CVPR 2025 Report [Link]の日本語版資料

CVPR 2025 速報

Hirokatsu Kataoka, Yoshihiro Fukuhara,

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

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

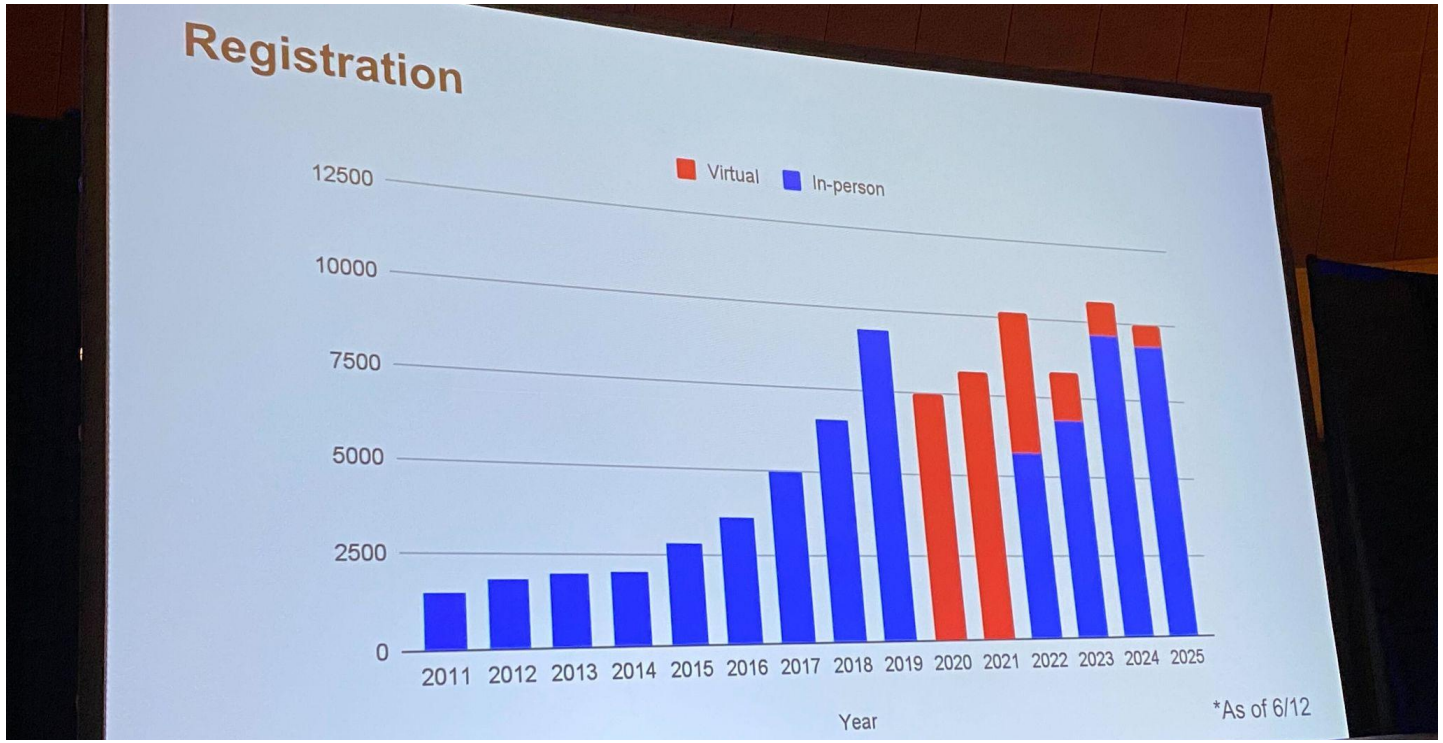
CVPR 2025 の動向・気付き

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

CVPR 2025 の動向・気付き (1/181)

Opening slideより

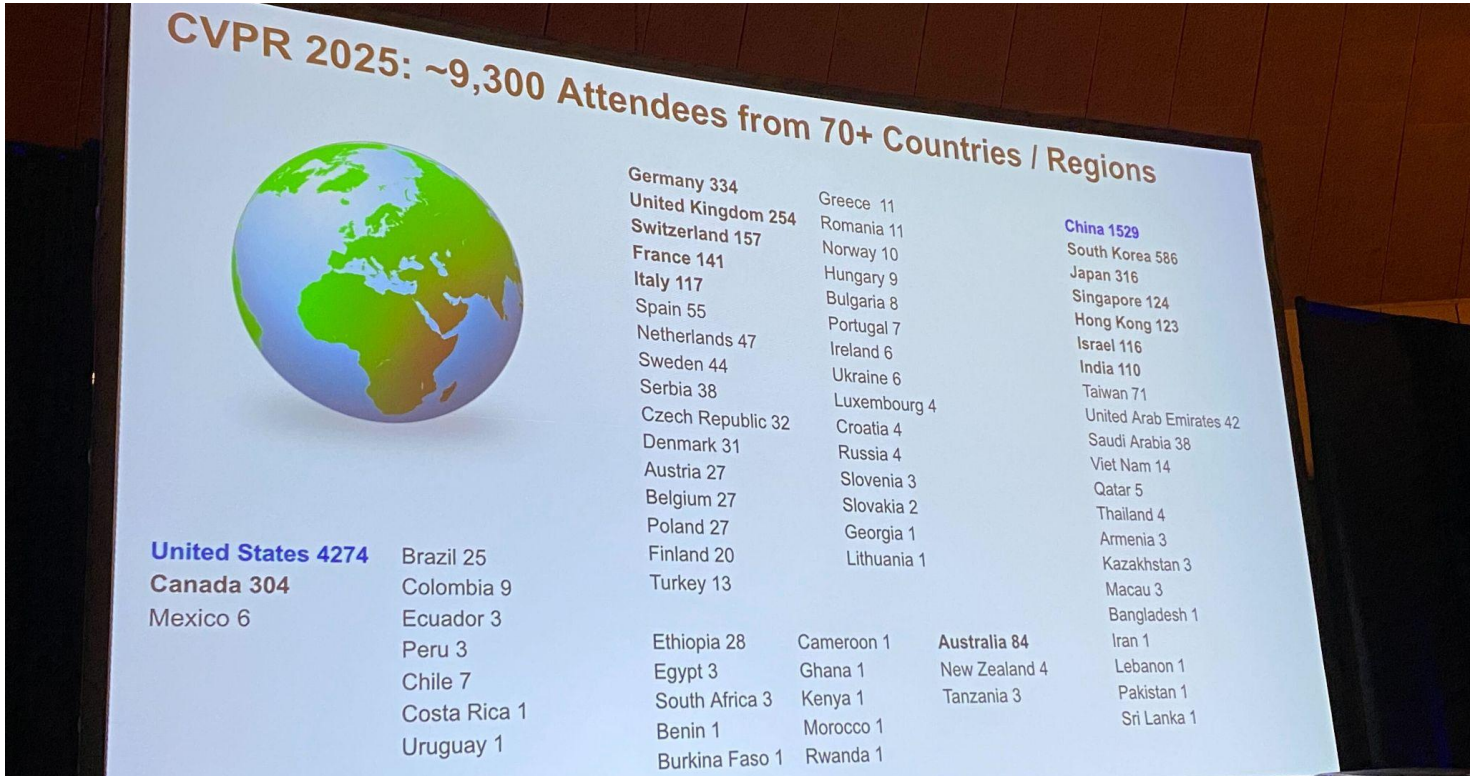
- 今年は参加者が少し減少
  - CVPR 2024の出席者数との比較
- バーチャル参加が許可されていた



CVPR 2025 の動向・気付き (2/181)

Opening slideより

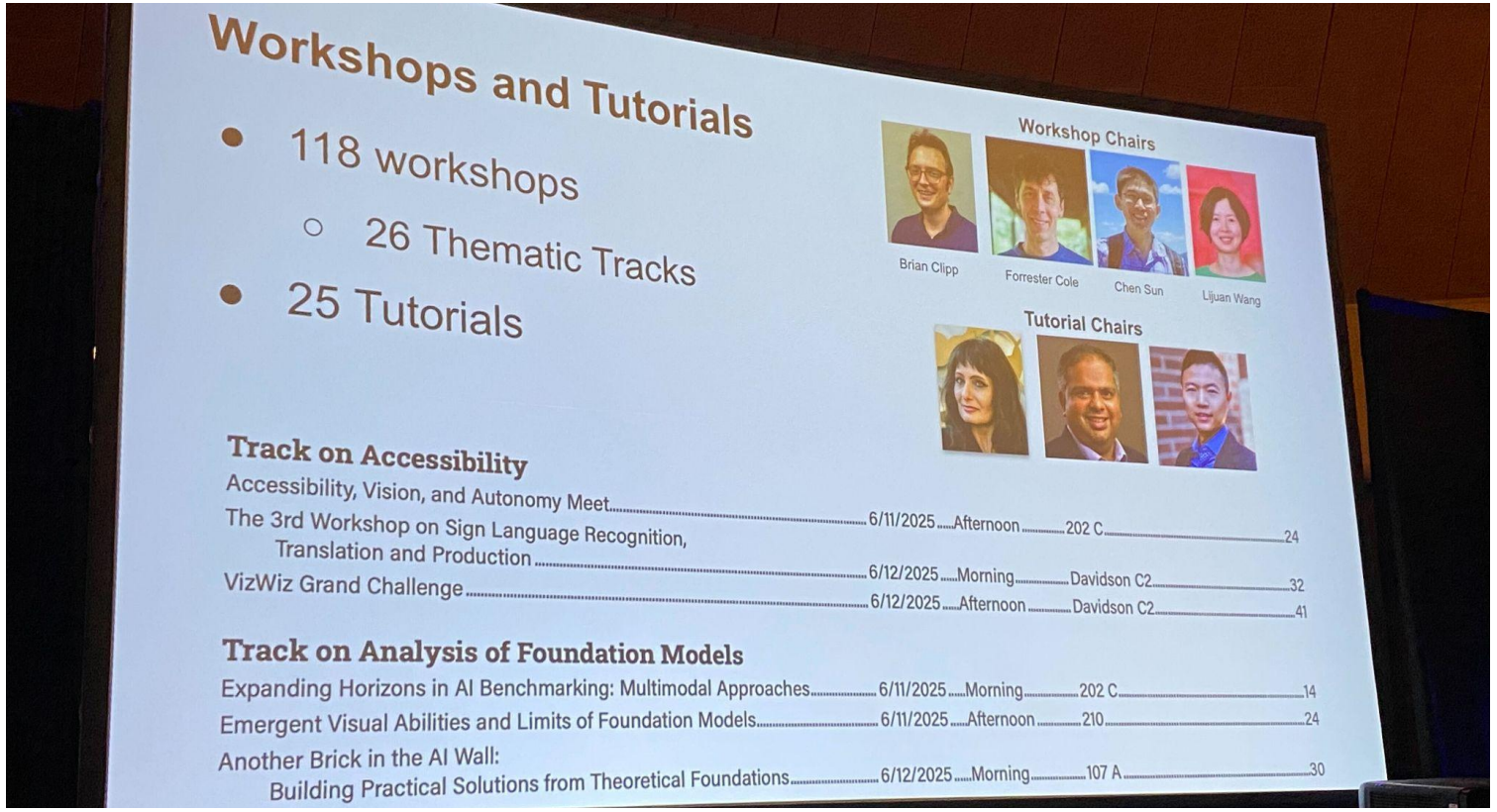
- 大陸ごとに掲載
  - 米国: 4,274、中国: 1,529、韓国: 586、ドイツ: 334、日本: 316、カナダ: 304、英国:254など
  - 合計70以上の国 / 地域



CVPR 2025 の動向・気付き (3/181)

Opening slideより

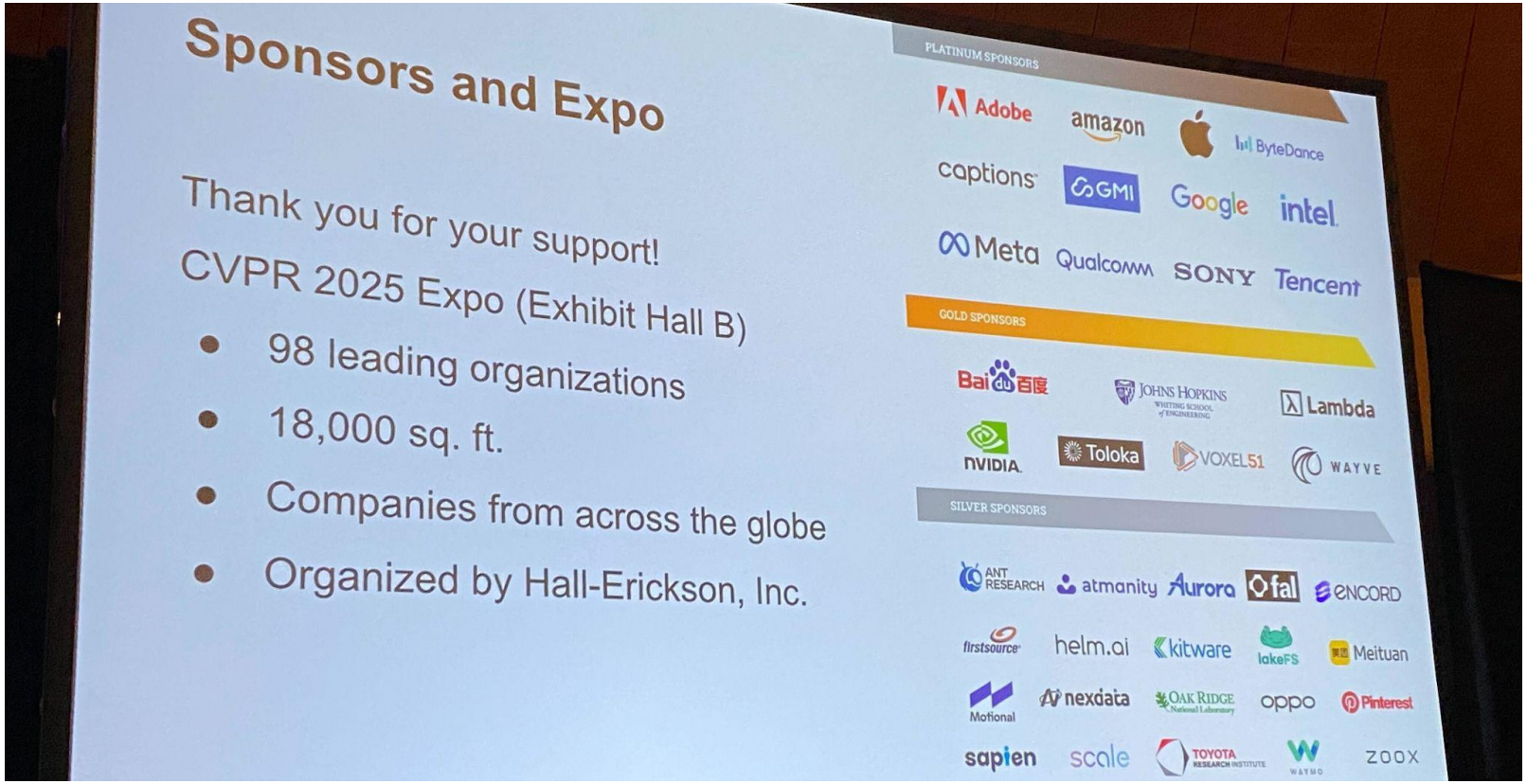
- ワークショップ数は前年度よりわずかに減少
  - 2024: 123 vs 2025: 118 ワークショップ
- 118のWSは26のトラックに分割



CVPR 2025 の動向・気付き (4/181)

Opening slideより

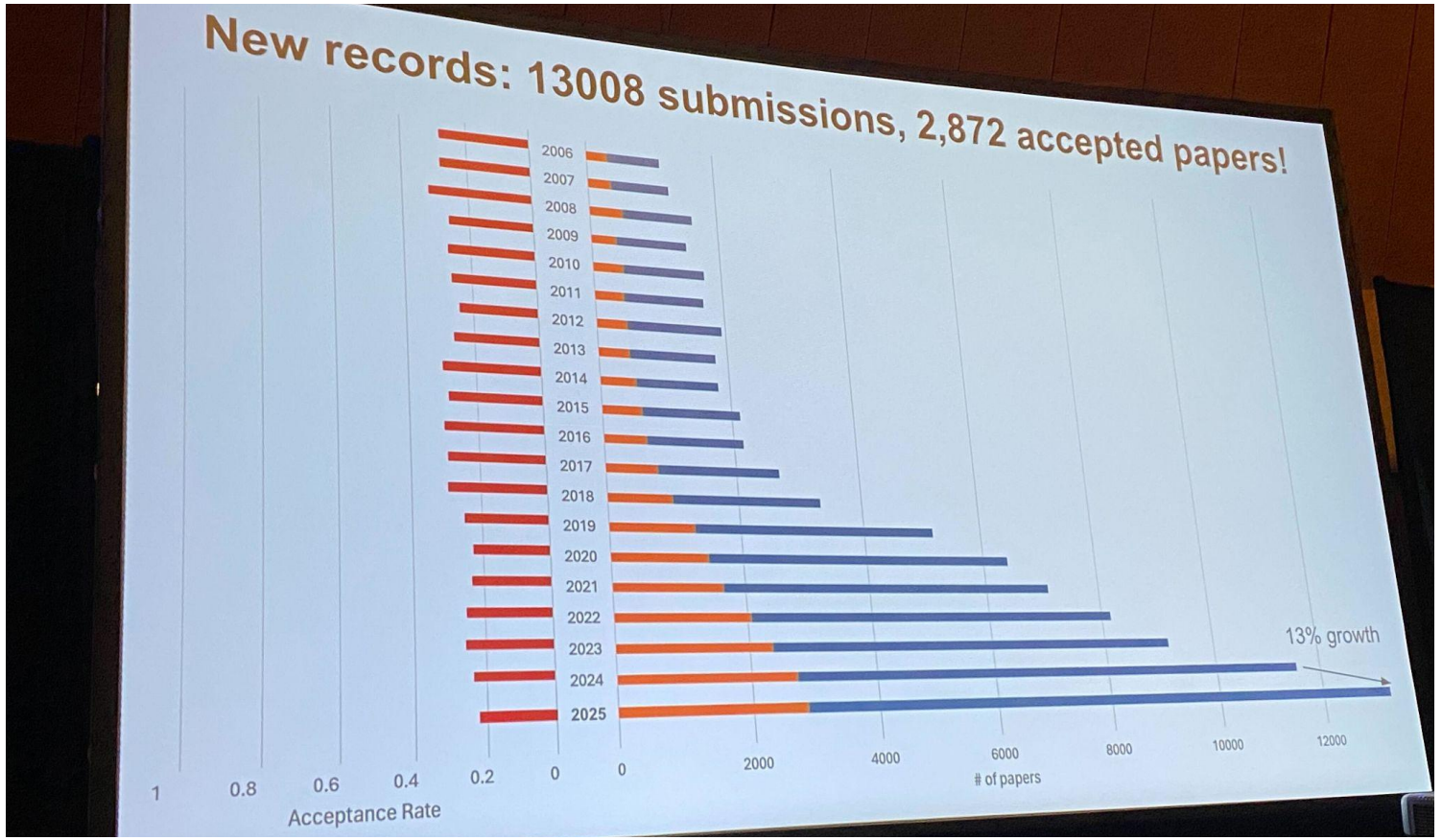
- CVPRをリードしているスポンサー企業一覧
- 企業の貢献もあり、CVPRコミュニティは加速し続けている



CVPR 2025 の動向・気付き (5/181)

Opening slideより

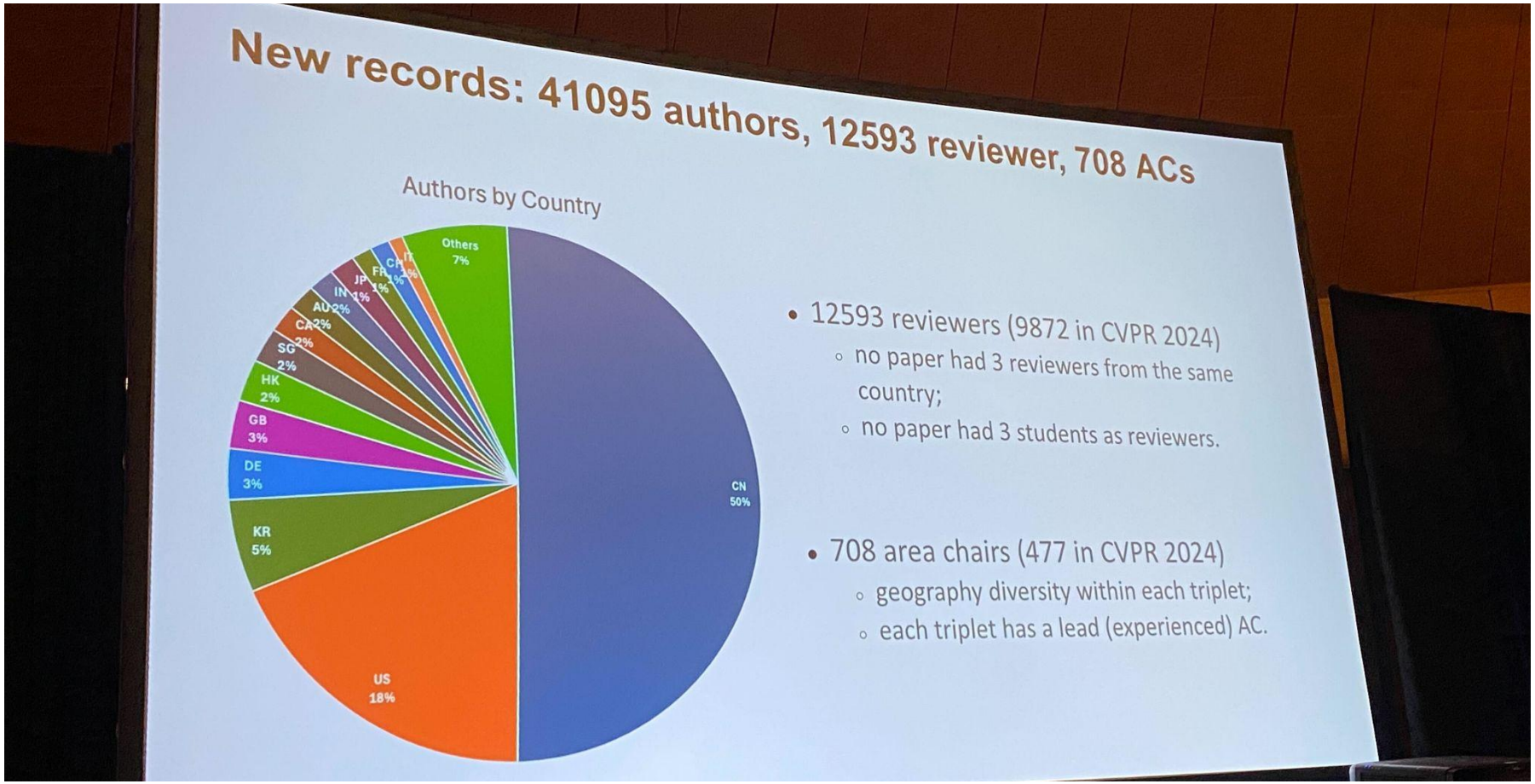
- 論文投稿数 – 2024: 11,532 vs. 2025: 13,008 (12.8% 増加)
- 採択率 – 2024年: 23.55% vs. 2025年: 22.08% (1.47 ポイント減少)



CVPR 2025 の動向・気付き (6/181)

Opening slideより

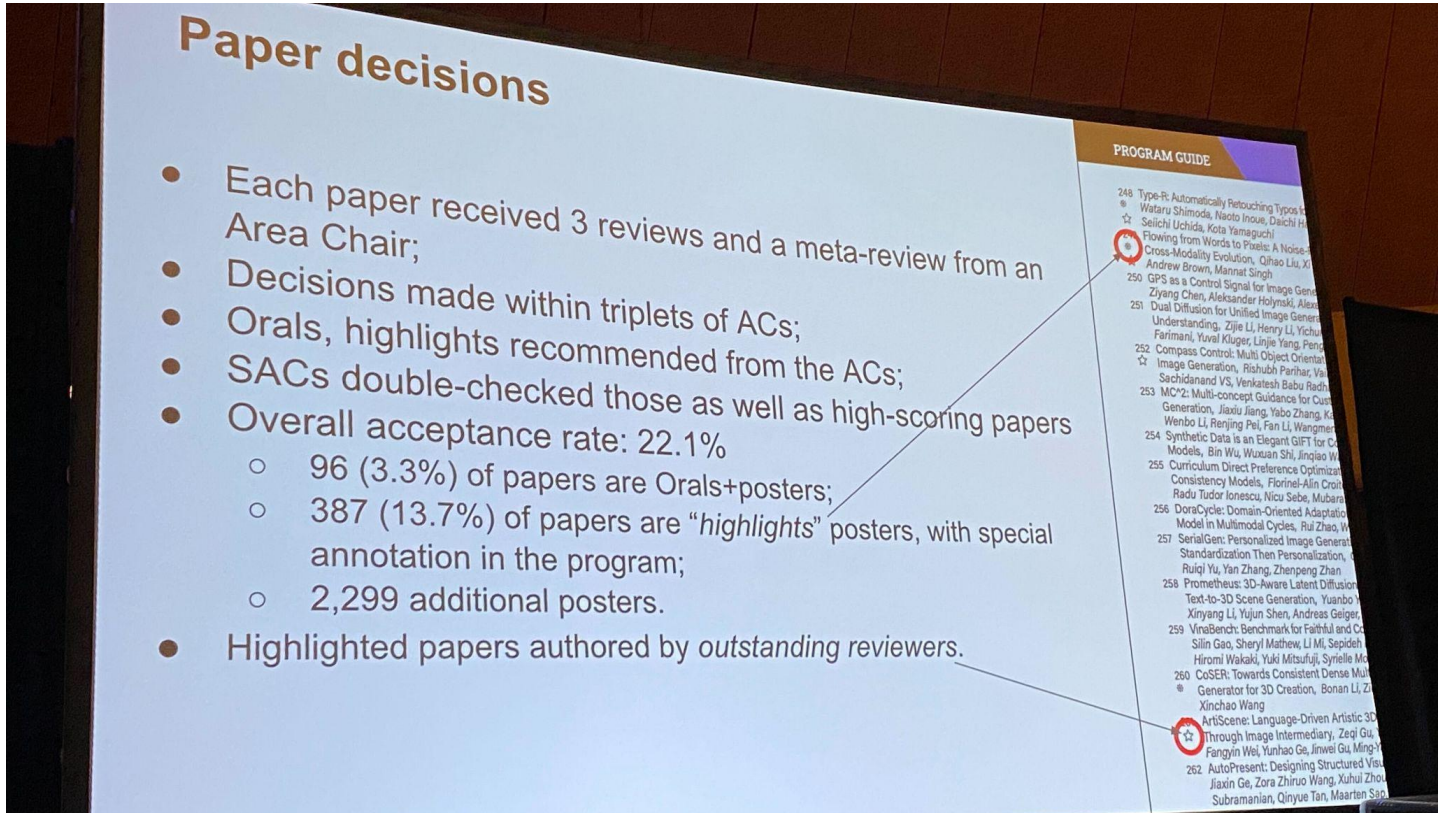
- 急速成長するCVPR community！
  - 著者数、査読者数、エリアチェア数から



CVPR 2025 の動向・気付き (7/181)

Opening slideより

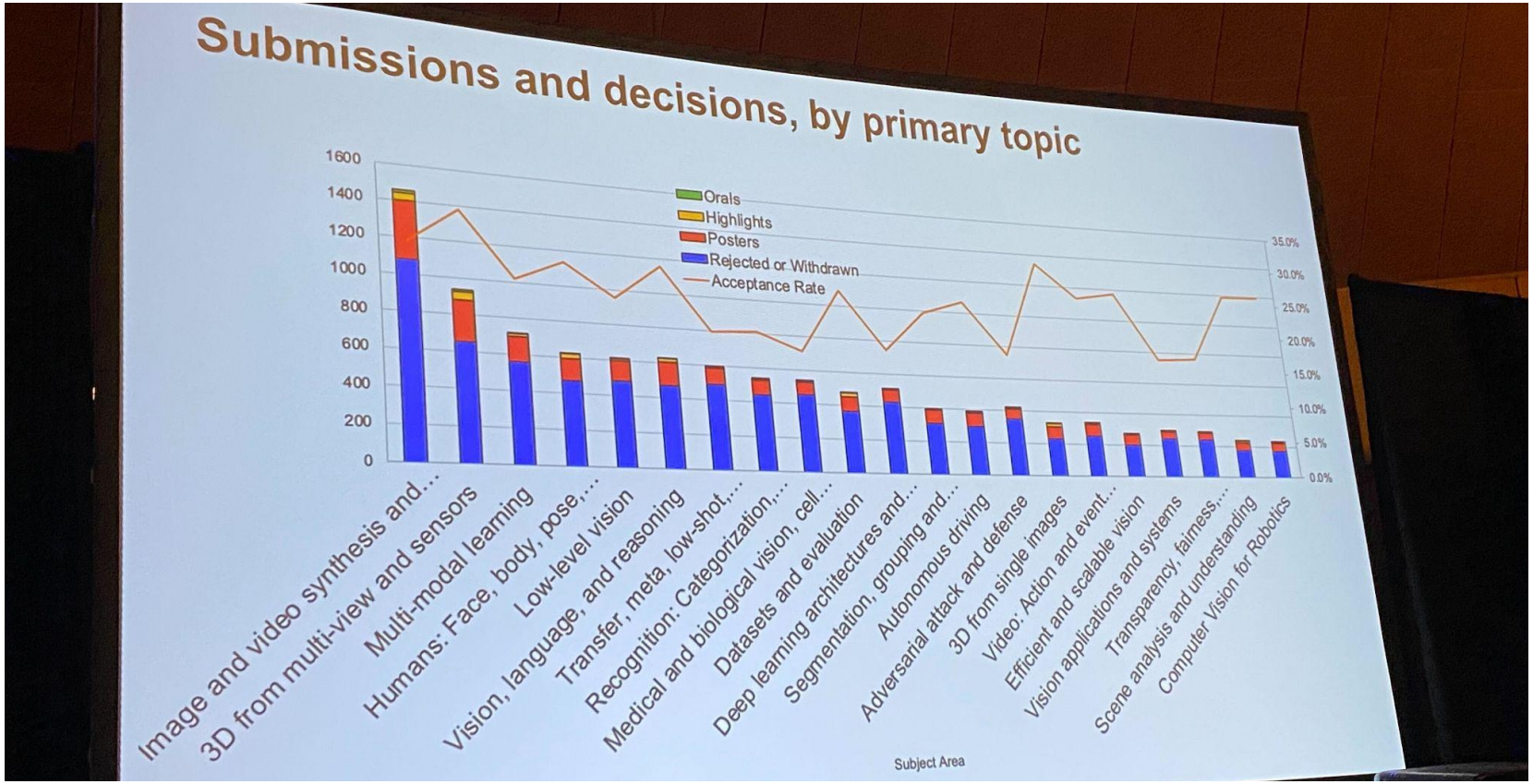
- 査読貢献によるCVPR論文への影響
  - Outstanding reviewer が(共)著者として含まれる論文はハイライトされる
  - 良い査読貢献は採択論文に影響を与える可能性がある



CVPR 2025 の動向・気付き (8/181)

Opening slideより

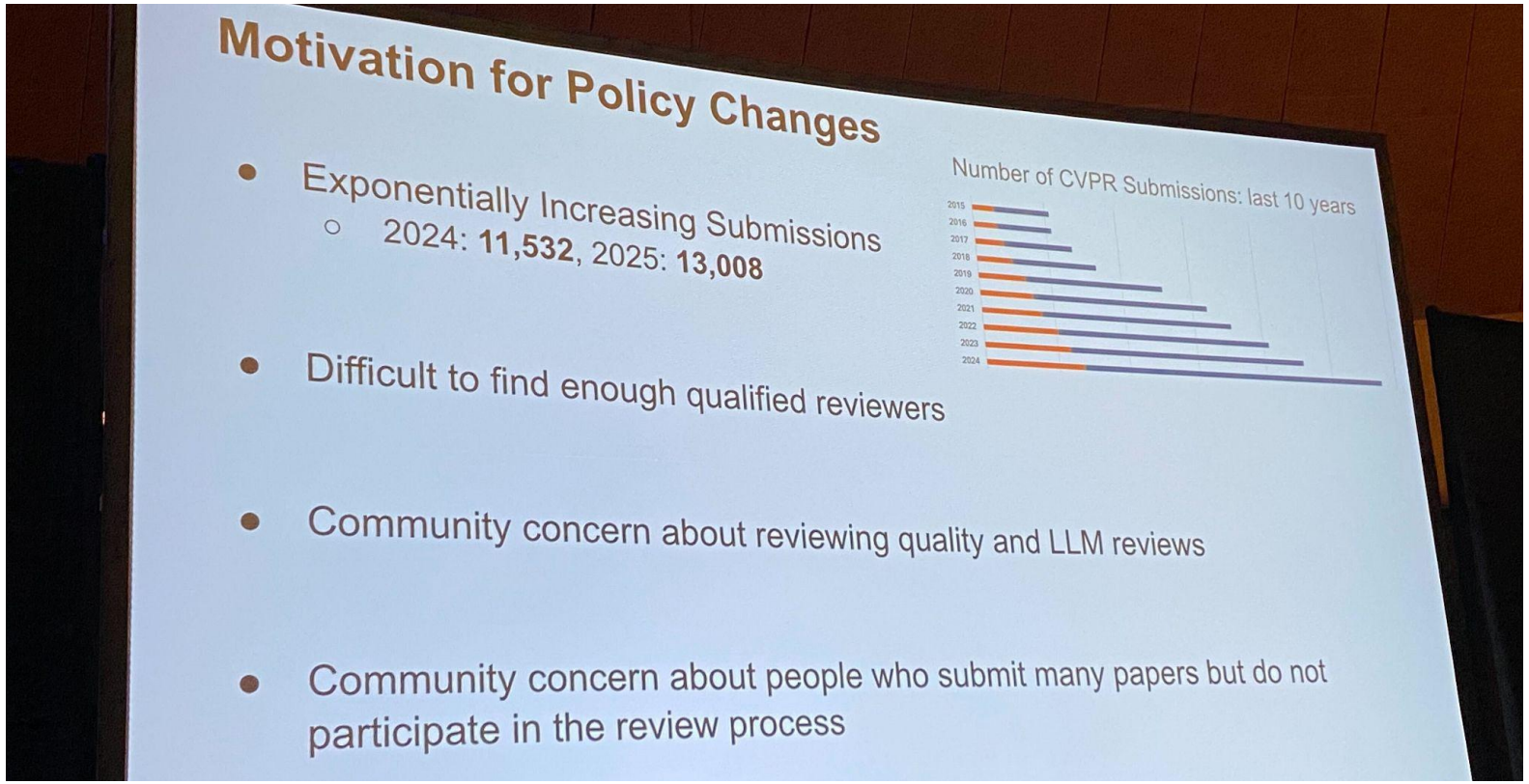
- CVPR 2025のキーワードからの動向
  - 次のトレンドは…？



CVPR 2025 の動向・気付き (9/181)

Opening slideより

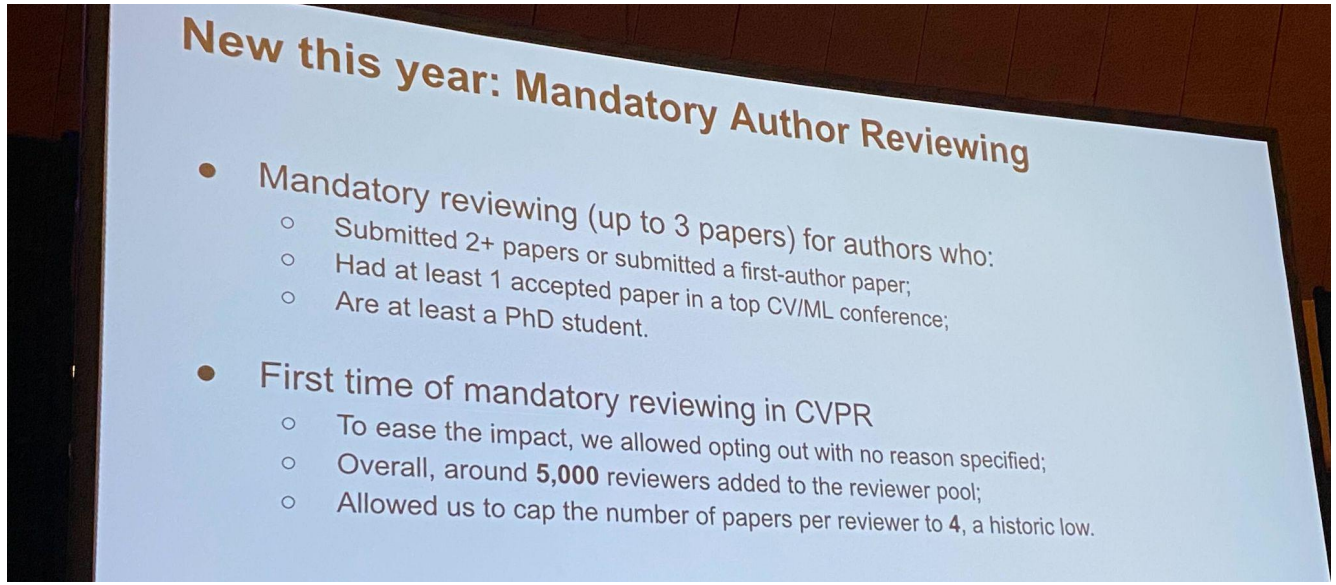
- Program committee は査読の質を維持するために検討中
  - 審査プロセスに関するいくつかの決定
    - 質を保つための査読者 / 査読者教育が必要

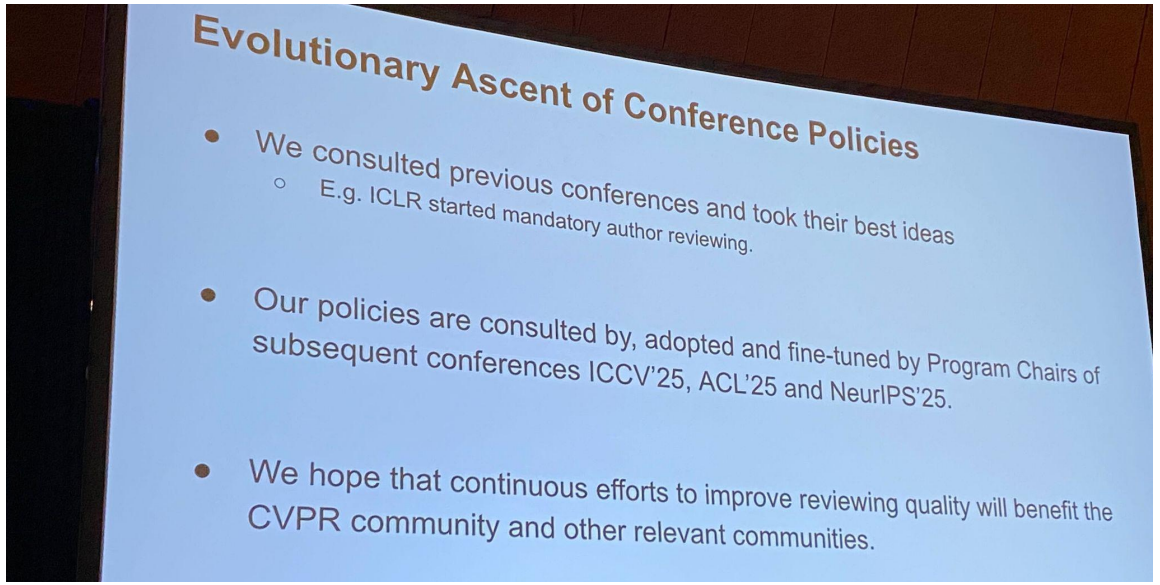


CVPR 2025 の動向・気付き (10/181)

Opening slideより

- 筆頭著者 / 複数論文投稿者は論文査読が必須化
  - 現状、論文査読は義務化されている
  - ボランティアのみでは査読数の確保が困難になりつつある

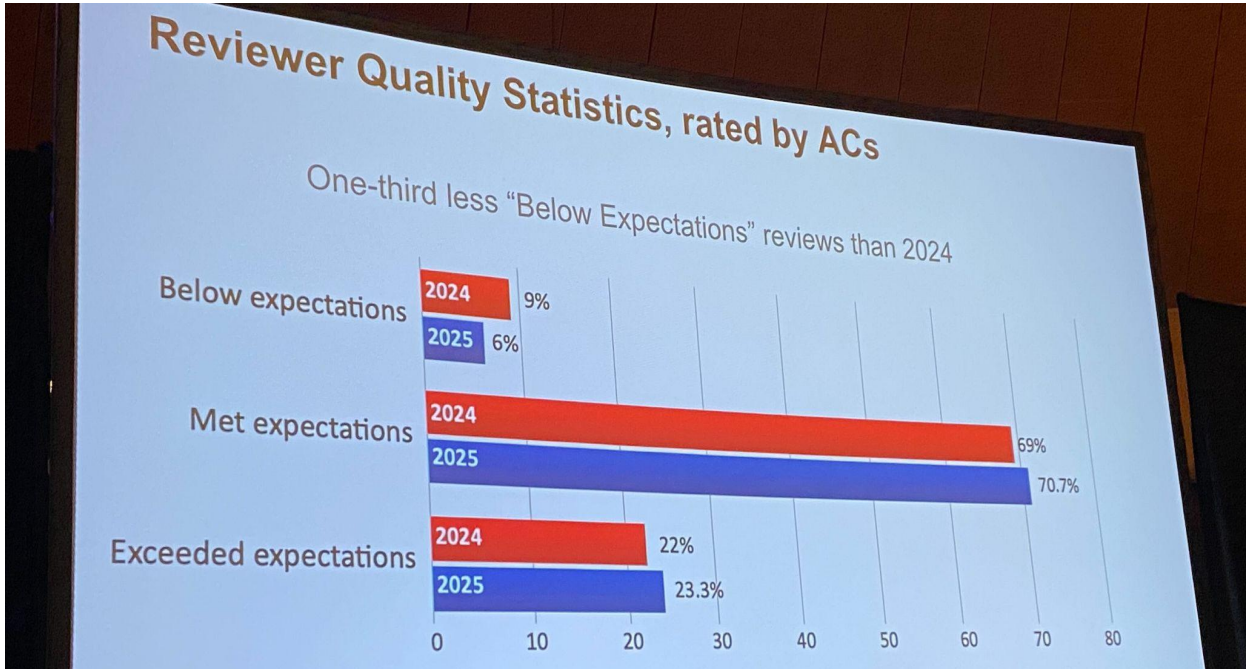


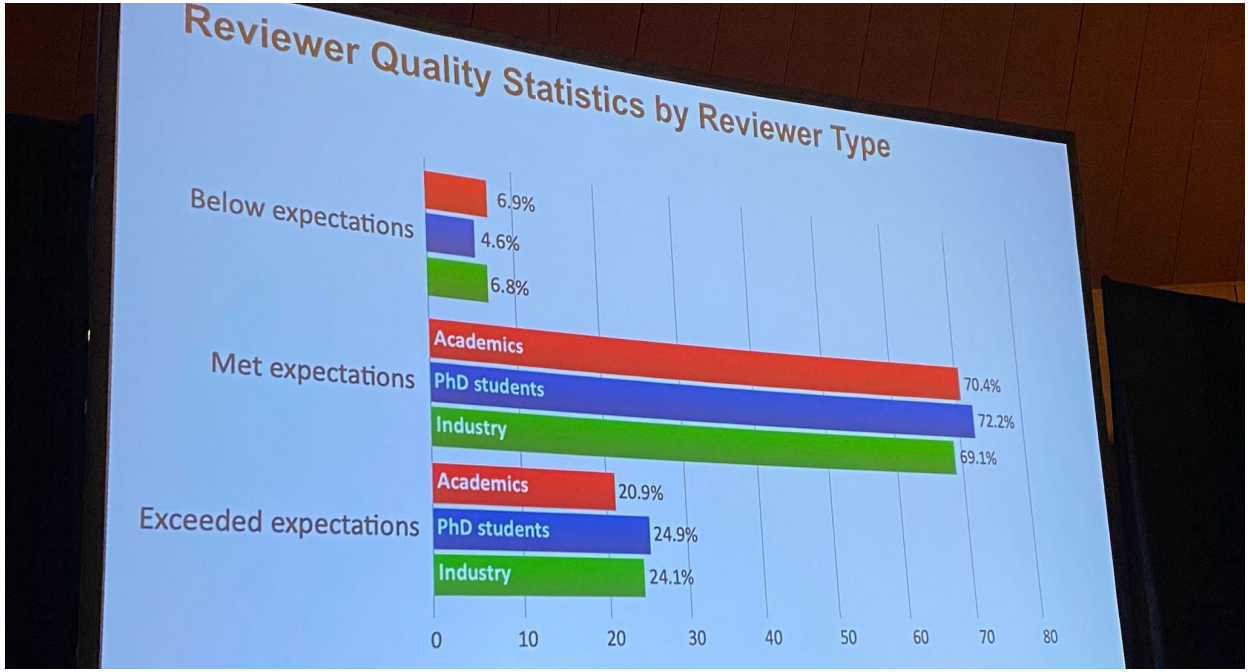


CVPR 2025 の動向・気付き (11/181)

