Image Super-Resolution via Neural Operator Diffusion
    - employ neural operator diffusion for continuous-scale super-resolution in satellite imagery, achieving high-fidelity restoration across varying spatial resolutions and complex terrain conditions
  - Medical Imaging
    - MedVSR: Medical Video Super-Resolution with Cross State-Space Propagation
    - introduce a specialized state-space-based framework for medical video super-resolution, overcoming motion blur, noise, and misalignment to recover diagnostically reliable tissue details

STAR: Spatial-Temporal Augmentation with Text-to-Video Models for Real-World Video Super-Resolution

- LIEM (Local Information Enhancement Module)
  - Focuses on local degradation removal before global aggregation, suppressing artifacts and improving details.
- Dynamic Frequency Loss
  - The diffusion process restores low-frequency (structure) in early stages and high-frequency (details) in late stages.
  - Uses Fourier transform to dynamically adjust the loss weights between low-frequency and high-frequency components based on the denoising step t.

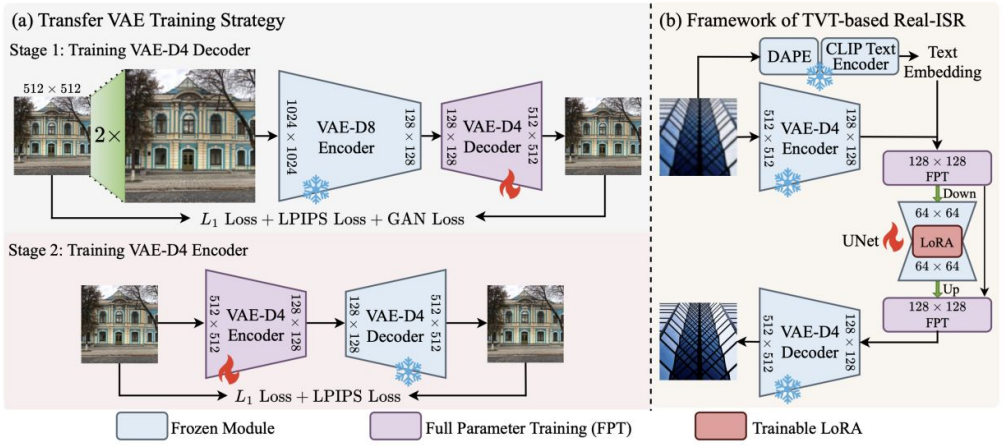


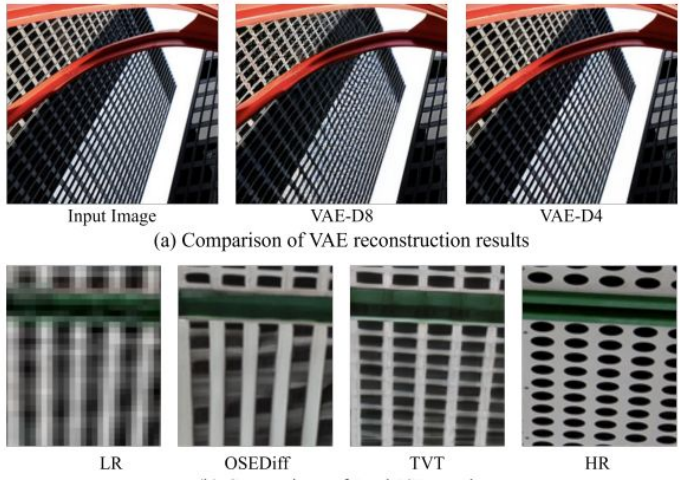




Fine-structure Preserved Real-world Image Super-resolution via Transfer VAE Training

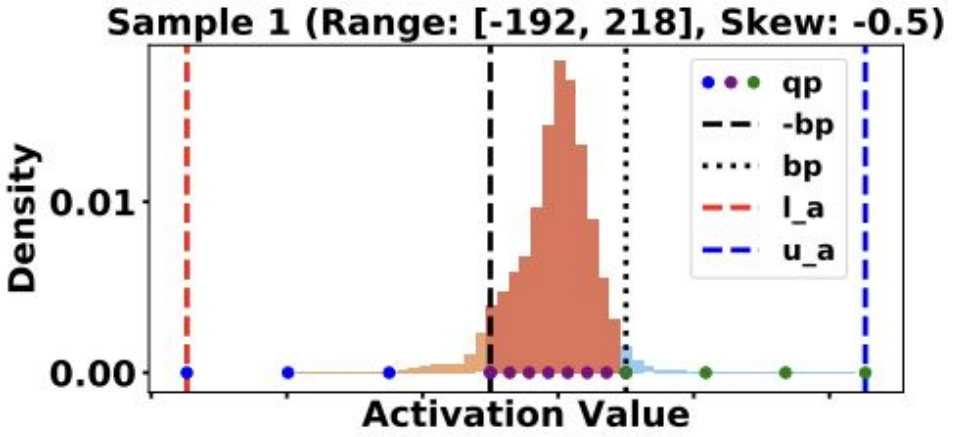
- Standard VAE in diffusion models irreversibly loses fine structures (like text) due to excessive 8x compression.
- Adopts a 4x compression VAE to retain fine structures in the latent space.
- A 2-stage transfer learning process to align the VAE-D4's latent space with the pre-trained, D8-based UNet.

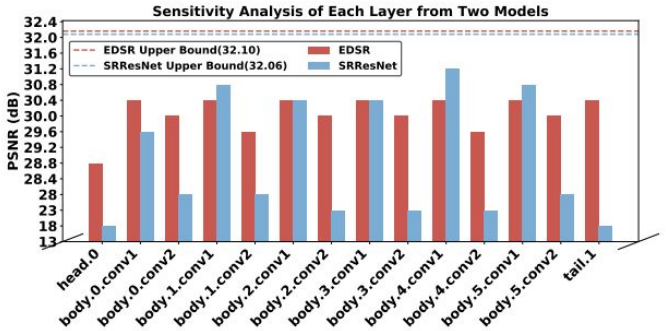




Outlier-Aware Post-Training Quantization for Image Super-Resolution

- Activation outliers strongly correlate with color information, and simple quantization causes color distortion and accuracy loss.
- PLQ (Piecewise Linear Quantizer)
  - Splits the activation distribution into a dense region and an outlier region.
  - Applies independent quantization to each, preserving both color info (outliers) and accuracy (dense region).
- SAFT (Sensitivity-Aware Fine-tuning)
  - Identifies per-layer sensitivity (activation variance) and focuses optimization on reconstructing features in sensitive (high-variance) layers.