Opening slideより

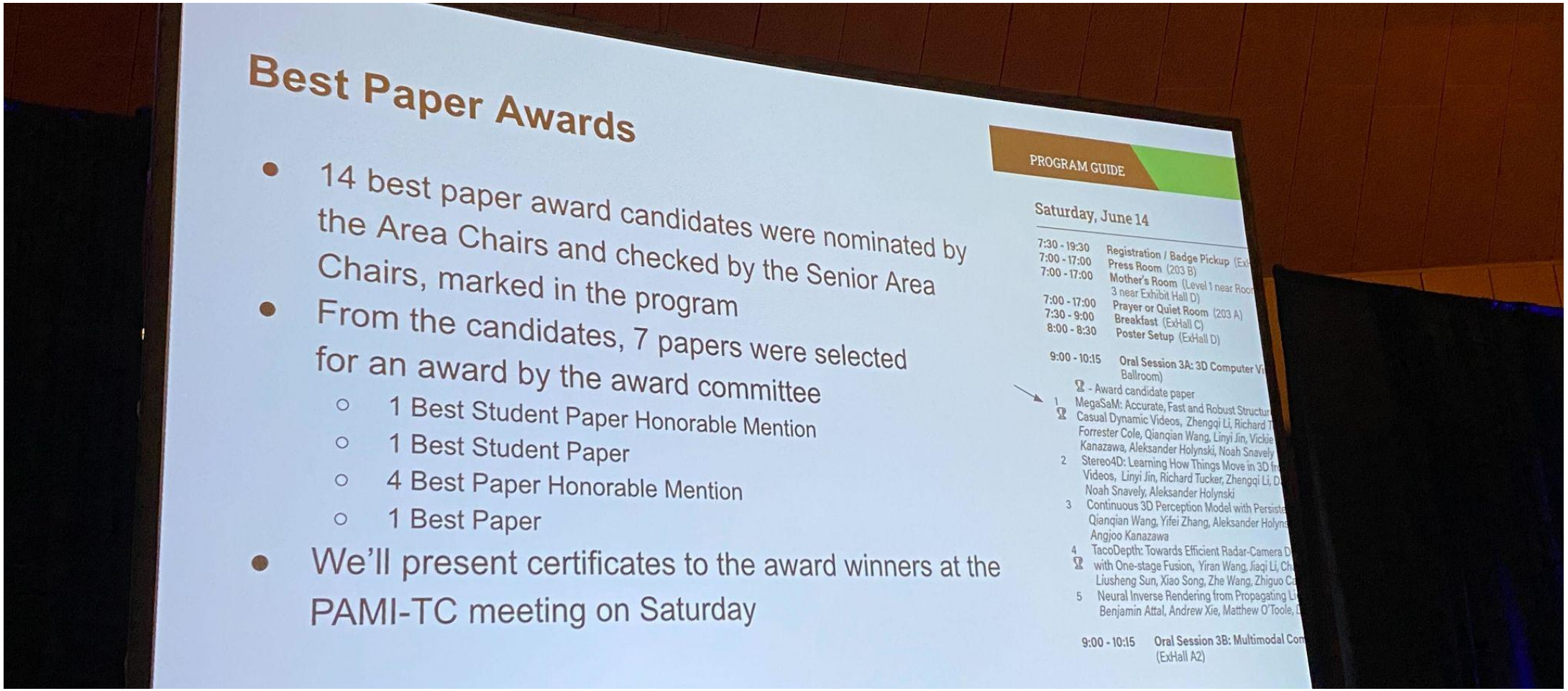
- エリアチェアの判断によると、
  - CVPR 2025では、前年比で良いレビューが集まった
  - 博士課程学生は優れた査読者になることができる！





CVPR 2025 の動向・気付き (12/181)

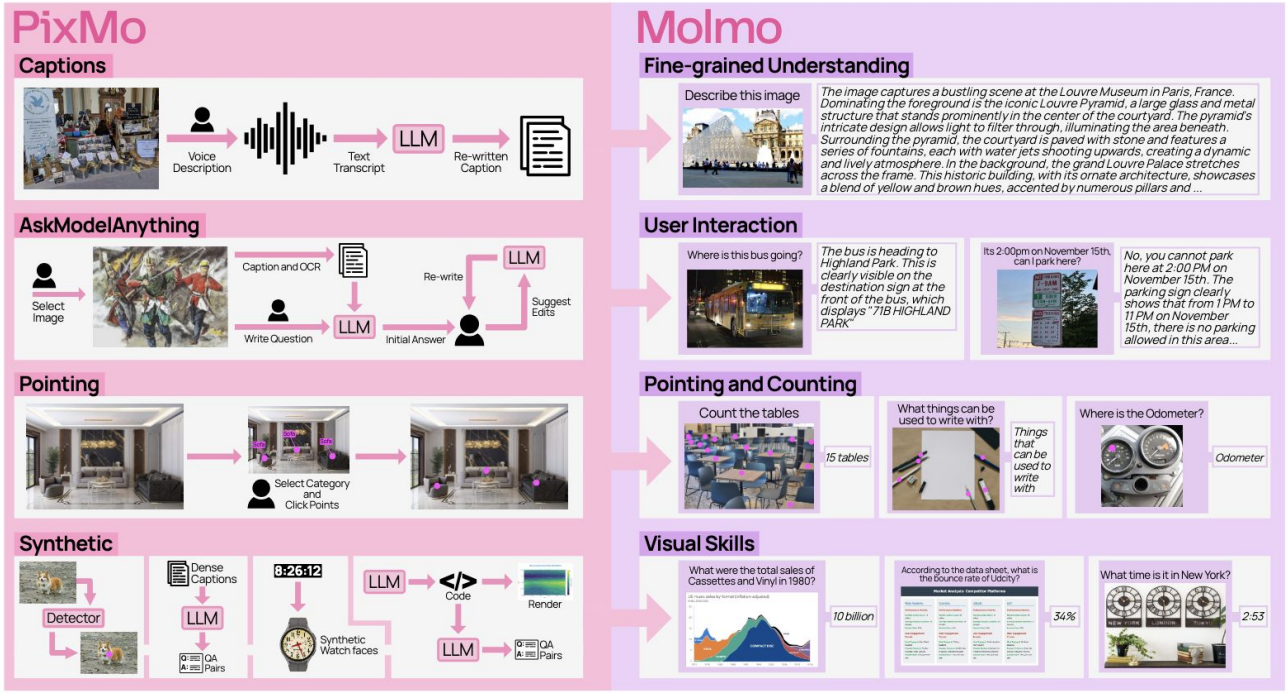
Award Ceremonyより



CVPR 2025 の動向・気付き (13/181)

Best paper候補論文 1: Molmo and PixMo: Open Weights and Open Data for State-of- the-Art Vision-Language Models

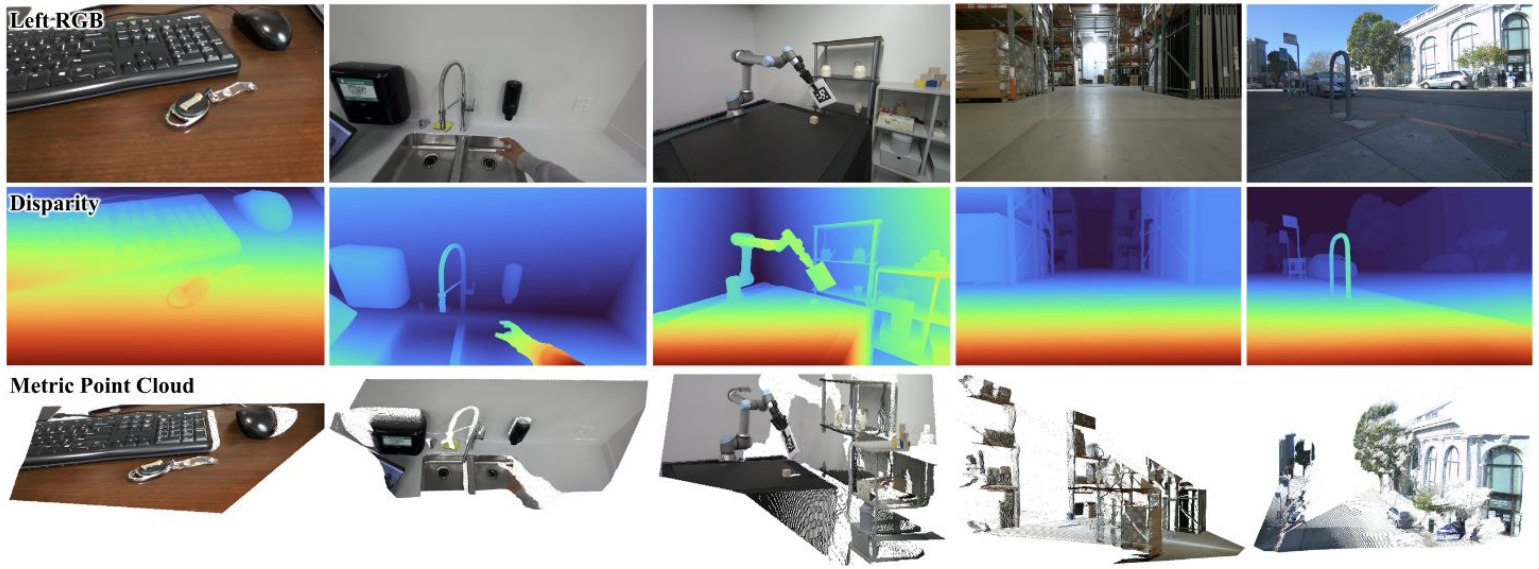
- Multimodal AI（特にVLM）に関して、最先端で用いられているものはGPT / Geminiなど、必ずしもモデルやデータが必ずしも公開されていないことが学術的な問題となっている。本稿では、データやモデルが公開されている状態で最先端のモデルが構築できるのか、について正面から取り組み、実際に論文投稿時のSotAモデルであるGPT-4oやGemini-1.5proなどと同等以上のVLM構築に成功した。もちろん、データ・コード・パラメータ・評価法などが一般的に公開されている。



CVPR 2025 の動向・気付き (14/181)

Best paper候補論文 2: Foundation Stereo: Zero-shot Stereo Matching

- 合成データ(FoundationStereo Dataset; FSD)やDepthAnythingV2のみでステレオマッチングによる実世界の距離画像推定を実現するという技術。ほぼ合成データでゼロショット距離画像生成ができて、DepthAnythingV2 のネットワーク微調整(論文中では Side-Tuning Adapter (STA))でゼロショット距離画像に成功。合成データについてはNVIDIA Omniverseを用いて1Mの高精細画像を生成して学習に使用。実験では、ゼロショットの距離画像推定でほぼSotA、fine-tuningすると尚良いという結果が得られている。



CVPR 2025 の動向・気付き (15/181)

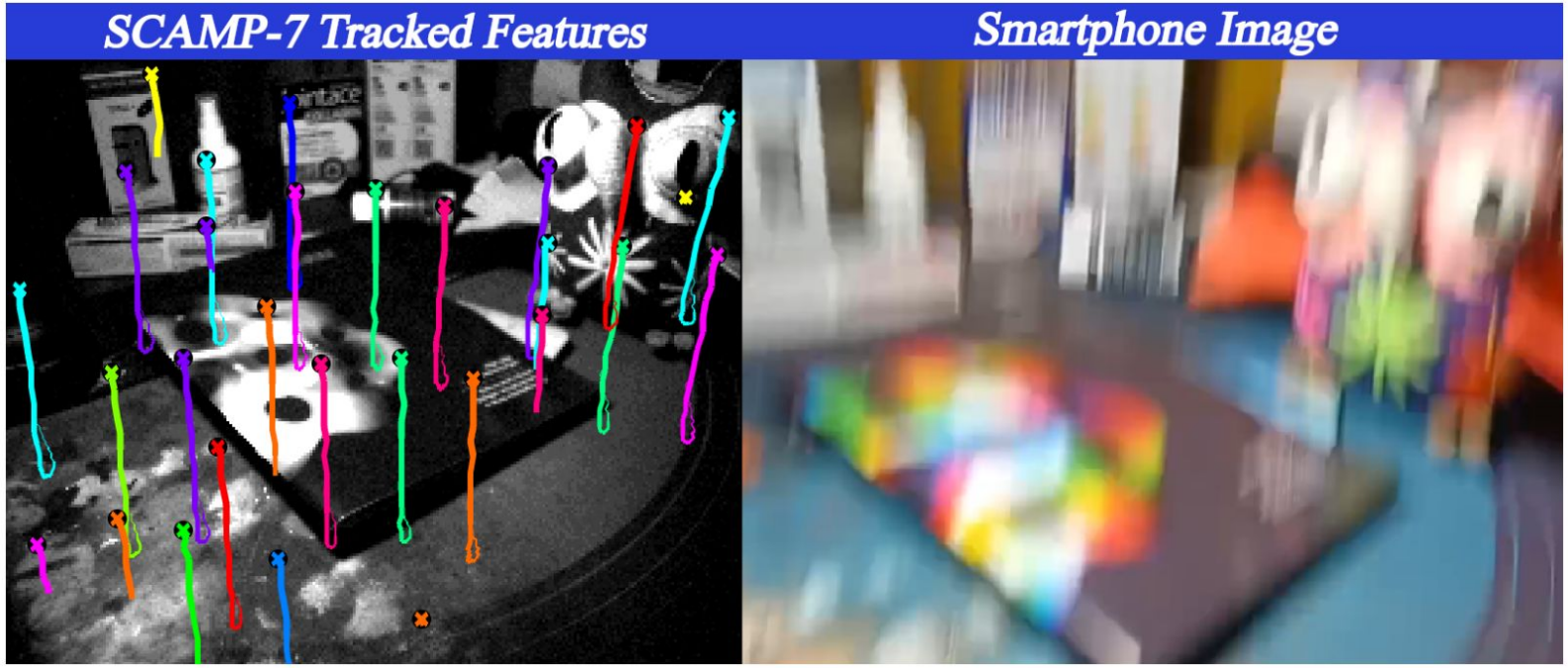
Best paper候補論文 3: VGGT: Visual Geometry Grounded Transformer

- 現実世界におけるシーン内の複数画像入力から3D再構成するTransformerであるVGGTを提案。単一Transformer内にてカメラ姿勢推定、深度マップ、特徴点対応関係推定を1回のfeed-forwardで高速に推定することでシーン内を3D再構成。実験的にも、多くの3次元構成や特徴点対応関係推定、新規視点推定のベンチマークでSotAを達成。モデルとしてはフレーム内と全体の self-attention を交互に実行するなどの機能が加えられている。画像内のローカル情報と複数視点間のグローバル情報の整合性を取ることも実現しており、3D再構成結果を後から修正する必要がない。

CVPR 2025 の動向・気付き (16/181)

Best paper候補論文 4: Descriptor-In-Pixel: Point-Feature Tracking For Pixel Processor Arrays

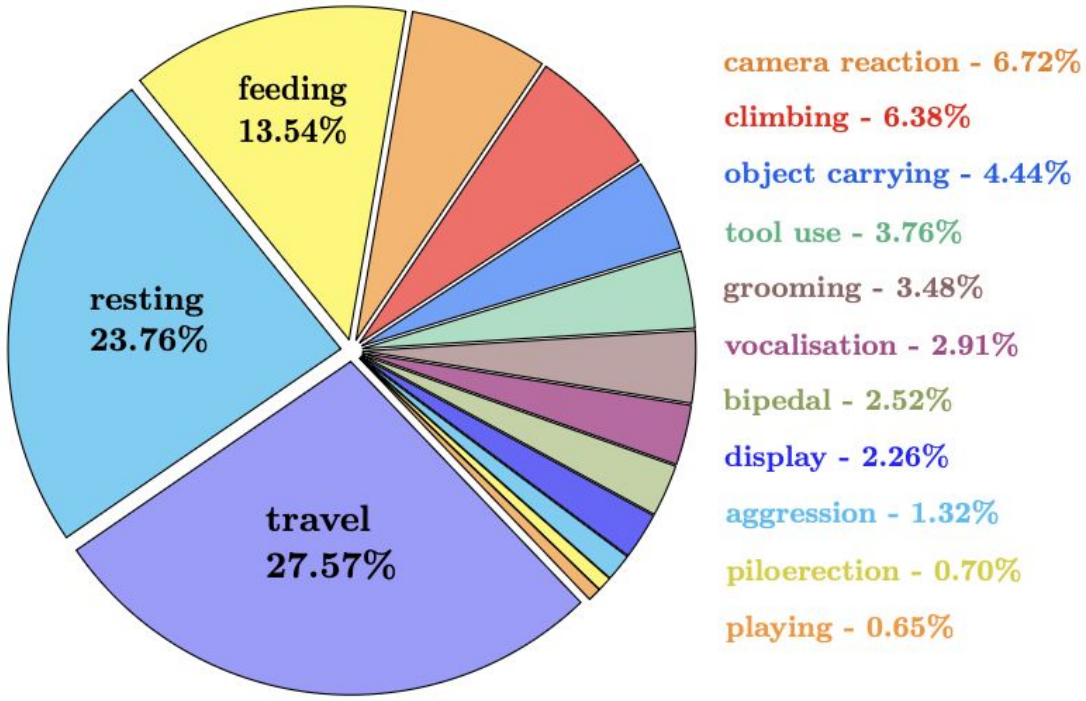
- まとまったピクセルを同時処理するセンサーを開発したもの(らしい)。センサー内で特徴点検出・追跡を同時処理できるため、超高速(3,000+FPS)でトラッキングできる。消費電力も低く、カメラ外に情報が漏れないので、セキュリティ面でも優れる。動画内では激しいカメラモーションがあったとしても動的なトラッキングに成功している。

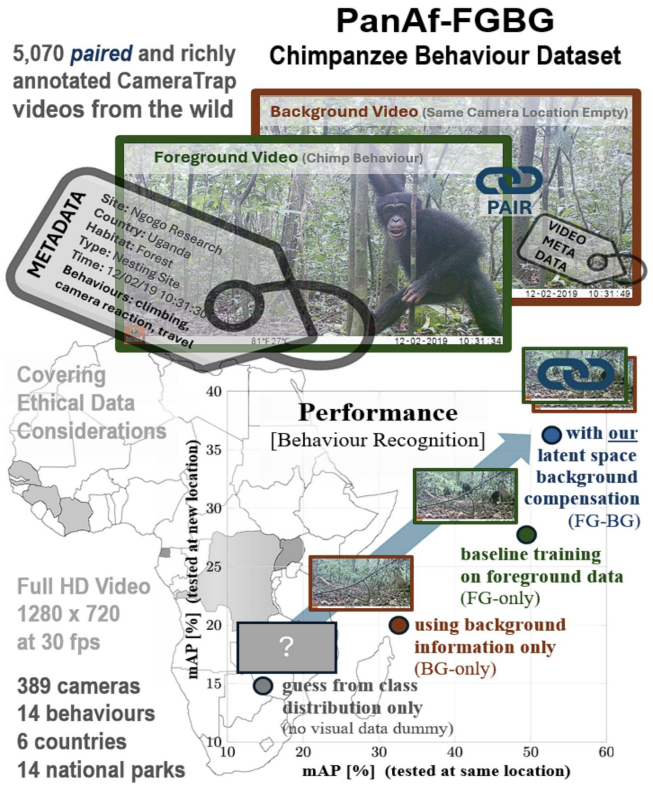


CVPR 2025 の動向・気付き (17/181)

Best paper候補論文 5: The PanAf-FGBG Dataset: Understanding the Impact of Backgrounds in Wildlife Behaviour Recognition

- 自然保護におけるカメラトラップのデータセット構築研究。同データセットでは、チンパンジーの保護や観測を目的としてアフリカ6カ国各地に設置されたカメラから20時間以上のチンパンジーの行動データを収録している。カメラ設置方法を工夫することで、in-distribution / out-of-distributionな行動観測を実現、生態学的にも貢献が見られる。

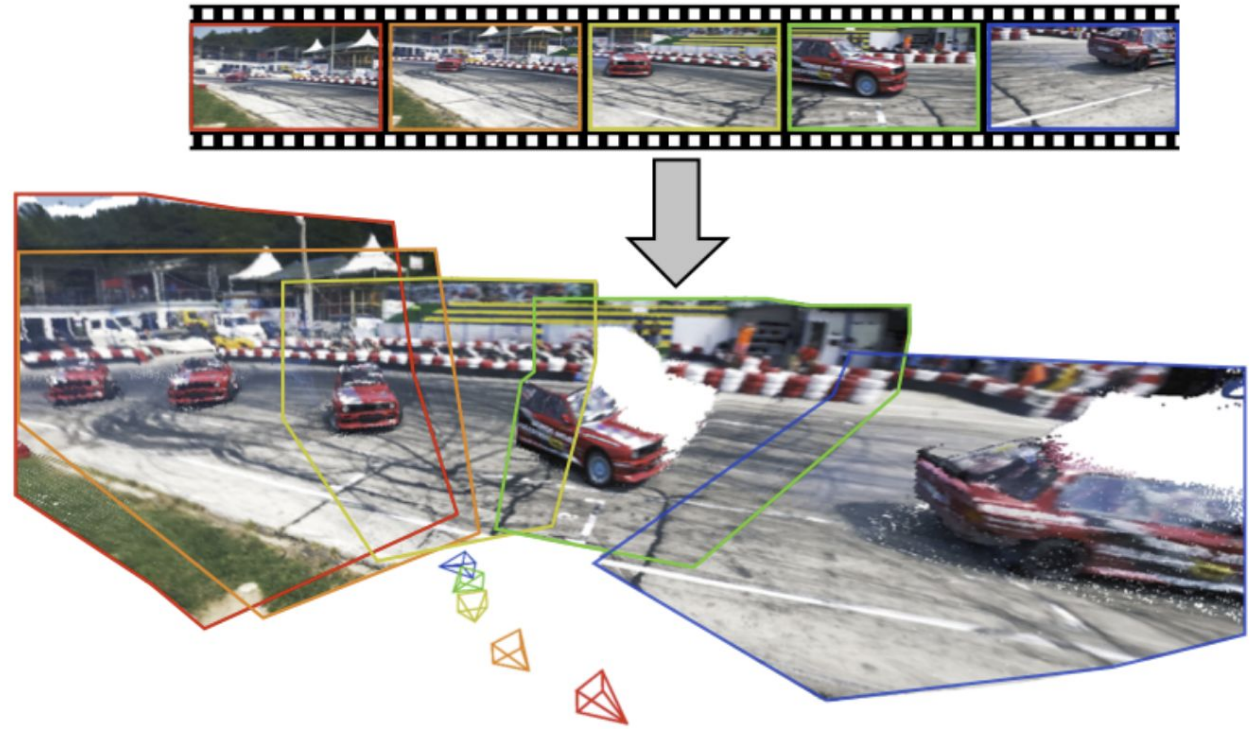




CVPR 2025 の動向・気付き (18/181)

Best paper候補論文 6: MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos

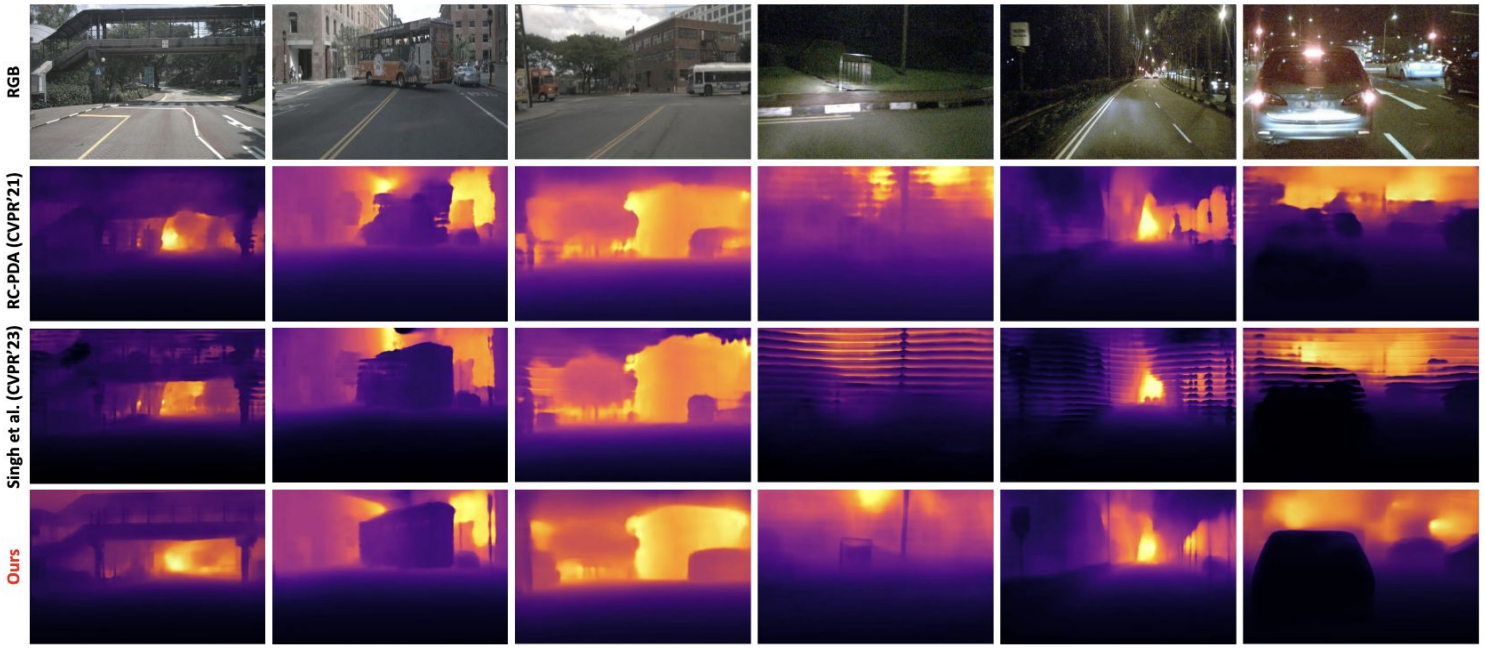
- 4D再構成(動的シーンの3D再構成)は深層学習をベースとしたVisualSLAMが主流となっているが、本稿では単一動画入力から高性能・高速・ロバストなカメラパラメータ推定や距離画像推定を実現している。かなり複雑なシーンにも対応できる。Monocular depth priors、motion probability maps、不確実性の考慮などさまざまな制約を加えている。

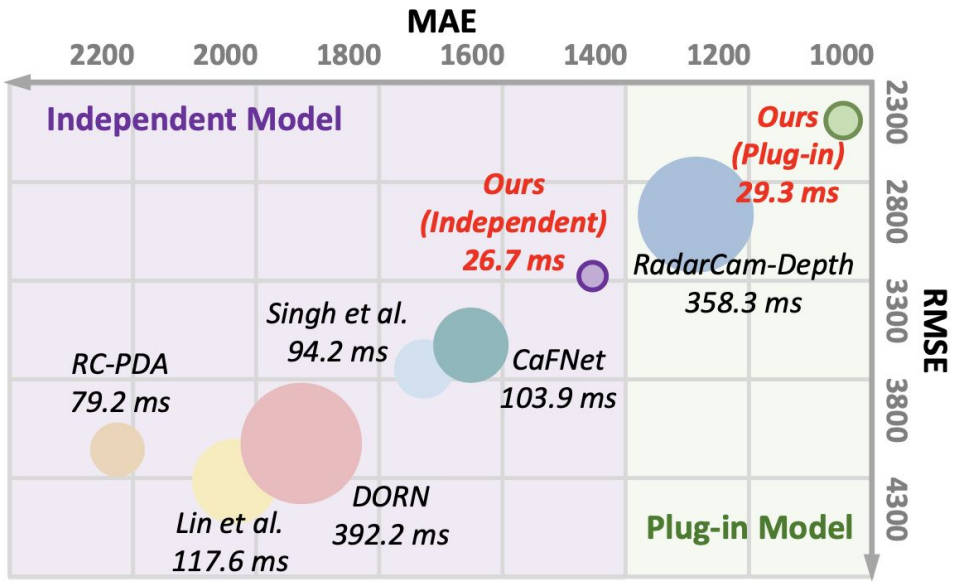


CVPR 2025 の動向・気付き (19/181)

Best paper候補論文 7: TacoDepth: Towards Efficient Radar-Camera Depth Estimation with One-stage Fusion

- 画像とレーダーセンサ(LiDARなど3D点群をスキャン)の入力から距離画像を構成する問題において、圧倒的な性能向上と高速化を実現。従来、スパースかつノイズを多数含むような推定結果だったが、ステレオ視で得られるようなdenseな推定結果を得ることに成功した。

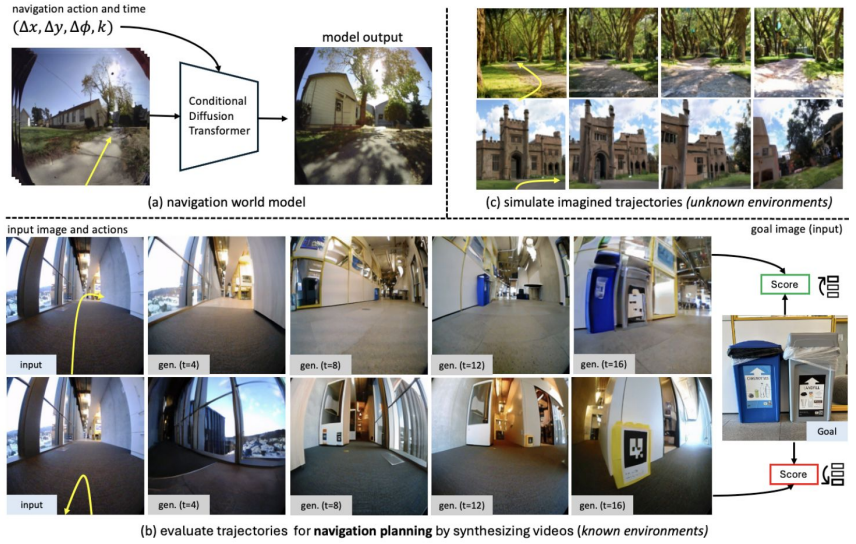




CVPR 2025 の動向・気付き (20/181)

Best paper候補論文 8: Navigation World Models

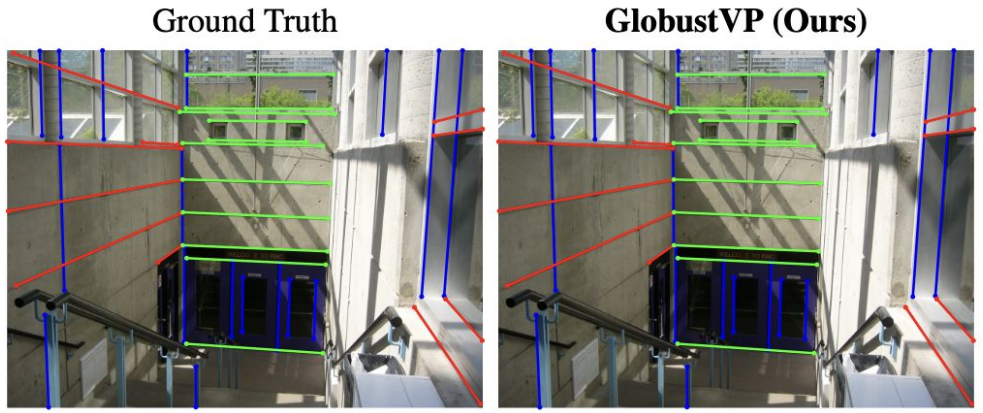
- 画像として与えられた環境から目的地までの案内(visual navigation)を、生成動画や環境中の方向指示により実現。この問題設定に対し、conditional diffusion transformer(CDiT)を1Bパラメータまでスケーリング、与えられた初期位置と目的地の画像をベースに、実世界環境を補完するように画像生成を実施しつつも鳥瞰図や初期位置画像中に矢印を描画する。従来のvisual navigationとは異なり、初期位置のみの入力から、目的地までの動画生成ができること、環境中や初期位置画像中に矢印としても案内ができる。World ModelとしてDiTを学習し、実世界を理解しつつ、初めての環境でもある程度道案内ができることから、研究としてのvisual navigationやロボット・自動運転などアプリケーションとしての幅が広がることが考えられる。



CVPR 2025 の動向・気付き (21/181)

Best paper候補論文 9: Convex Relaxation for Robust Vanishing Point Estimation in Manhattan World

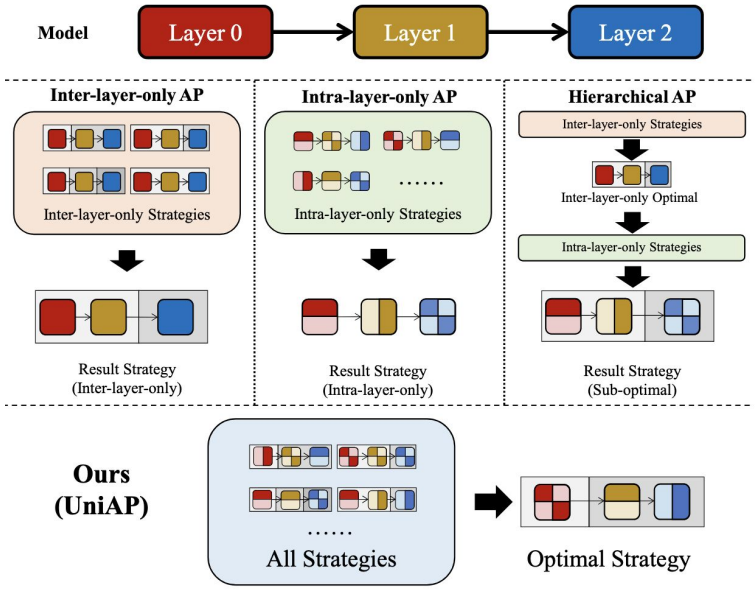
- マンハッタンワールドにおける消失点推定(Vanishing Points; VP)を、線分–VP対応と位置推定を同時最適化する問題を設定。従来、マンハッタンワールドにおける消失点推定の同定問題は、SLAM/カメラ校正/3次元再構成などの重要技術であったが、線分–VP対応と位置推定では局所解が一意に定まらないという課題があった。本論文では、Truncated multi–selection error (loss)を導入、線分–VP対応とVP位置を soft に同時最適化する。このlossに従い、定式化をquadratically constrained quadratic programming (QCQP) により実施、semidefinite programming (SDP; 半正定値計画問題)への凸緩和(convex relaxation)を初めて導入した。提案手法であるGlobustVP反復法で各消失点と対応線分をグローバル最適に検出した後に、Manhattan Post–Refinementで直交性を保証。実験では、合成画像・実画像を用いたデータセットで従来手法を上回る高精度、ロバスト性、処理速度を示した。

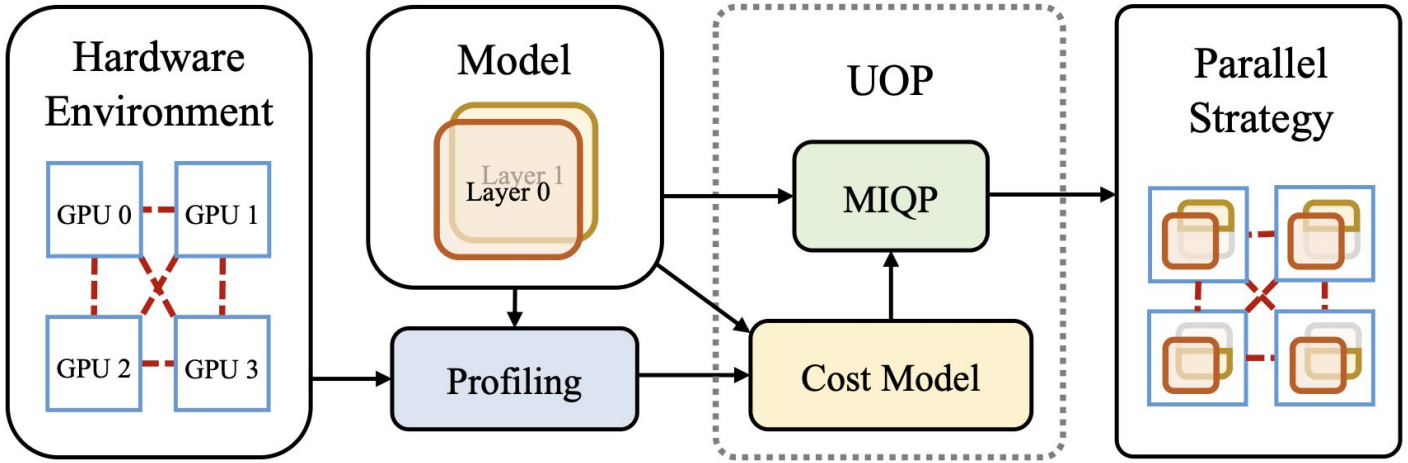


CVPR 2025 の動向・気付き (22/181)

Best paper候補論文 10: UniAP: Unifying Inter- and Intra-Layer Automatic Parallelism by Mixed Integer Quadratic Programming

- 分散学習 / 並列計算に関する研究。もともとAIのモデルを並列学習する際には、層内、層間を別々に最適化しつつ順次統合するために全体最適化する手法がなかったが、この問題を解決したことが大きな貢献。提案手法であるUniAPは、MIQPと呼ばれる最適化手法を用いて、層内、層間を同時に最適化することに成功。最適化に必要なコストも最大100倍程度削減、BERT, ViT, SwinTransformer, LLaMAなど画像や言語モデルなど多様なモデルの最適化ができる。

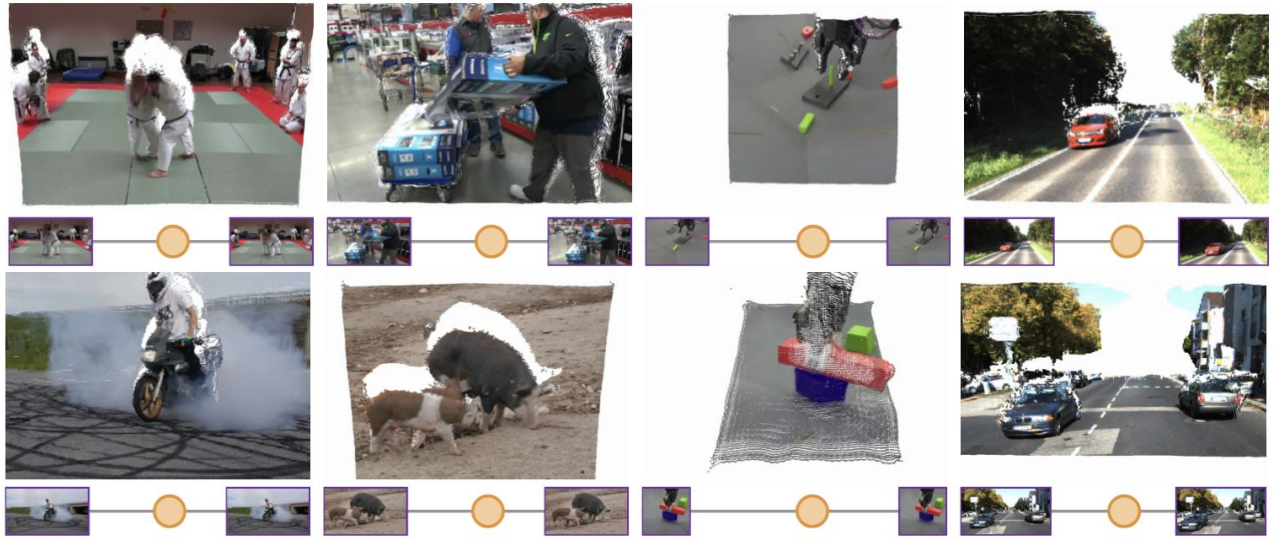




CVPR 2025 の動向・気付き (23/181)

Best paper候補論文 11: Zero-Shot Monocular Scene Flow Estimation in the Wild

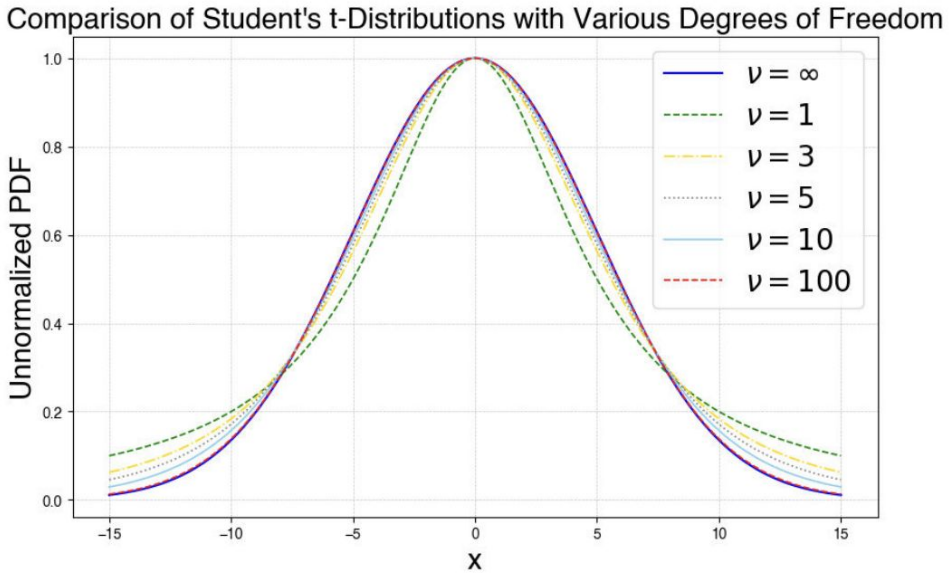
- Scene Flow（SF）は「距離画像推定」と「オプティカルフロー」を同時に解いて高度に統合する必要があり、なかなか解決できない問題であった。データセットは用意されるものの、比較的小規模や特定問題に対して解いているに過ぎなかった。本稿では、合成データによる大規模SFデータセットを構築、3D point mapや3D motion vectorの導入により、zero-shot scene flowの導入を実現した。学習に含まれていないin-the-wildな環境においても、SF推定することができ、SF推定の可能性を大きく押し上げたと言える。



CVPR 2025 の動向・気付き (24/181)

Best paper候補論文 12: 3D Student Splatting and Scooping

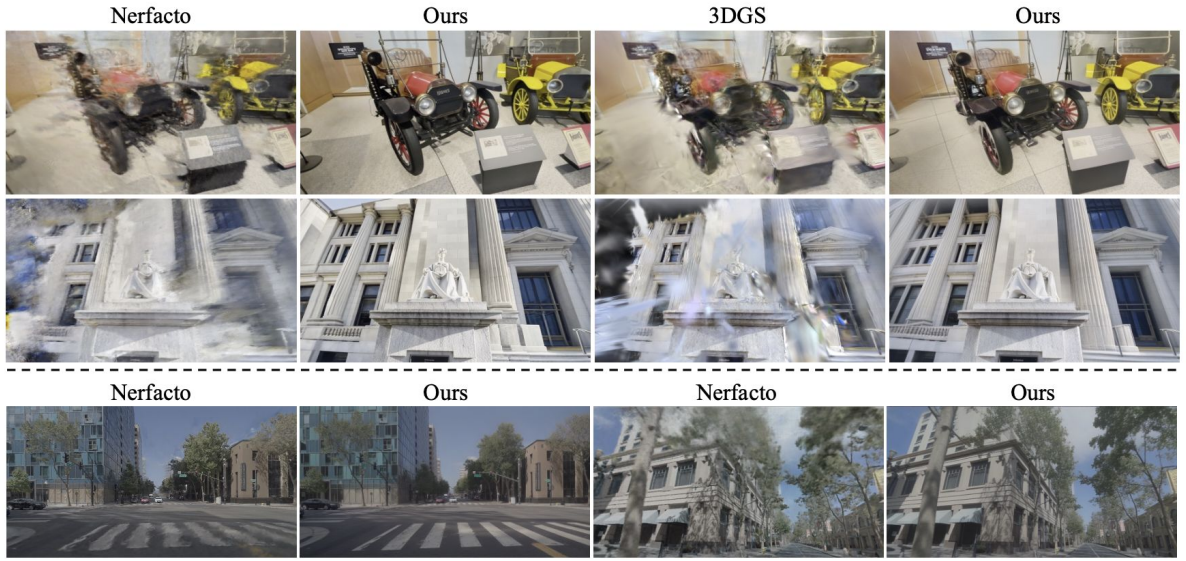
- 3DGSの問題設定において、ガウシアン分布の代わりにStudent’s t分布のsplattingとscoopingの組み合わせにより空間表現を強化・拡張して高い表現力を持ちパラメータ効率も改善。最適化には確率的勾配ではなくサンプリングベースの手法を取り入れている。正の密度空間だけでなく負の密度空間もとらえるようにすることで少ないパラメータで効率的に表現可能。細かい部分まで再構成に成功。同一空間の再構成において、既存手法より82%表現に必要な分布のcomponent数を減らすことに成功。



CVPR 2025 の動向・気付き (25/181)

Best paper候補論文 13: DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models

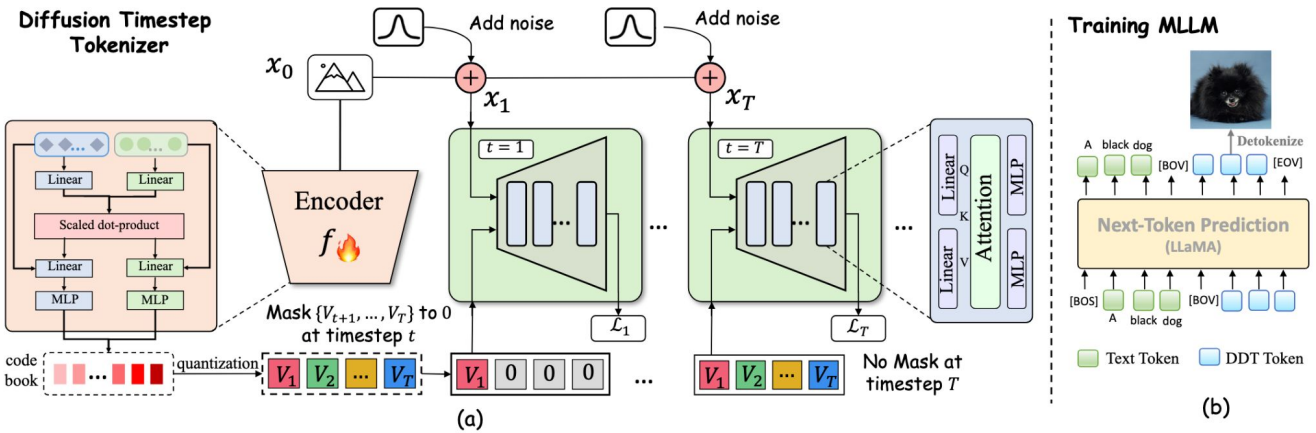
- NeRFや3DGSで生成されたアーティファクト等のノイズを除去する拡散モデルを提案。さらに3D再構成結果からある視点を切り取って取得された画像に対して生成ミスを修正。その結果を再度3Dに起こして繰り返しアップデートして綺麗な3Dを完成させる(DIFIX)。推論時においてはその拡散モデルをそのまま使って、リアルタイムに新規視点のアーティファクトを除去した綺麗な画像を生成できる(DIFIX3D+)。Refinementでの競合はないが、単一モデルで3D再構成の修正とリアルタイムのさらなる改善をできる。

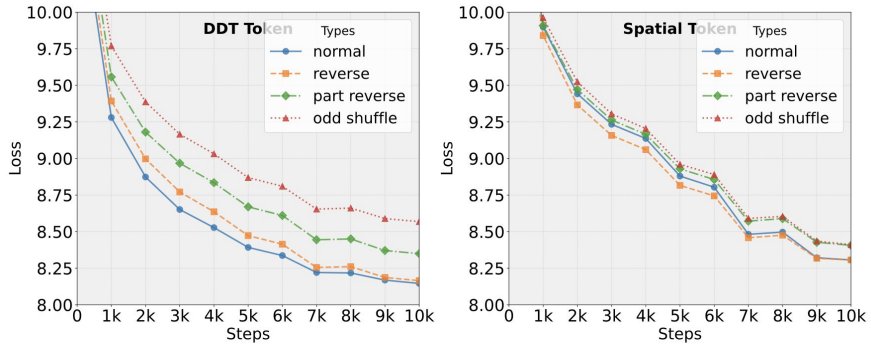


CVPR 2025 の動向・気付き (26/181)

Best paper候補論文 14: Generative Multimodal Pretraining with Discrete Diffusion Time-step Tokens

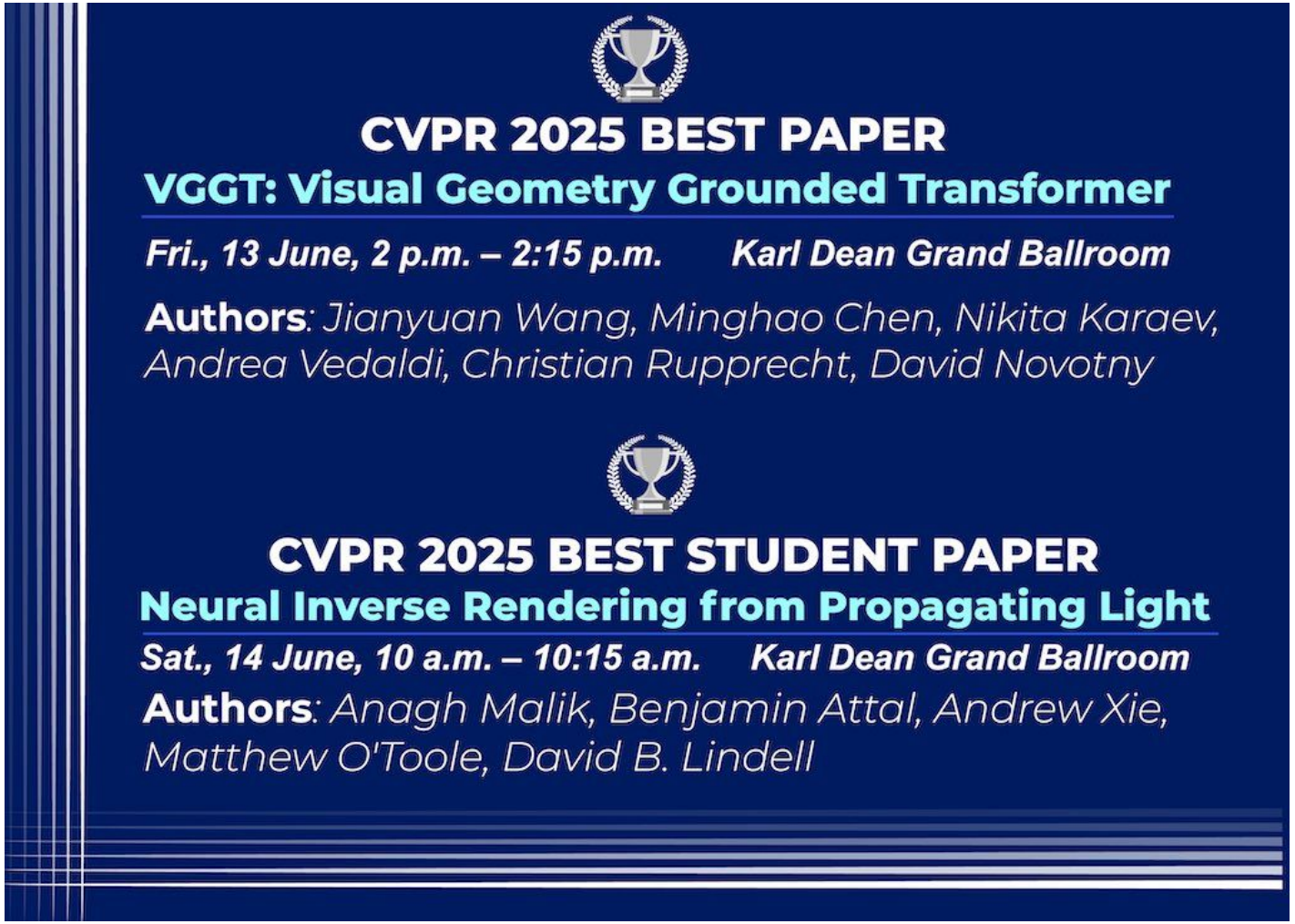
- 画像と言語を統合して理解や生成ができるAIモデルを実現するために、DiffusionとLLMを自然に融合させる手法を提案。既存手法は画像を単純に画像パッチに分割して並べる空間的トークンを使っていたが、言語のような再帰的構造がないため、LLMにとって扱いにくいという課題があった。視覚情報を「意味のあるトークン列」に変換する方法としてdiffusionで画像に少しずつノイズを加えていくプロセスを利用し、「ノイズで失われた情報を補うトークン」を順に作り出すことで、言語のように再帰的構造を扱える画像トークン列を得る。これによりtext2img、画像編集、画像キャプションにおいて高精度を達成。





CVPR 2025 の動向・気付き (27/181)

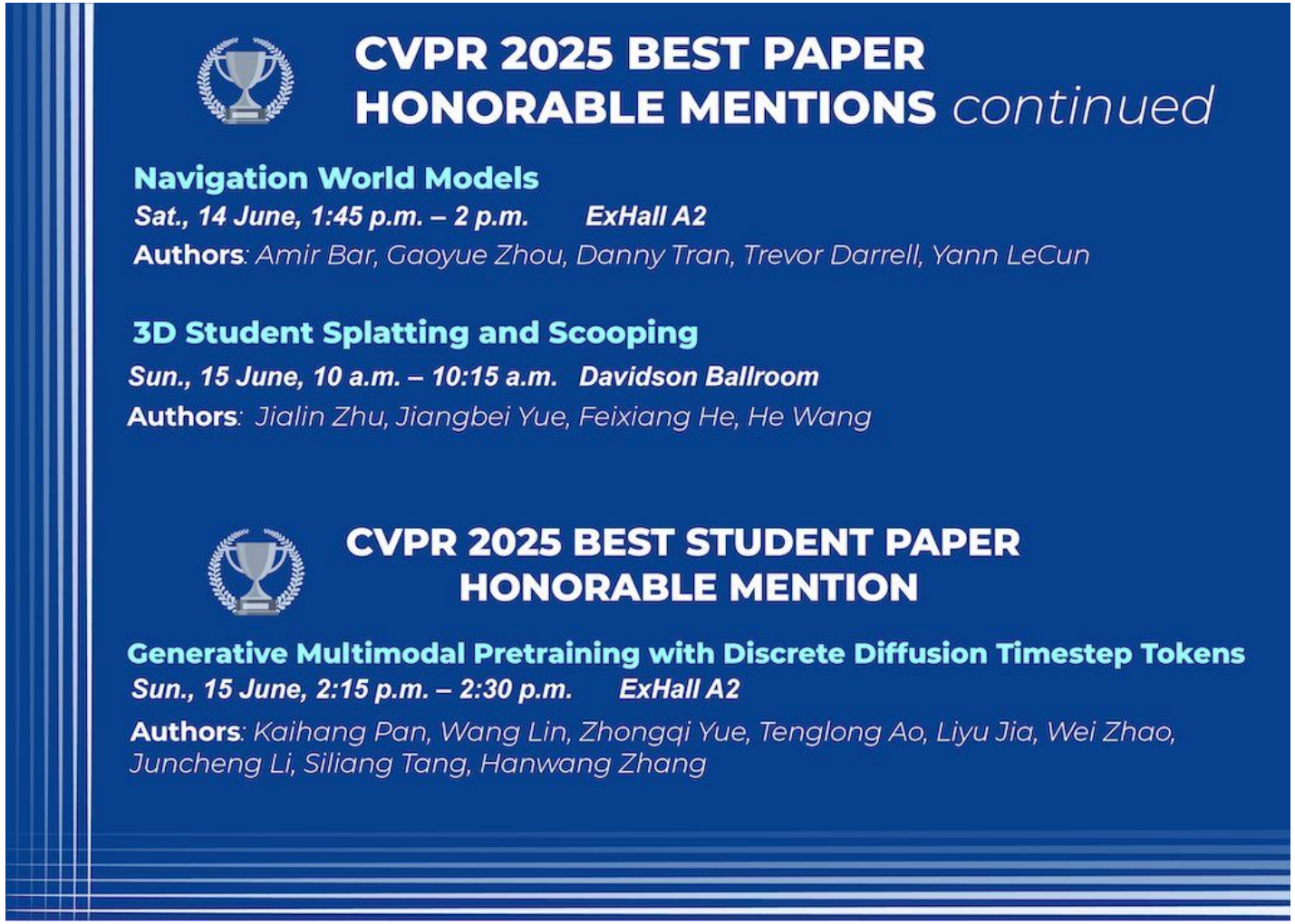
Award Ceremonyより



リファレンス：CVPR X ウェブサイト

CVPR 2025 の動向・気付き (28/181)

Award Ceremonyより



リファレンス：CVPR X ウェブサイト

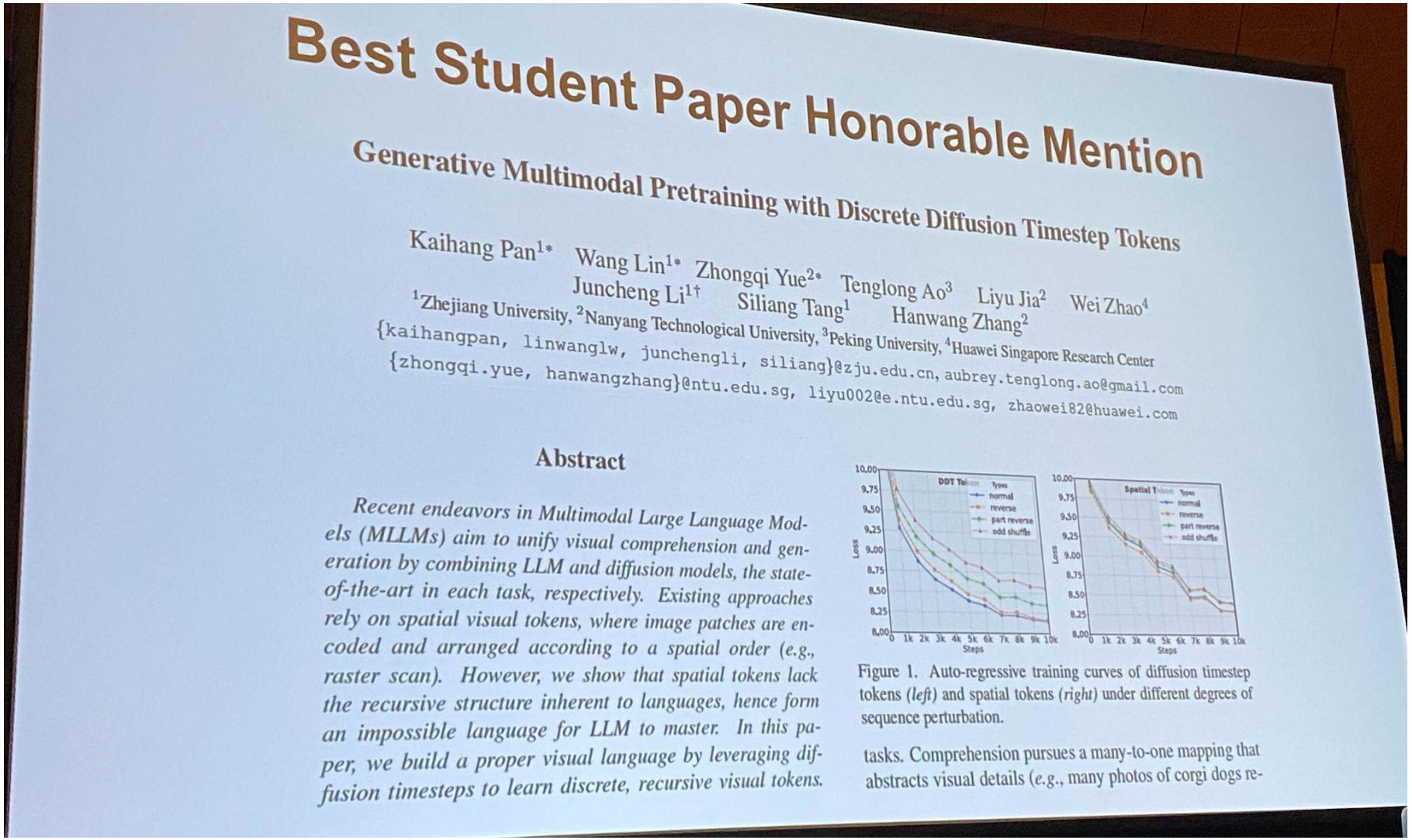
CVPR 2025 の動向・気付き (29/181)

Award Ceremonyより

リファレンス：CVPR X ウェブサイト

CVPR 2025 の動向・気付き (30/181)

Award Ceremonyより



Best Student Paper Honorable Mention

Generative Multimodal Pretraining with Discrete Diffusion Timestep Tokens

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

Abstract

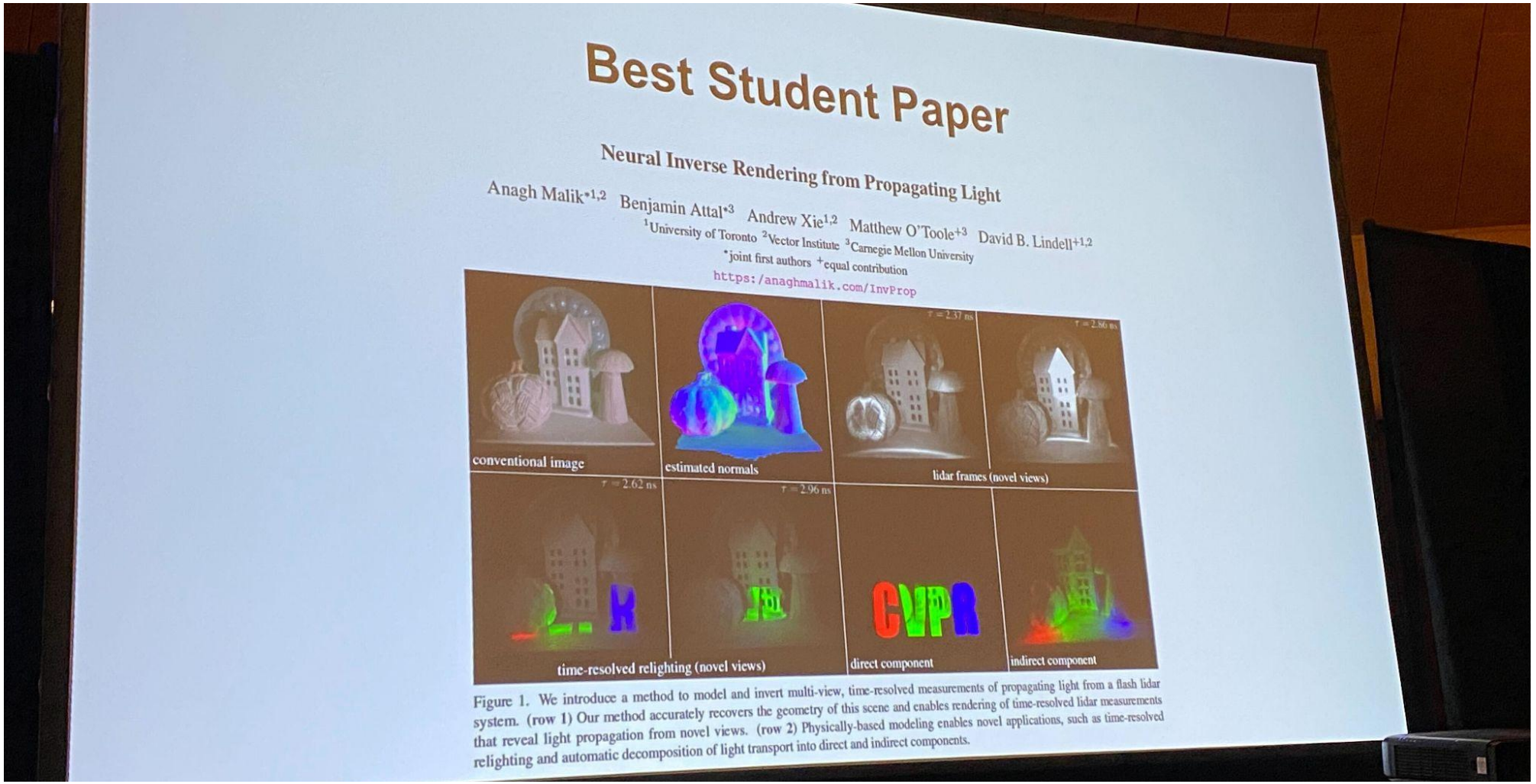
Recent endeavors in Multimodal Large Language Models (MLLMs) aim to unify visual comprehension and generation by combining LLM and diffusion models, the state-of-the-art in each task, respectively. Existing approaches rely on spatial visual tokens, where image patches are encoded and arranged according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

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

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

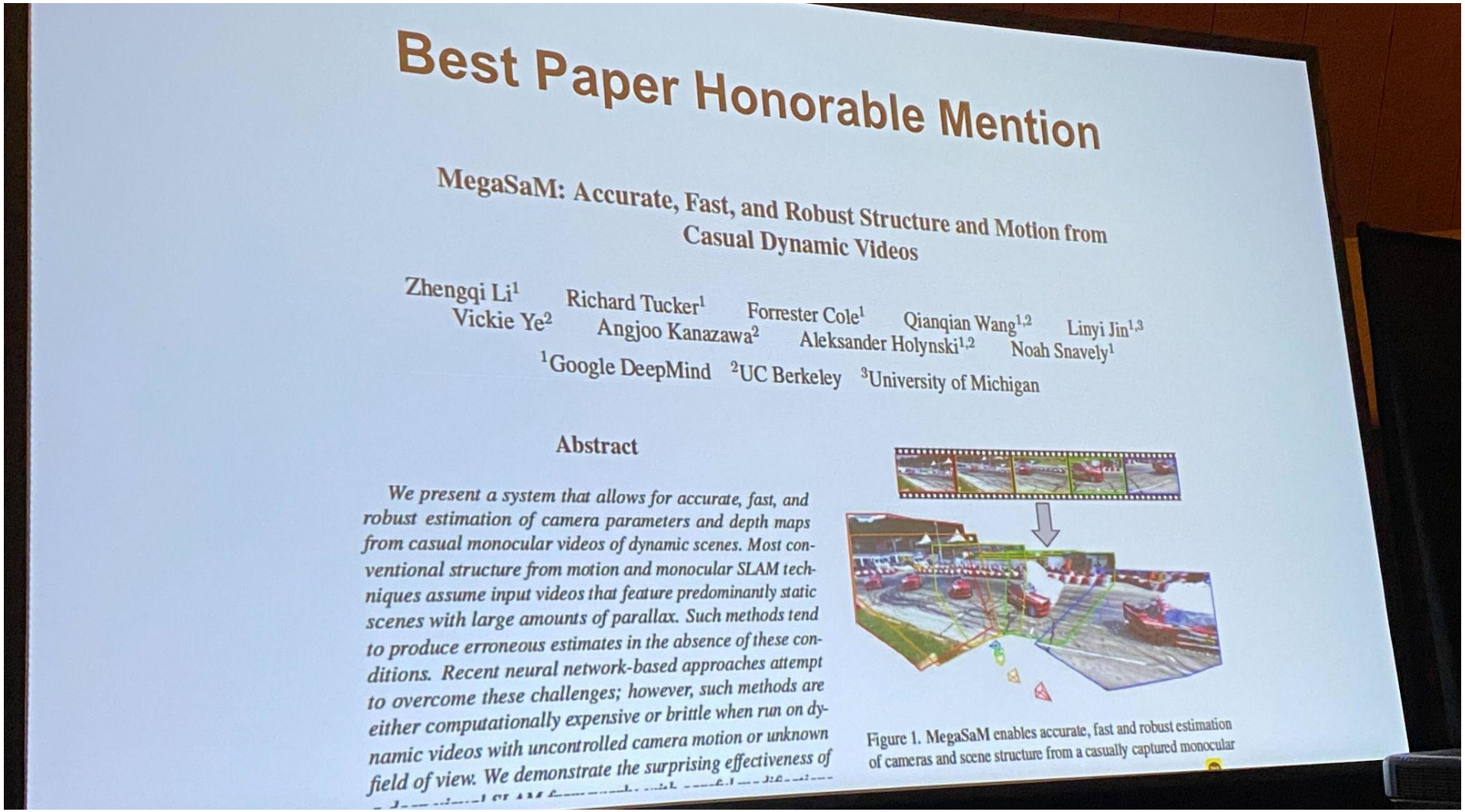
CVPR 2025 の動向・気付き (31/181)

Award Ceremonyより



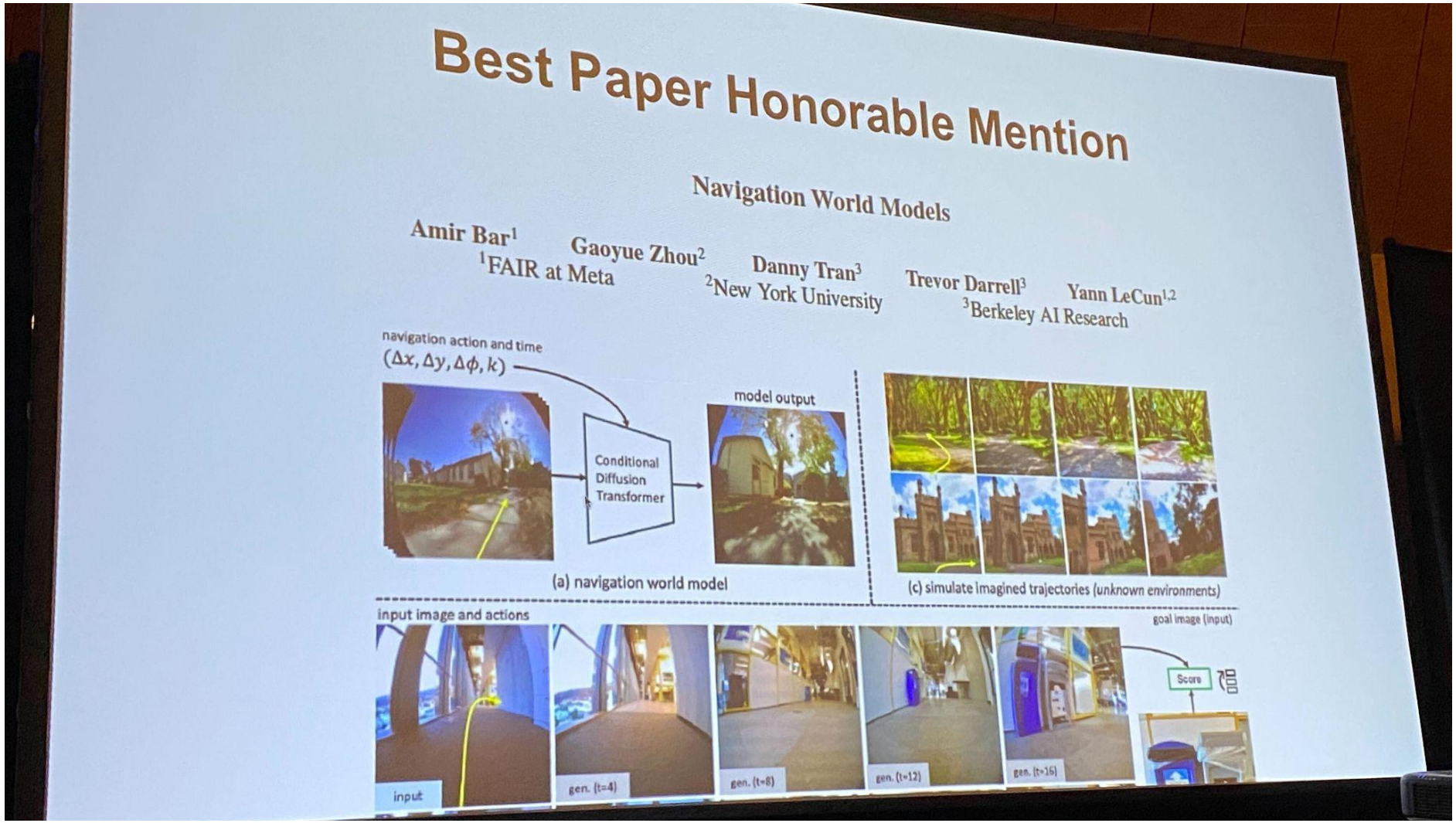
CVPR 2025 の動向・気付き (32/181)

Award Ceremonyより



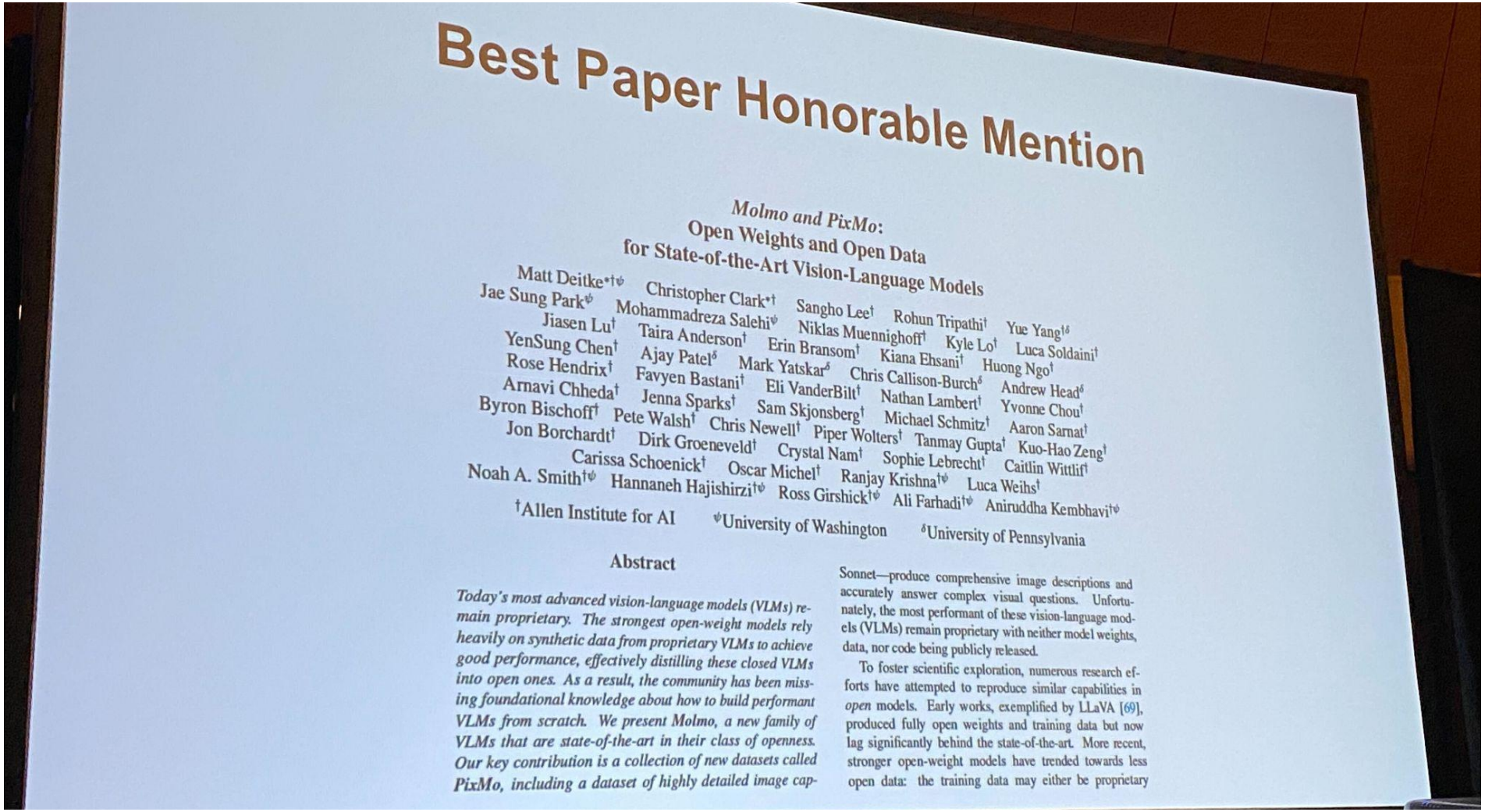
CVPR 2025 の動向・気付き (33/181)

Award Ceremonyより



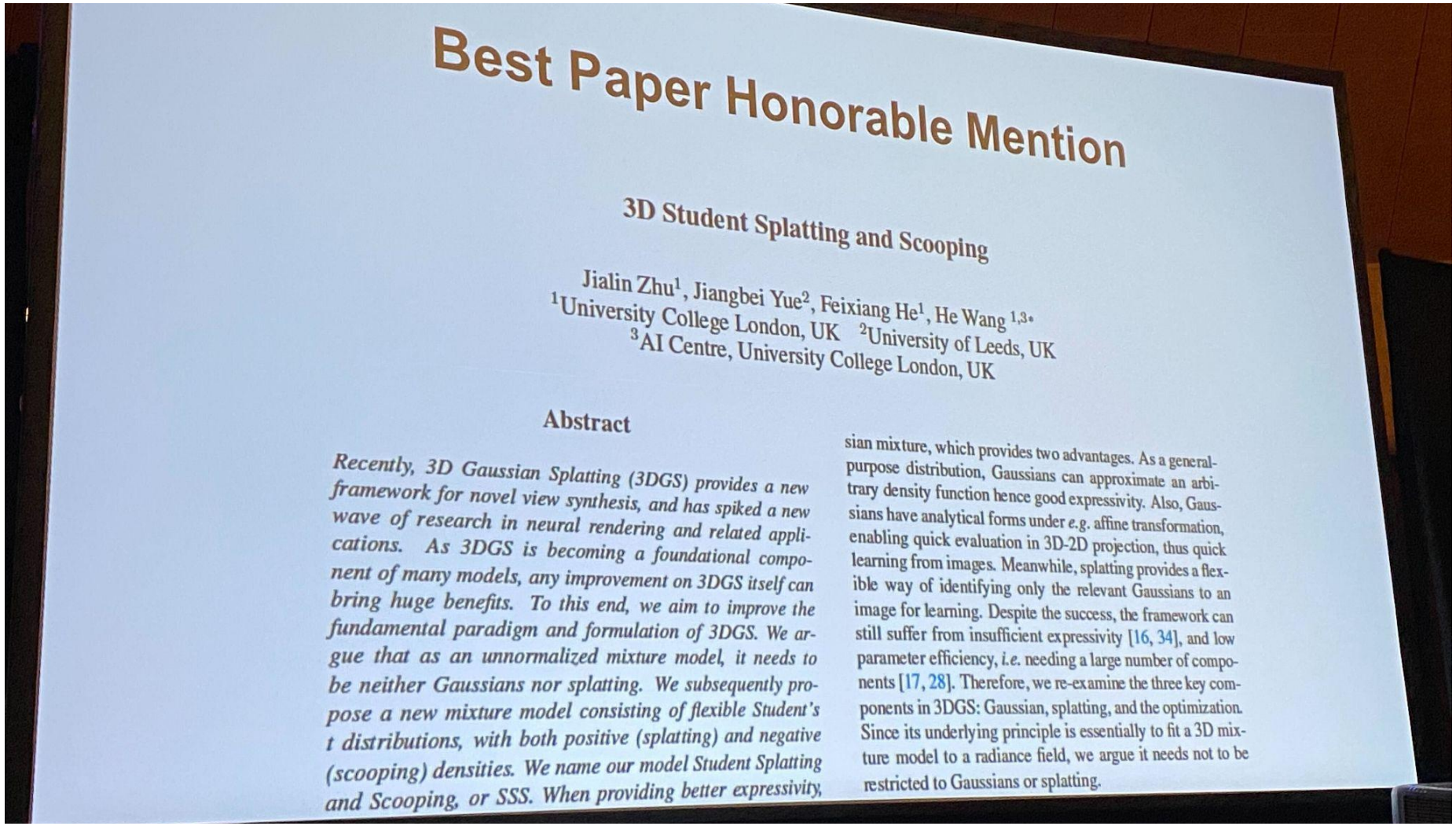
CVPR 2025 の動向・気付き (34/181)

Award Ceremonyより



CVPR 2025 の動向・気付き (35/181)

Award Ceremonyより



CVPR 2025 の動向・気付き (36/181)

Award Ceremonyより

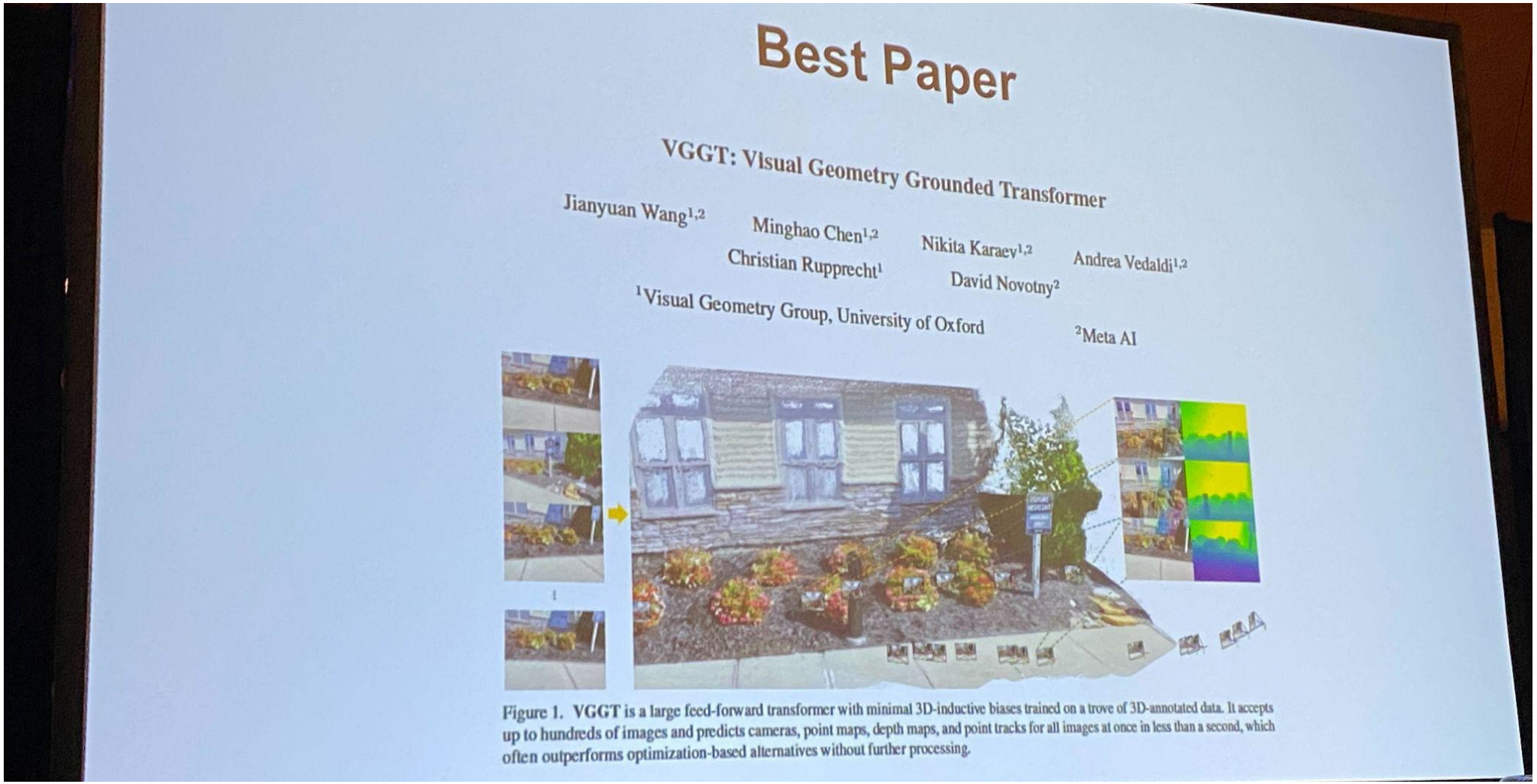
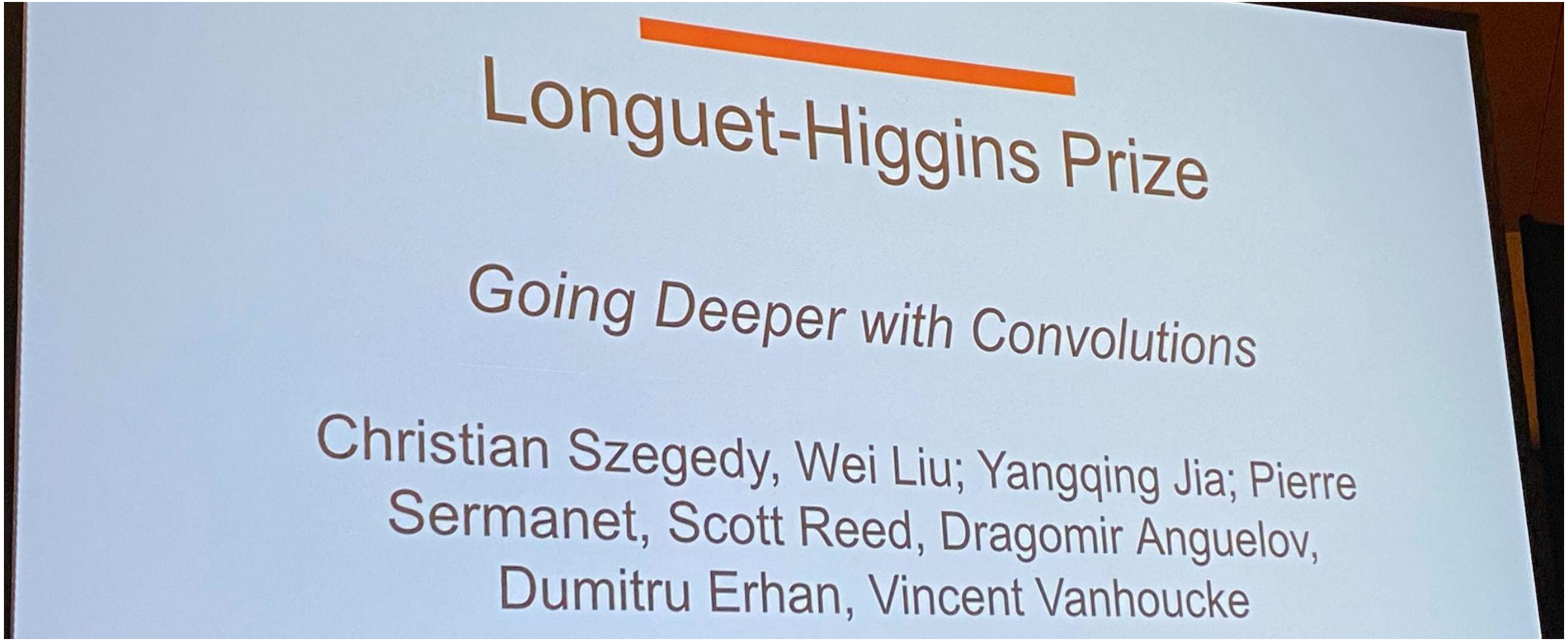


Figure 1. **VGGT** is a large feed-forward transformer with minimal 3D-inductive biases trained on a trove of 3D-annotated data. It accepts up to hundreds of images and predicts cameras, point maps, depth maps, and point tracks for all images at once in less than a second, which often outperforms optimization-based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