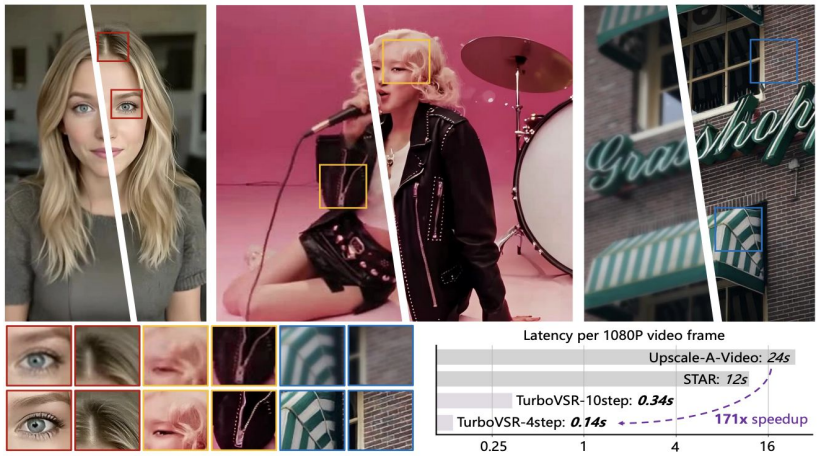


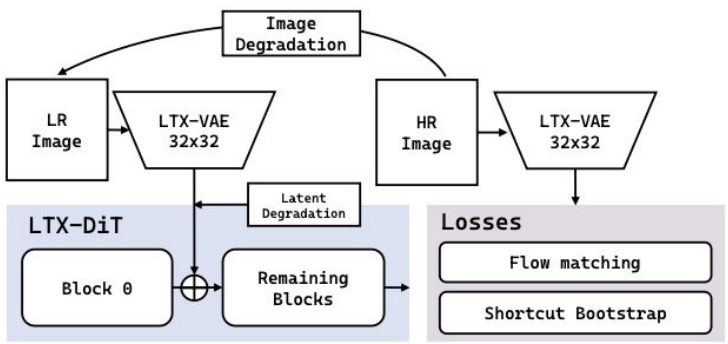


ICCV25: Meta Insights into Trends and Tendencies (137/153)

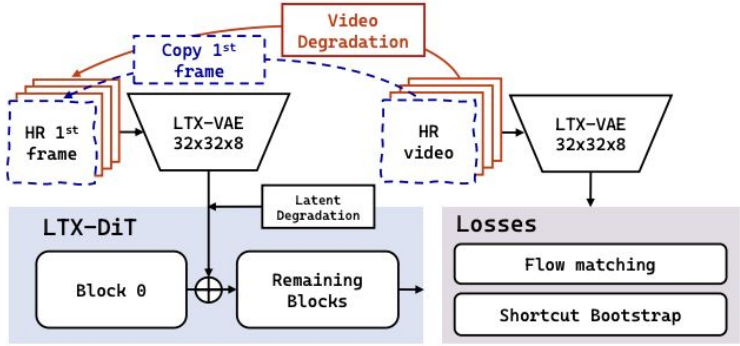
TurboVSR: Fantastic Video Upscalers and Where to Find Them

- Maintain the realism of diffusion VSR while dramatically improving computational efficiency (over 100x faster than existing methods).
- Reduces token length by 1/32 using 32x spatial and 8x temporal compression.
- Accelerates convergence by decomposing the difficult VSR task into Initial Frame SR and Video SR.
- Reduces sampling steps from 10 to 4 by biasing samples toward high-noise initial steps to suppress errors.





(a) Image/Initial frame super-resolution



(b) Video super-resolution with factorized conditioning

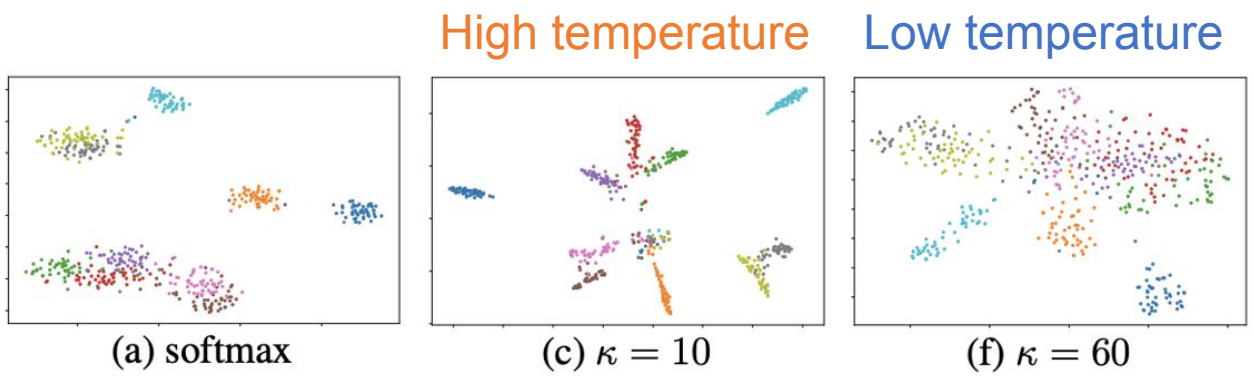
Temperature in Cosine-based Softmax Loss

- The method is proposed for dynamically adjusting the temperature parameter of the cosine-based softmax loss during training.
- The study revealed that the temperature parameter affects model characteristics.
  - Low temperature : produces diverse feature representations with strong generalization ability. → **High performance on downstream tasks via transfer learning.**
  - High temperature : produces highly class-discriminative feature representations
    → **Useful for detecting missing data or out-of-distribution (OOD) data.**

👍 **Model characteristics can be effectively controlled through the temperature parameter.**

- A model pretrained with a standard softmax loss can be re-trained using a cosine-based softmax loss, enabling temperature-based control of model characteristics.

Feature distribution of the first 10 classes in ImageNet.(κ= 1/τ)



Takumi Kobayashi, “Temperature in Cosine-based Softmax Loss”, in ICCV 2025.

ICCV25: Meta Insights into Trends and Tendencies (139/153)

LOTS of Fashion! Multi-Conditioning for Image Generation via Sketch-Text Pairing

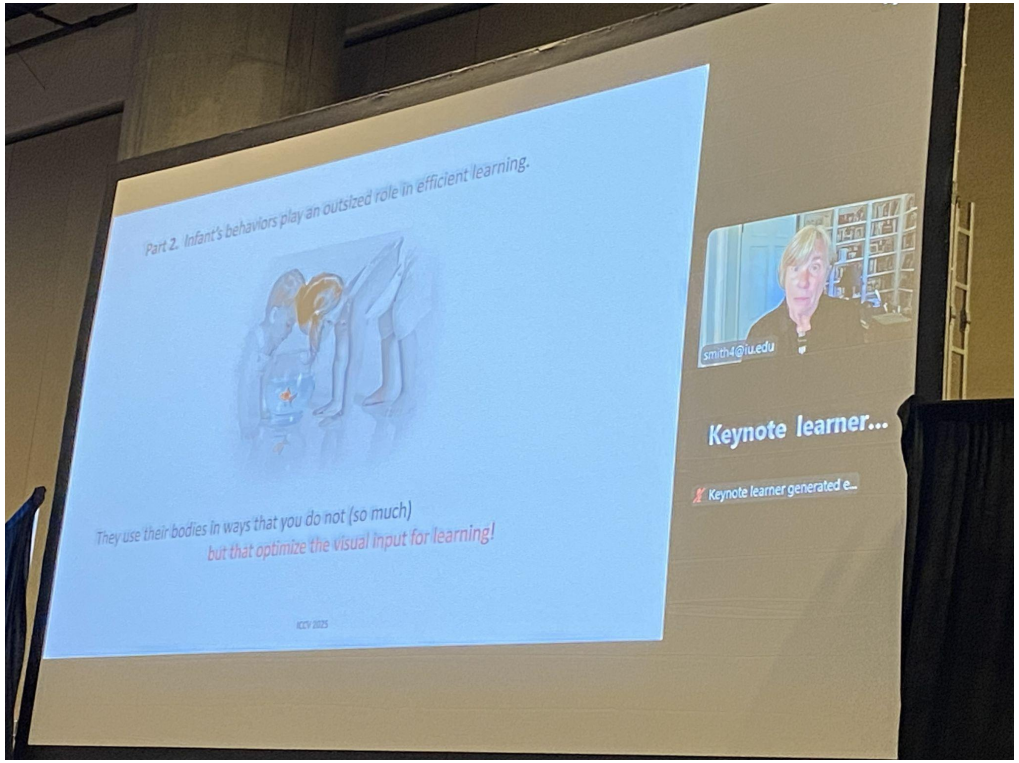
- Challenge:
  - attribute confusion (e.g., the pattern for a shirt is incorrectly applied to the pants)
  - Lack of Local Control: Difficulty in specifying fine-grained details for individual items using only global text.
- Key idea:
  - A new model that uses separate "sketch + text" pairs for each fashion item.
  - Sketchy: A new dataset created for this task, providing multiple localized pairs per image.
  - Pair Guidance: A method to integrate these multiple inputs correctly during the diffusion process.