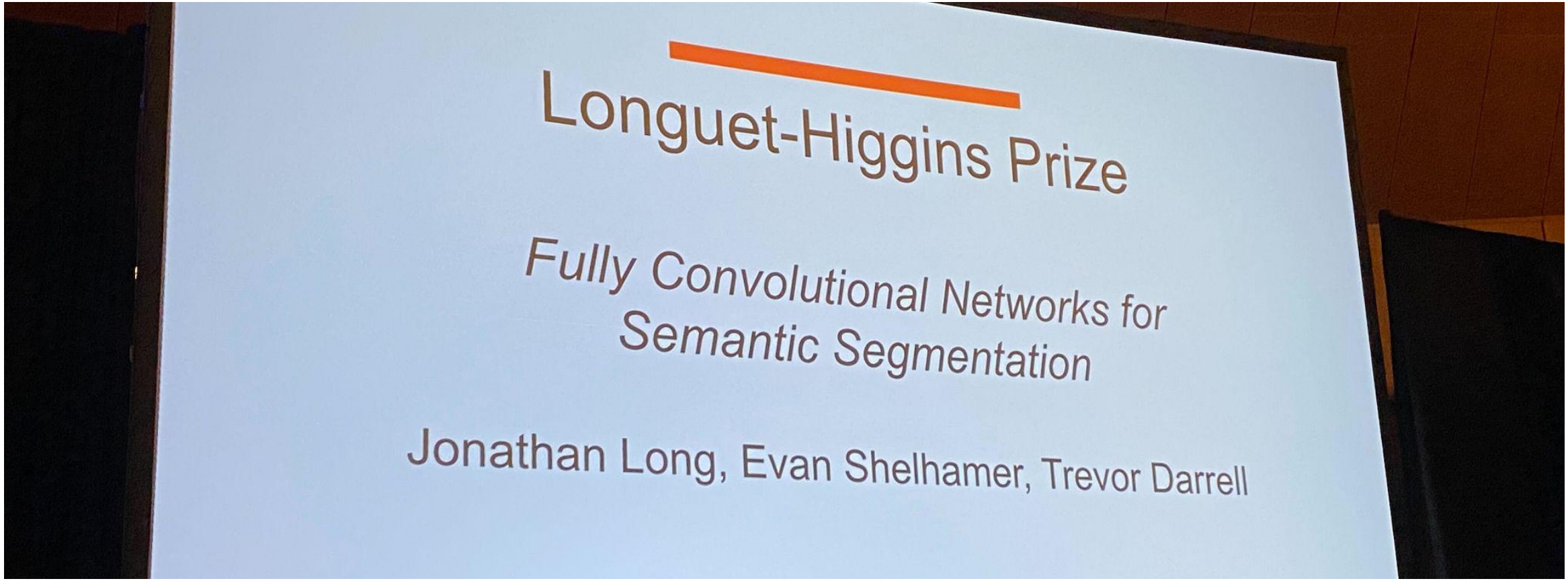
CVPR 2025 の動向・気付き (37/181)

Award Ceremonyより



CVPR 2025 の動向・気付き (38/181)

Award Ceremonyより



CVPR 2025 の動向・気付き (39/181)

Award Ceremonyより

CVPR 2025 の動向・気付き (40/181)

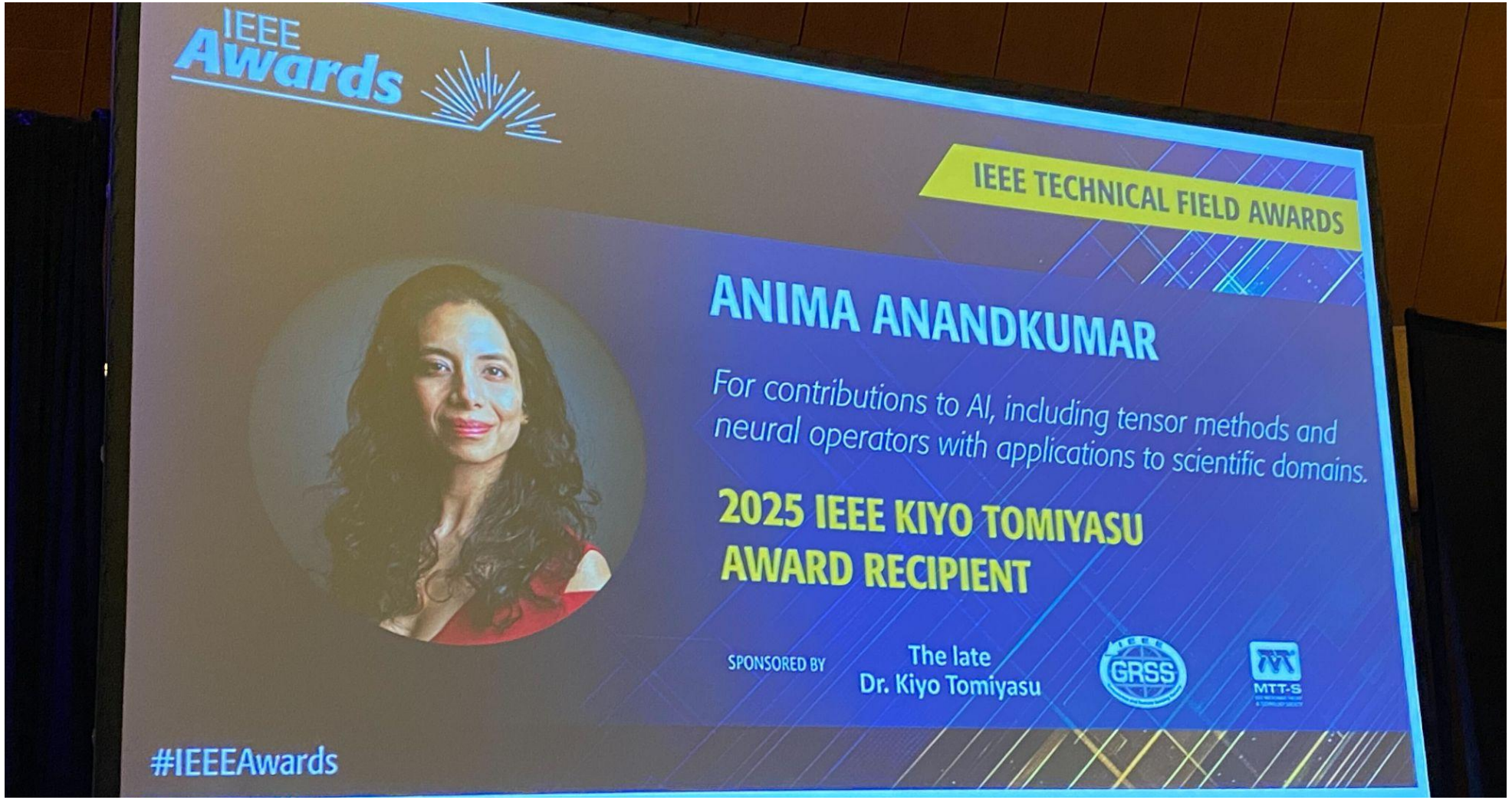
Award Ceremonyより

CVPR 2025 の動向・気付き (41/181)

Award Ceremonyより

CVPR 2025 の動向・気付き (42/181)

Award Ceremonyより



CVPR 2025 の動向・気付き (43/181)

LIMIT.Lab における非公式のbest paper予測

- CVPR 2025 best paper 候補論文の発表前、有志メンバーによって組織
- 正式な審査プロセスとは別に、国際研究コミュニティにおいて、どの論文が受賞するかを予測
- ノミネートされた論文を注意深く読み、単純な予測にとどまらず、その新規性、影響、将来の可能性を評価することで、論文を査読・評価するための理解力を深化
- この取り組みは、議論や意見交換を通じ、我々の視野を広げ、研究を評価する基準を洗練させるのにも役立つ

CVPR 2025 の動向・気付き (44/181)

LIMIT.Lab における非公式のbest paper予測

- 国際研究コミュニティ内の選考プロセス
  - (非公式の)CVPR 2025 Award Committee の結成のためSlackで呼びかけ、専用Slackチャンネルを作成
  - 全論文について、少なくとも1人の委員がそれを徹底して読み込み要約を作成、メンバー全員がこれらの要約をレビューし、判断を下す前に少なくともすべての論文に関して目を通して基本的な理解があることを確認
  - 各メンバーがそれぞれの受賞リストを共有、委員会で審議して微調整
  - 受賞リストを作成した後、選ばれた論文は最終決定を下す前に時間をおいて再審査
- 選択基準
  - 評価はCVPR規制に従い、この分野への貢献度とその将来に大きな影響を与える可能性に焦点を当てた
  - もちろん、著者の知名度や既存の引用数などの要素は考慮されない

CVPR 2025 の動向・気付き (45/181)

LIMIT.Lab における非公式のbest paper予測

- 厳選された各賞論文
  - 合計: CVPR 2025から9件の論文 リスト (CVPR 2024で受賞した論文10件)
  - Best Student Paper Honorable Mention（2）
  - Best Student Paper Award（2）
  - Best Paper Honorable Mention（3）
  - Best Paper Award（2）
  - 注意：Best paper 候補リスト（ https://cvpr.thecvf.com/virtual/2025/events/AwardCandidates2025 ）とCVPR awards のリストが少し違う論文を示していたことに注意

CVPR 2025 の動向・気付き (46/181)

LIMIT.Lab における非公式のbest paper予測

- 厳選された各賞論文
  - 合計: CVPR 2025から9件の論文 リスト (CVPR 2024で受賞した論文10件)
    - Best Student Paper Honorable Mention (2)
      - DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models
      - 3D Student Splatting and Scooping
    - Best Student Paper Award (2)
      - Zero-Shot Monocular Scene Flow Estimation in the Wild
      - Convex Relaxation for Robust Vanishing Point Estimation in Manhattan World
    - Best Paper Honorable Mention (3)
      - Navigation World Models
      - MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos
      - Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models
    - Best Paper Award (2)
      - Generative Multimodal Pretraining with Discrete Diffusion Time-step Tokens
      - VGGT: Visual Geometry Grounded Transformer

CVPR 2025 の動向・気付き (47/181)

LIMIT.Lab における非公式のbest paper予測

- 厳選された各賞論文 と現実
  - 合計: CVPR 2025から9件 → 7 の論文 リスト (CVPR 2024で受賞した論文10件)
  - Total: 9 → 7 papers from the CVPR 2024 list (10 awarded papers in CVPR 2024)
  - Best Student Paper Honorable Mention (2→ 1)
    - ❌DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models
    - ✅ 3D Student Splatting and Scooping → But, it's Best Paper Honorable Mention
  - Best Student Paper Award (2→ 1)
    - ❌Zero-Shot Monocular Scene Flow Estimation in the Wild
    - ❌Convex Relaxation for Robust Vanishing Point Estimation in Manhattan World
  - Best Paper Honorable Mention (3→ 4)
    - ✅ Navigation World Models
    - ✅ MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos
    - ✅ Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models
  - Best Paper Award (2→ 1)
    - ✅ Generative Multimodal Pretraining with Discrete Diffusion Time-step Tokens → But, it's Best Student Paper Honorable Mention
    - ✅ VGGT: Visual Geometry Grounded Transformer

CVPR 2025 の動向・気付き (48/181)

メタな洞察とBP選択の議論ポイント

- 論文を全く読まずタイトルだけからの予測でもかなり正確で7つ挙げたうち6つは最終的な委員の予測と一致した　→　ある程度貢献度やインパクトを明確に伝えるタイトルを作成することが重要であることを示唆
- 狙っても中々 CVPR award は取れないが、狙わないと絶対に手に入らない！
- VGGTは、広範囲にわたるメソッドの探索やパラメーターのチューニング、VGGSfM や CoTracker などのコンポーネントを組み合わせたラボ内での密接な研究連携など、長年にわたる努力の積み重ねの成果

CVPR 2025 の動向・気付き (49/181)

メタな洞察とBP選択の議論ポイント

- 今年も 3D vision が大きな注目ポイント
- 3D vision → 再構成は高速・高品質な結果に！
- VLM（視覚言語モデル）→ オープンソースの勢い、柔軟性向上！
- ロボティクス → 世界モデルの画期的な応用！

CVPR 2025 の動向・気付き (50/181)

メタな洞察とBP選択の議論ポイント

- CVPR Best paper candidateの研究はどれも、SF小説を読んでいるようなワクワク感があった
- 単なる技術向上ではなく、未来の道具のプレゼンみたいに感じた
- 「こんなすごいことができるなら、自分ならこういうものを作る」みたいな、インスピレーションを与えてくれる研究
- 綿密な実験によりアイデアの実現性を確実に証明
- SFもリアリティがある方が面白いのと同じで、丁寧な実験が面白い研究に繋がると感じた

CVPR 2025 の動向・気付き (51/181)

Backstory of best paper award – VGGT （Jianyuan Wangに裏話を共有してもらえました！ありがとうございます🙏）

- VGGSfMから研究がスタート
  - VGGSfMは複雑すぎて新規研究参入が難しい
  - 従来のSfMパイプラインに従う、つまり“images – correspondences – camera poses & points – bundle adjustment – refined camera poses & points”というプロセス
- VGGT著者陣は DUSt3Rに着目
  - 3D再構成においても大規模なデータ駆動型パラダイムに移行
  - そしてVGGTへ – Transformerのフォワード実行のみで誰でも使用できるように
- （研究室としての）VGGは同時期に正確な特徴点追跡器 CoTracker も技術提案

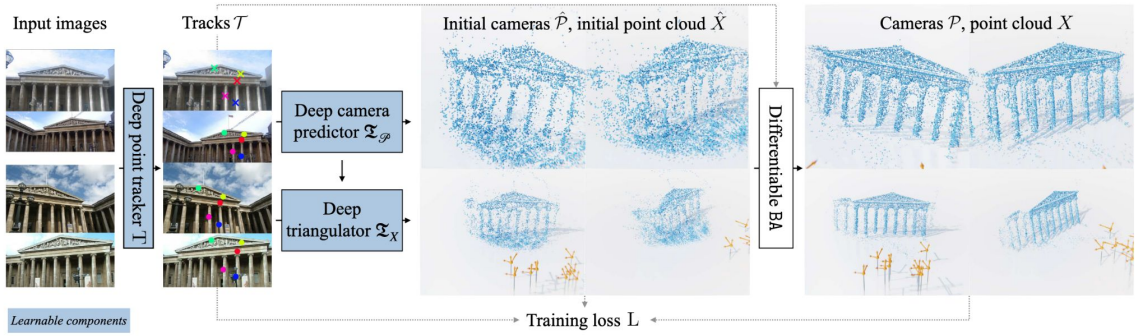


Figure 2. **Overview of VGGSfM.** Our method extracts 2D tracks from input images, reconstructs cameras using image and track features, initializes a point cloud based on these tracks and camera parameters, and applies a bundle adjustment layer for reconstruction refinement. The whole framework is fully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

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

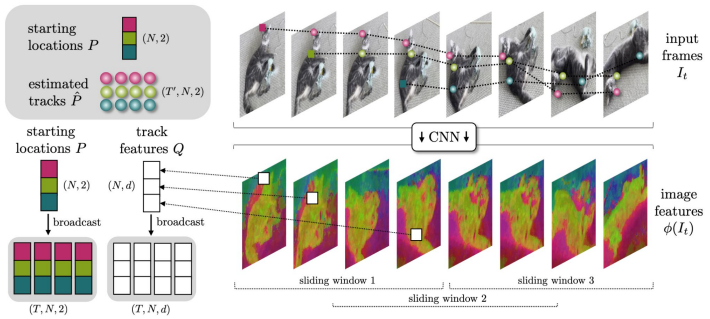


Fig. 3: CoTracker architecture. We compute convolutional features $\phi(I_t)$ for every frame and process them with sliding windows. To initialize track features Q, we bilinearly sample from $\phi(I_t)$ with starting point locations P. Locations P also serve to initialize estimated tracks $\hat{P}$.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

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

CVPR 2025 の動向・気付き (52/181)

Backstory of best paper award – VGGT （Jianyuan Wangに裏話を共有してもらえました！ありがとうございます🙏）

- 英国🇬🇧 から米国🇺🇸 へのフライト
  - 主著のJianyuanは前日までCVPR現地入りできるか不明な状態
  - 出発直前にパスポート入手、そのままロンドンの空港へ
  - Award ceremony直前の深夜にNashville着
- CVPR award ceremony
  - Program chairsから受賞についての事前通知はなく、今年のCVPRは Oscar のようだった
  - Best paper award 一番最後に読み上げ（つまり、honorable mention発表までに読みあげられていなかったのでBP発表直前は “all or nothing” の状態）
  - Jianyuan からのコメント:

    I was extremely nervous. I still don’t know how to fully describe the feeling. Even now, it all feels kind of unreal.（ものすごく緊張しました。その気持ちをどう表現したらいいのかまだ分かりません。今でも、全て現実には感じられないです。）

CVPR 2025 の動向・気付き (53/181)

Backstory of best paper award – VGGT

- VGGNet から VGGTransformer への華麗なる切り替え！
  - VGGNet（ICLR 2015）
    - Convolutional Neural Networks（CNNs）
    - 画像認識のbackbone networkとしての利用
  - VGGTransformer / VGGT（CVPR 2025）– VGGNet の10 年後
    - Transformer
    - 単一 Transformer モデルで3D再構成が完結
    - 3D vision における backbone networkとしての期待
  - Visual Geometry Group（VGG）の大まかな歴史
    - 2010年代以前：3D geometry が主なフォーカスだが、画像認識に関する研究も行われてきた
    - 2010年代以降：Deep learning が主なフォーカスだが、3D vision研究も行われてきた
    - VGGT: 3D geometry x deep learning の強力なクロスポイント
    - 2025年以降？: 3D vision x deep learning は誰でも参画できる状態になるか？

CVPR 2025 の動向・気付き (54/181)

医療 × CVの動向

a. 医療分野のためのVision and Language
   i. Medical VQA with noisy labels and diffusion [Guo+, CVPR 2025]
   ii. VLMs alignment with CoOP [Koleilat+, CVPR 2025]
   iii. VQA with visual reference [Chen+, CVPR 2025]
   iv. VLM with soft label [Ko+, CVPR 2025]
   v. Pre-training of VLM [Ziyang+, CVPR2025]

b. 医療向けSemi-supervised learning
   i. 以前から注目されていたが、未だに残る問題
   ii. 医療の基盤モデルがあっても、ラベル付きデータ不足が解決されていないことを示す？
   iii. Deal with annotation ambiguity [Kumari+, CVPR 2025]
   iv. Depth guided segmentation [Li+, CVPR 2025]
   v. Find overconfidence prediction of foundation model [Ma+, CVPR 2025]
   vi. Unsupervised prompting for SAM by DPO-inspired loss [Konwer+, CVPR 2025]
   vii. Uncertainty-aware consistency and contrastive [Assefa+, CVPR 2025]

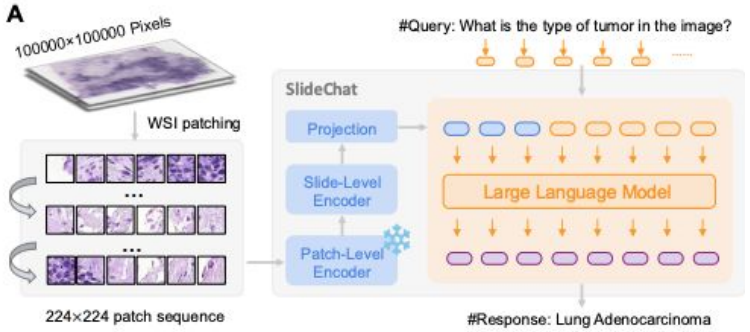
CVPR 2025 の動向・気付き (55/181)

病理学 × CV研究の動向

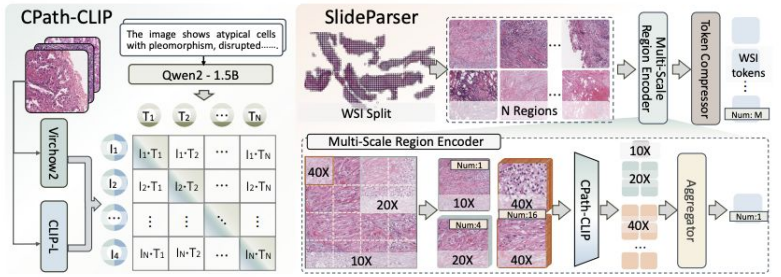
a. 病理学のためのVision and Language

i. Patch-level VLLM -> multi-resolution (Patch-level、Slide-level含む) VLLM

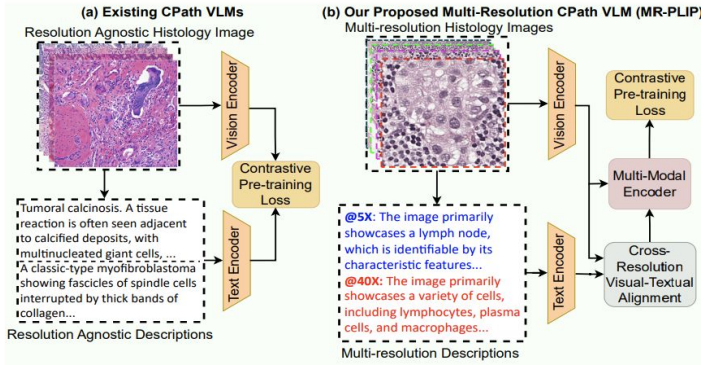
[Chen+、CVPR 2025]



[Sun+、CVPR 2025]



[Albastaki+、CVPR 2025]



b. スライド画像全体のMambaベースモデル

i. MambaにPatch-levelの情報を集約するための手法

[Chang+、CVPR 2025]

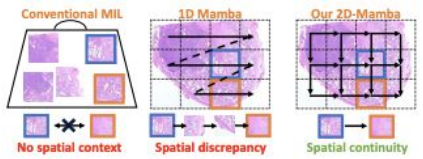
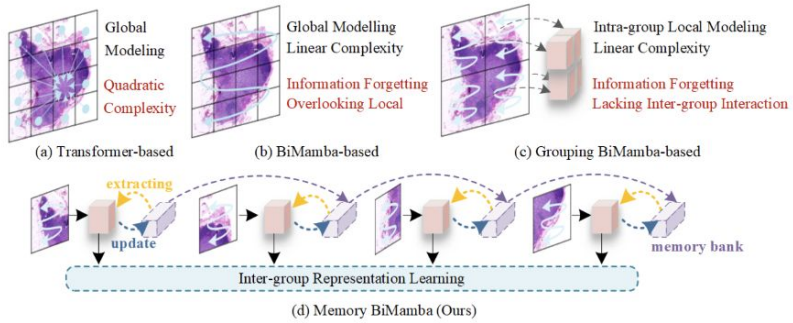


Figure 1. **Left: Conventional MIL Bagging** of patches adopts *no spatial context*. **Middle: 1D Mamba-based methods** flatten a WSI into a 1D sequence and lose the 2D structure. The adjacent blue and orange patches are far away in the sequence. We call this "*spatial discrepancy*". **Right: 2DMamba**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

[Zheng+、CVPR 2025]



CVPR 2025 の動向・気付き (56/181)

Direct Preference optimization (DPO) for Vision and Language

SymDPOによるin-context learningの促進
[Jia+、cvpr 2025]

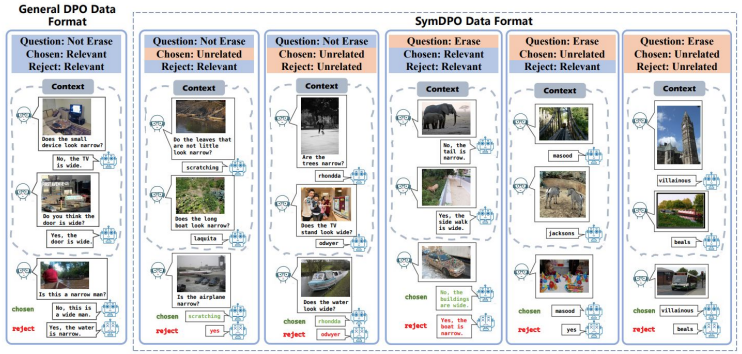


Figure 2. Comparison of General DPO and SymDPO Formats: General DPO relies solely on standard text for Questions, Answers, Chosen, and Rejected Answers, focusing on text-based training. In contrast, SymDPO replaces textual Answers with symbolized text to boost multimodal understanding, requiring models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

Noise-aware preference optimizationによるVLMのバイアス解消
[Zhang+、cvpr 2025]

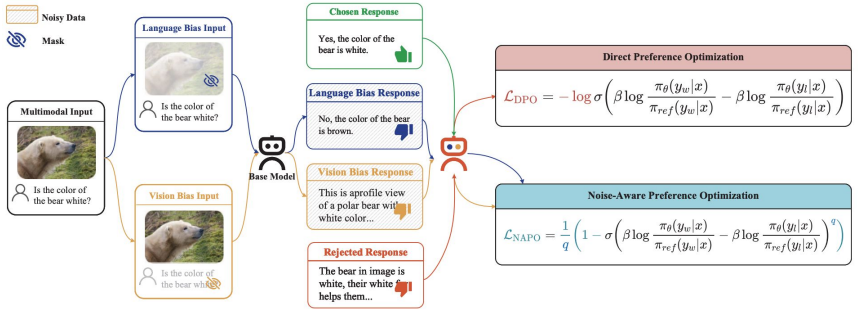


Figure 2. **Method details.** First, biased responses are constructed by using masking to guide the model toward over-relying on prompts and generating responses based on the base model. Next, NaPO is applied for noise-robust preference optimization to counteract noise in automatically constructed data,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

タスク固有のトークンの導入によるタスク優先度最適化によるVLMの改善
[Yan+、cvpr 2025]

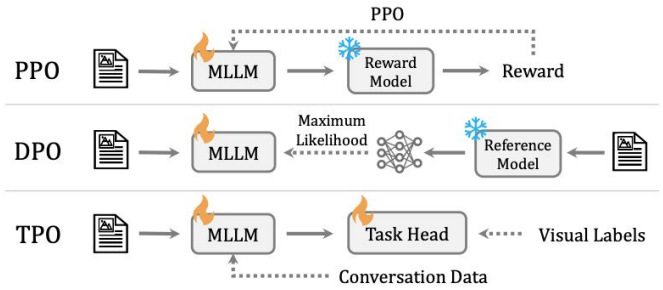


Figure 2. **Comparison of Learning Method.** A solid line indicates data flow, and a dotted line represents feedback. ❄ and 🔥 denote modules that are frozen and unfrozen.

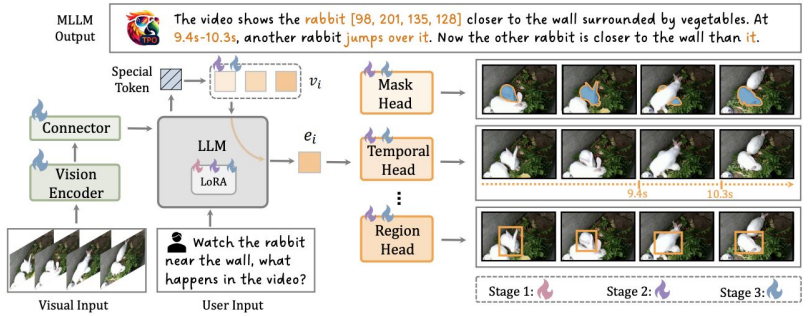


Figure 3. **Overall Pipeline of TPO.** The architecture of Task Preference Optimization (TPO) consists of four main components: (1) a vision encoder, (2) a connector, (3) a large language model, and (4) a series of visual task heads. Differently colored flame symbols indicate which components are unfrozen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

CVPR 2025 の動向・気付き (57/181)

拡散モデルの直接優先最適化 (DPO)

段階的なプリファレンスの最適化
各ステップに PO を使用する [Lang+、cvpr 2025]

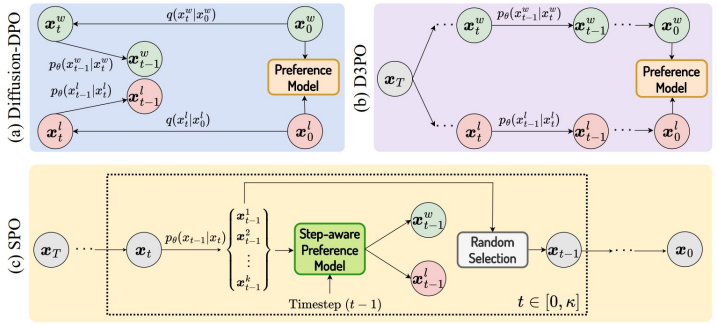


Figure 3. Comparing frameworks of SPO, Diffusion-DPO, and D3PO approaches. SPO does not adopt direct preference propagation as other DPO methods do. In SPO, a pool of samples are generated at each step, from which a proper win/lose pair is selected and used to fine-tune the diffusion model. Then, a single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

キャリブレーションによるプリファレンスの最適化 [Lee+, cvpr 2025]

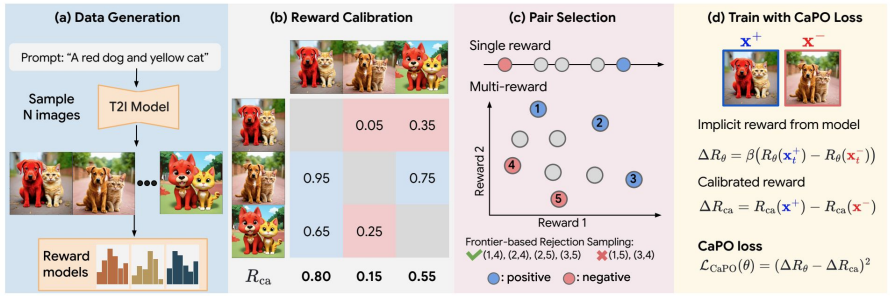
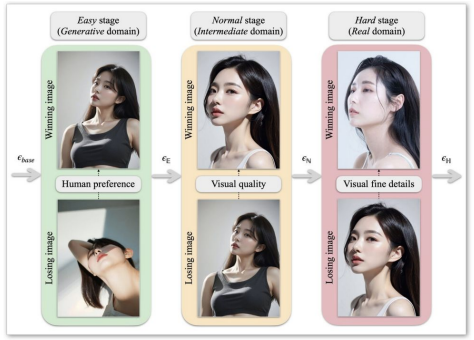


Figure 2. **Overview**. (a) We generate N images using pretrained T2I diffusion model using the prompt dataset, and infer the scores from reward models. (b) Then, we calibrate the rewards by making pairwise comparison between images. For each image, we compute the win-rates between other $N-1$ images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

カリキュラム学習付きDPO
[MA+、cvpr 2025] [Croitoru+、Cvpr 2025]



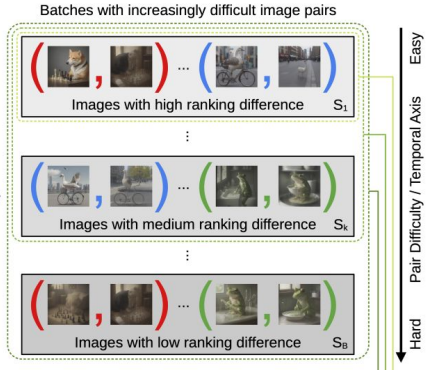


Figure 2. **Three-stage training of HG-DPO.** It progressively enhances the model's human image generation capabilities.

インバージョンプリファレンスの最適化
と 再パラメータ化 DDIM [Lu+、cvpr 2025]

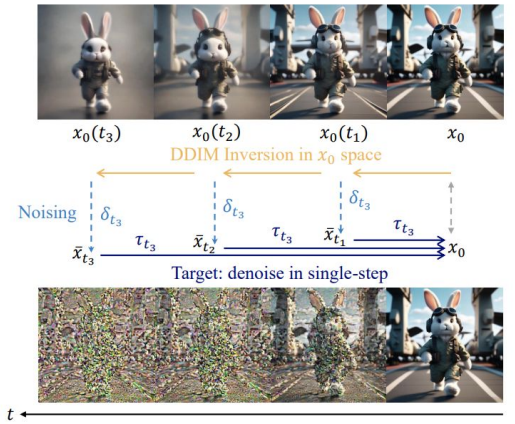


Figure 2. Illustration of Inversion for Preference Optimization.

CVPR 2025 の動向・気付き (58/181)

下流タスクにおけるDirect Preference Optimization (DPO)
DPOベースのfine-tuningは次のトレンドになるのか？

効率的なプロンプトとプリファレンスの最適化によるSAMの強化
半教師あり医療画像セグメンテーション用
[Koner+、cvpr 2025]

- 効率的な教師なしのプロンプト戦略
  セグメンテーションのパフォーマンスを向上させる
- ステップ1: 候補者に好みの評価を追加
  視覚と言語モデルによって推定
- ステップ 2: DPO によるモデルトレーニング
  DPOのラベル効率の高いデータセット生成に
  取り組む必要性あり
- 労働効率
  - 基盤モデルの利用
- データの品質を確保
  - ラベリング効率化
- データのバイアス
  - 複数モデルの統合

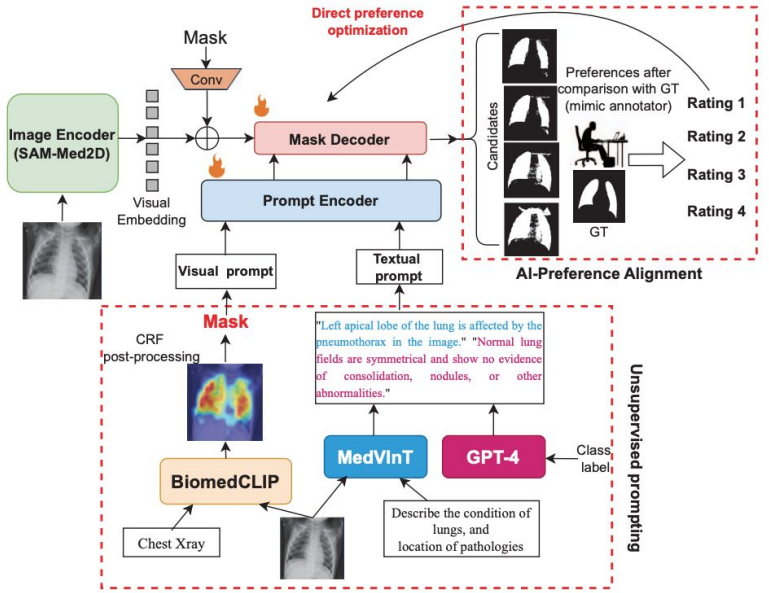
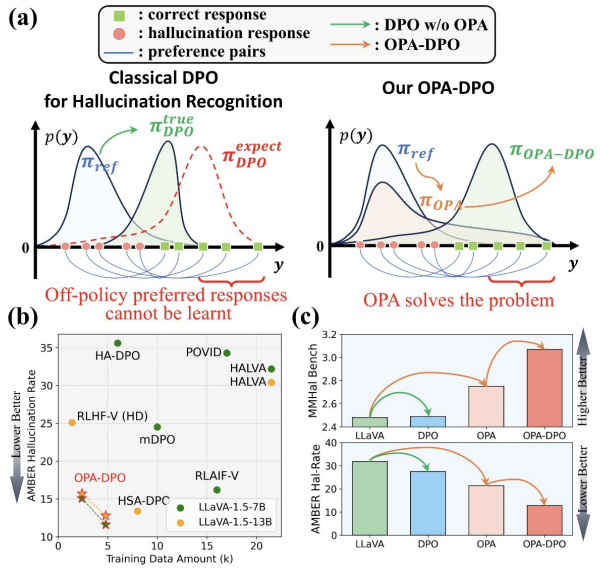


Figure 2. **Illustration of the proposed framework for semi-supervised segmentation:** Unsupervised geometric and text prompts, obtained from pretrained BiomedCLIP, MedVInT, and GPT-4 models, are fed into the prompt encoder for finetuning the framework on a small fraction of annotated data. In the next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

CVPR 2025 の動向・気付き (59/181)

DPOによる大規模視覚言語モデルにおける幻覚の軽減：ポリシーに沿ったデータがキー

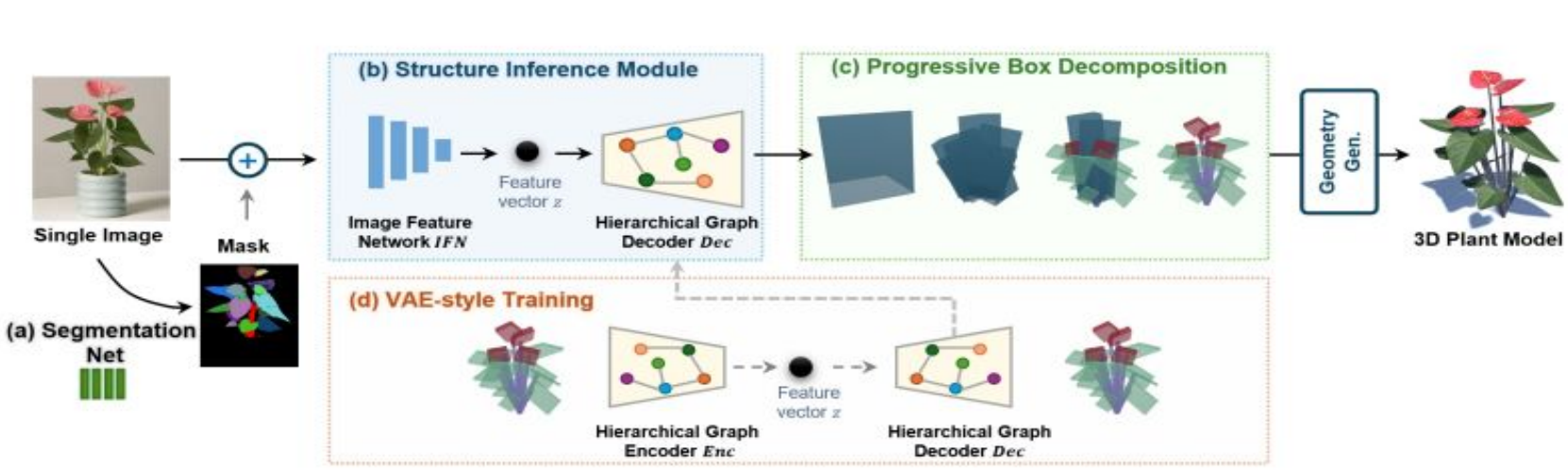
- DPO はポリシーから完全に外れた優先回答を学習できない (reverse-KL → ∞)
- DPOの前に、ポリシーに基づいて専門家が修正した回答を行う
- パイプライン全体:
  - 1. 収集の上で評価
    - ベース LVLM → レスポンス候補
    - GPT-4V タグセンテンスレベルの幻覚重症度 / エラータイプと最小限の修正テキスト (4.8k ペア)
  - 2. オン・ポリシー・アライメント (OPA)
    - 元の回答と修正された回答のLoRa-SFT → モデルのサポートに修正を加える
  - 3. OPA-DPO ファインチューン (3 つのプリファレンスペア)
    - 言語補正 (幻覚加重)
    - イメージフォーカス (クリーンなイメージと 30% マスクされたイメージ)
    - アンカープリファレンス (優先プロブがドリフトしないようにする)

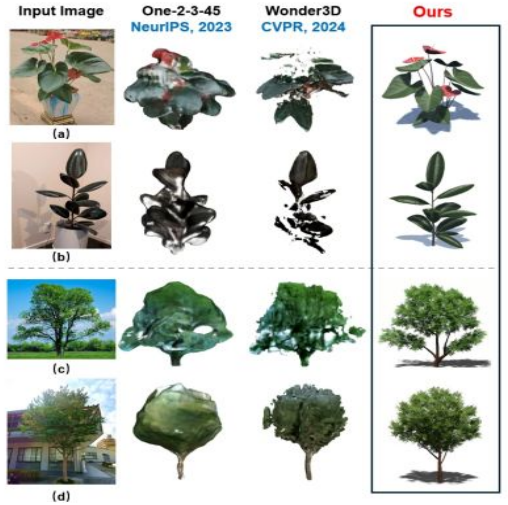


CVPR 2025 の動向・気付き (60/181)

Neural Hierarchical Decomposition for Single Image Plant Modeling

- 要約:
  - 植物の高品質な3Dモデルを構築することは困難
  - そこで、植物の画像1枚から3Dモデルを構築
  - セグメンテーション → 植物の構造を推論 → パーツごとに分割 → VAE
- キーポイント:
  - このアプローチは、屋内と屋外の両方の樹木に汎用可能



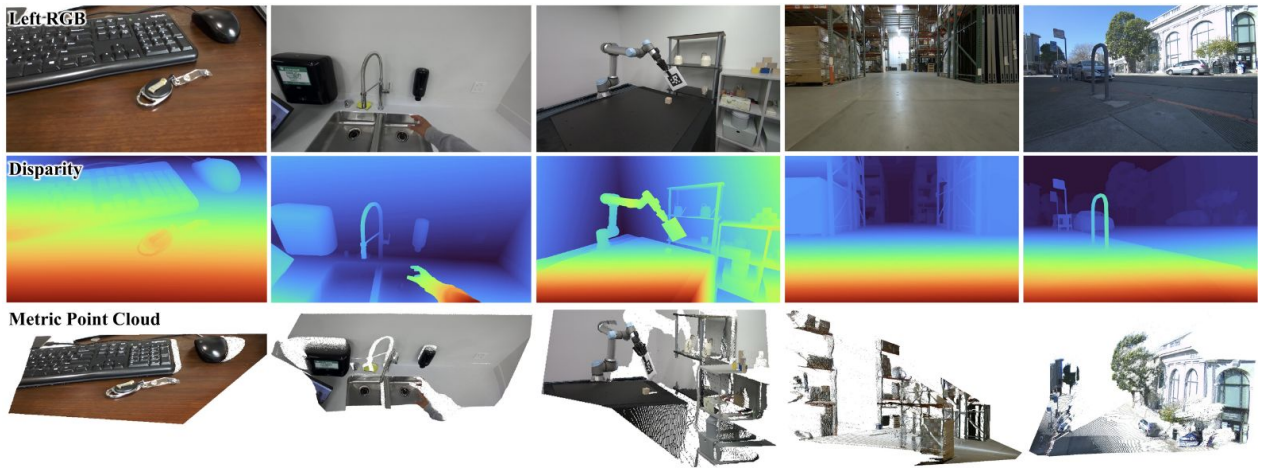


Zhihao et al., “Neural Hierarchical Decomposition for Single Image Plant Modeling”, CVPR 2025, [Link]

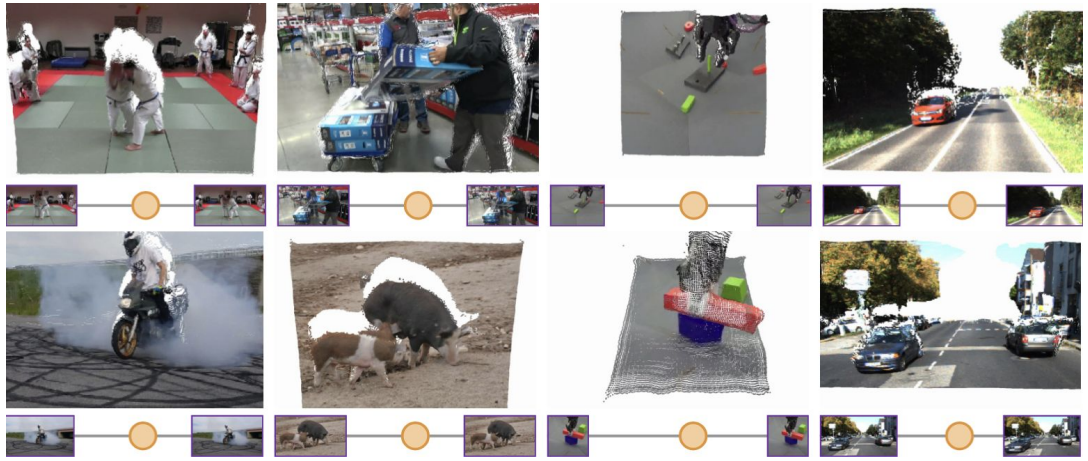
CVPR 2025 の動向・気付き (61/181)

合成データ (synthetic data) のポテンシャル

- Best paper candidates の視点から:
  - 1. 合成データ-実世界でのシミュレータ、生成モデル、その他リソースを使用したデータ生成
  - 2. 言語に依存しない画像についての基盤モデル構築:深度推定、オプティカルフロー推定、マッティング、セグメンテーション
  - 3. Zero-shot / in-the-wild 認識:追加のファインチューニングを行わなずとも、合成画像の事前学習は現実世界のデータに汎用可能であることが示された