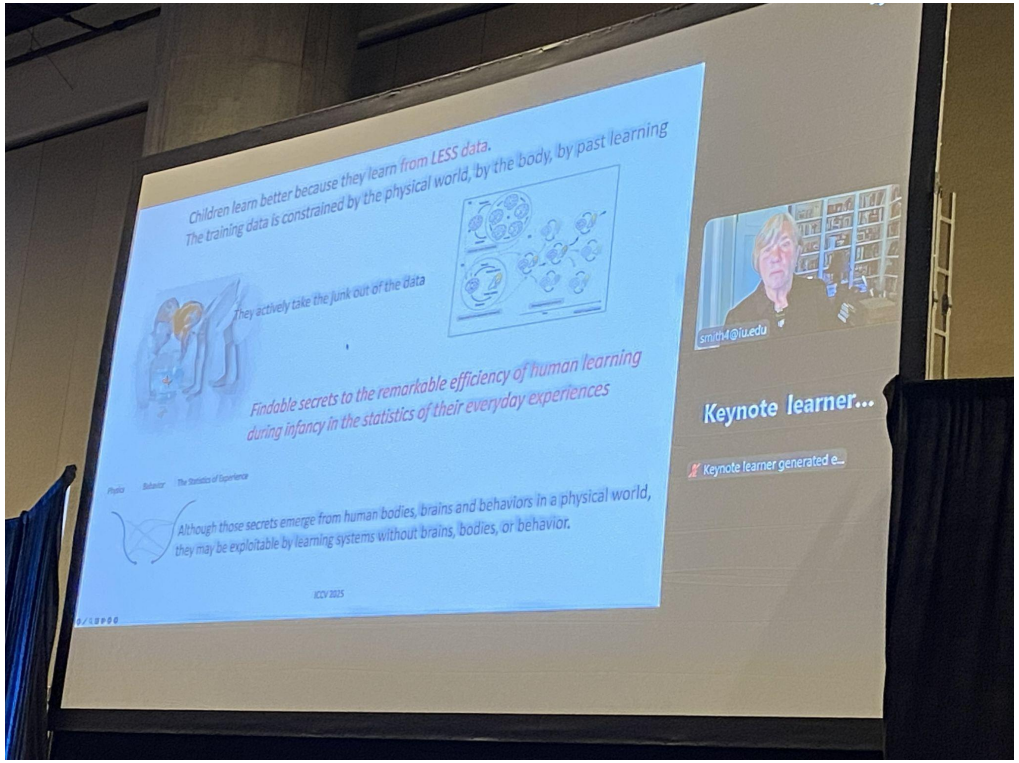
- Meta insights:
  - This approach could also be applied to domains beyond clothing (furniture and interior decoration, characters and avatars, industrial products)

ICCV25: Meta Insights into Trends and Tendencies (140/153)

Efficient learning (1/5)

- Learning from human infants' behaviors
  - keynote





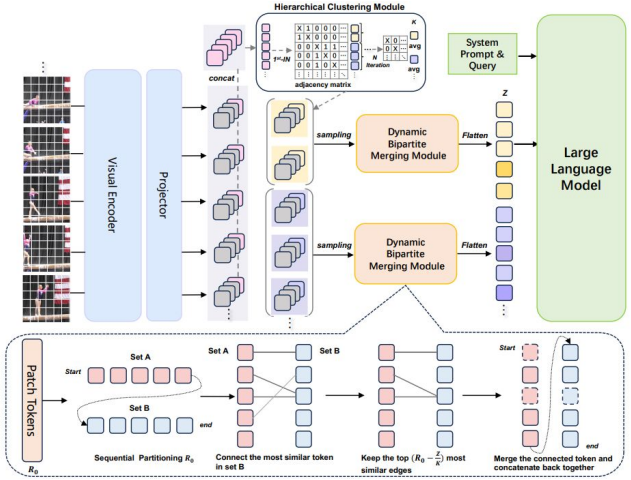
Linda B Smith, Distinguished Professor, Indiana University, Secrets in the training data:
The visual statistics of infants and children's everyday interactions with objects

ICCV25: Meta Insights into Trends and Tendencies (141/153)

Efficient learning (2/5)

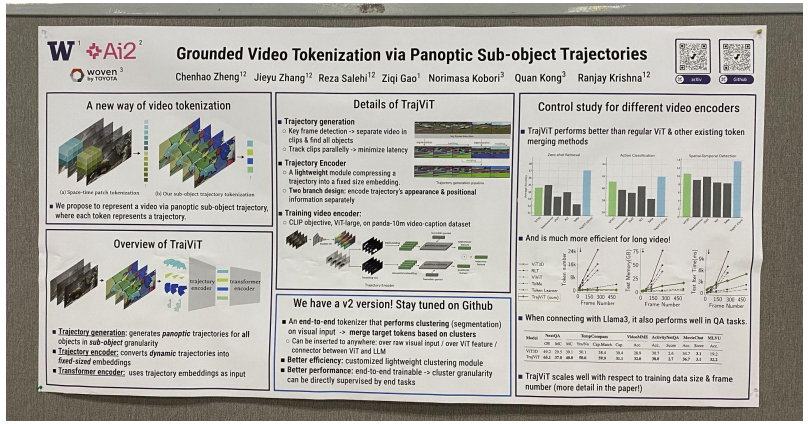
- video tokenization
  - token merge

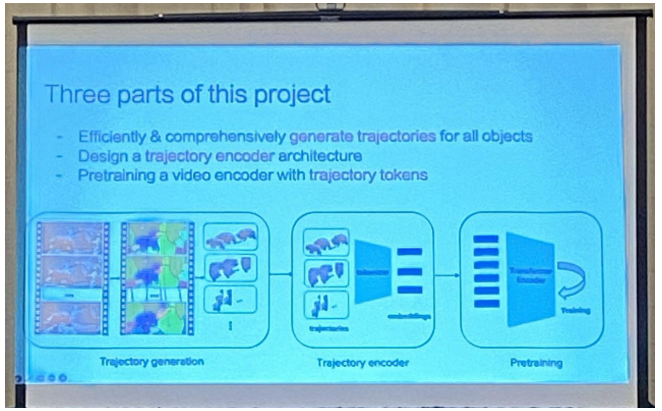
Beyond Training: Dynamic Token Merging for Zero-Shot Video Understanding, Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zenghui Ding, Xianjun Yang, Yining Sun



  - using trajectory

One Trajectory, One Token: Grounded Video Tokenization via Panoptic Sub-object Trajectory, Chenhao Zheng, Jieyu Zhang, Mohammadreza Salehi, Ziqi Gao, Vishnu Iyengar, Norimasa Kobori, Quan Kong, Ranjay Krishna

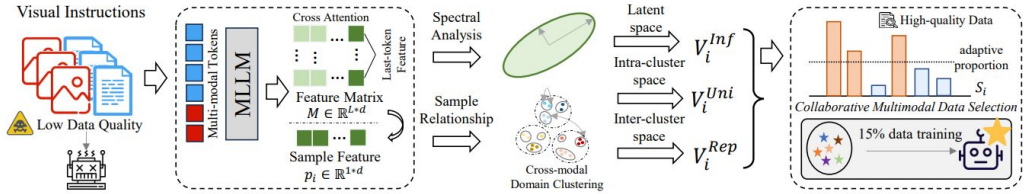




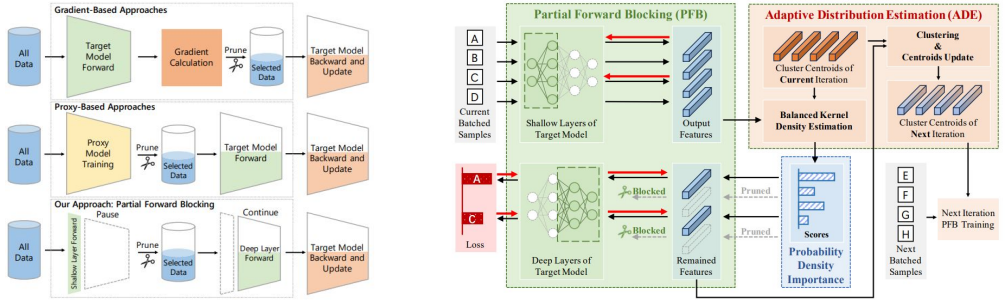
ICCV25: Meta Insights into Trends and Tendencies (142/153)

Efficient learning (3/5)

- Data cleaning
  - Select important data

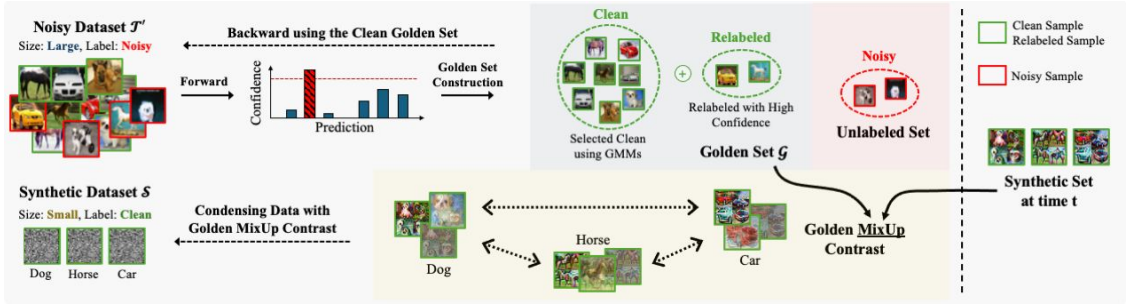


Mastering Collaborative Multi-modal Data Selection: A Focus on ❇ Informativeness, Uniqueness, and Representativeness, Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Bosheng Qin, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, Yueting Zhuang



Partial Forward Blocking: A Novel Data Pruning Paradigm for Lossless Training Acceleration, Dongyue Wu, Zilin Guo, Jialong Zuo, Nong Sang, Changxin Gao

- Robust dataset compression

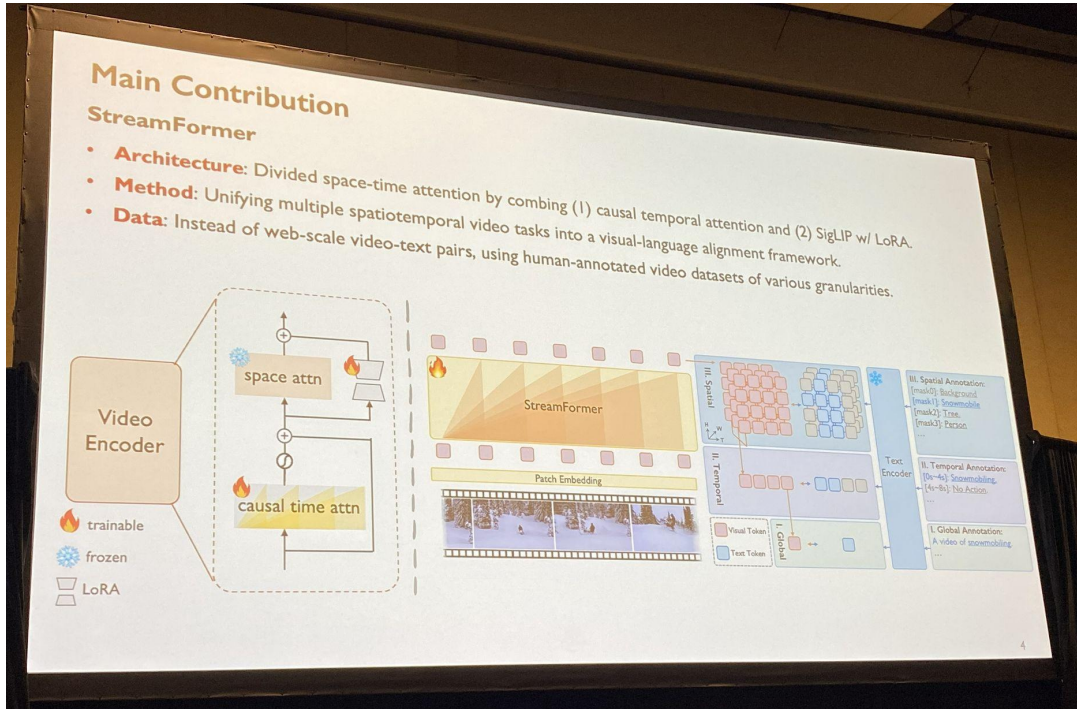


Robust Dataset Condensation using Supervised Contrastive Learning, Nicole Hee-Yeon Kim, Hwanjun Song