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

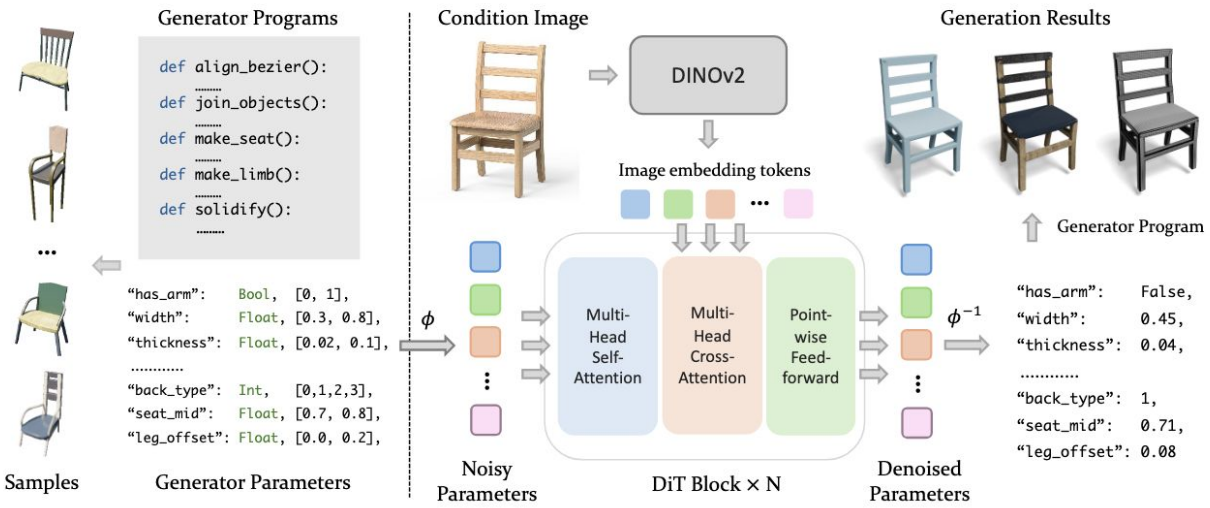


Y. Liang et al. “Zero-Shot Monocular Scene Flow Estimation in the Wild,” CVPR 2025. [Link]

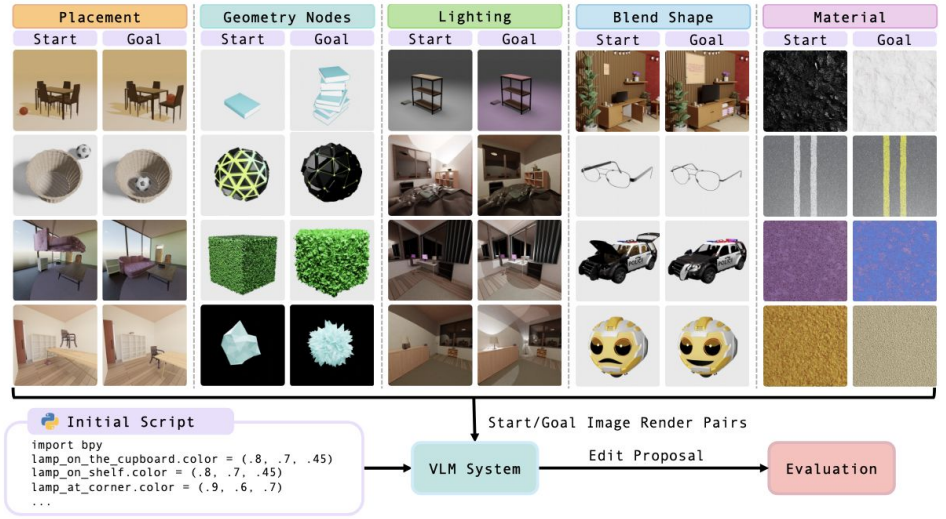
CVPR 2025 の動向・気付き (62/181)

合成データ (synthetic data) のポテンシャル

- CVPR 2025 論文の視点から:
- Infinigen / Infinigen Indoorsが下記論文に適用
  - DI-PCG: 拡散モデルのパラメータ推定により、画像から3Dへの生成を実現するための逆PCG(手続き型コンテンツ生成)フレームワーク
  - BlenderGym:3D画像を最初から目標の状態まで再構築するVLMを用いることで、現実世界のグラフィックを編集するタスクのための包括的なベンチマーク



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



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

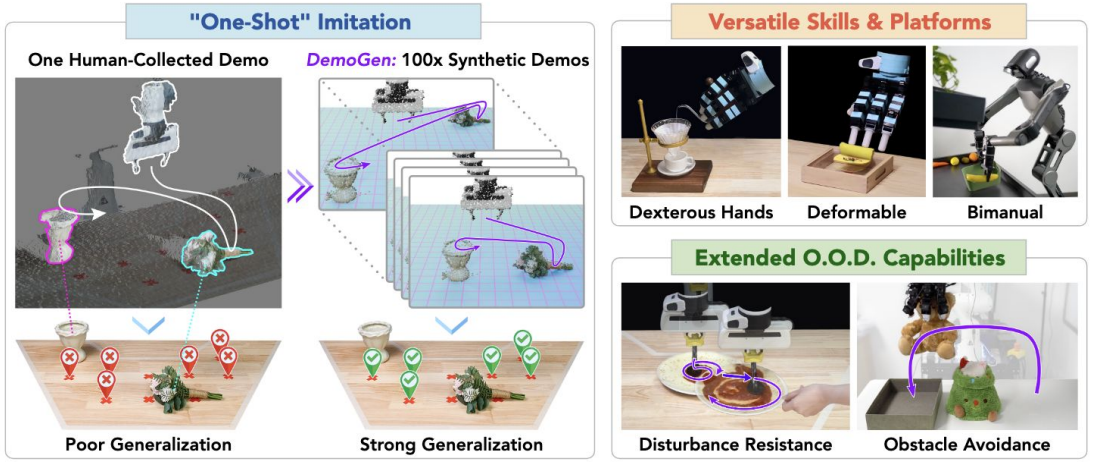
CVPR 2025 の動向・気付き (63/181)

合成データ (synthetic data) のポテンシャル

- CVPR 2025 SynData4CV の視点から:
- 視覚 / VLモデルの学習に合成 / 生成データが使用可能
  - SynData4CV Workshopには60以上のポスターが掲載
  - 分類、検出、セグメンテーション、3D認識、異常検出、動画、ロボット操作などの多くのタスクにおいて有効性が示されている

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

Workshop overview



Best Long Paper: Synthetic data for robot policy learning

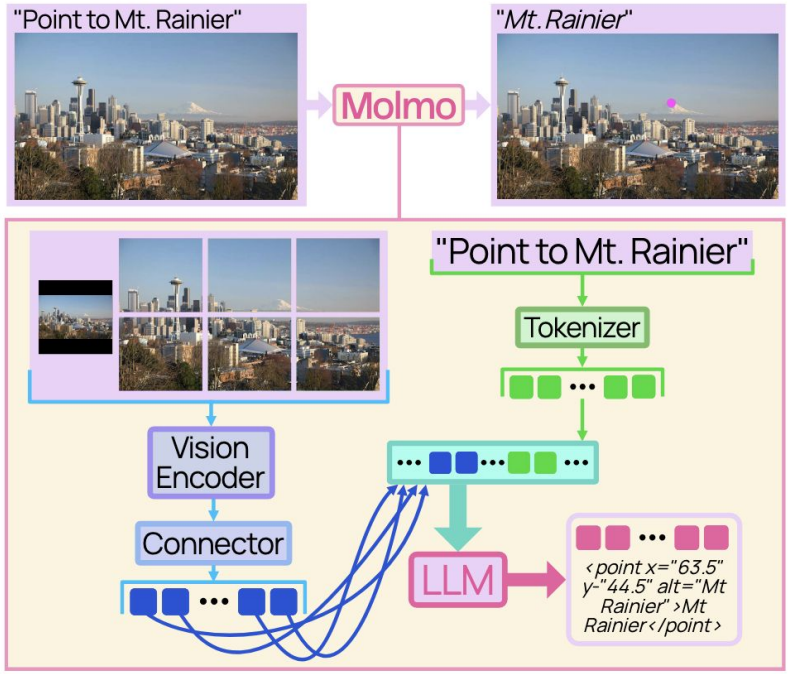
Z. Xue et al. “DemoGen: Synthetic Demonstration Generation for Data-Efficient Visuomotor Policy Learning,” CVPR 2025 SynData4CV Workshop. [Link]

CVPR 2025 の動向・気付き (64/181)

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

- MolmoとPixMoが示したこと:
  - 単純なモデルスケールではなくデータ品質 (改めて強調された)
  - Molmoは標準アーキテクチャとシンプルなVL接続をするモデル
    - オープン戦略の観点から、PixMoは公開データと合成データで構成されている

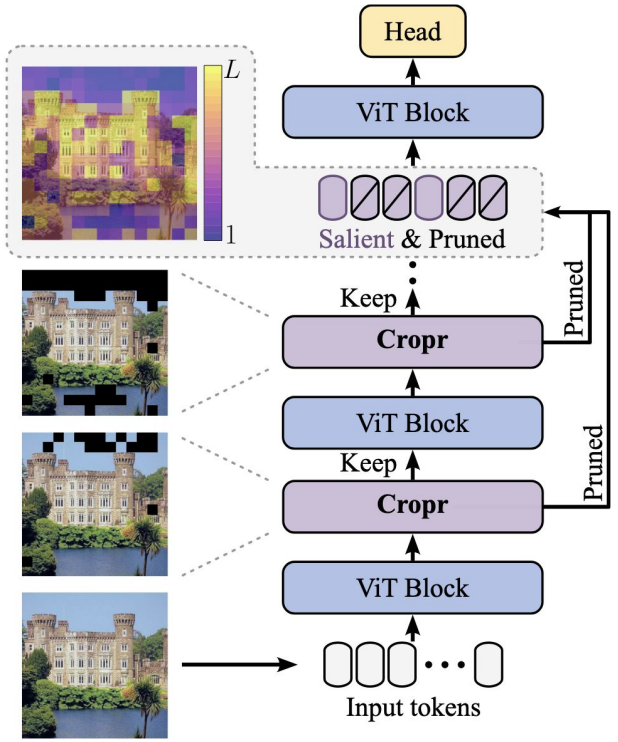
https://arxiv.org/abs/2409.17146

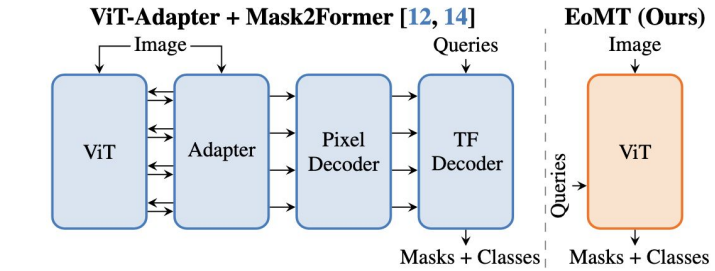


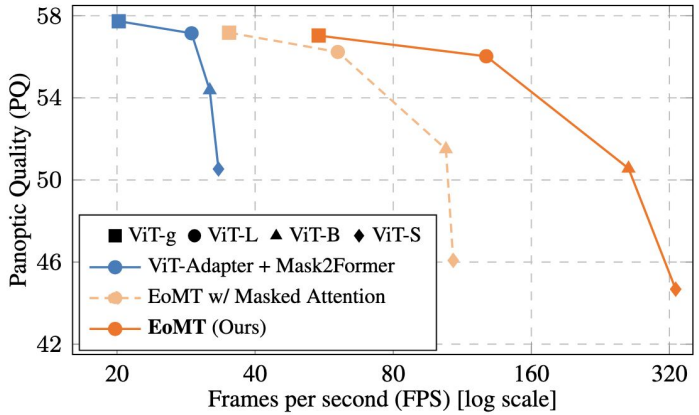
CVPR 2025 の動向・気付き (65/181)

ViT (Vision Transformer) 自体の改善

- Token Cropr ViT https://arxiv.org/html/2412.00965v1
  - 補助予測ヘッドを使用するトークンプルーニング
- Your ViT is Secretly an Image Segmentation Model https://arxiv.org/abs/2503.19108
  - 学習可能なquery + mask logitを使用すると、単純なViTでも画像セグメンテーション可能



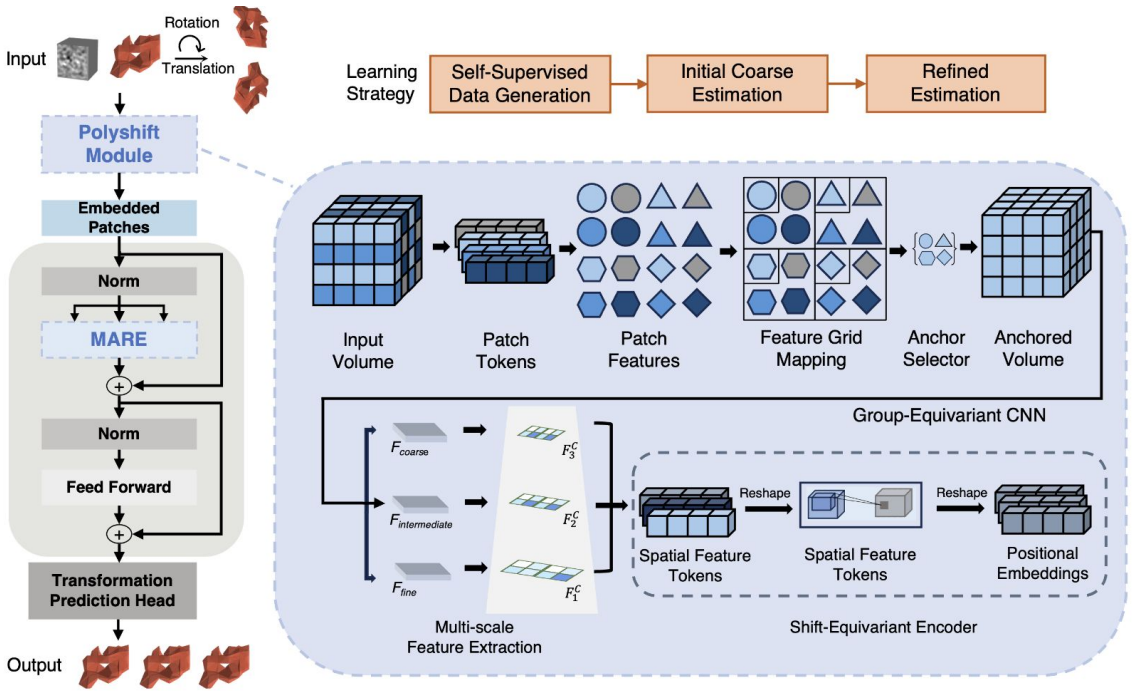


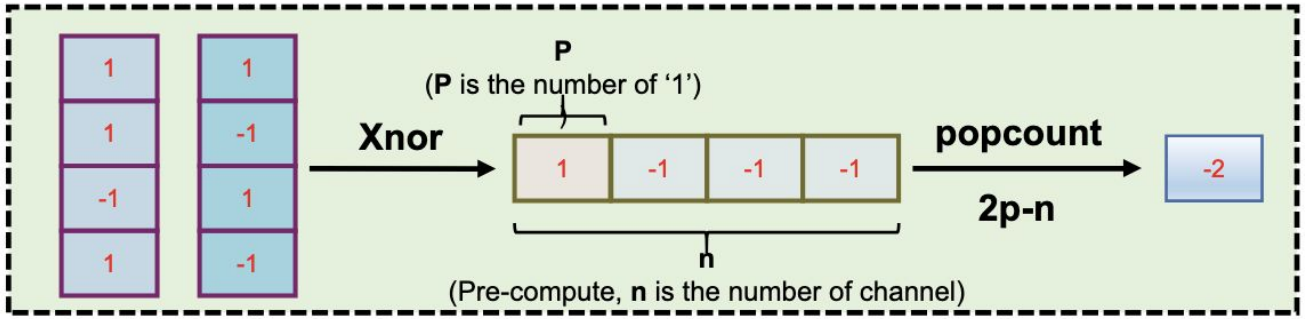


CVPR 2025 の動向・気付き (66/181)

ViT (Vision Transformer) 自体の改善

- BOE-ViT (Boosting Orientation Estimation ViT)
  - シフト/ 回転推定による3Dサブトモグラムアライメント用のViT
- BHViT (Binarized Hybrid ViT)
  - 完全二値化 ViT モデル / Xnor およびポップカウント演算により構成





... その他いくつかの方法 vIT アーキテクチャについて

CVPR 2025 の動向・気付き (67/181)

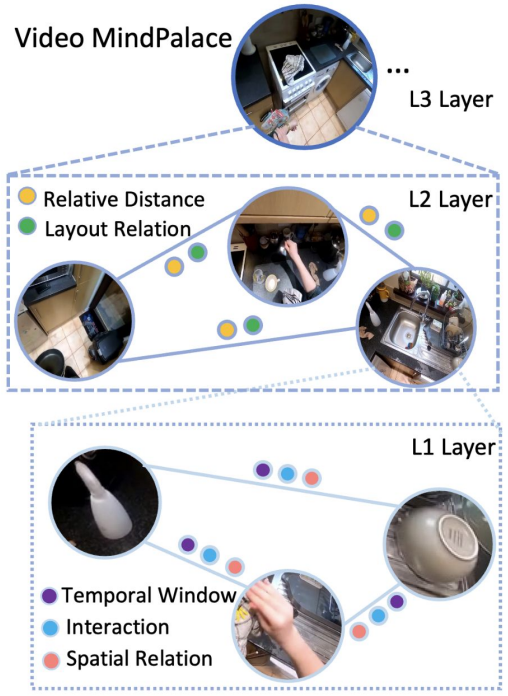
コンピュータビジョンの合成データ

- **CLIPAsso: Semantically-Aware Object Sketching**
- スケッチ画像の生成タスクの重要性
  - 多くの芸術的、科学的なアイデアはスケッチから始まる
- 拡散モデルによるスケッチ画像生成の課題
  - キーポイント: 既存のスケッチは数に限りがあり、著作権の制限を受けることが多いため、大規模なスケッチデータセットの収集が困難
- アプローチ: GControlNetを使用して、大規模な実画像データセットからスケッチ画像を生成
  - その後、合成スケッチを使用して新しい拡散モデルのトレーニングを行うことができるため、著作権フリーで大規模なスケッチを生成可能
- https://clipasso.github.io/clipasso/

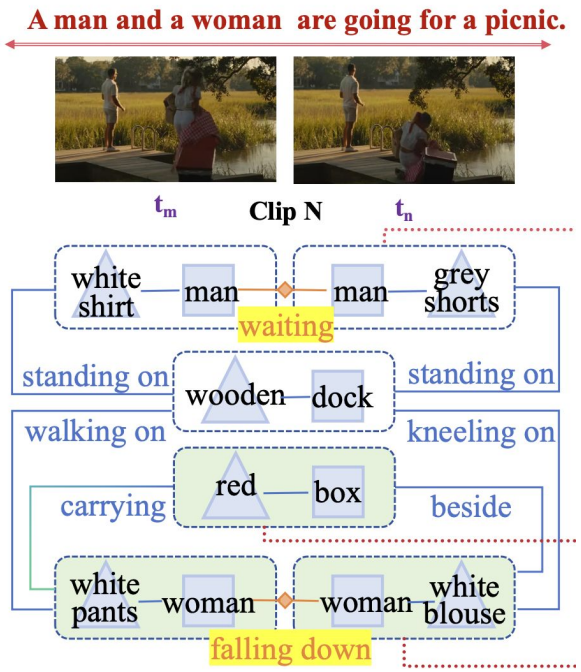
CVPR 2025 の動向・気付き (68/181)

VLMを使った動画理解のためのグラフ構造

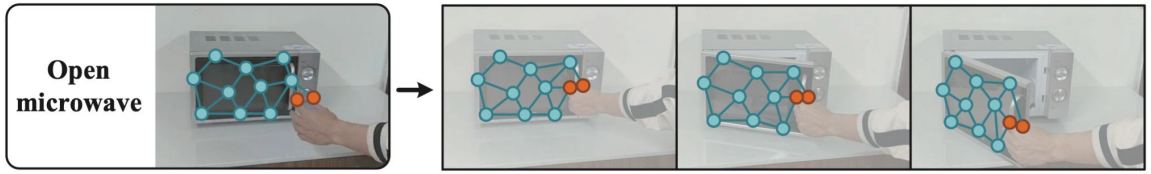
- [1] 時空間シーングラフにより、VLMは詳細な多段階推論データを自己生成、構成的推論能力を向上 (Qui et al.[Link])
- [2] トポロジカルなセマンティックグラフにより、VLMは複雑な長編動画の3Dコンテキストを深く理解し、人間のような推論を行う(Huang et al.[Link])
- [3] 物体・行動のグラフ構造により、ロボットは将来の状態と行動を予測 (Chen et al.[Link])



[1]



[2]



[3]

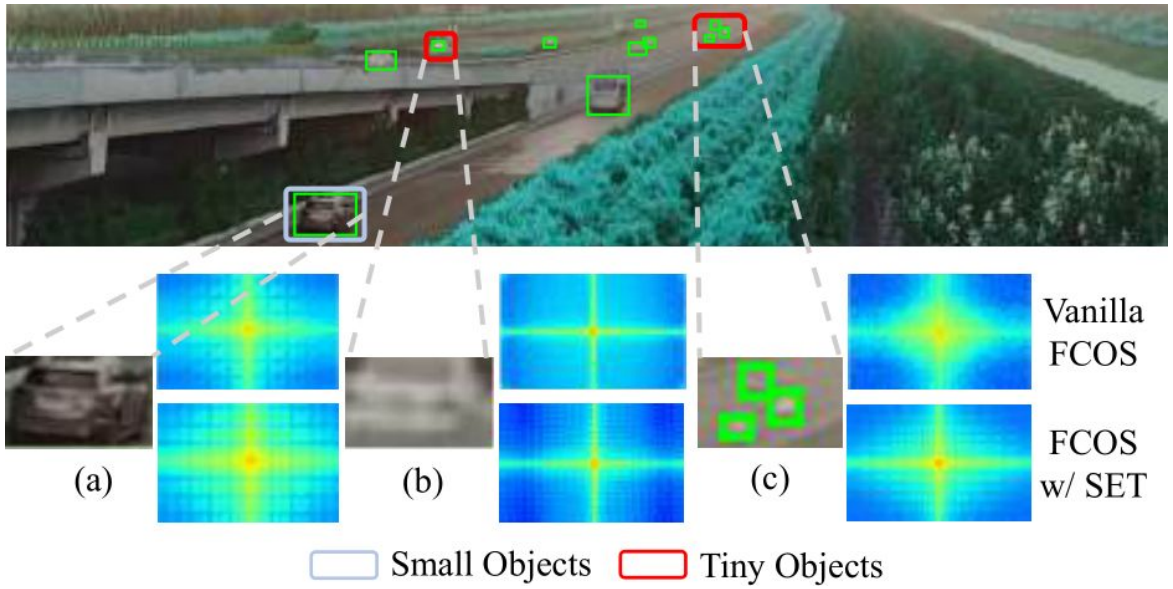
CVPR 2025 の動向・気付き (69/181)

顕微鏡画像に取り組む地球科学者 (1)

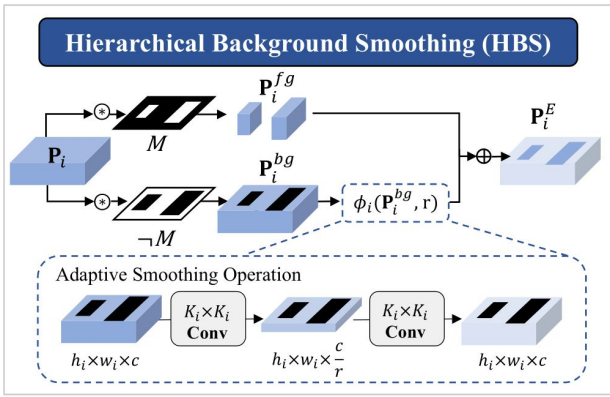
SET: 微小物検出のためのスペクトル強化

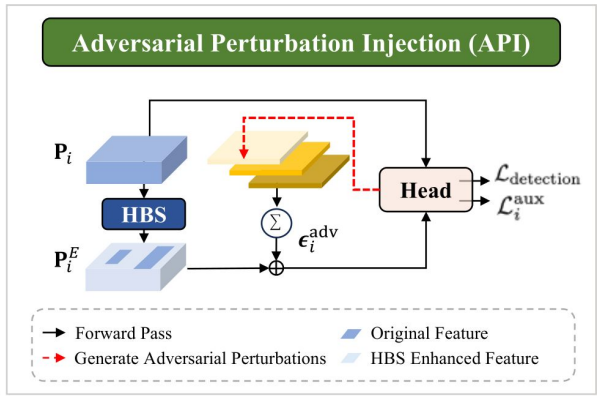
By Huixin Sun et al.
https://cvpr.thecvf.com/virtual/2025/poster/34394

ターゲット:検出 小さな オブジェクト



提案法: HBS + API





バックグラウンドの高周波ノイズを抑制することにより、小さな特徴の識別性を高めます

特徴マップに敵対的な摂動を追加して、トレーニング中の小さなオブジェクトの特徴を強化

結果: AI-TOD (微小オブジェクト検出) データセットでの性能 +3.2%

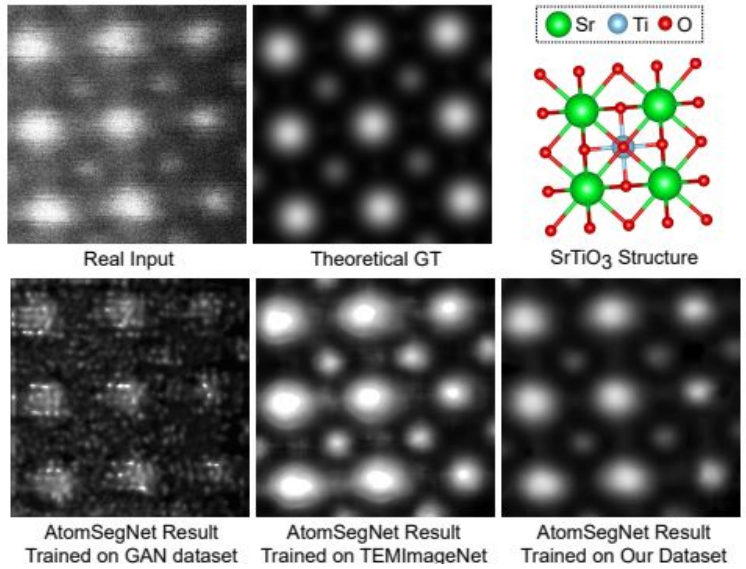
CVPR 2025 の動向・気付き (70/181)

顕微鏡画像に取り組む地球科学者 (2)

STEM画像強調のためのノイズキャリブレーションと空間周波数インタラクティブネットワーク

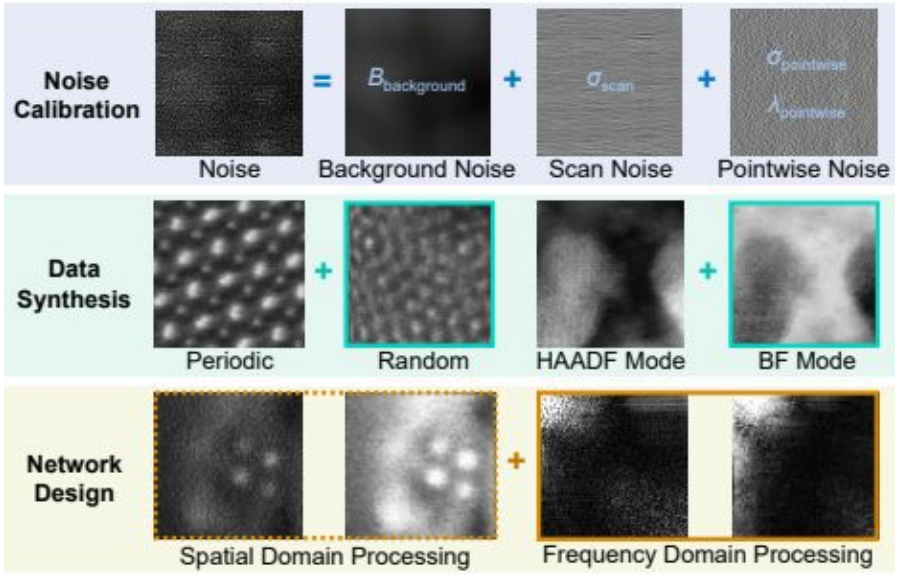
By Hesong Li et al. CVPR 2025 Open Access Repository

目標：STEM*による明確な観測



*STEM: 走査型透過電子 **顕微鏡**

提案法:



データセット作成

結果：微細構造物のより現実的な視覚化

CVPR 2025 の動向・気付き (71/181)

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

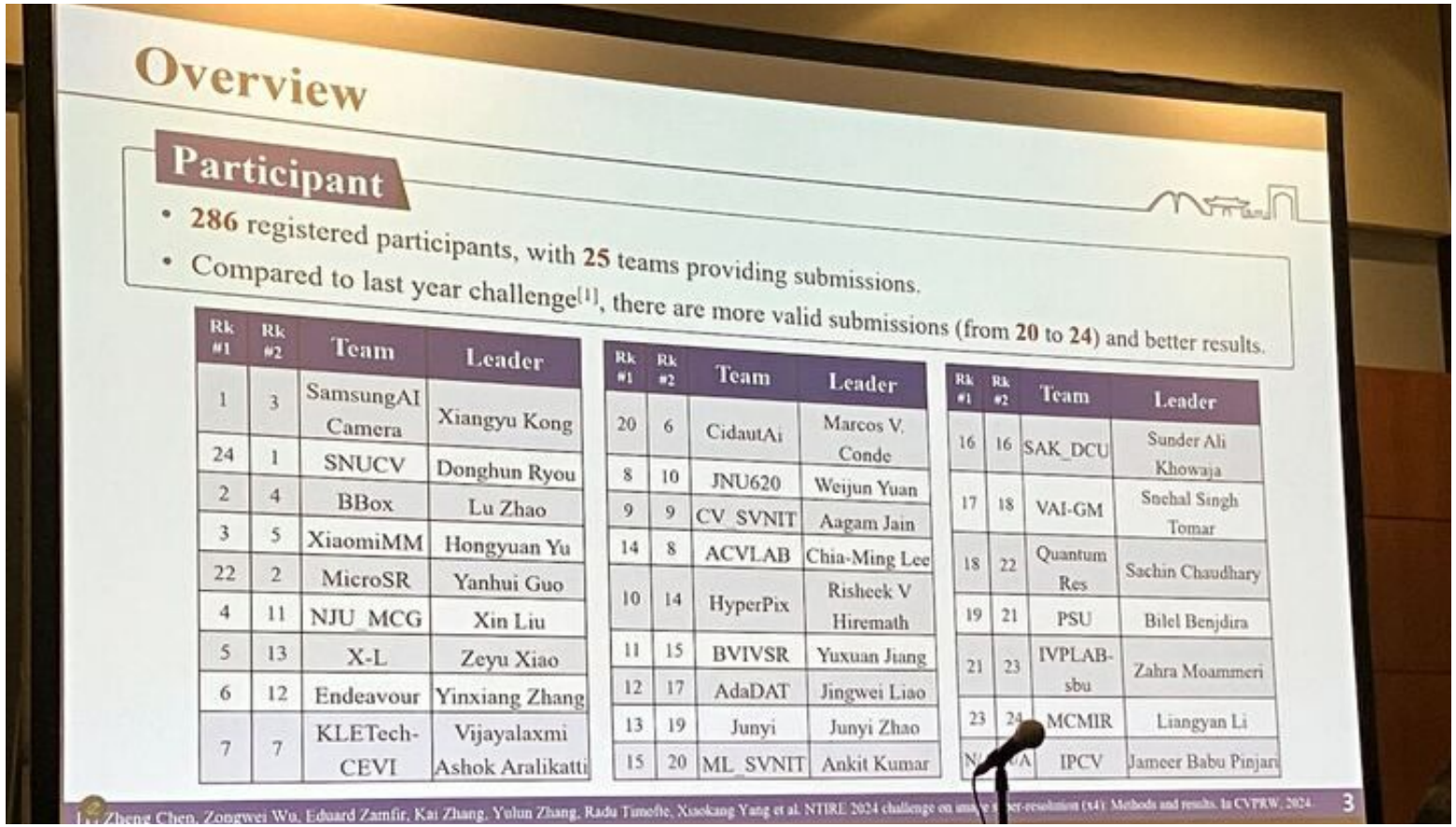
NTIRE 2025 画像の超解像に関するチャレンジ (x4)

- 4倍縮小された単一の低解像度入力(LR)から元の高解像画像(HR)を復元
- Dataset
  - チャレンジには、DIV2K、Flickr2K、LSDIRの3つの公式データセットが用意
  - LR-HR イメージペアは、bicubic down sampling を使用して生成
- Track
  - このチャレンジでは、次の2つのトラックを含むDIV2Kテストセットでのパフォーマンスを評価
  - レストレーショントラック:YチャンネルのPSNRとSSIMによるランク付け
  - パーセプチュアルトラック:7つのIQA指標の複合スコアでランク付け

CVPR 2025 の動向・気付き (72/181)

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

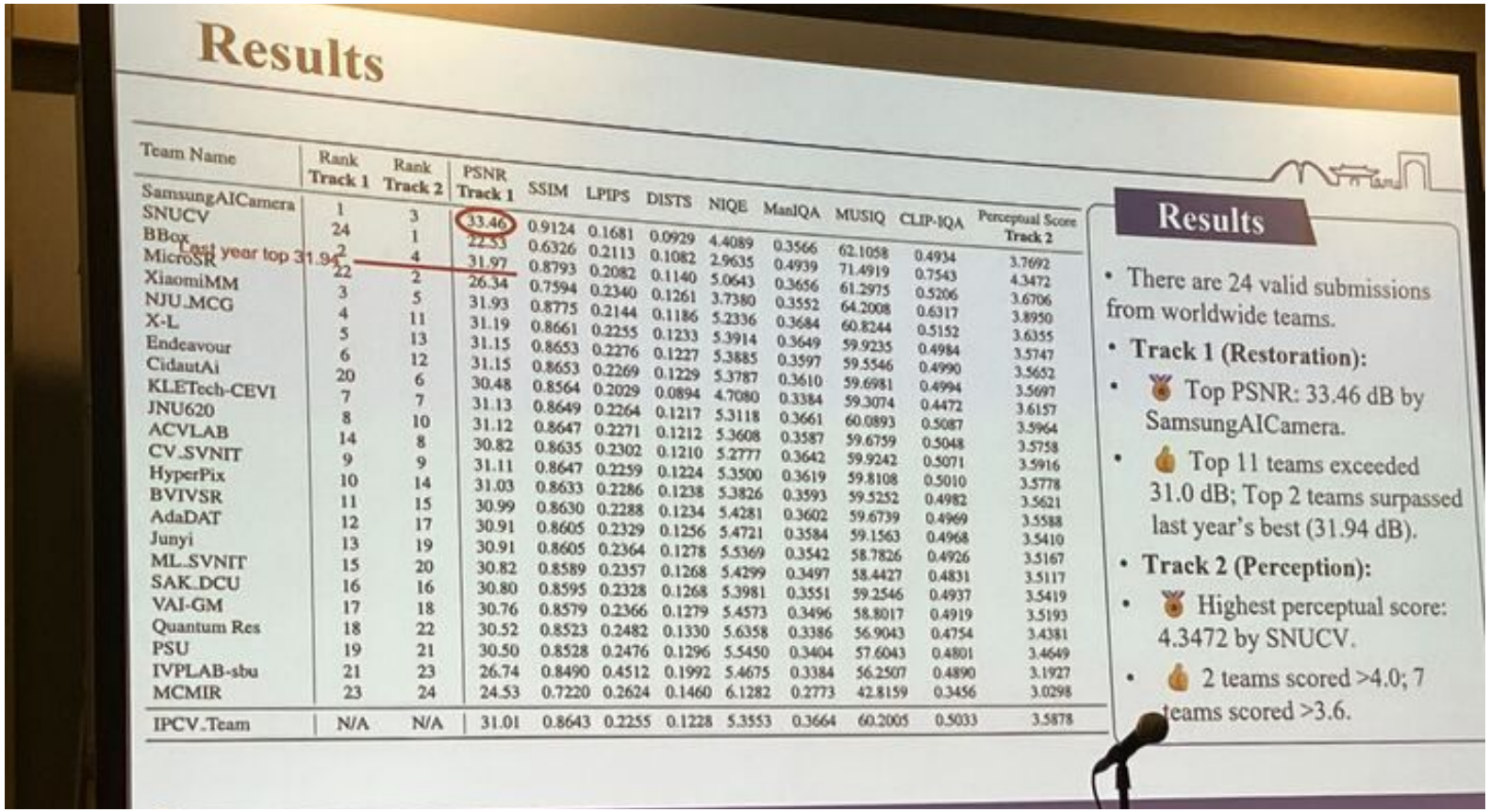
NTIRE 2025 画像の超解像に関するチャレンジ (x4)



CVPR 2025 の動向・気付き (73/181)

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

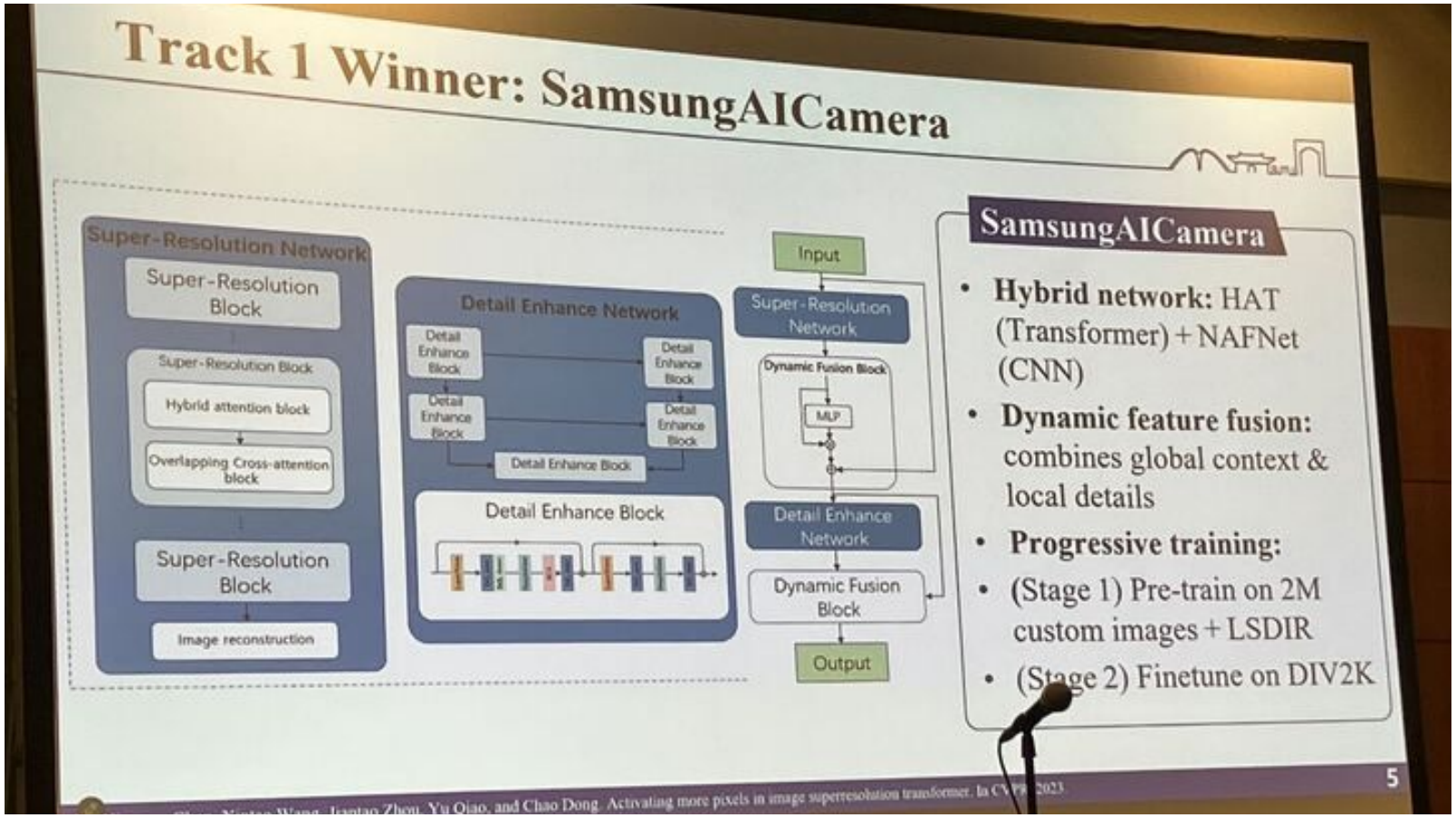
NTIRE 2025 画像の超解像に関するチャレンジ (x4)



CVPR 2025 の動向・気付き (74/181)

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

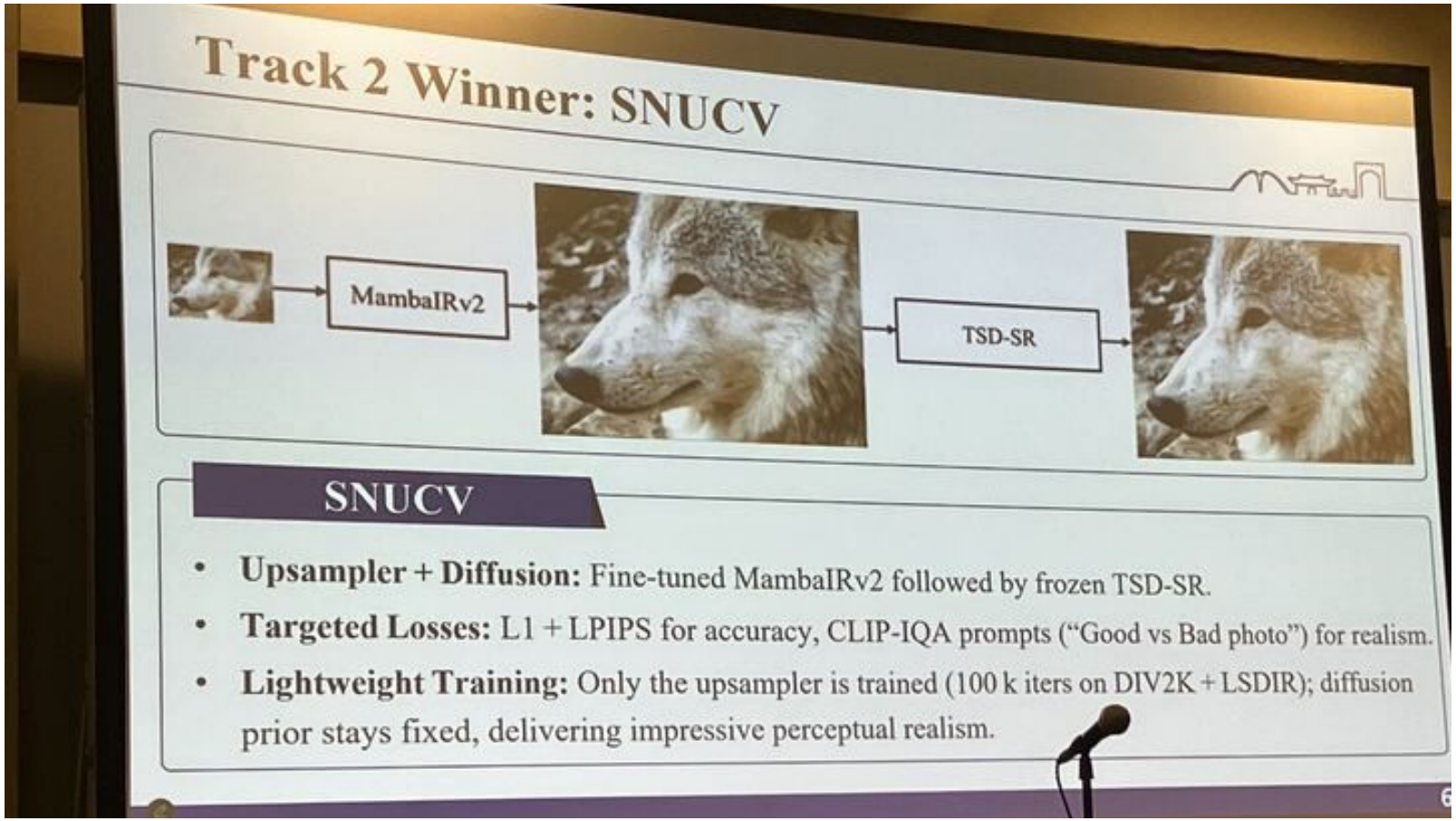
NTIRE 2025 画像の超解像に関するチャレンジ (x4)



CVPR 2025 の動向・気付き (75/181)

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

NTIRE 2025 画像の超解像に関するチャレンジ (x4)



CVPR 2025 の動向・気付き (76/181)

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

NTIRE 2025 画像の超解像に関するチャレンジ (x4)

超解像技術の動向

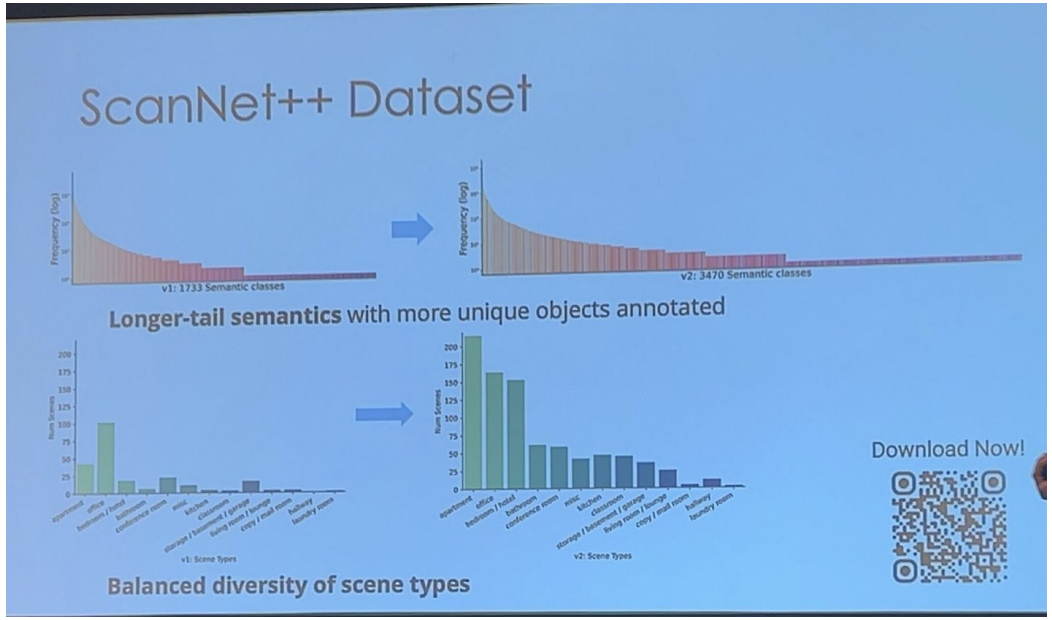
- Transformer ベースのアーキテクチャが依然として主流のアプローチ
- Mamba アーキテクチャの統合によるグローバル・コンテキスト・モデリングの改善
- マルチステージパイプライン、プログレッシブパッチトレーニング、CLIPベースのセマンティックフィルタリングなどの高度なトレーニング戦略により、モデルの汎化性とロバスト性が向上
- Generativeな事前確率、特に事前にトレーニングされた拡散モデルとCLIPベースの知覚損失を組み合わせることで、最小限のトレーニングで優れた知覚品質を実現

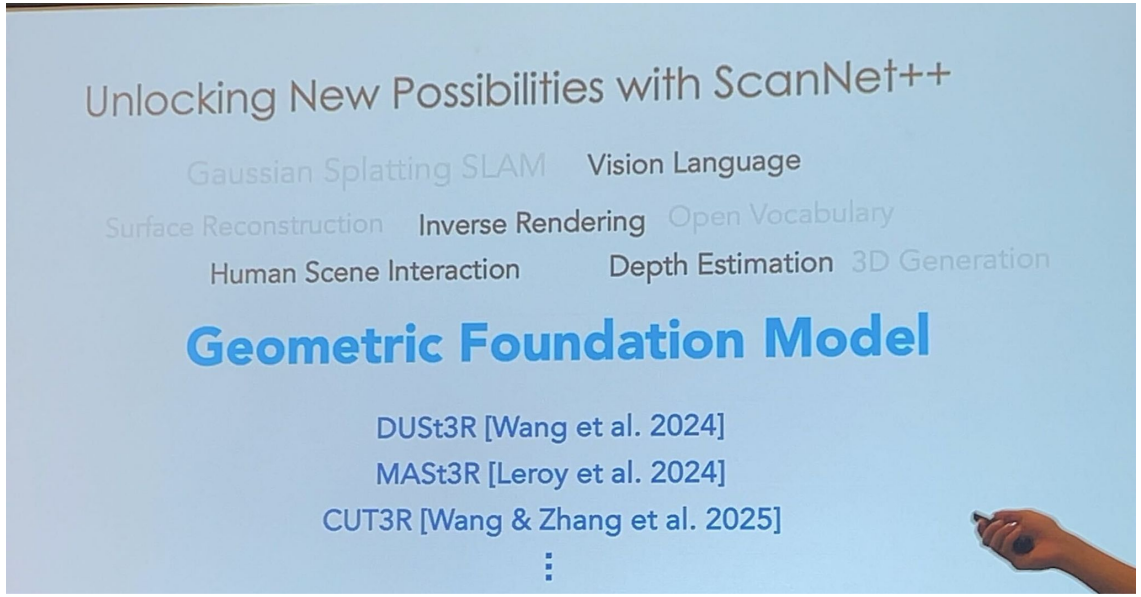
CVPR 2025 の動向・気付き (77/181)

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

同ワークショップでは下記の点にフォーカス

- ❑ 忠実度が高くボキャブラリーの多い3Dセマンティックシーン理解
- ❑ 大規模3Dシーンにおける新規視点合成（Novel View Synthesis）



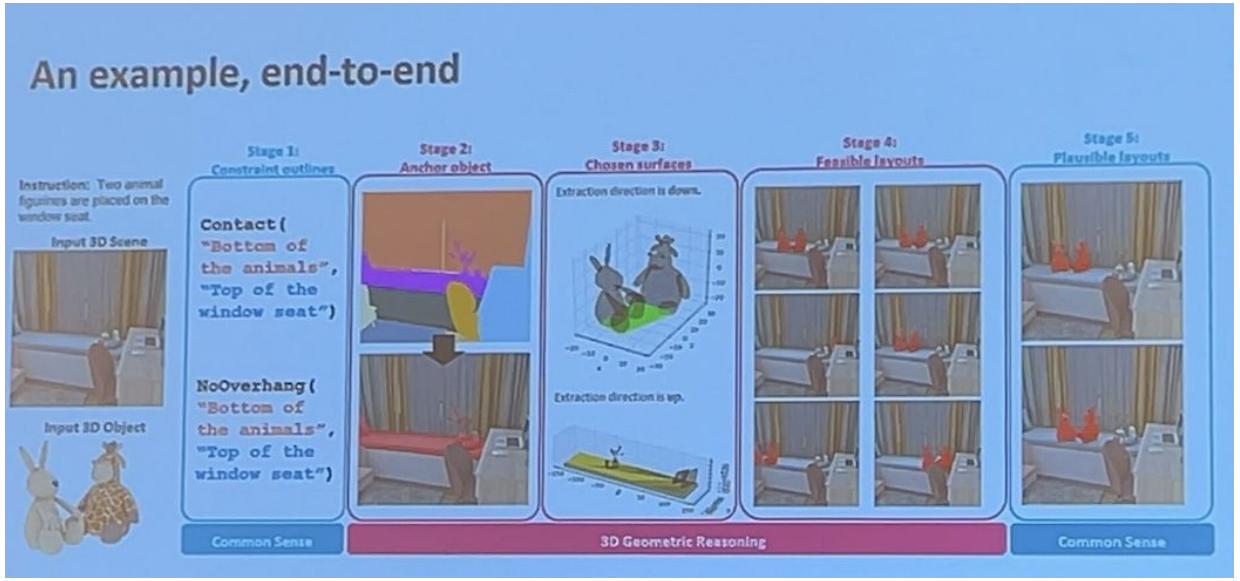


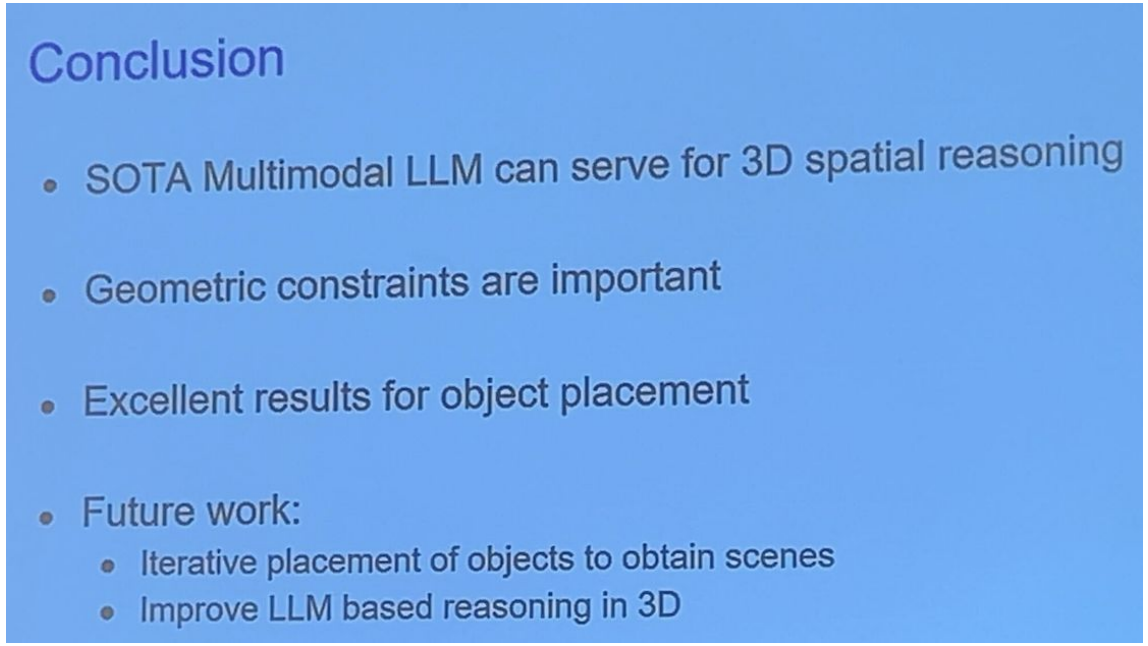
CVPR 2025 の動向・気付き (78/181)

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

LLMと空間的推論の統合による3Dシーンレイアウト生成

- ❏ 3Dシーン生成技術を改善してロボットシミュレーションやゲームアセットなどの分野に応用

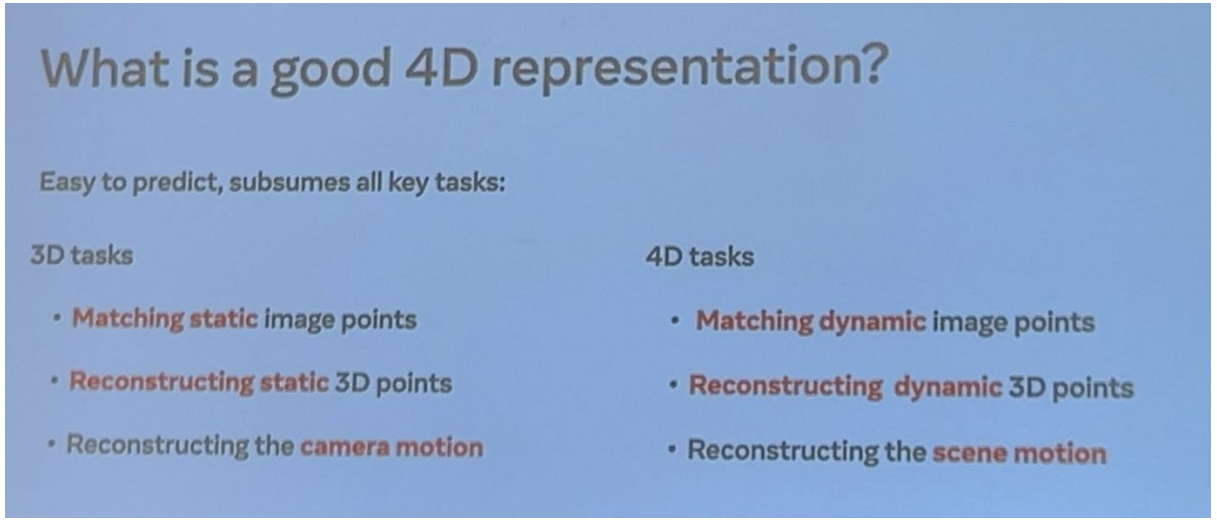


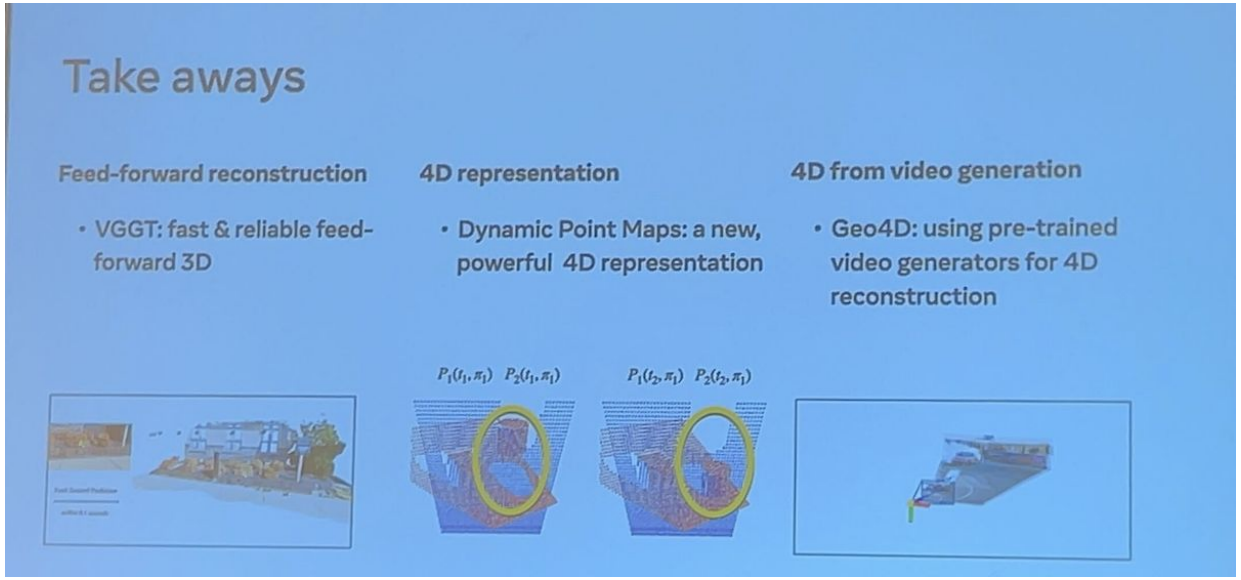


CVPR 2025 の動向・気付き (79/181)

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

今後のトレンドは、3D x 動画統合による4Dにシフトか





CVPR 2025 の動向・気付き (80/181)

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

- CAV-MAE Sync: Improving Contrastive Audio-Visual Mask Autoencoders via Fine-Grained Alignment 本会議論文
  - 視聴覚表現用のMasked AutoEncoderで時間的対応が可能。詳細な時間分解能によるCAV-MAEの向上、ImageBindと比較してオーディオとビジュアルの対応が向上
- Diagnosing and Treating Audio-Video Fake Detection
  - オーディオ動画のディープフェイクであり、ベンチマークデータセット、簡単なベースライン、評価プロトコルを提案
- UWAV: Uncertainty-weighted Weakly-supervised Audio-Visual Video Parsing 本会議論文
  - タスク：視聴覚動画解析、提案モデル：より大きなトレーニングセットによる疑似ラベルモデルのトレーニング、不確実性加重機能の取り違えによるトレーニング、評価データセット：オーディオイベント、ビジュアルイベント、オーディオビジュアルイベントの両方を含むLLPデータセット
- STM2DVG: Synthetically Trained Music to Dance Video Generation leveraging Latent Diffusion Framework
  - ポーズ条件付きLDM、合成データセット生成パイプライン、ミュージック・トゥ・ポーズ・エンコーダー
- Seeing Speech and Sound: Distinguishing and Locating Audio Sources in Visual Scenes 本会議論文
  - タスク:話し言葉と非音声の重複を含む、オーディオとビジュアルの同時グラウンディング

CVPR 2025 の動向・気付き (81/181)

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

- AVS-Net: Audio-Visual Scale Net for Self-supervised Monocular Metric Depth Estimation
  - コンセプト: エコーを使用して深度推定を強化、特にエコーを使用してより高い精度を獲得
- SAVGBench: Benchmarking Spatially Aligned Audio-Video Generation
  - コンセプト：空間的に調整されたオーディオ動画生成、人間のスピーチと楽器音を含むデータセットの提案（アンビソニックスオーディオ、360度動画を含む）、方法：MMディフュージョンベース
- BGM2Pose: Active 3D Human Pose Estimation with Non-Stationary Sounds
  - コンセプト：非定常音によるアクティブな3D人間の姿勢推定を支援
- Visual Sound Source Localization: Assessing Performance with Both Positive and Negative Audio
  - 画像からの音源位置特定問題
  - 動機：オフスクリーンやホワイトノイズに対するロバスト性向上、ネガティブオーディオの追加によるデータセット拡張、複数の評価指標導入
- VGGSounder: Audio-Visual Evaluations for Foundation Models
  - 包括的なAVベンチマーク導入
  - 人間がラベル付けしたもので、各動画には各モダリティのクラスに注釈

CVPR 2025 の動向・気付き (82/181)

Workshop: Sight and Sound (invited speaker)

- 社会的コミュニケーションにおける視聴覚的注意の推測法(James M. Rehg)
  - 一人称視点からの認識は社会的コミュニケーションにおいて重要
    - Paper: Listen to Look into the Future: Audio-Visual Egocentric Gaze Anticipation (ECCV2024)
      - タスク：視線予測
      - データセット (Ego4D Social、Aria)
      - 方法：動画とオーディオ、空間と時間の統合
  - 視線推定のためのビジョン基盤モデル
    - Paper: Gaze-LLE: Gaze Target Estimation via Large-Scale Learned Encoders (CVPR 2025)
      - タスク：動画内のすべての人の人間の視線を認識
      - 手法：基盤モデル + 微調整デコーダーパーツ活用
  - グループ会話認識
    - Paper: The Audio-Visual Conversational Graph: From an Egocentric-Exocentric Perspective (CVPR2024)
      - 動機：Ego-Exo Centric、どちらも重要
      - タスク：Ego-Exo Centric 会話グラフ予測
      - モデル：複数のエンコーダー（画像、音声）、プラスクロス、セルフアテンション
  - コンピュータビジョンタスクと組み合わせた社会的ジェスチャー、およびその他の社会的行動
    - Paper: SocialGesture: Delving into Multi-person Gesture Understanding (CVPR2025)
      - ジェスチャーを認識してローカライズするためのベンチマーク
      - 現在のマルチモーダルLLM、基盤モデルは、複数の人の社会的相互作用を認識するにはまだ不十分
    - Paper: Werewolf Among Us: A Multimodal Dataset for Modeling Persuasion Behaviors in Social Deduction Games
      - 説得戦略予測タスク
    - その他のデータセット：スピーキングターゲットの識別

CVPR 2025 の動向・気付き (83/181)

Workshop: Sight and Sound (invited speaker)

- 大規模言語モデルによる視覚と音 : 動画ダビングと空間音声理解への応用 (David Harwath)
- 2 つのアプリケーションにおけるマルチモーダル LLM の機能の拡大
- 動画吹き替え
  - Paper: VOICECRAFT: Zero-Shot Speech Editing and Text-to-Speech in the Wild
    - タスク: ボイスクローニング TTS
    - 問題: トーキングヘッドのデータセット不足
    - アプローチ: 言語ベースのデータを使用してテキスト読み上げのバックボーンをトレーニング、 2つ目は動画データを使用してファインチューニング
    - その他のアイデア : ニューラルコーデックによる音声のトークン化
  - Paper: VoiceCraft-Dub: Automated Video Dubbing with Neural Codec Language Models
    - タスク: 吹き替えにおいて、人物、テキスト、動画に合わせた音声を生成
    - 方法: VoiceCraft モデルからの初期化、リップ /フェイスエンコーダーの追加、動画ベースのデータセットへの微調整
- 空間音の理解
  - Paper: BAT:Learning to Reason about Spatial Sounds with Large Language Models
    - タスク: 空間音声認識; サウンドイベント検出 ; 方向; 距離,...
    - データセット: サウンドスペース 2.0 シミュレーター
    - 方法: LLM + 事前学習に空間サウンドエンコーディングを追加(Spatial-ASTモデル)

CVPR 2025 の動向・気付き (84/181)

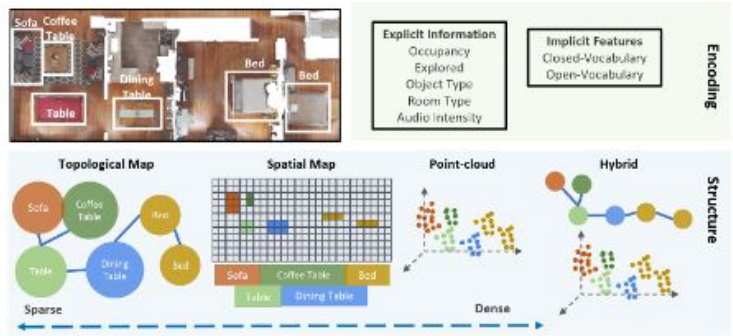
Workshop: Sight and Sound (invited speaker)

- ジェネレーティブ・モデルによる視覚と音の学習(Ziyang Chen)
  - Paper: Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models
    - 動画からのサウンド生成用の LDF
  - Paper: Video-Guided Foley Sound Generation with Multimodal Controls (CVPR 2025)
    - マルチモーダル条件 : 動画+テキスト+オーディオ/テキスト+オーディオデータのペアを使用した同時トレーニング
    - 非常に印象的な生成結果
  - Paper: Composing Images and Sounds on a Single Canvas (NeurIPS 2025)
    - 画像表現とスペクトログラムの交点
  - アイデア：聞こえる画像（画像は画像のキャプションのように見え、音声のキャプションのように聞こえる画像）
  - 方法：拡散モデルを使用

CVPR 2025 の動向・気付き (85/181)

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

- Embodied AI のビジョン言語マップの作成 (Angel Chang)
  - Semantic Mapping in Indoor Embodied AI (survey paper)



- 物体中心のナビゲーションのためのマップ作成
  - マルチオブジェクトナビゲーション(MultiON) タスク、LangNaviBench、GOAT-Bench
  - 多層機能マップ
    - 2フェーズ：探索と探索によるオブジェクト情報の取得（マップの作成）-> 探索のみ（マップ内の目標の検索）
  - 論文: ゼロショット・物体中心のインストラクション・フォロー：　基盤モデルと従来のナビゲーションの統合
    - 基盤モデルの使用：指示に従うための推奨言語推論因子グラフ、グラフ構造により、より構造化されたナビゲーションと優れた精度を実現

CVPR 2025 の動向・気付き (86/181)

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

- Hierarchical Action Models for Open-World 3D Policies (Dieter Fox)
  - Perceiver Actor
    - アクションを実行する前に 3D のポーズ/位置を予測
      - Paper: RVT: Robotic View Transformer for 3D Object Manipulation
        - 複数の視点情報を組み合わせてロボット観測を生成することによる明示的な3D表現の構築
      - Paper: RVT2: RVT-2: Learning Precise Manipulation from Few Demonstrations
        - 次のキーフレームポーズの予測
  - ビジョン言語モデルを活用して3Dポリシーを導く
    - 入力 -> Vision-Language-Action Model -> 低レベルポリシーのコンテキスト -> コントロール
      - Paper: HAMSTER: Hierarchical Action Models For Open-World Robot Manipulation (ICLR 2025)
        - VLM がポリシー学習用の 2D トラジェクトリスケッチを生成
        - 3D ポリシートレーニング用の 3D ポリシーモデル
        - 実世界評価の汎化性において有望な生成能力を示す

CVPR 2025 の動向・気付き (87/181)

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

- Genesis: An Unified and Generative Physics Simulation for Robotics (Chuang Gan)
  - How far are we from an embodied generalist agent?
    - Generalist agents: マルチモーダル、マルチタスク、マルチ環境。
    - 少なくとも、世界モデルが必要
  - 物理世界をモデル化する方法
    - ジェネレーティブAI +微分可能な物理エンジン
    - 微分可能な物理を使ったフォワードシミュレーション
    - マルチビュー動画からの関節形状と物理推定
    - アクション付きフォワードシミュレーション
    - 結果はRLベースの学習よりも良い
  - アプリケーション
    - ソフトオブジェクトアニメーション, Thin-shell manipulation
  - 上記のアプローチをスケールアップできるか？
    - Genesis: 完全微分可能な汎用物理シミュレータ
      - データ生成をスケールアップするために、著者らはタスク提案（言語を使用、言語をプロンプトとして使用）、シーン生成、トレーニング教師生成が可能なRoboGenを提案
      - ヒューマンツールがロボット工学に適さないという問題を解決するには？-> ロボット工学用ジェネレーティブツール

CVPR 2025 の動向・気付き (88/181)

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

- The One RING: A Robotic Indoor Navigation Generalist
  - 全ての異なるロボット工学のためのユニバーサルナビゲーションポリシー? — ロボットが異なれば、観察や動作も異なる
  - 著者らは、データとモデル両方のOne Ringを提案、方法: 大規模トレーニング + ランダムな実施形態でのRLファインチューニング、シミュレーションでトレーニングされたが実際の環境に適応できる。
- Manual2Skill: Learning to Read Manuals and Acquire Robotic Skills for Furniture Assembly Using Vision-Language Models
  - 視覚言語モデルを活用した操作スキルの学習、2D画像と言語を入力として使用し、アセンブリ用の階層グラフを生成
- エージェント言語に基づく適応型ロボットアセンブリ
  - 課題: アセンブリアプリケーションでは、さまざまなモデル形状や構造の変化にそれに応じて適応する能力が不可欠
  - 解決策:LLMを使用して、さまざまな部品やモデルの変更にアセンブリスキルを適応させるための階層ガイドを生成
- ZeroMimic: Distilling Robotic Manipulation Skills from Web Videos
  - 課題: ロボット訓練用のデータセット
  - 解決策:オンライン動画から一般的なスキルポリシーを学ぶ、モデル:ステップ1:人間の手が何をしているのかを知る、ステップ2:人間の手を操作するためのBCポリシーのトレーニング、ステップ3:ロボットアームにヒューマンアームポリシーを適用、基盤モデルは、人間の動画をロボット・ポリシーに抽出することを容易に

CVPR 2025 の動向・気付き (89/181)

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

- After Diffusion Models (Bill Freeman)
  - 拡散モデルの問題点
    - 遅い、制御困難、編集困難 ...
  - うまくいく可能性のある方法
    - 従来型グラフィックス、ヒューマンアーティストの採用、 シンプルで説明可能な生成アルゴリズム、 エンジニアリング、より良い推論モデル、...
  - 従来型グラフィックス
    - なぜそうしない？: 時間がかかり、複雑で、手作業が多い
  - シンプルで説明可能な生成アルゴリズム
    - Paper: Infinite Images: Creating and Exploring a Large Photorealistic Virtual Space
      - モーションをアンラップして無限のイメージにする
    - Paper: WonderWorld: Interactive 3D Scene Generation from a Single Image (CVPR 2025)
  - 丁寧な質問によるコンピュータービジョン
    - 物理世界に関する十分な知識がある基盤モデル
      - Paper: Alchemist: Parametric Control of Material Properties with Diffusion Models (CVPR 2024)
    - あなたが求めることをするように LLMを説得する、腹立たしいが時には力強い
  - より優れた拡散モデルのエンジニアリング
    - Paper: Improved Distribution Matching Distillation for Fast Image Synthesis (NeurIPS 2024)
    - 拡散モデルの蒸留

CVPR 2025 の動向・気付き (90/181)

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

- Controllable, Intuitive Generation of 4D Objects and Scenes (1/2) (Jiajun Wu)
  - 2D ピクセルを超えた知覚と生成
    - 物理オブジェクトの組み込み関数へのde-render – 制御性が向上
  - アイデア1: SSL の誘導バイアスとしての組み込み関数
    - 組み込み関数にデレンダリングしてから画像に再レンダリングする
    - Phisical intrinsics (照明、3D形状、カメラ、..)
      - Paper: Seeing a Rose in Five Thousand Ways (CVPR 2023)
        - 1つの画像、複数のオブジェクトインスタンス(花束など)– すべてのバラには同様の本質がある
        - 次のことを可能にする可制御性
      - Paper: Learning the 3D Fauna of the Web (CVPR 2024)
        - 単一カテゴリから複数のカテゴリへ、上記を拡張
  - アイデア2:Vision-only FMの蒸留ターゲットとしての内因性
    - Paper: Birth and Death of a Rose (CVPR 2025)
      - 事前にトレーニングされた2D基盤モデルから、時間的オブジェクト固有要素(オブジェクトのジオメトリ、反射率、花が咲くバラなどのテクスチャの時間的に変化するシーケンス)を生成
    - Paper: PhycsDream: From Object Motion to Action (ECCV 2024)
      - 3D オブジェクトのアクション条件付きダイナミクス予測

CVPR 2025 の動向・気付き (91/181)

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

- Controllable, Intuitive Generation of 4D Objects and Scenes (2/2) (Jiajun Wu)
  - Idea2A: FM（視覚と言語）の抽出ターゲットとしての内因性
    - VLMの観察から再構築まで
      - Paper: WonderJourney: Going from Anywhere to Everywhere (CVPR 2024)
        - モデルイラスト: LLMを使用して長いシーンの説明を生成し、テキスト駆動型のポイントクラウド生成パイプラインを使用して3Dシーンを合成し、VLMを使用して生成された例を検証する
      - Paper: WonderWorld: Real-Time, Interactive 3D Scene Generation (CVPR2025)
        - ユーザーは、生成する場所と内容を対話形式で指定できる
      - Paper: WonderPlay: Dynamic 3D Scene Generation from a Single Image and Actions
        - 4D 方式での編集と操作を許可
      - Paper: The Scene Language: Representing Scenes with Programs, Words, and Embeddings (CVPR 2025)
        - チェス盤の生成を例に挙げると、表現力豊かなシーンの生成が可能になり、高度なインタラクティブ編集が可能になります
        - 構造: プログラム関数依存
        - セマンティクス:自然言語単語
        - ビジュアルアイデンティティ: ニューラル 埋め込み
        - テキストから3D/4Dへの生成と編集を可能にする
- What's the next:
  - 疑問: 何をモデル化する必要があるか?ビジュアルの生成にどのように役立つのか ?FM の役割とは？
  - 因果的、物理的、およびユニバーサルオブジェクト組み込み関数に基づくレンダリングコンディショニング
  - 現場のデータを効果的に活用し、解釈可能性向上させる

CVPR 2025 の動向・気付き (92/181)

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

- Language as a Visual Format (Phillip Isola)
  - ビジョンモデルが世界を見る方法は、言語モデルが世界を見る方法とどの程度似ているか
    - Paper: The Platonic Representation Hypothesis (ICML, 2024)
      - カーネルを使った表現のキャラクタライズ
      - 類似性比較 (DINO、Llama)
      - DINOのモデルが大きいほど、類似度は高くなる
      - より詳細なキャプションは、ビジョンの表現と相性が良くなります。
    - Paper: Cycle Consistency as Reward: Learning Image-Text Alignment without Human Preferences
      - 画像と言語のサイクルを一貫させることで画像とテキストの配置を改善
      - 軽量、高速、差別化可能
        - Dataset: Cycle consistency preference collection (CyclePrefDB)
          - 画像とテキストの配置の評価と改善に便利
          - CyclePrefDB - T2Iを使用したdirect preference optimization (DPO)
          - 使用法: pip install cyclereward
- What's after diffusion
  - ピクセルの代わりとなる言語、ただし視覚的に説明的な言語が必要

CVPR 2025 の動向・気付き (93/181)

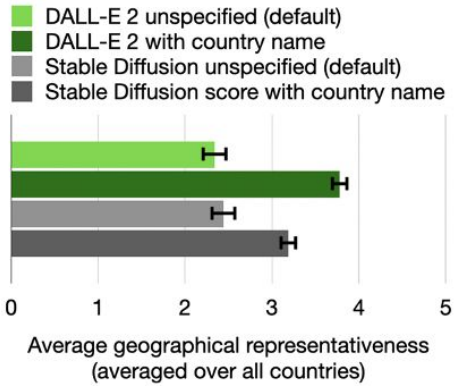
Workshop: VLMs-4-All 2025

- Re-scaling cultural knowledge in a UK heritage and museum context (Maya Indira Ganesh)
  - 文化遺産ハブ (ArCH) 向け人工知能
    - 重要課題：現在のAIを使って人々が文化遺産の認識にもっと関与できるようにするにはどうすればよいか
  - GLAMセクターと文化遺産収集の課題
    - 課題：アクセシビリティ、断片的で分散したオブジェクトの再構築、専門知識が必要、複数のプロバイダーにわたる遺産技術システム、植民地主義の遺産、アーティファクトや文化的知識の出所に関する争い
    - その他課題：AIに対する否定的な認識、不確実性、エラー、デジタル化、デジタル化されていないバックログ、レガシーインフラストラクチャ, provisional semantics
  - AIによる日常の街頭観測所
    - 道路認識や道路での具体化されたタスクには、未解決問題が山積

CVPR 2025 の動向・気付き (94/181)

Workshop: VLMs-4-All 2025

- Richer Outputs for Richer Countries: Geographical Disparities in Language and Image Generation (1/2) (Danish Pruthi)
  - モノニム (単語の名前)
    - 一部の現代社会（インド、ミャンマーなど）では一般的
    - デジタルフォームで単一名が使用されることはほとんどない
    - 問題: 最近のLLMでは認識されない可能性あり
    - その他の AI システムの問題: 性別、人種、地理に関する偏見
  - 代表生成
    - さまざまなアプリケーションや実際のシナリオで重要な、視覚的な概念を表す画像 / 画像の生成
  - 課題: 現在のモデルは地理的な代表者を生み出しているのか?
    - Paper: Inspecting the Geographical Representativeness of Images from Text-to-Image Models (ICCV 2023)
      - 現在のモデルでは、地理的代表性が低い

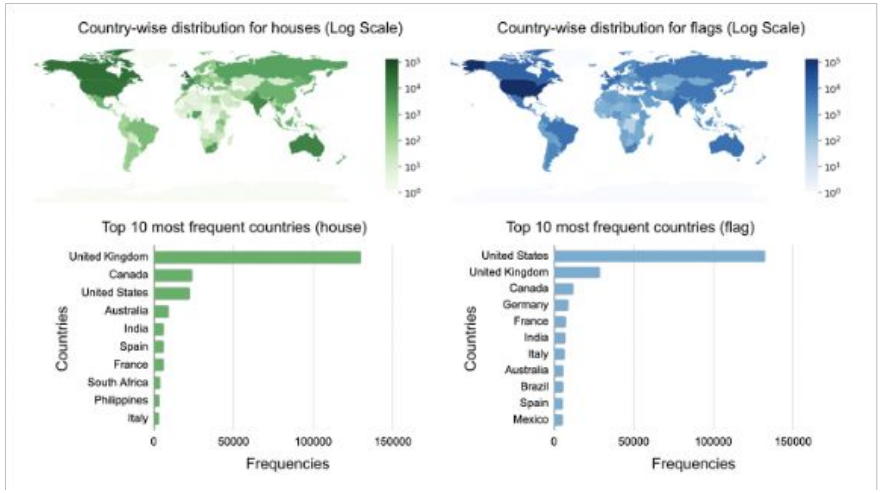


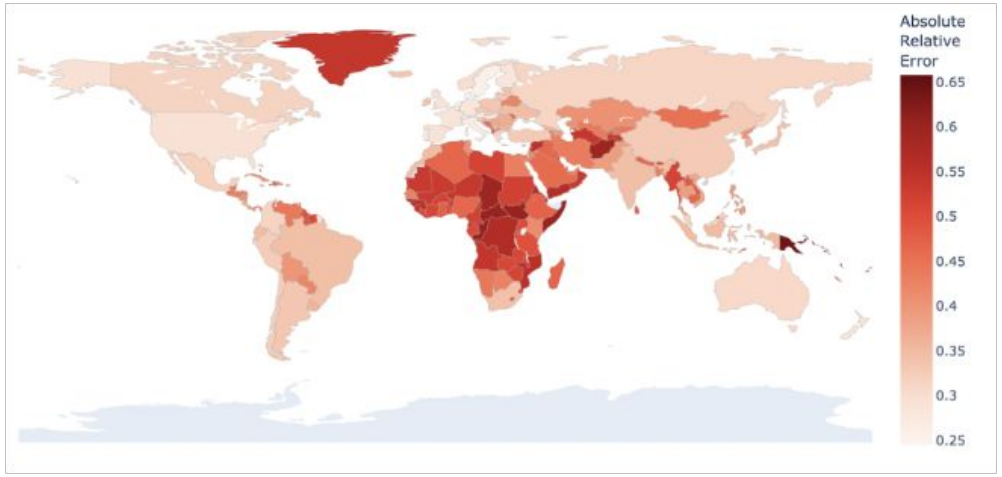
    - Paper: Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale (FAccT 2023)
      - 特性、記述子、職業、または目的を単に言及するだけのプロンプトなど、さまざまな通常のプロンプトがステレオタイプを生み出す
      - プロンプトでアイデンティティや人口統計学的言語が明示的に言及されているか、そのような言葉を避けているかに関係なく、ステレオタイプが存在

CVPR 2025 の動向・気付き (95/181)

Workshop: VLMs-4-All 2025

- Richer Outputs for Richer Countries: Geographical Disparities in Language and Image Generation (2/2) (Danish Pruthi)
  - 課題: 現在のモデルは代表的地理を生成しているのか? - NO
    - Paper: Where Do Images Come From? Analyzing Captions to Geographically Profile Datasets (left image)
      - タスク: データセットを地理的にプロファイリング、画像とキャプションのペアを指定して、データセットを特定の場所にマッピング
      - LAION2B-EN の地理的プロファイリングにより、地理的偏りが大きく、富裕国の方がデータ量が多い地域ではデータ分布が大きく偏っていることが判明





  - 言語抑圧: 少数民族言語の消去
    - Paper: Geographical Erasure in Language generation (EMNLP 2023)
      - 多くの国が地理的消去に苦しんでいる
    - Paper: WorldBench: Quantifying Geographic Disparities in LLM Factual Recall (right image)
      - 大規模言語モデル (LLM) が特定の国に関する事実情報を思い出す能力を評価
    - Paper: Richer Output for Richer Countries: Uncovering Geographical Disparities in Generated Stories and Travel Recommendations
      - 旅行の推薦やストーリー生成などのアプリケーションに関する地理的消去の問題を明らかにする

CVPR 2025 の動向・気付き (96/181)

Workshop: VLMs-4-All 2025

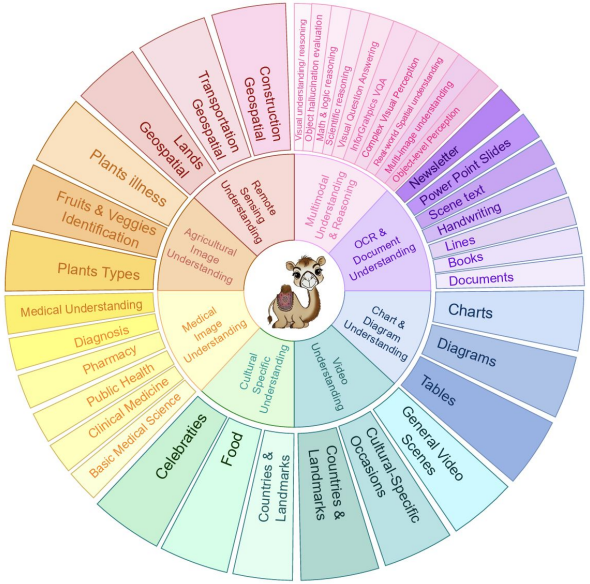
- Building Culturally Aware Multilingual LMM Benchmarks (1/2) (Fahad Shahbaz Khan)
  - 取り組み1: 多言語マルチモーダル理解ベンチマーク
    - Paper: All Languages Matter: Evaluating LMMs on Culturally Diverse 100 Languages (CVPR2025) (left image)
      - 文化に配慮した多言語画像 LMM ベンチマーク
      - リソースの少ない言語を含む100言語、ネイティブ言語の専門家(800時間の注釈)
      - ベンチマークの対象となるのは、遺産、習慣、建築、文学、音楽、スポーツなど、13の文化的側面
      - 人による検証による AI 生成による QA による生成パイプライン
    - Paper: A culturally-diverse multilingual multimodal video benchmark and model (right image)
      - 課題: ほとんどの動画データセットは英語に集中しており、モデルはリソースの少ない言語に苦労している
      - 提案: 文化に配慮した多言語動画 LMM ベンチマーク

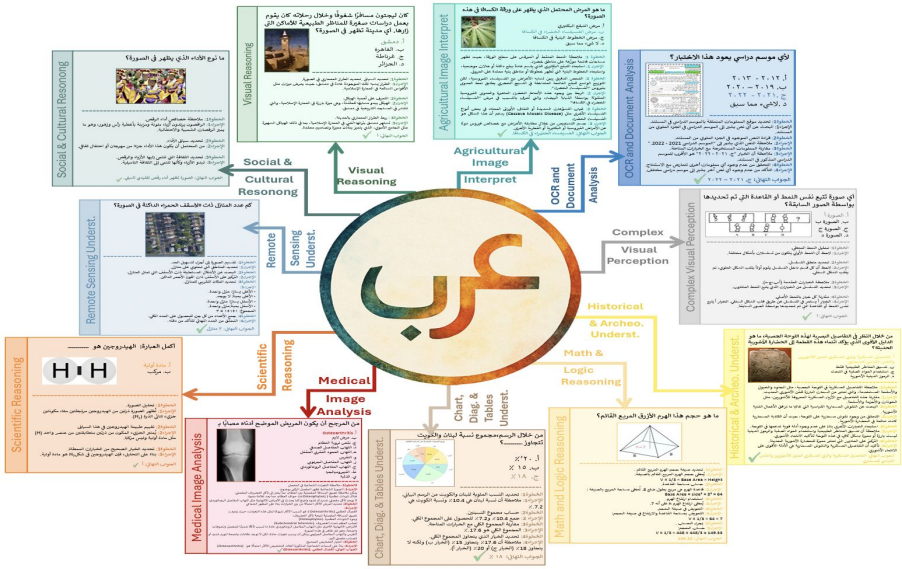
Benchmark	Multilingual	# of Languages	# of Scripts	Total Domains	Total Samples	Question Types	Question Forms	Annotation Type	Cultural Specific Content	Open-Source	Bias Correction	Diversity
Henna [4]	✗	1	1	5	1132	SVQA, LVQA	-	Auto+Manual	✓	✗	✓	✗
CulturalVQA [31]	✗	1	1	5	2378	SVQA	-	Manual	✓	✗	-	✗
GD-VCR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

CVPR 2025 の動向・気付き (97/181)

Workshop: VLMs-4-All 2025

- Building Culturally Aware Multilingual LMM Benchmarks (2/2) (Fahad Shahbaz Khan)
  - 取り組み2: アラビア語LMMベンチマーク
    - Paper: CAMEL-Bench: A Comprehensive Arabic LMM Benchmark (left image)
      - CAMEL-Bench は、マルチ画像理解、複雑な視覚認識、手書き文書理解、動画理解、医療画像処理、植物病害、リモートセンシングに基づく土地利用理解など、8つの多様なドメインと38のサブドメインで構成
      - アラビア語のユーザー数は4億人を超えていますが、クローズドソースのGPT-4oでも総合スコアは 62% に到達
    - Paper: ARB: A Comprehensive Arabic Multimodal Reasoning Benchmark (right image)
      - テキスト形式と視覚的方法の両方にわたって、アラビア語の段階的な推論を評価するために設計された最初のベンチマーク
      - ARBは、視覚的推論、文書理解、OCR、科学的分析、文化的解釈など、11の多様な領域にまたがる





CVPR 2025 の動向・気付き (98/181)