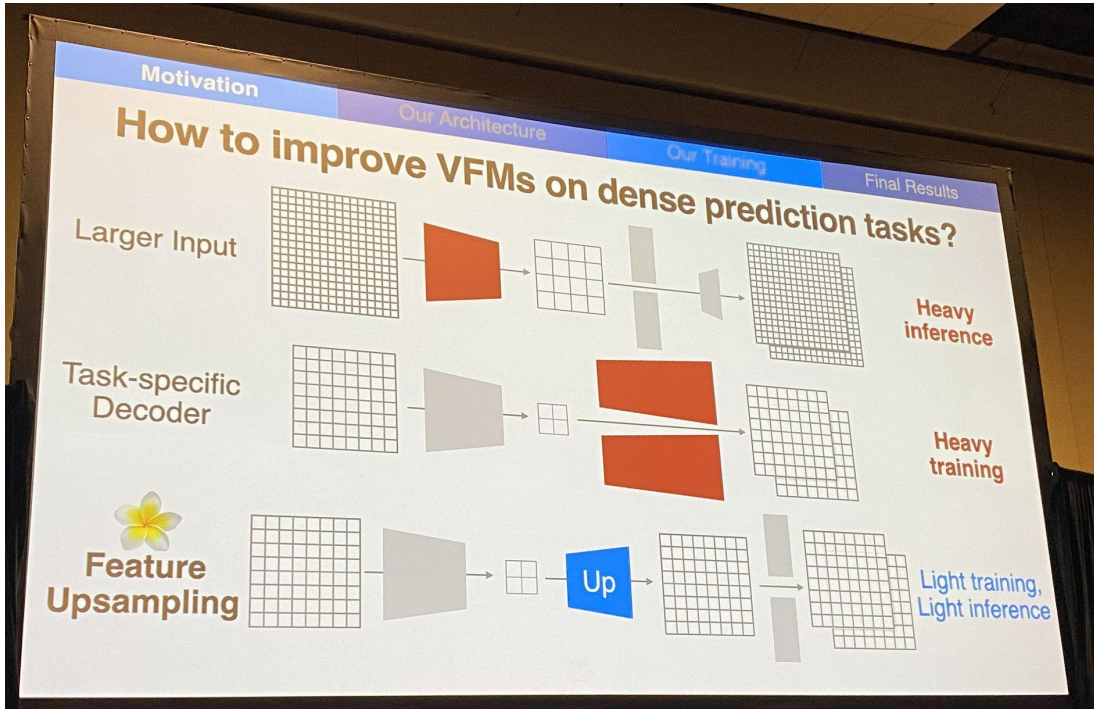
ICCV25: Meta Insights into Trends and Tendencies (143/153)

Efficient learning (4/5)

- Learning pipeline/algorithm
  - Adding light weight module



Learning Streaming Video Representation via Multitask Training, Yibin Yan, Jilan Xu, Shangzhe Di, Yikun Liu, Yudi Shi, Qirui Chen, Zeqian Li, Yifei Huang, Weidi Xie



LoftUp: Learning a Coordinate-Based Feature Upsampler for Vision Foundation Models, Haiwen Huang, Anpei Chen, Volodymyr Havrylov, Andreas Geiger, Dan Zhang

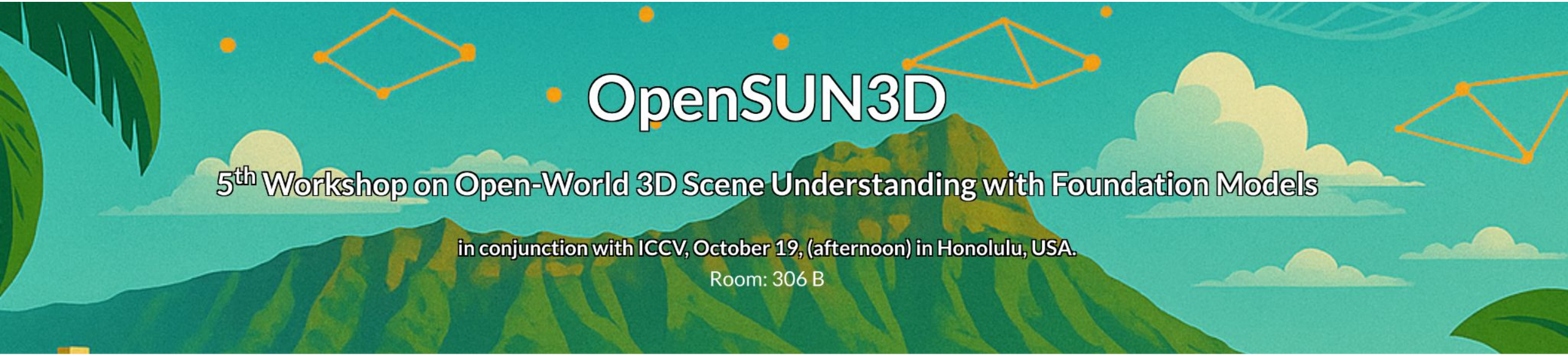
Efficient learning (5/5)

- Summary
  - Video → Heavy data
    - Extracting useful information
  - Multimodal Data → Including noisy data
    - Cleaning data
  - Algorithm → High performance, but too costly
    - Adding light weight module
    - Referring human infants' behaviors

ICCV25: Meta Insights into Trends and Tendencies (145/153)

Open-World 3D Scene Understanding with Foundation Models (Workshop)

- The goal of this workshop is that how do we lift to open-vocabulary recognition from closed data & label in 3d scene understanding
  - **New!! Benchmark Challenge Track**
    - **SceneFun3D benchmark and Articulate3D dataset** (Details on next page)



https://opensun3d.github.io/

ICCV25: Meta Insights into Trends and Tendencies (146/153)

Open-World 3D Scene Understanding with Foundation Models (Workshop)

- Benchmark challenge track
  - The challenge focuses on fine-grained functionality, affordance, and interaction understanding in 3D indoor environments
    - Track 1: Functionality Segmentation [SceneFun3D] Benchmark
    - Track 2: Open-Vocabulary 3D Affordance Grounding [SceneFun3D] Benchmark
    - Track 3: Interaction Understanding [Articulate3D]

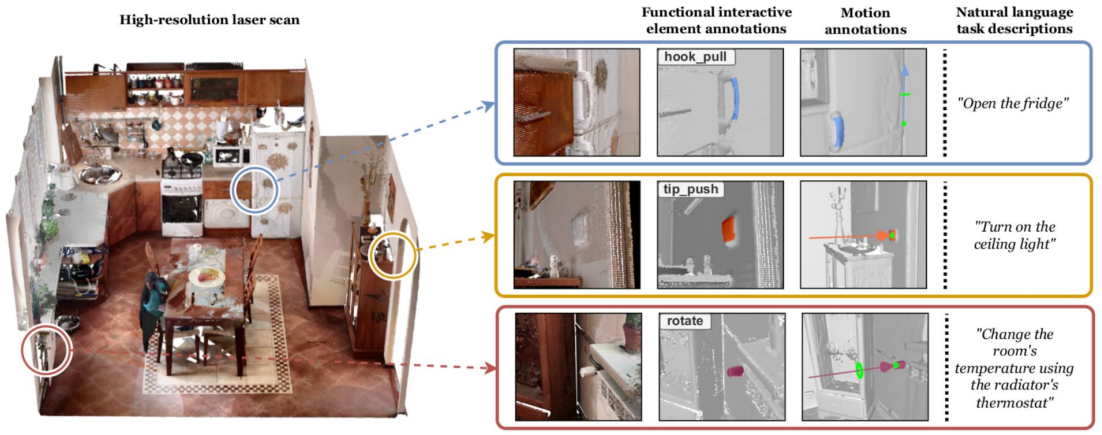
🏆 Prize per challenge winner team: €1,000

https://opensun3d.github.io/

ICCV25: Meta Insights into Trends and Tendencies (147/153)

Open-World 3D Scene Understanding with Foundation Models (Workshop)

- **Track 1: Functionality Segmentation [SceneFun3D] Benchmark**
  - What's task?: Segment the functional interactive elements in a 3D scene
    - Given a 3D point cloud of a scene, the goal is to segment the functional interactive element instances (e.g., handles, knobs, buttons) and predict the associated affordance labels.



https://scenefun3d.github.io/

Functionality Segmentation (test split)

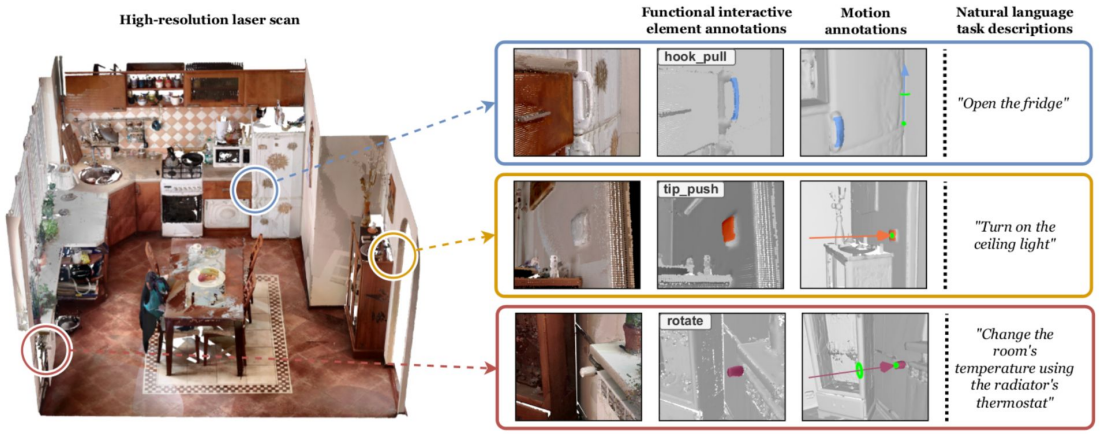
Team/Method	AP ↑	AP_50 ↑	AP_25 ↑
pico-mr	**6.54**	**13.97**	**27.82**
RPL	2.91	6.46	16.91
MaNET	0.76	3.76	13.92
SegFunCT	0.00	0.00	10.69

https://scenefun3d.github.io/benchmark

ICCV25: Meta Insights into Trends and Tendencies (148/153)

Open-World 3D Scene Understanding with Foundation Models (Workshop)

- **Track 2: Open-Vocabulary 3D Affordance Grounding [SceneFun3D] Benchmark**
  - What's task?: Segment the functional interactive elements in a 3D scene
    - Given a text-based description of a task, the aim is to localize and segment the functional interactive elements that an agent needs to interact with to successfully accomplish the task



https://scenefun3d.github.io/

Task-Driven Affordance Grounding (test split)

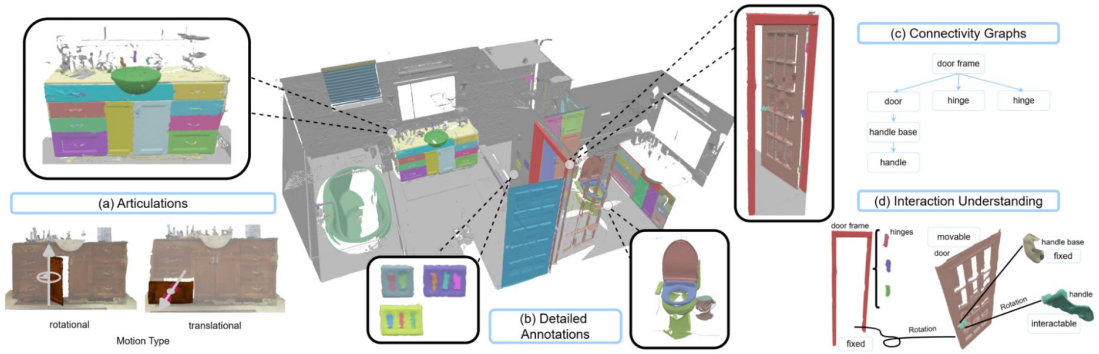
Team/Method	AP_50 ↑	AP_25 ↑
pico-mr	**11.31**	**23.30**
affordance_test_on_jane	6.03	16.41
Fun3DU	3.56	8.80
J. Corsetti, F. Giuliari, A. Fasoli, D. Boscaini, F. Poiesi. "Fun3DU: Functionality Understanding and Segmentation in 3D Scenes". CVPR 2025 Highlight		
MaNET	0.97	6.43
SegFunCT	0.46	5.99

https://scenefun3d.github.io/benchmark

ICCV25: Meta Insights into Trends and Tendencies (149/153)

Open-World 3D Scene Understanding with Foundation Models (Workshop)

- **Track 3: Interaction Understanding [Articulate3D]**
  - What's task?: Segmentation of all movable (articulated) parts.
    - Given a 3D indoor scene, the objective is to identify all movable parts and predict their interaction specifications. These include the part's motion characteristics—such as axis, origin, and motion type (rotation or translation)—as well as the specific graspable region that enables interaction (e.g., a door knob or window handle).



https://insait-institute.github.io/articulate3d.github.io/

Leaderboard

Public Leaderboard

Phase: Movable Prediction Test, Split: Test Split

Order by metric

B - Baseline * - Private V - Verified

Visible Metrics

Rank	Participant team	AP50 (↓)	AP50_axis (↓)	AP50_origin (↓)	AP50_axis_origin (↓)	Last submission at
1	Aaaa	0.00	0.00	0.00	0.00	14 days ago
2	XT (ORI)	0.31	0.26	0.20	0.16	1 month ago
3	ytttt	0.32	0.29	0.20	0.18	23 days ago
4	3DV	0.36	0.32	0.28	0.20	1 month ago
5	IDN (GUS)	0.41	0.35	0.28	0.21	1 month ago
6	KKing	0.44	0.38	0.32	0.26	11 days ago
7	submission_test	0.44	0.38	0.32	0.26	1 month ago
8	Host_5966_Team (USDNet-Baseline) V B	0.44	0.38	0.32	0.26	2 months ago

https://art3dchallenge.jumpingcrab.com/web/challenges/challenge-page/22/leaderboard/86

ICCV25: Meta Insights into Trends and Tendencies (150/153)

SUPERDEC: 3D Scene Decomposition with Superquadric Primitives (1/3)

Ultimate goal of the paper is to:

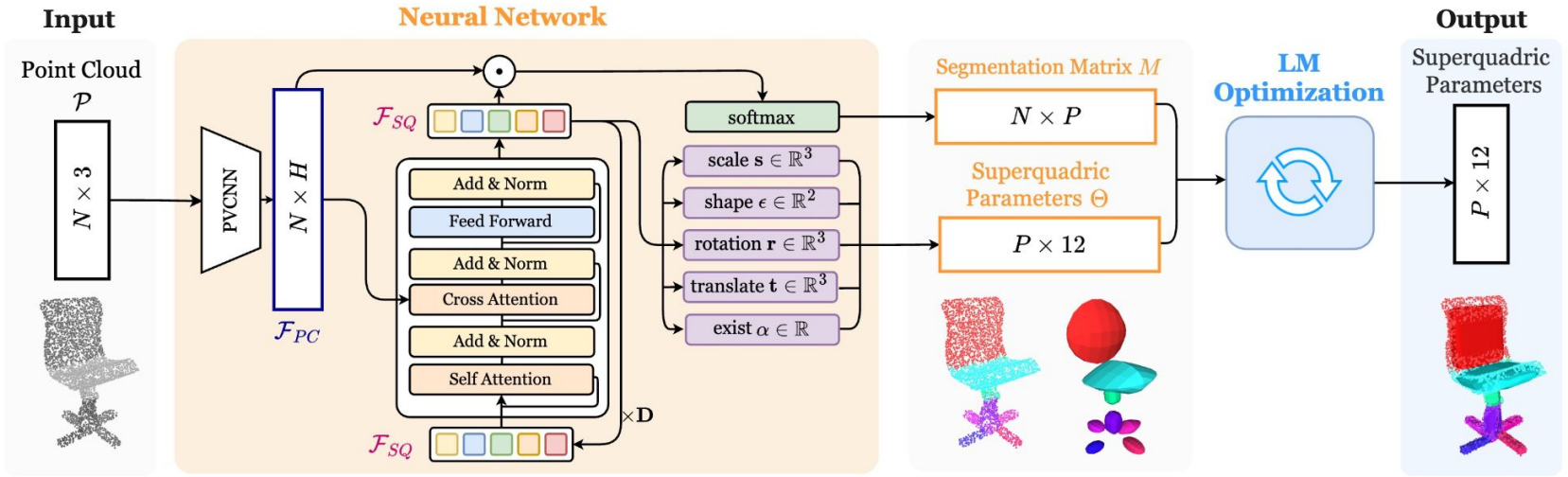
A 3D scene decomposition using superquadric primitives

How do we lift to a 3D scene??