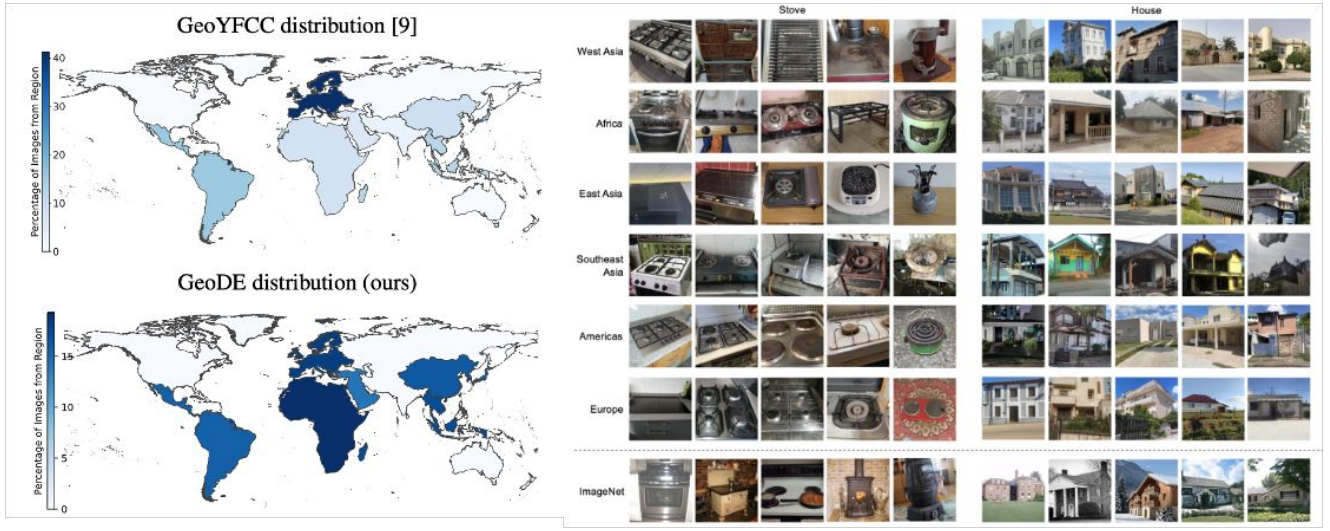
Workshop: VLMs-4-All 2025

- Challenges with Geo-Cultural Understanding and Generation (Roopal Garg)
  - 地質文化マルチモーダルコンテンツの理解と生成
  - 課題：現在のモデルでは、真にグローバルなユーザーベースでは不十分
    - 影響：世界中の何十億ものユーザーのアクセシビリティ、公平性、有用性、信頼に影響
  - データ品質ギャップ — 包括性と文化的特異性
    - 質の高いスケーリングされたデータセットが不足
    - データ生成 / 収集：食べ物、祭りなど、文化的特徴を浮き彫りにするデータを収集
    - ロケール参照を使用して、直接翻訳するのではなく、ロケール対応の人間による注釈を生成
    - インターネットから検索して正しいラベルを取得し、現地の人の注釈や検証を使用
  - 地理的文化理解を深めるための生成モデル
    - 問題：画像生成は文化に関連している可能性があり、英語翻訳からのテキストから画像への生成には複数の問題があり、文化的背景が異なる概念では精度が低い
    - モデル構造: 生成前の迅速な展開 (ローカル言語の取得とターゲット言語のコンテキストに基づく変更)

CVPR 2025 の動向・気付き (99/181)

Workshop: VLMs-4-All 2025

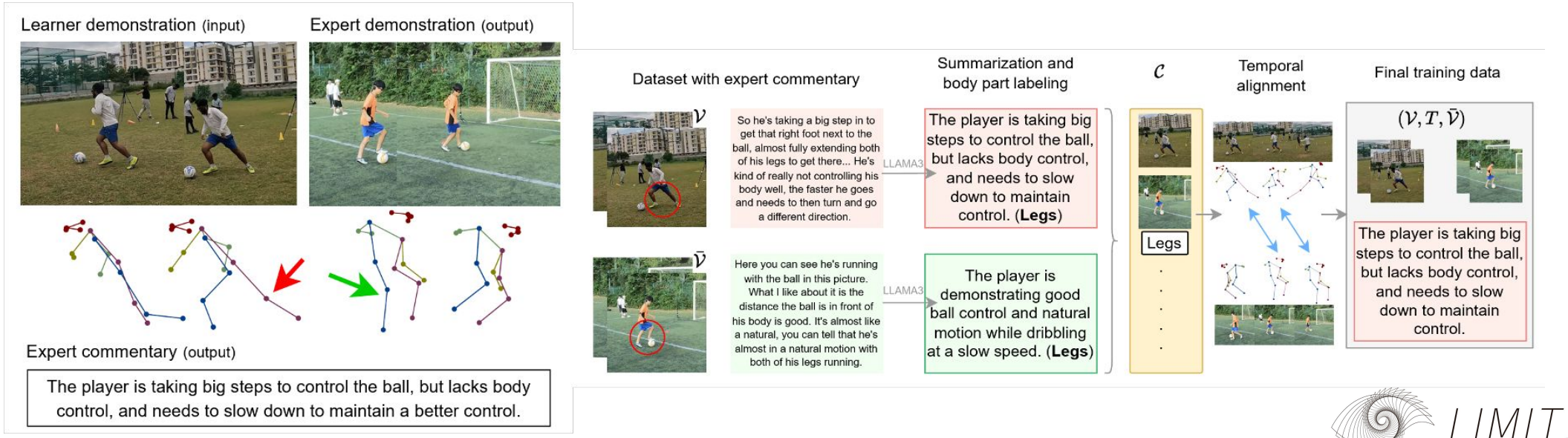
- Building geo-diverse model (Olga Russakovsky)
  - 課題: データセットにおける文化的多様性の欠如
    - GeoDE: Geographically Diverse Evaluation Dataset (NeurIPS2023) (左の画像)
      - Appenと提携して世界中の人々から写真を募集
      - 6つの異なる地域からの61,940枚の画像
      - クラウドソーシングから画像を収集、同意ありの画像か認識できる人物がいない画像の場合に生成
      - 調査結果:
        - クラウドソースの画像は地域によって見え方が異なる(右の画像)
        - クラウドソース画像とWeb スクレイピング画像では特徴表現が異なる
        - GeoDEはモデル内のギャップを見つけることができる、たとえばCLIPは西洋以外の物体、建物の認識におけるステレオタイプやエラーを表示
        - GeoDE でのトレーニングはパフォーマンスを向上



CVPR 2025 の動向・気付き (100/181)

ExpertAF: Expert Actionable Feedback from Video

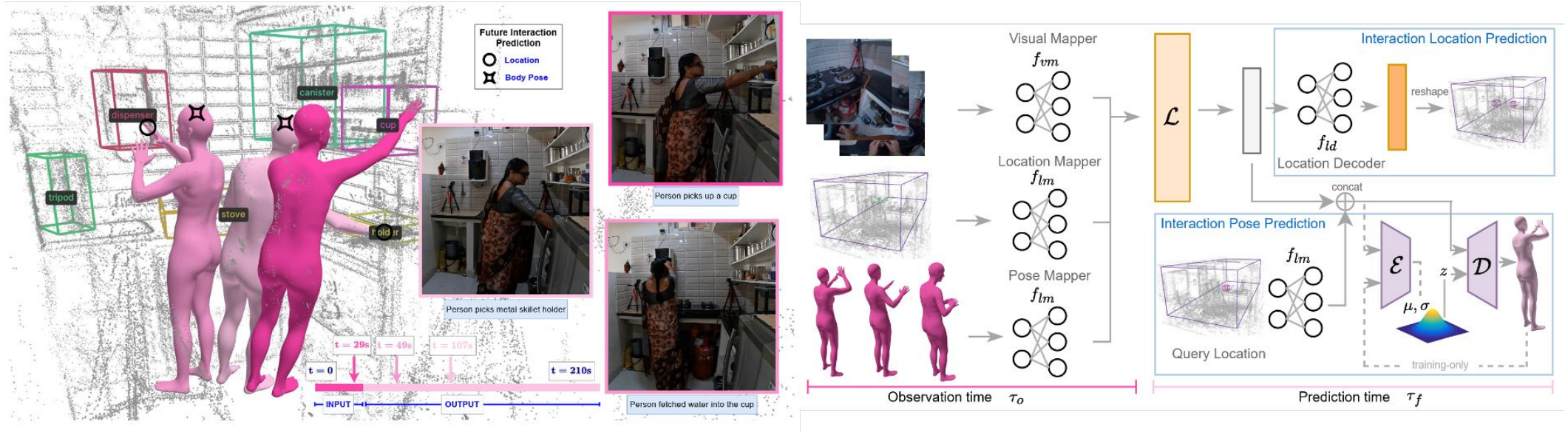
- **概要:** アマチュア動画(サッカーなど)とそれに対応するポーズ(骨格)データから、動画解説、動きの指示(テキスト)、正しい姿勢(骨格)を生成するタスクを提案。スキル学習用の Ego-Exo4Dデータセットから、類似したサンプル(アマチュア動画 +ポーズ、エキスパート動画 +ポーズ)をペアリングし、LLMを使用して動きのガイダンスを生成、それによって半自動的にデータセットを構築。シンプルな手法で、各入力用のエンコーダー、LLM ベースのモーションガイダンスジェネレーター、 Retrieval-Augmented Generator(RAG)ベースのポーズジェネレーターという個別のモジュールで構成される。
- **新規性:** タスク定義とデータセット。
- **気付き:** ポーズ出力部分がおもしろい。高品質で視覚的に意味のあるフィードバックを生み出すことを目指すタスクには、まだ多くの課題がある。メソッドとデータセットの両方に改善の余地がある。また、大量のデータを使ったトレーニングで正確なフィードバックが得られるかどうかも検討する価値がある。グループアクティビティのデモンストレーションを作成することは特に難しい場合がある。



CVPR 2025 の動向・気付き (101/181)

FICTION: 4D Future Interaction Prediction from Video

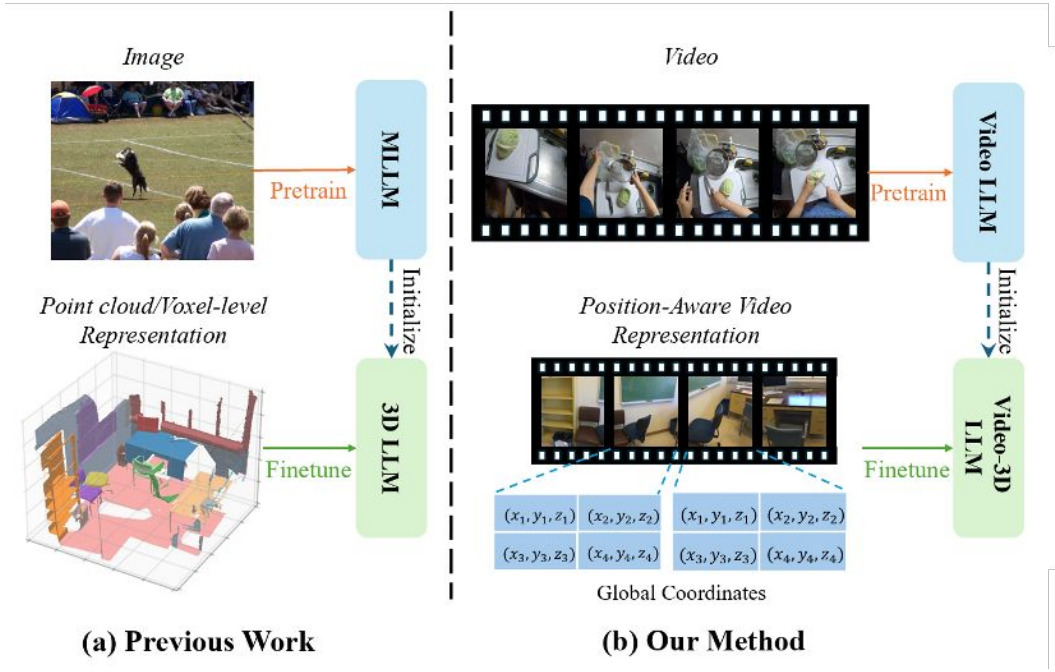
- **概要:** 動画から人物とオブジェクト操作を when, where, whichの観点で4D推定の方法を提案。FICTION では 2 つの VAE 構造を使用して、過去の観測データを埋め込むことから位置と姿勢をそれぞれ推定 (右図参照)。
- **新規性:** これまでの研究では動画から動画シーングラフを推定する方法が検討されてきたが、本研究ではそのタスクを　4D推定にまで拡張。
- **気付き:** 動画と3Dを組み合わせた研究が増加。Kristen Grauman Groupの論文では、特にタスク設定作業では、単純でわかりやすい方法が用いられることが多い。そのような場合は、まずわかりやすいアプローチでベースラインを確立する傾向がある。また、同じグループが他の　4D 生成タスクに取り組んでいるため、動画のみの表現から、動画 + 3D (つまり 4D) 表現へと移行しつつある。

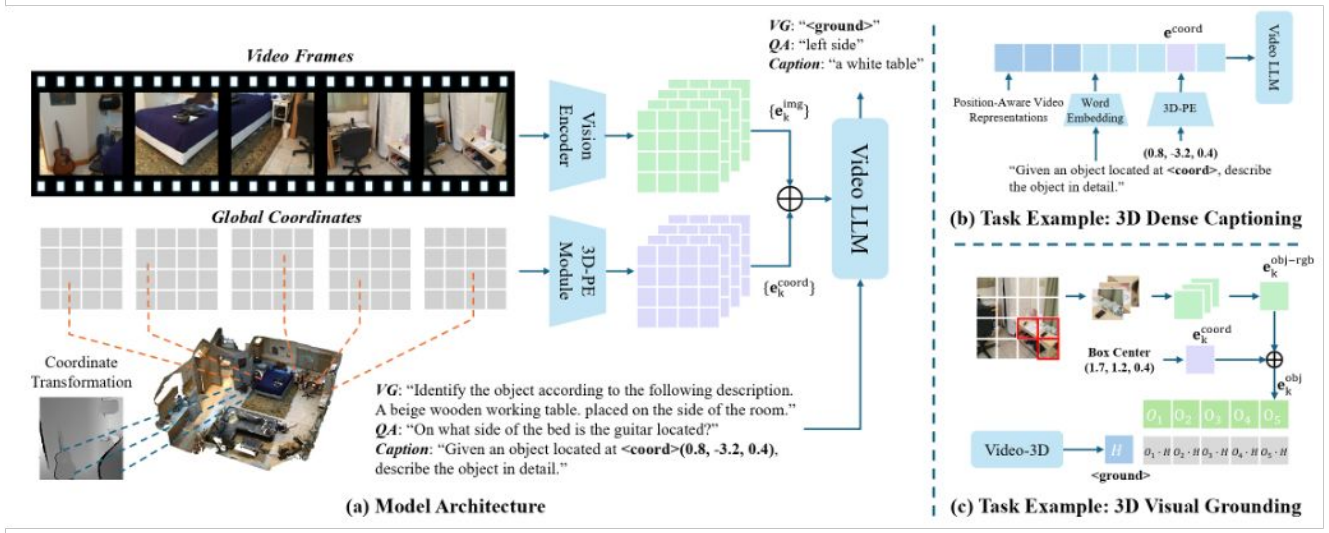


CVPR 2025 の動向・気付き（102/181）

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

- **概要**：本研究では *動画3D LLM*を提案。3D認識が可能な新しいMLLM（マルチモーダル大規模言語モデル）の手法である。動画表現を 3D 表現として扱う。このモデルは、複数の 3D 認識ベンチマークで高い精度を達成した。
- **新規性**：主な目新しさは、動画を3D情報として扱い認識し、MLLMがこのアプローチを使用して3Dを理解できるようになったこと。
- **気付き**：動画のみで3Dを知覚できることは非常に実用的。実際のアプリケーションでは、システムが動画からの 3D を認識できれば、明示的な 3D 入力は必ずしも必要ではない。動画のみで 3D を認識できるようになるので、この方法ははるかに使いやすく、適用しやすい。

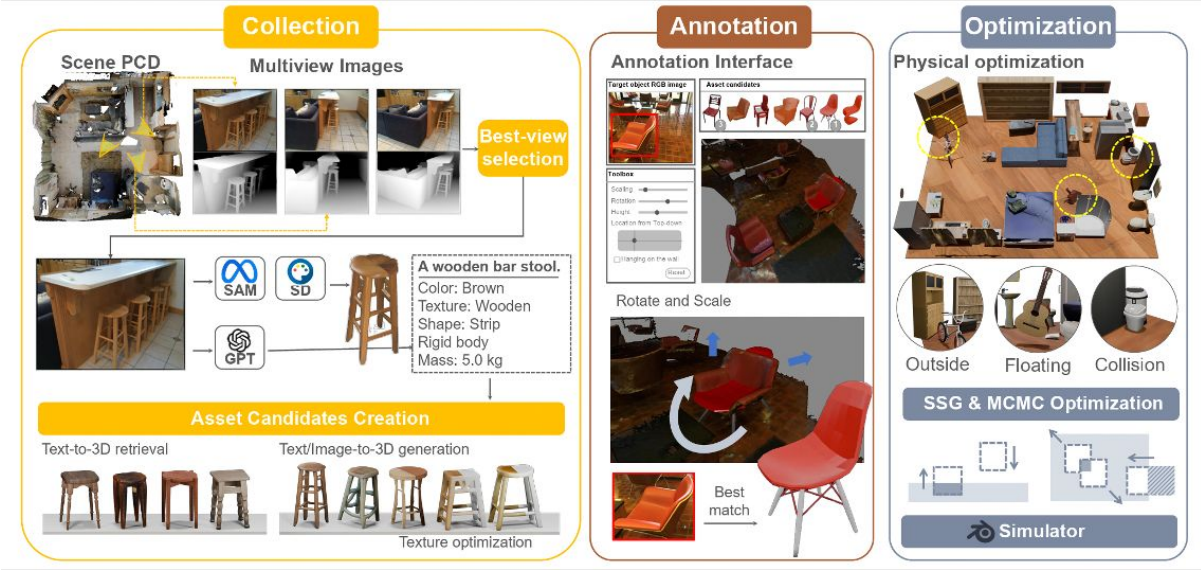


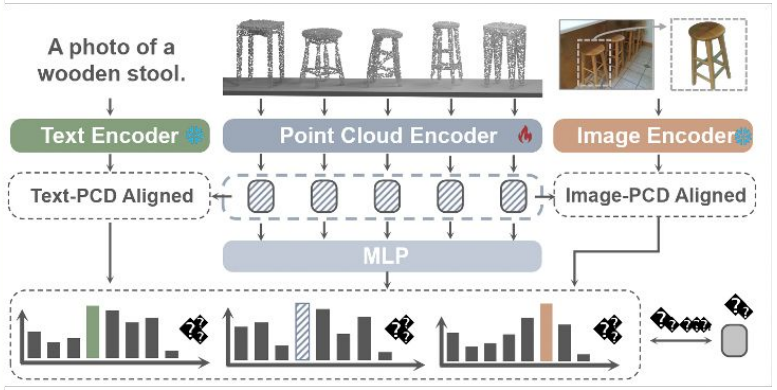


CVPR 2025 の動向・気付き (103/181)

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

- **概要:** 新規データセット METASCENES とScan2Simと呼称される検索ベースの手法を提案。METASCENES は ScanNet をベースに構築され、人によるアノテーションで補強、ScanNet アセットを CAD モデルに置き換えることで物理的な最適化によって強化。 METASCENES が精度の向上に有効であることは、実験によって実証。
- **新規性:** 本研究では、実際のスキャンデータを簡単にシミュレーションデータに変換するパイプラインが導入されている。屋内環境の　3D データセットを生成するコストが大幅に削減される。
- **気付き:** 屋内の3Dシミュレーションデータを自動的に作成する他の方法との比較を見るのは興味深い。テキストベースの生成方法の方が使いやすい場合もあると報告。

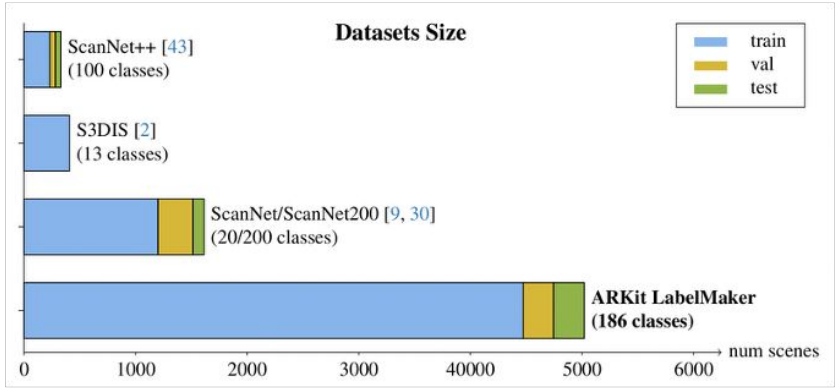


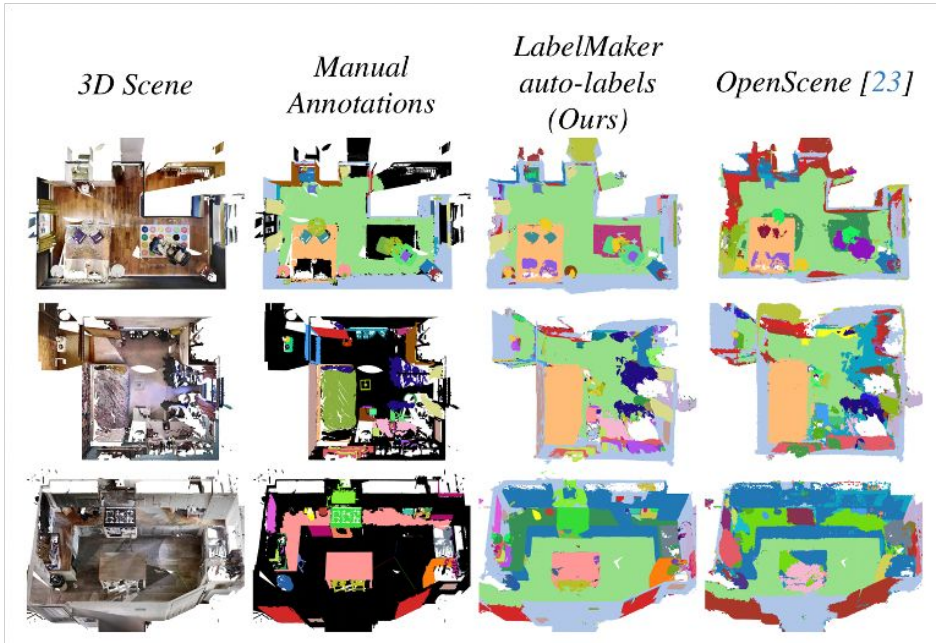


CVPR 2025 の動向・気付き（104/181）

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

- **概要：** *ARKit LabelMakerを提案*、3D空間におけるラベル付を効率化して既存のデータセットの 3倍を超える大規模3Dセマンティックセグメンテーションデータセットを構築。さらに、複数の高性能モデル（ OVSeg、Grounded-SAM、InternImage、Mask3D）を活用して、ARKit LabelMaker を生成するための自動パイプラインも導入。ARKit LabelMakerで事前にトレーニングされた方法は、既存のデータセットでトレーニングされた方法よりも優れる。
- **新規性：** データ生成のさらなるスケーリングを可能にする自動パイプラインと並び、データセット規模も重要なファクターであることを実証。
- **気付き：** このアプローチで言語と3Dデータを一緒にスケーリングできれば、パフォーマンスはさらに向上する可能性がある。

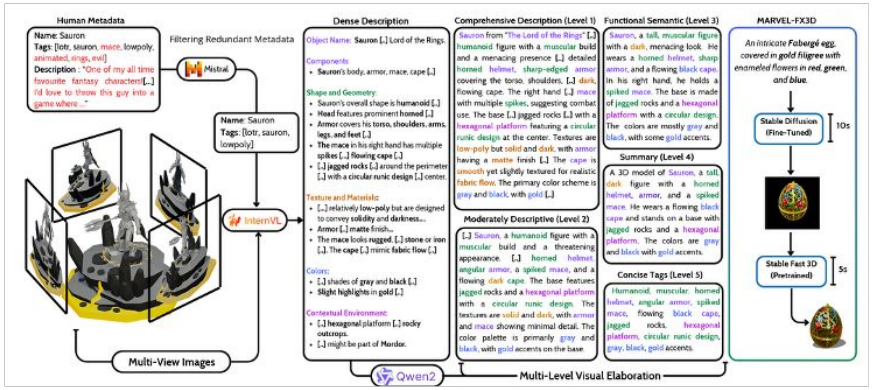




CVPR 2025 の動向・気付き (105/181)

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

- **概要:** 本論文では*MARVEL-40M+*、テキストと3Dモデルの大規模なデータセット(4,000万のテキストエントリと890万の3D形状)を提案。マルチモーダルLLMを使用して各3Dモデルのタグと詳細な説明を生成する自動データセット生成パイプラインが導入されている。さらに2 ステージの手法を提案、ステージ1:text-to-image、ステージ2:image-to-3D。
- **新規性:** 詳細な3Dアノテーションと3Dモデル生成の両方を実行できること。強力な実験結果が、このアプローチをさらに裏付けている。
- **気付き:** 3D prompts (バウンディングボックス、セグメンテーション、テキスト) と 3D 認識を組み合わせたタスクが興味深い。

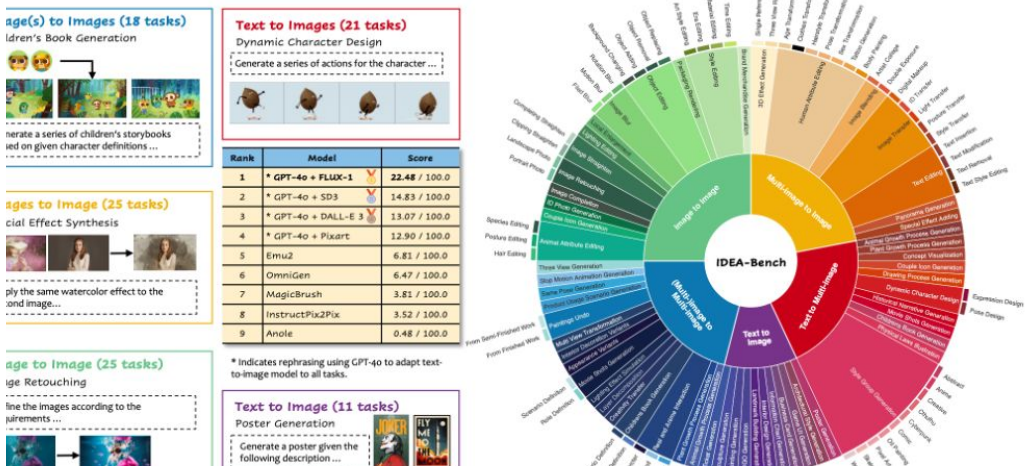


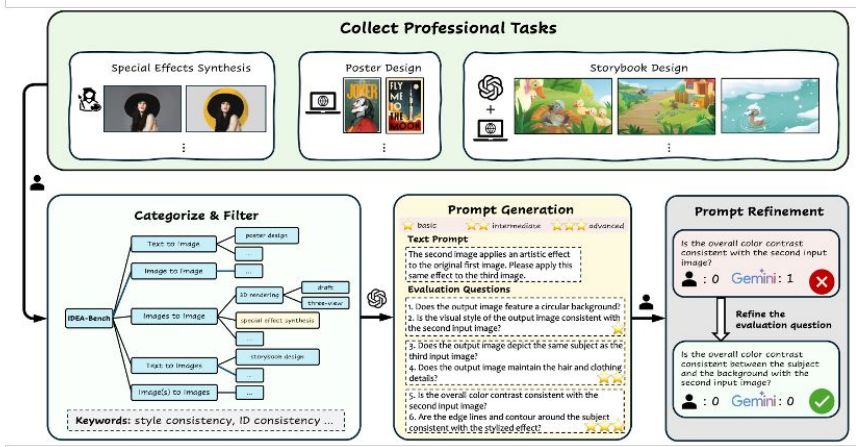
CVPR 2025 の動向・気付き（106/181）

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

- **概要：** IDEA-BENCHと呼ばれる新しいベンチマークを提案。これは、現在の生成モデルが「プロのデザイナー」環境で設計作業をどの程度厳密に実行できるかを調べるためのもの。このベンチマークには、テキストから画像、画像から画像、画像への変換、画像編集の各シナリオにわたる100のタスクが含まれ、いくつかの著名なモデルの評価に使用された。この調査では、設計上の問題を解決するように求められた際に、これらのモデルが示す欠点を詳細に分析している。

- **新規性：** IDEA-BENCHは、実際の設計タスクを複数の軸に沿って評価する。 DALL-E 3やInstructPix2Pixのような強力なシステムでも、このベンチマークでは100点満点中約20点しか得られず、完全には文書化されていなかった体系的な弱点が明らかになっている。
- **気付き：** モデルが人間の専門知識にどれだけ近いかを測定する評価タスクが一般的になりつつある。データセットが重大な欠陥を確実に明らかにできるのであれば、適度なデータセットサイズで十分。生成モデルのロバスト評価スキームを開発することは、重要な研究。

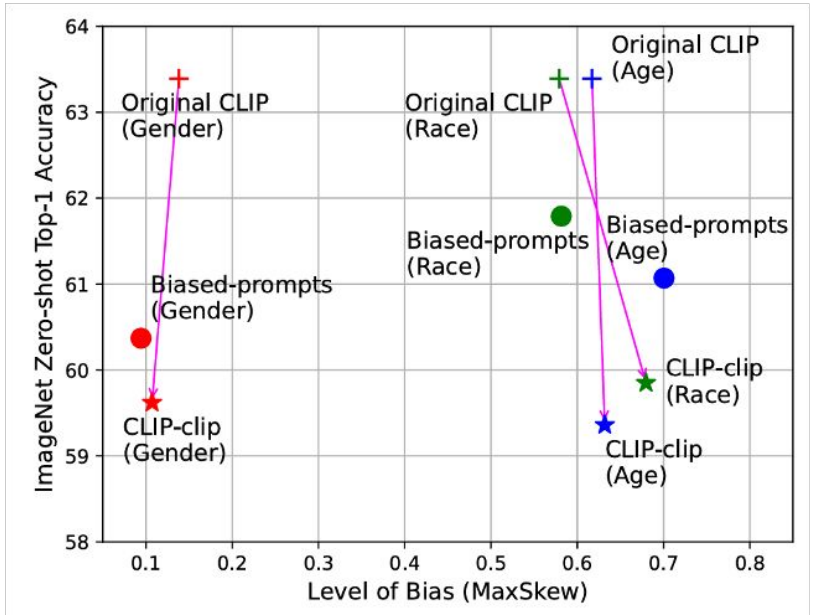


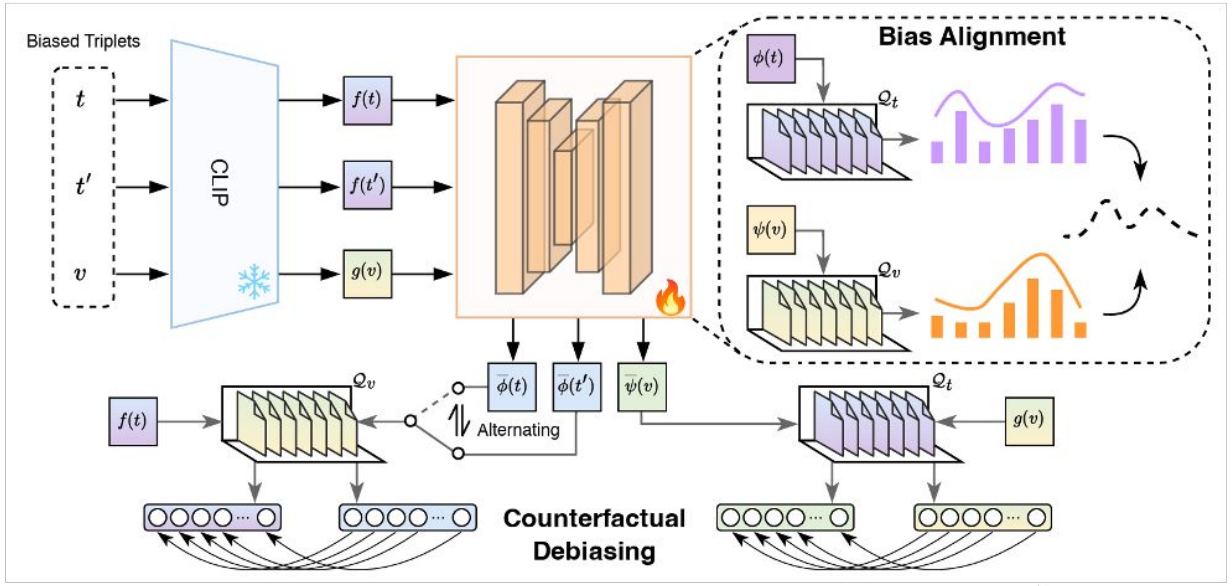


CVPR 2025 の動向・気付き (107/181)

Joint Vision-Language Social Bias Removal for CLIP

- **概要:** CLIP 埋め込みから性別、人種、年齢のバイアスを取り除く方法を導入。埋め込みから偏った属性を取り除く従来のアプローチでは、下流のタスクパフォーマンスが低下することがよくある。包括的分析を通じて、CLIPの画像ブランチとテキストブランチの両方でバイアスを取り除き、下流のタスクの精度を大幅に維持しながら強力なバイアスを軽減する手法を提示。
- **新規性:** CLIPのバイアス除去は、その重要性にもかかわらず、まだ比較的十分に検討されていない。本論文では、既存の方法を徹底的に検討、その限界を克服し、公平性とタスクパフォーマンスのバランスをとるアプローチを提案している。
- **気付き:** 大規模な(マルチ)モーダル言語モデルにおけるバイアスの調査は喫緊の課題であり、この貢献によりターゲットを絞ったバイアス除去戦略の難しさと可能性の両方を浮き彫りにした。

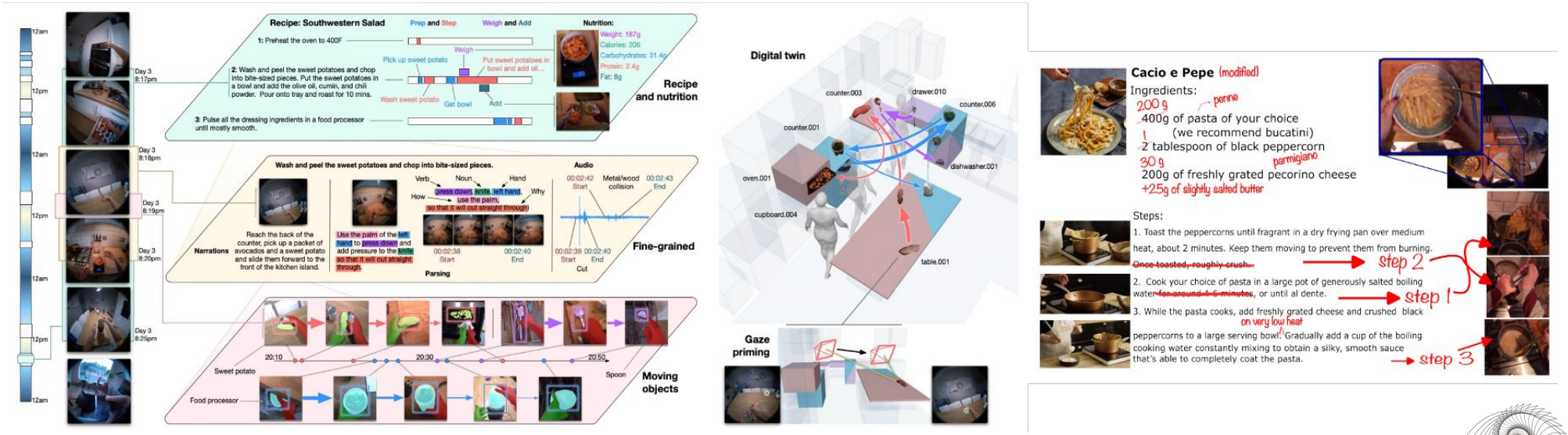




CVPR 2025 の動向・気付き（108/181）

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

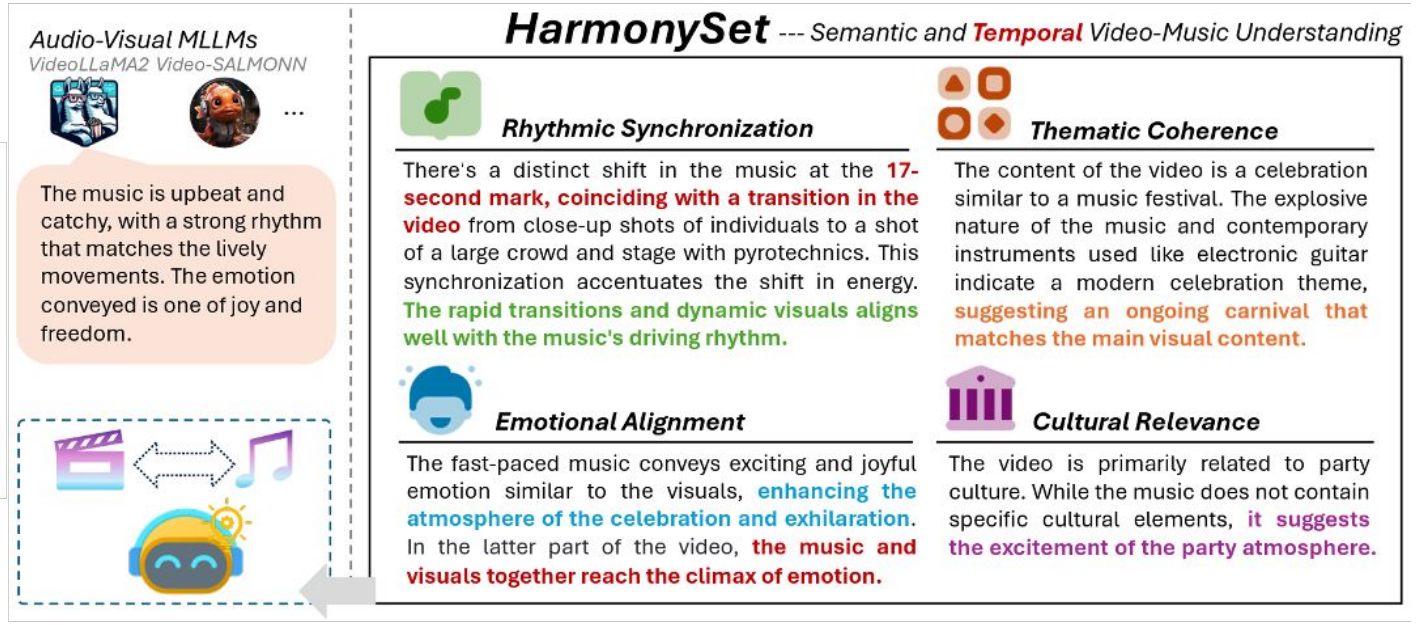
- **概要:** 本研究では、レシピステップ、アトミックアクション、栄養価を含む成分リスト、オブジェクト追跡ラベル、音声注釈、および完全に調整された3Dデジタルツインデータが豊富なきめ細かな動画データセットであるHD-EPICが提示された。強力なGemini Proモデルでさえ、付属のベンチマークでは37.6%の精度しか達成しておらず、データセットの難しさを浮き彫りにしている。
- **新規性:** HD-EPICは、きめ細かな動画注釈と明示的な3Dツインを組み合わせることで既存のリソースを拡張し、視覚言語研究と体現AI研究のための統一されたテストベッドを構築する。2 次元動画と 3 次元幾何の緊密な連携は、重要な貢献である。
- **気付き:** この結果は、現在のマルチモーダル言語モデルがいまだに詳細な動画理解には不十分であることを明らかにしている。詳細な 3D アライメントされたラベルの作成には多大な労力がかかるが、結果として得られるデータセットは、動画、ロボティクス、および現実世界の推論といった幅広いタスクにて使用可能な状態になる。



CVPR 2025 の動向・気付き (109/181)

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

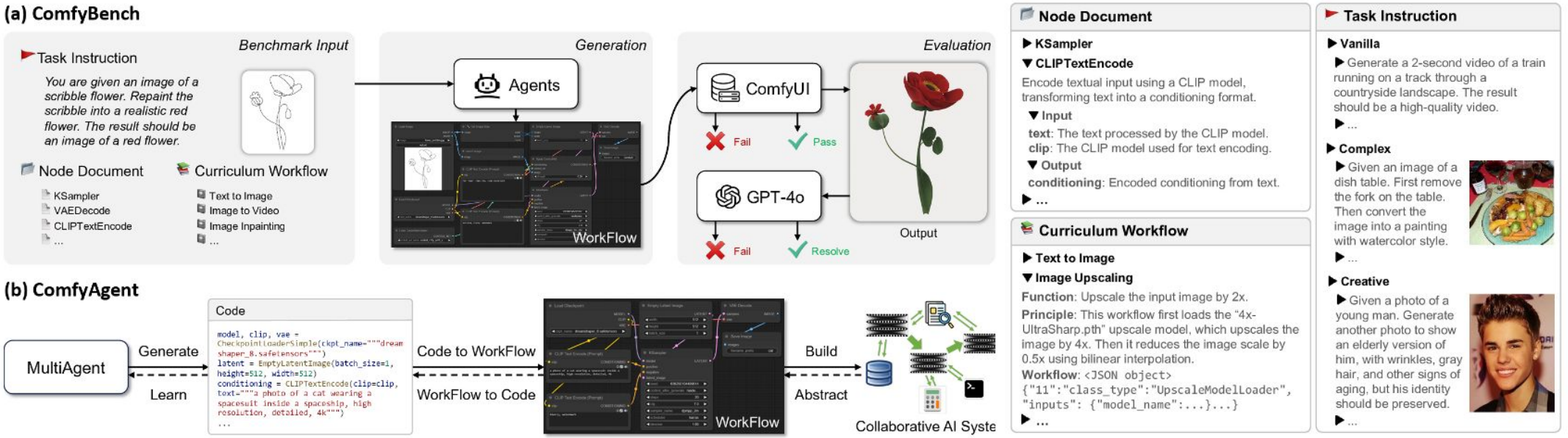
- **概要**: HarmonySetは、音楽と動画の共同理解に関するマルチモーダル言語モデルのトレーニングと評価のためのインストラクションチューニングデータセットとして導入された。このデータセットは、リズム同期、感情の調整、テーマの一貫性、文化的関連性の評価に使用される。すべての注釈は既存のMLLMを利用して自動的に生成されるため、広範囲にカバー可能。
- **新規性**: MLLMの音楽的意味論と視覚的意味論の融合を明示的に対象としたベンチマークはほとんどない。 HarmonySetはこのギャップを解消し、従来のオーディオビジュアル・アライメントを超える新しい評価軸を提供。
- **気付き**: 音楽やデザインなど、主観的な判断が評価を困難にする難しい分野のデータセットや指標を構築する傾向がある。 HarmonySetは、その方向への第一歩である。



CVPR 2025 の動向・気付き (110/181)

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

- **概要:** ComfyBenchと呼称されるベンチマークとComfyAgentというデータに付随する手法が導入され、高次のタスクが与えられると、必要なツール、ステップバイステップワークフロー、および中間プロンプトが自動的に生成される。生成された「フロー」は編集しやすいコード形式で表現され、ComfyAgent は O1-Preview システムに匹敵する精度に達する。
- **新規性:** ビジュアルプログラミングのアイデアを新しいツールエコシステムに拡張するもの。コンセプトの飛躍は控えめだが、 O1-Preview のパフォーマンスに匹敵することは注目に値する。
- **気付き:** 事前定義済みのツールセットとLLMで生成されたプロンプトを組み合わせてエージェントを構築することは、重要なトレードオフであり、現代のビジュアルプログラミングパラダイム。



CVPR 2025 の動向・気付き (111/181)

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

- **概要:** 本論文では、顕微鏡ベースの視覚的質問応答データセットであるMicroVQAを紹介する。従来のVQAコーパスとは異なり、MicroVQAは観察、仮説生成、実験的検証といった科学ワークフロー全体を中心に構成されており、専門の生物学者によってキュレーションされている。ベンチマークでは複数のマルチモーダル言語モデルが評価されており、大規模なモデル構築できる点が利点である。
- **新規性:** MicroVQAは、科学的な推論ステップをデータ構造に組み込むことで、単なる質問と回答のペアに止まらない。評価の結果、ほとんどのエラーは誤った視覚認識に起因することが示されており、顕微鏡画像の解釈が依然としてボトルネックであることがわかる。
- **気付き:** 「AI for Science」のリソースは増加傾向にあり、MicroVQAなどの研究プロセス全体を反映したデータセットは特に価値がある。その結果、現在のモデルでは、ドメイン固有の画像(顕微鏡、リモートセンシングなど) の事前トレーニングが依然として不十分であることが示唆されている。

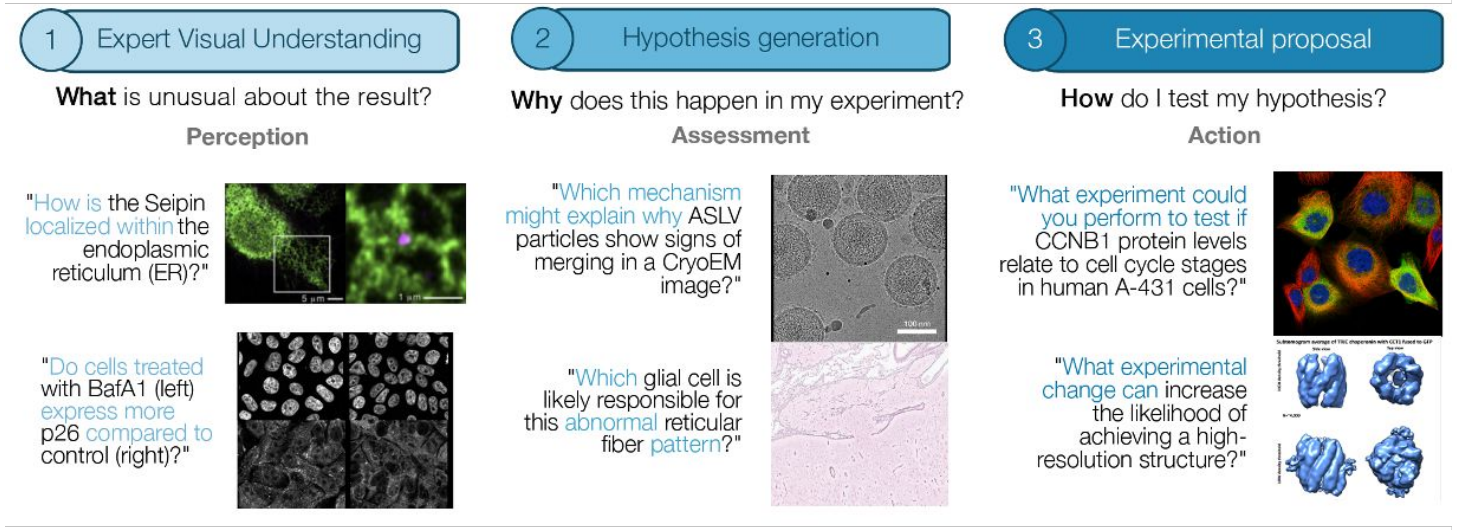


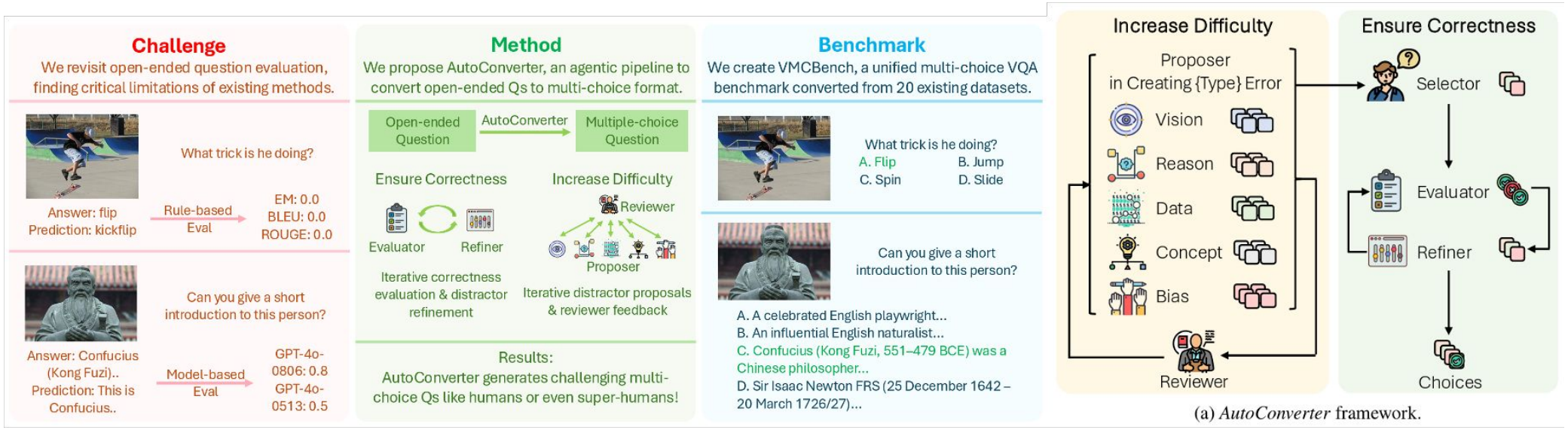
Table 1. MicroVQA benchmark attributes.