Combing the single-object decomposition & 3D instance segmentation

SUPERDEC: 3D Scene Decomposition with Superquadric Primitives (2/3)

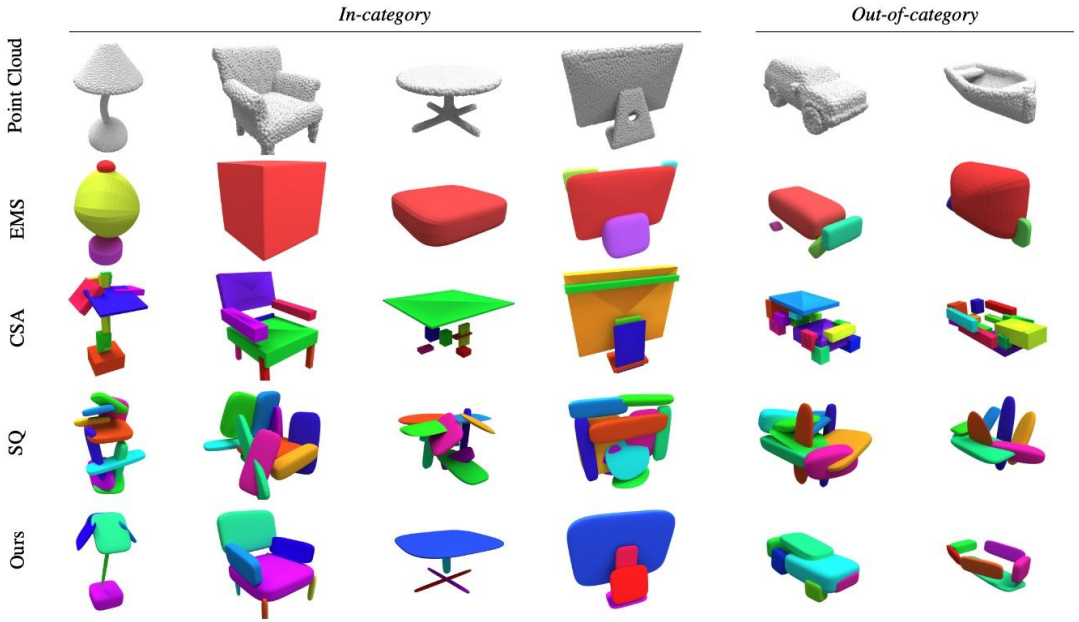
1. Single object decomposition
   a. consists of two main components: superquadric parameters & a segmentation matrix associating points to superquadrics
2. Decomposition of full 3D scenes
   a. extract 3D object instance masks using Mask3D
   b. predict the superquadric primitives for each object individually

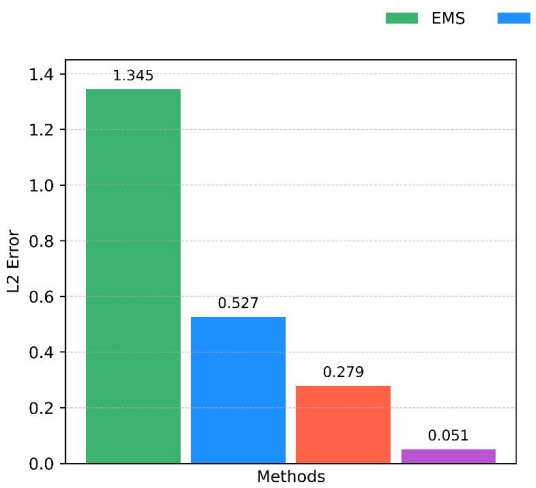


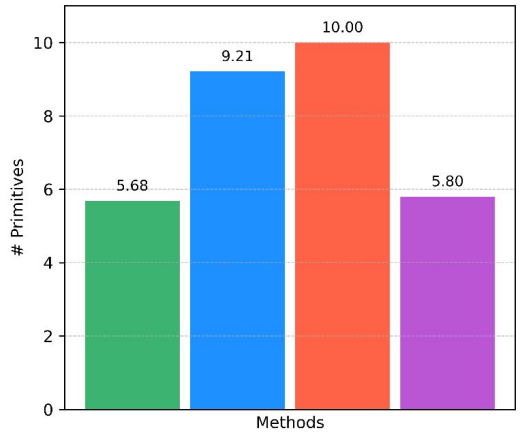
ICCV25: Meta Insights into Trends and Tendencies (152/153)

SUPERDEC: 3D Scene Decomposition with Superquadric Primitives (3/3)

SUPERDEC significantly outperforms both learned and non-learned baselines



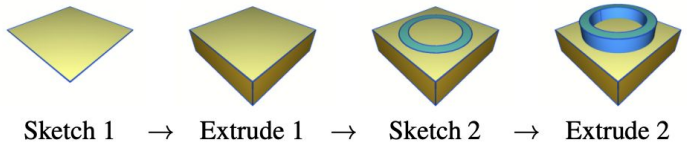




ICCV25: Meta Insights into Trends and Tendencies (153/153)

How to represent and infer geometric information when only edge data is available (like CAD data)?

- Textureless 3D CAD models are crucial for industrial application, yet how to best represent them for current 3D recognition models remains an open question.



(a) Sketch-Extrude sequence

```
SOL Line -50 -50 Line 50 50 Line 50 -50 Line -25
-50 Ext. 0 0 1 0 0 0 14 0 NewBody SOL Circle 0 0 40
SOL Circle 0 0 30 Ext. 0 0 1 0 0 0 -10 0 Union
```

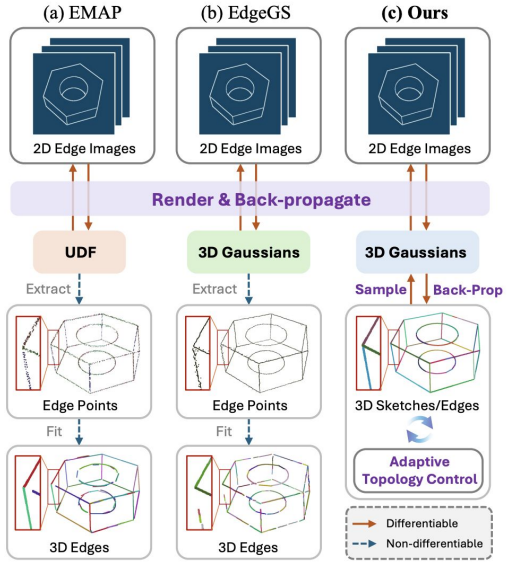
(b) DeepCAD representation

```
import cadquery as cq
w = cq.Workplane('XY')
w.box(100,100,14).union(
  w.sketch().circle(40).circle(30,mode='s')
  .finalize().extrude(-10))
```

(c) Our CadQuery representation

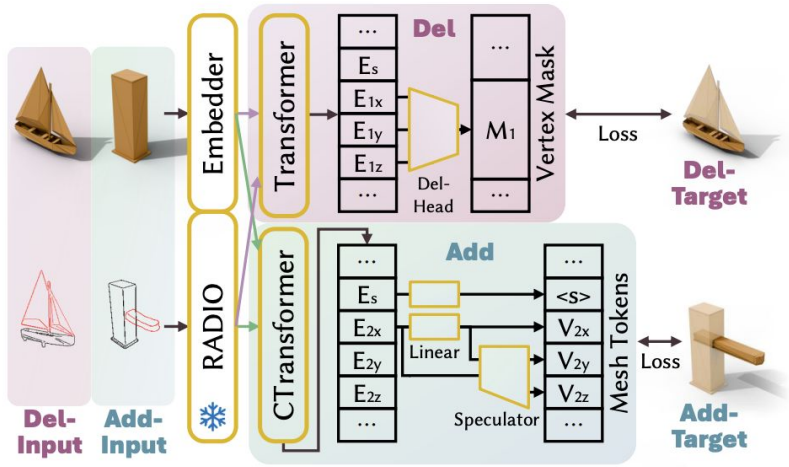
[CAD-Recode]

(Python Code as representation)



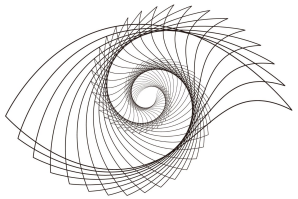
[SketchSplat]

(parametric lines as representation)



[MeshPad]

(3D Mesh as representation)



LIMIT.Lab

A collaboration hub

for building multimodal AI models under limited resources

Core members at LIMIT.Lab

Our Members

Hirokatsu Kataoka
AIST / Oxford VGG

Yoshihiro Fukuhara
AIST

Rintaro Yanagi
AIST

Ryousuke Yamada
AIST / UTN FunAI Lab

Daichi Otsuka
AIST

Partha Das
UvA

Risa Shinoda
Osaka University / AIST

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

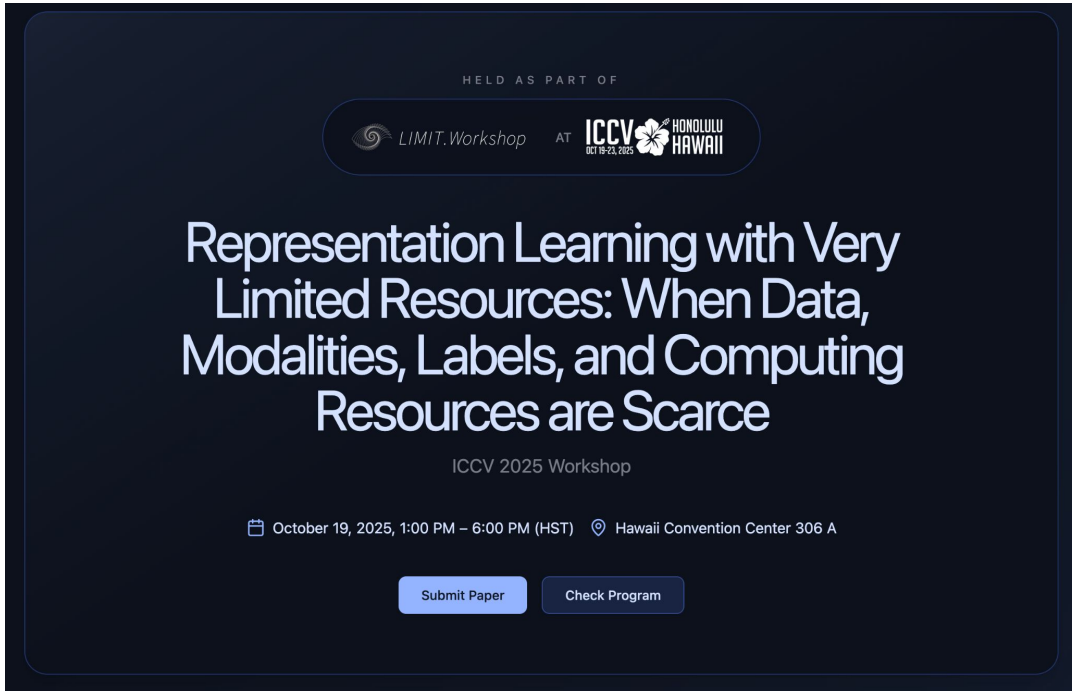
Daniel Schofield
Oxford VGG

Martin R. Oswald
UvA

Limited Resources, Unlimited Impact with Multimodal AI Models

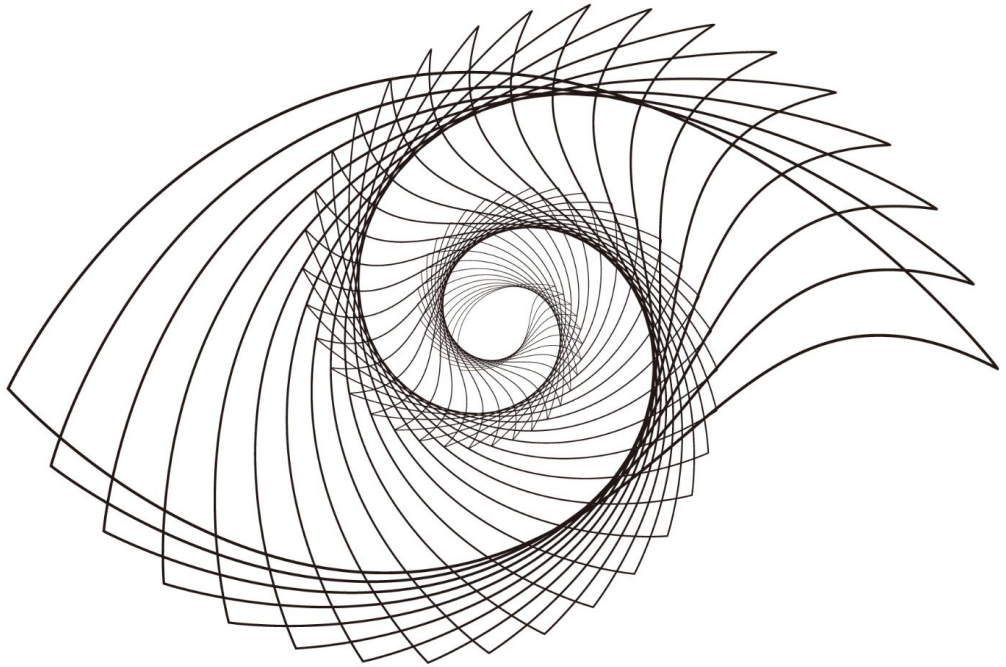
AI foundation models are increasingly dominating various academic and industrial fields, yet the R&D of related technologies is controlled by limited institutions capable of managing extensive computational and data resources. To counter this dominance, there is a critical need for technologies that 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.

2 accepted workshops at ICCV 2025



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

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



LIMIT.LAB

Join us! -> Slack invitation [Link]