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

CVPR 2025 の動向・気付き (112/181)

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

- **概要:** 本論文では、既存のオープンエンド VQAデータセットを複数選択形式に自動的に変換する MLLMベースの方法 AutoConverter を提案。変換されたデータにより、より大規模な多肢選択式の VQA ベンチマークを作成できるようになり、マルチモーダル言語モデルの大規模な評価が容易になる。
- **新規性:** AutoConverterの主な貢献は、自由形式の VQAデータを多肢選択式の評価に転用したことにある。概念の進歩は比較的控えめだが、有用な手法である。
- **気付き:** 同じ考え方を動画 VQA にも応用できる。画像や動画のデータセットは豊富ですが、 3 次元認識の包括的なベンチマークは未だに不足 している。長い目で見れば、自分の知識のギャップを特定し、関連データを取得し、自律的に学習を続けることができる AI システムを開発することは価値がある。

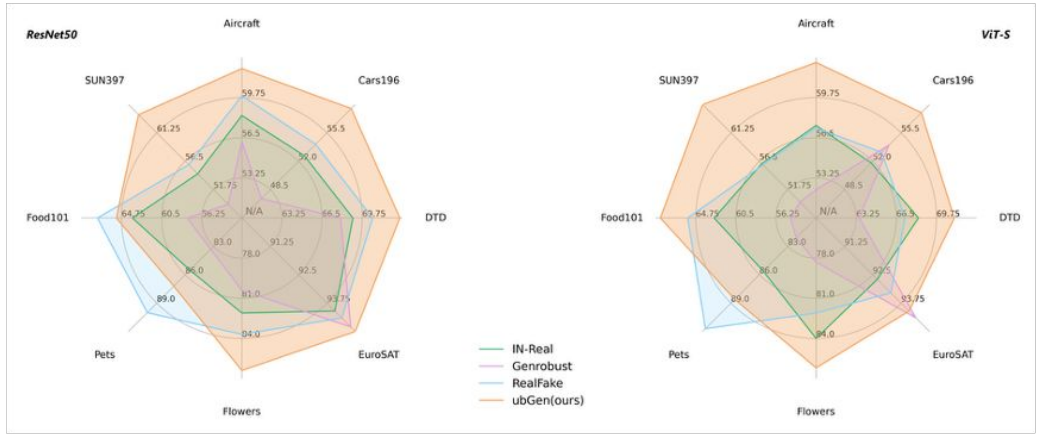


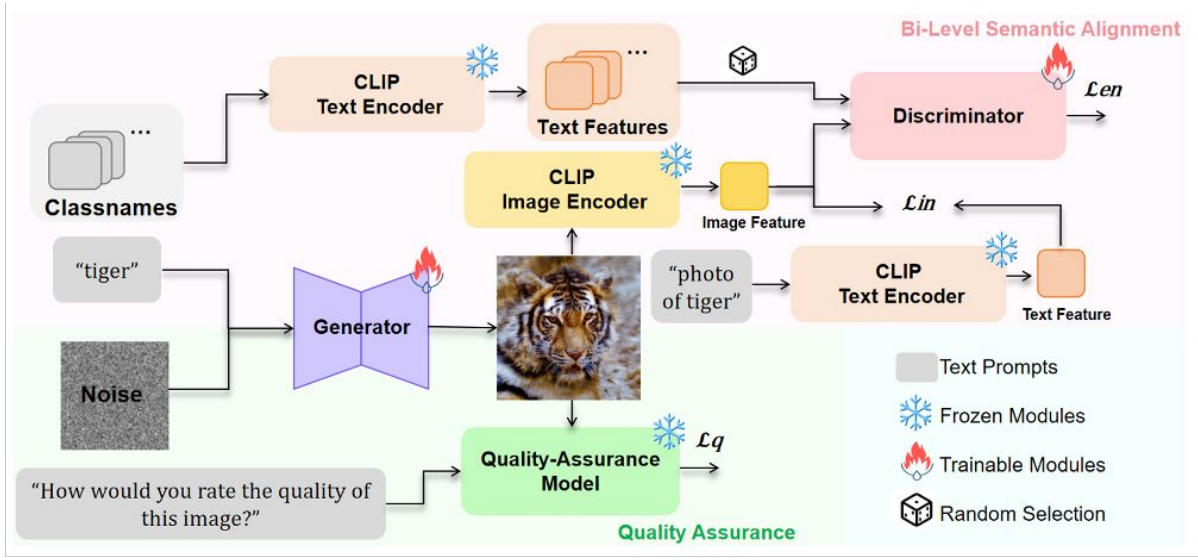
(a) *AutoConverter* framework.

CVPR 2025 の動向・気付き (113/181)

Unbiased General Annotated Dataset Generation

- **概要:** 画像認識モデルの一般化を改善するために、UBGenと呼ばれるデータ生成方法が導入された。このアプローチでは、CLIPのセマンティック空間で方向に沿って画像を生成し、品質–テキストアライメント機能を用いて忠実性と多様性の両方を確保する。UBGEN で拡張されたデータでトレーニングされたモデルでは、複数のベンチマークデータセットにわたって転送パフォーマンスが著しく向上する。
- **新規性:** UBGenは、カテゴリ名だけを使用して転送性の高いトレーニング画像を作成できるため、大規模な手動収集に代わる軽量な代替手段となる。
- **気付き:** 多くの場合、MLLMを活用して拡散モデルの表現力を高める取り組みが勢いを増しているようで、UBGenはこの新たなトレンドにうまく適合している。

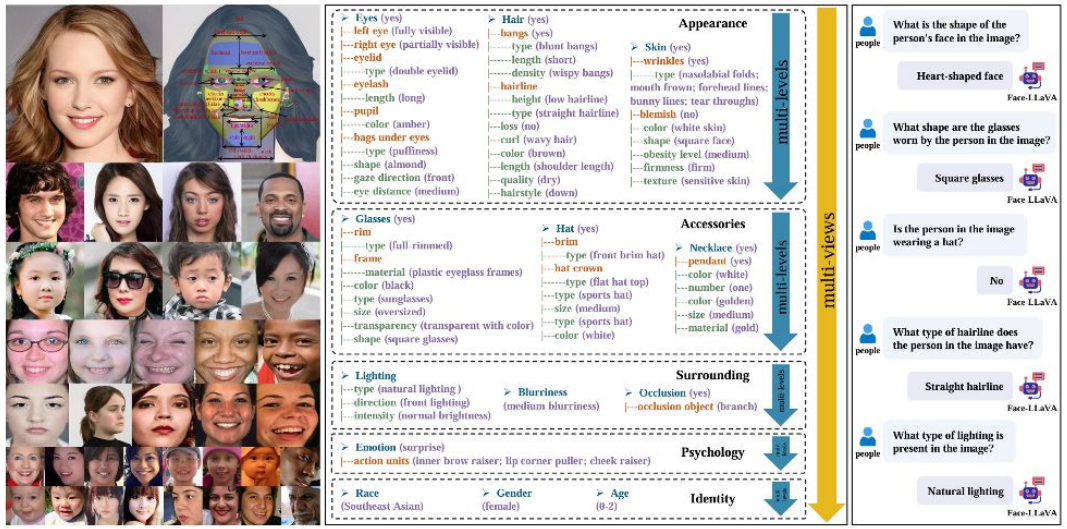


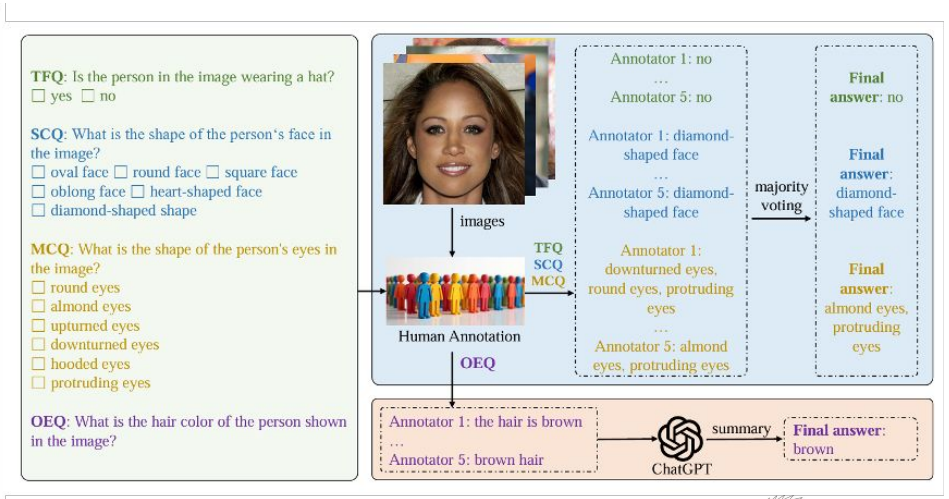


CVPR 2025 の動向・気付き (114/181)

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

- **概要:** FaceBenchという名前のデータセットが、マルチビューできめ細かな顔認識のために導入された。このデータには、210 組の注釈付き顔属性と、70,000 組を超える質問と回答が含まれている。FaceBenchで事前トレーニングされたモデルは、ベンチマークでGPT-4OやGeminiに匹敵する精度を達成する。
- **新規性:** FaceBenchは、以前のリソースよりも顔属性の範囲が大幅に細かくなり、より詳細な認識と分析が可能になった。
- **気付き:** FaceBenchは静止した顔画像をターゲットにしているが、微妙な表情を認識するための補完的な動画コーパスも貢献。多くの分野で詳細なデータセットに対する需要が高まっているが、FaceBenchでトレーニングされたモデルが未だGPT-4oを超えていない。

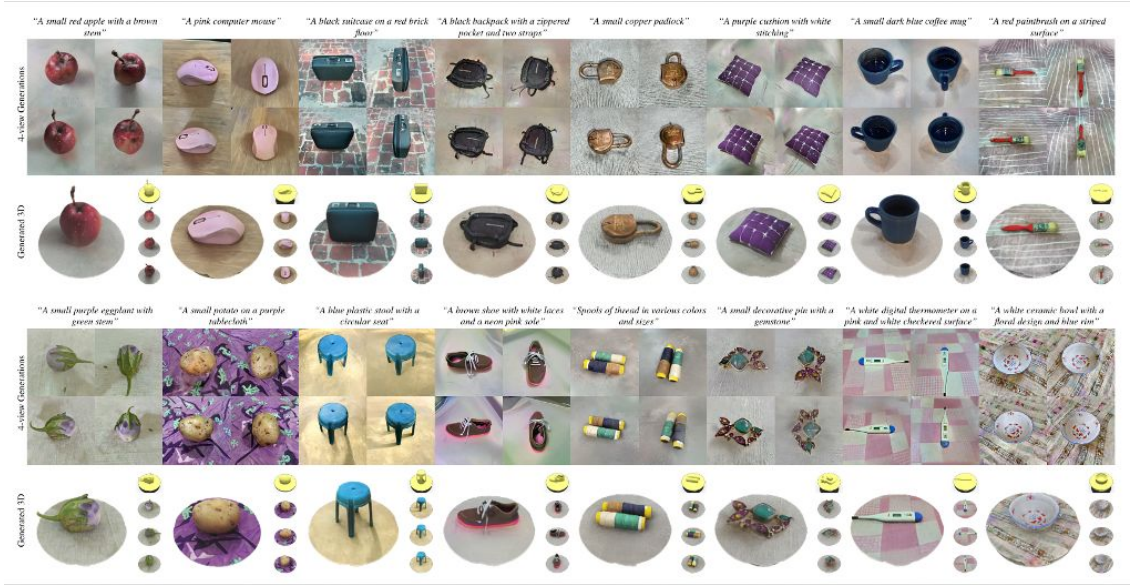




CVPR 2025 の動向・気付き (115/181)

UnCommon Objects in 3D

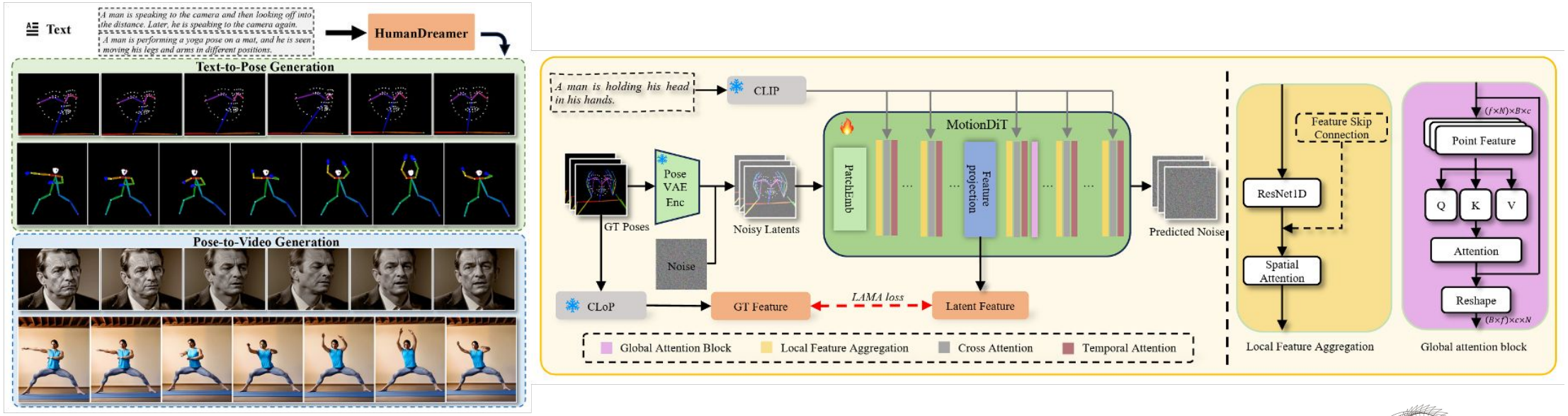
- **概要:** uCO3D という名前の大規模な3 次元物体データセットが導入された。1,000 を超えるオブジェクトカテゴリにまたがる各サンプルには、動きから見た正確な構造カメラポーズ、点群、3 次元のGaussian Splatting再構成が含まれている。実験によると、uCO3Dでトレーニングされたモデルは、MVImageNetやCo3Dv2でトレーニングした場合よりも、新規ビュー合成と3D再構築において著しく優れたパフォーマンスを発揮する。
- **新規性:** uCO3Dは、スケールとアノテーション品質の両方で既存のデータセットを上回り、以前のリソースよりも高密度のジオメトリとオブジェクトの多様性を実現した。
- **気付き:** 複雑な形状やテクスチャをテキストで記述することは依然として困難だが、このような忠実度の高いジオメトリと強力な言語教師を組み合わせることは、重要な次のステップになる可能性がある。



CVPR 2025 の動向・気付き (116/181)

HumanDreamer: Generating Controllable Human-Motion Videos via Decoupled Generation

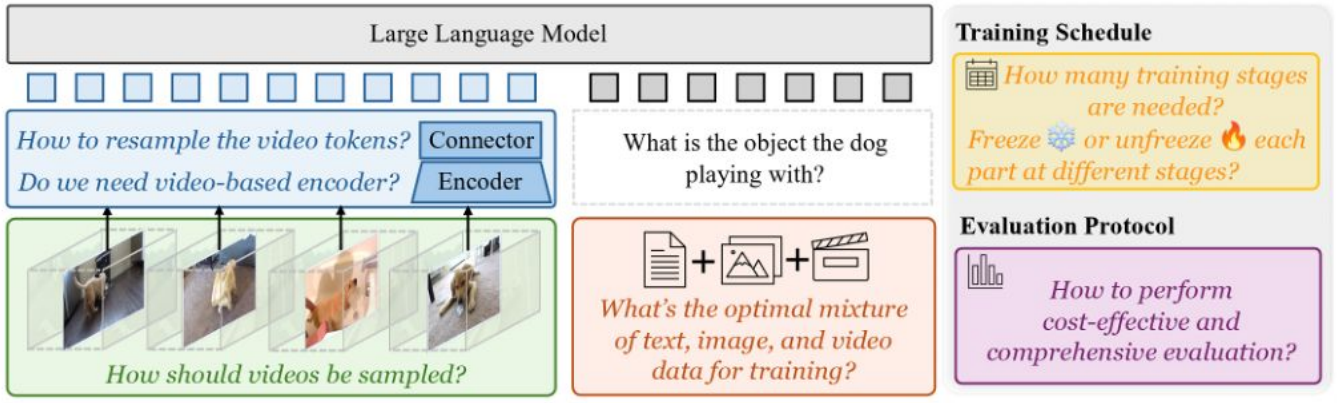
- **概要**: 本論文では、テキスト入力からポーズと対応する動画の両方を 2段階のプロセスで生成する拡散ベースの動画生成方法を提案する。最初にテキストからポーズを生成し、次に予測されたポーズから動画生成する。テキストからポーズまでの段階を可能にするために、半自動で構築された大規模なデータセットを紹介する。さらに、予測されるポーズとポーズの特徴をより正確に一致させるために、 text-to-pose モデル用の新しい損失関数 LAMA loss が提案されている。
- **新規性**: 主な貢献は、テキストからポーズまでの大規模なデータセットの構築プロセスと公開、およびテキストからのポーズ予測に合わせた LAMA lossの導入である。
- **気付き**: テキストからポーズへの生成は、幅広い用途がある。この取り組みは、既存のモデルや大規模言語モデル (LLM) を使用してデータセット構築を自動化する傾向が高まっている。

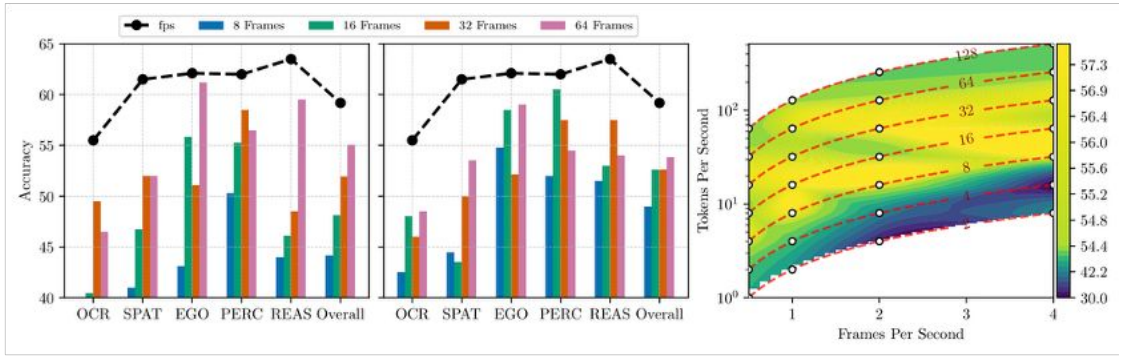


CVPR 2025 の動向・気付き (117/181)

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

- **概要:** この分析研究では、Video LLMの設計選択に関する包括的な実験的研究と比較を行い、最終的にApolloモデル・シリーズの導入に至った。Apollo は、計算コストが同等レベルのモデルよりも高い精度を実現している。
- **新規性:** 主な調査結果には、スケーリングの一貫性の観察が含まれる。比較的大規模なデータセットで見られるパフォーマンスの傾向は、さらに大きなデータセットでも引き続き維持される。さらにFPS ベースのフレームサンプリングは均一サンプリングよりも優れており、フレーム数に比例してパフォーマンスが向上することが示されるが、これは均一サンプリングには当てはまらない。出版時点では、InternVideo2 + SigLip SO400M 機能を使用したモデルが最良の結果を達成。
- **気付き:** この研究は広範囲にわたる実験を含み、動画認識モデルの設計に関する貴重な洞察を提供する。

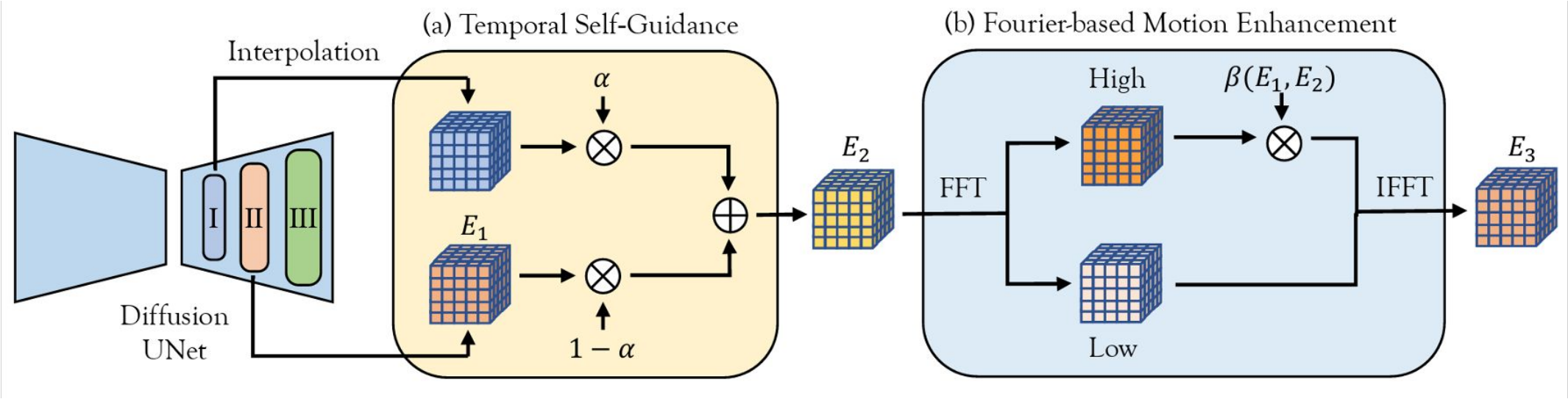




CVPR 2025 の動向・気付き (118/181)

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

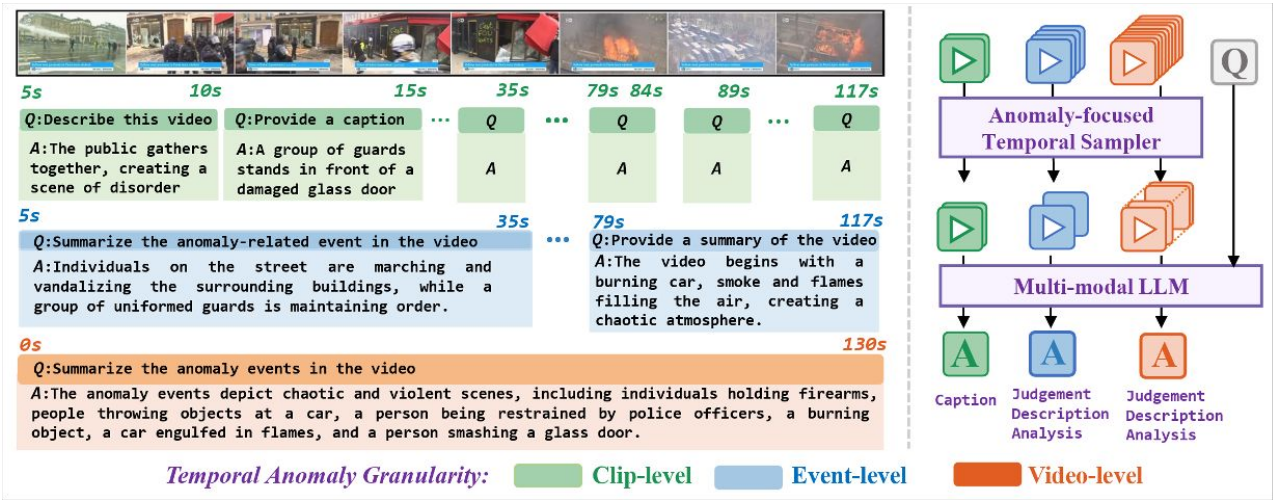
- **概要:** 本研究では、拡散ベースの動画生成における生成の一貫性と動きの大きさを改善するためのトレーニング不要の方法を紹介している。著者は時間的注意の分析を通じて、時間的注意マップ間の違いが時間的一貫性に強く影響することを突き止めた。これらのマップを調整して一貫性を高めることで、全体的な生成の安定性が向上する。さらに、時間的アテンションマップのエネルギーが運動のマグニチュードに影響することがわかっており、運動の強さを増幅する方法が開発されている。提案された手法は、 AnimationDiff や VideoCrafter2 などの既存のモデルの出力品質を大幅に向上させる。
- **新規性:** この研究では、時間的注意マップの包括的な分析が行われ、生成された動画の一貫性と動きのダイナミクスを高めるためのトレーニング不要の斬新な方法が紹介されている。
- **気付き:** 提案された方法が他の拡散ベースの動画生成モデルにも一般化できるかどうかを検証することは重要。

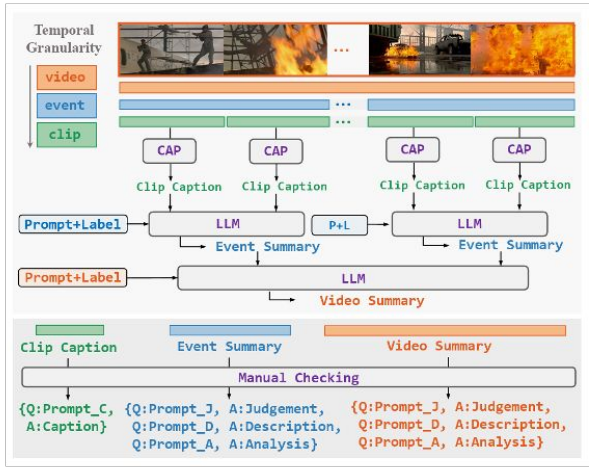


CVPR 2025 の動向・気付き (119/181)

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

- **概要: 複数の時間的粒度** をもち全域の動画の異常を認識するように設計された大規模なデータセット。データセットは、既存のマルチモーダル言語モデル (MLLM) を使用して半自動的に構築される。さらに、**temporal sampler** MLLMベースのモデルを使用して、さまざまな時間スケールでの異常検出を提案している。
- **新規性:** 主な貢献は、全体にわたる**さまざまな時間的粒度** の異常認識の設定にある。これは、標準的な固定スケールの動画分析とは比べて、新しい視点を提供する。
- **気付き:** さまざまな時間解像度で動画を分析するというアイデアは、明確に構造化されている。ただし、時間的ローカリゼーションのベンチマークのような関連データセットはすでに存在するため、新規性の程度は限られている可能性がある。

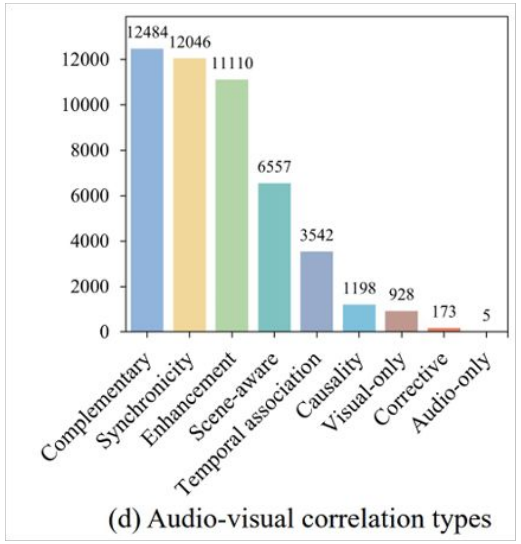




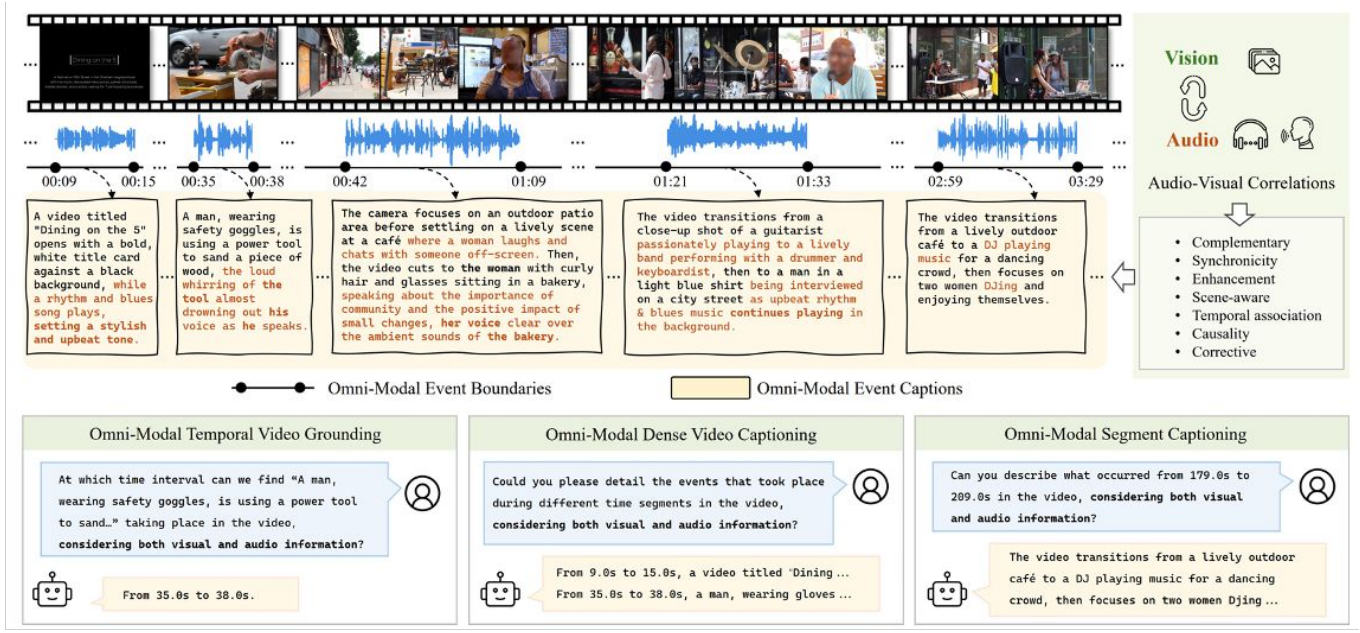
CVPR 2025 の動向・気付き (120/181)

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

- **概要:** 本論文では、視覚、音声、言語を統合するオムニモーダルモデルにおける包括的な理解を評価するために設計されたベンチマークであるLongValeを提案。LongValeで事前にトレーニングされたモデルは、オムニモーダル推論と詳細な推論タスクの両方で優れたパフォーマンスを発揮した。
- **新規性:** LongValeは、オムニモーダル認識の体系的な評価を可能にし、きめ細かな理解やクロスモーダル理解など、幅広い推論能力にわたる評価をサポートする。
- **気付き:** LongValeは、既存のモデルと自動化プロセスを組み合わせて構築された。最近の進歩にもかかわらず、画像や動画などの視覚データを詳細に理解することは、現在のモデルにとって依然として大きな課題である。



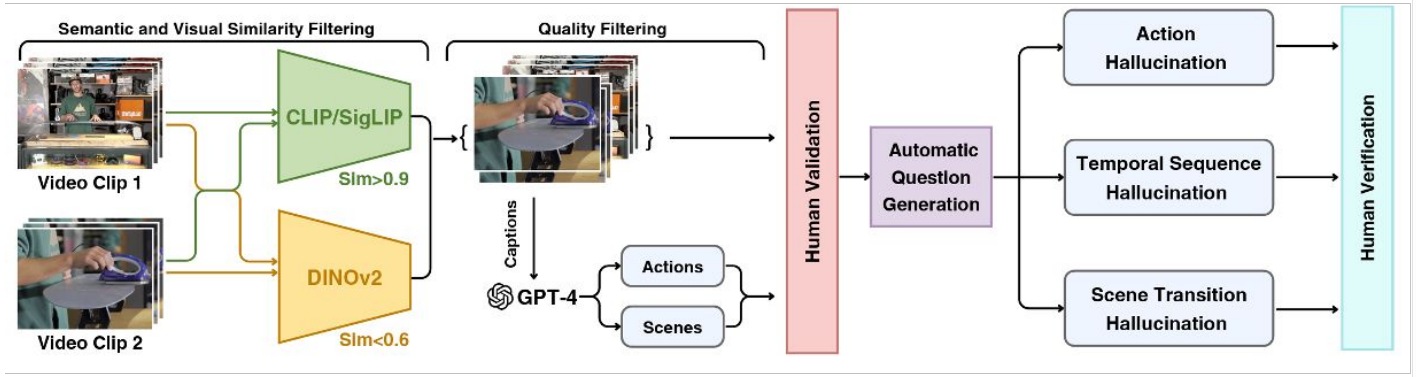
(d) Audio-visual correlation types

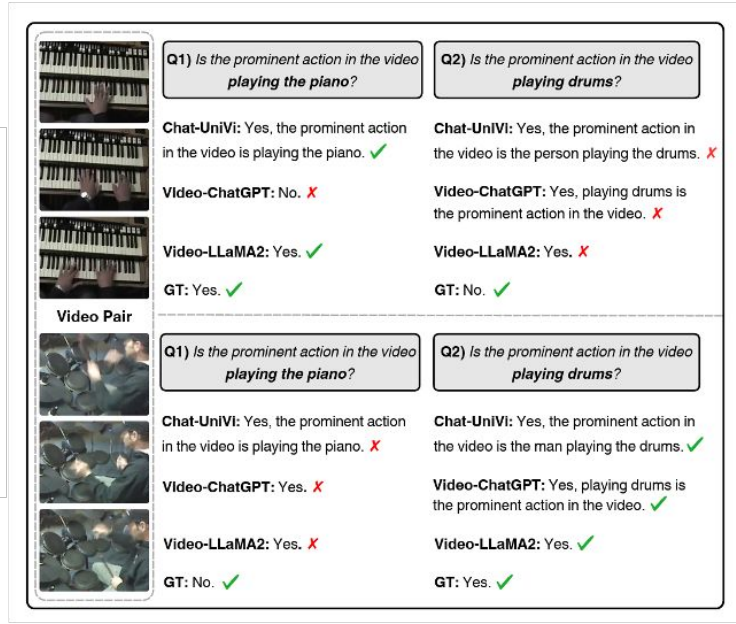


CVPR 2025 の動向・気付き (121/181)

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

- **概要:** 本論文では、動画大規模言語モデル（動画LLM）における時間的推論と時間的幻覚を評価するためのデータセットであるVIDHALLUCを提案する。データセットは、図 (左) に示すプロセスを通じて自動的に構築され、最終的な検証は人間のアノテーターが行う。
- **新規性:** VIDHALLUCを使用すると、現在の動画LLMが一時的な幻覚をどのように処理するかを包括的に評価できるため、人間の推論と比較して顕著なパフォーマンスギャップが明らかになる。
- **気付き:** 時間的推論と詳細な動画理解は依然として難しい分野であり、VIDHALLUCは、正確な時間的ダイナミクスを捉える上での既存のモデルの限界を浮き彫りにしている。

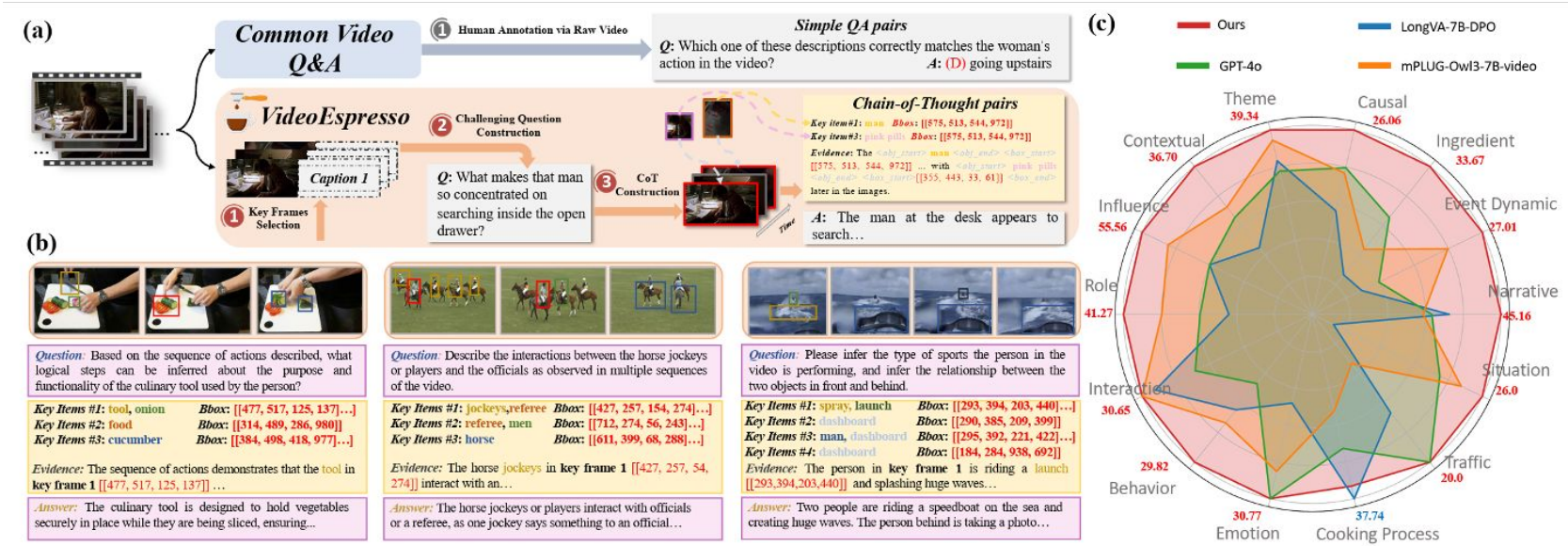




CVPR 2025 の動向・気付き (122/181)

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

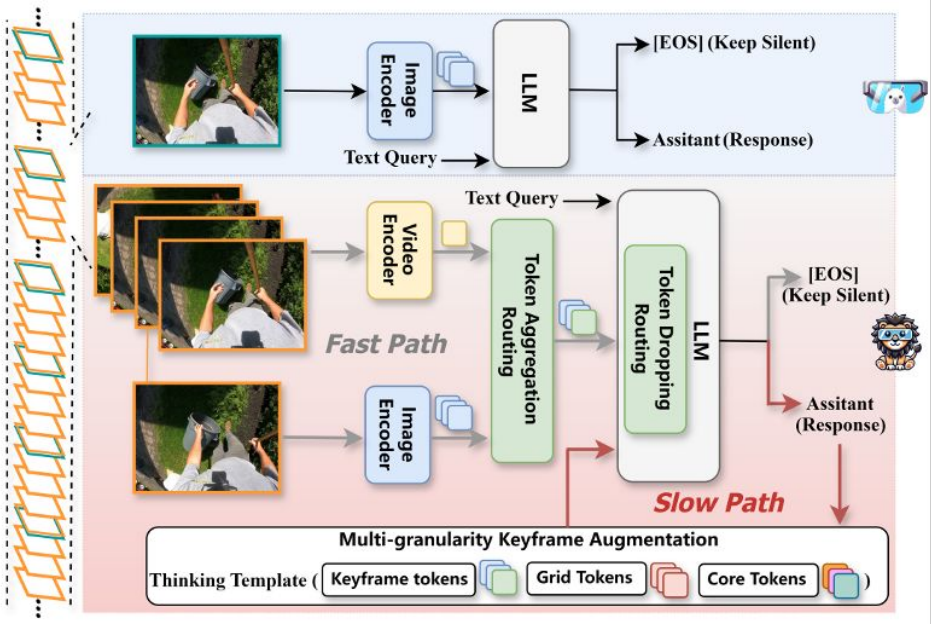
- **概要**: 本研究では、思考連鎖推論と既存のマルチモーダル言語モデルを組み合わせて自動的に構築された、きめ細かな動画理解を評価するためのベンチマークである VideoExpressoを紹介する。VideoExpressoでトレーニングされたモデルは、いくつかの既存のデータセットで最先端のパフォーマンスを実現した。
- **新規性**: 主な貢献として、詳細な動画認識への注力や、大規模なデータ生成における思考連鎖推論の使用などがある。これは、解釈可能性とトレーニングの質の両方を向上させるアプローチである。
- **気付き**: CVPRでは、この分野への関心の高まりを反映して、詳細動画認識のベンチマークがいくつか発表されている。高精度を達成するには、詳細なタスクに関する事前学習が不可欠であり、詳細な認識と思考の連鎖的推論の両方が未解決の課題である。

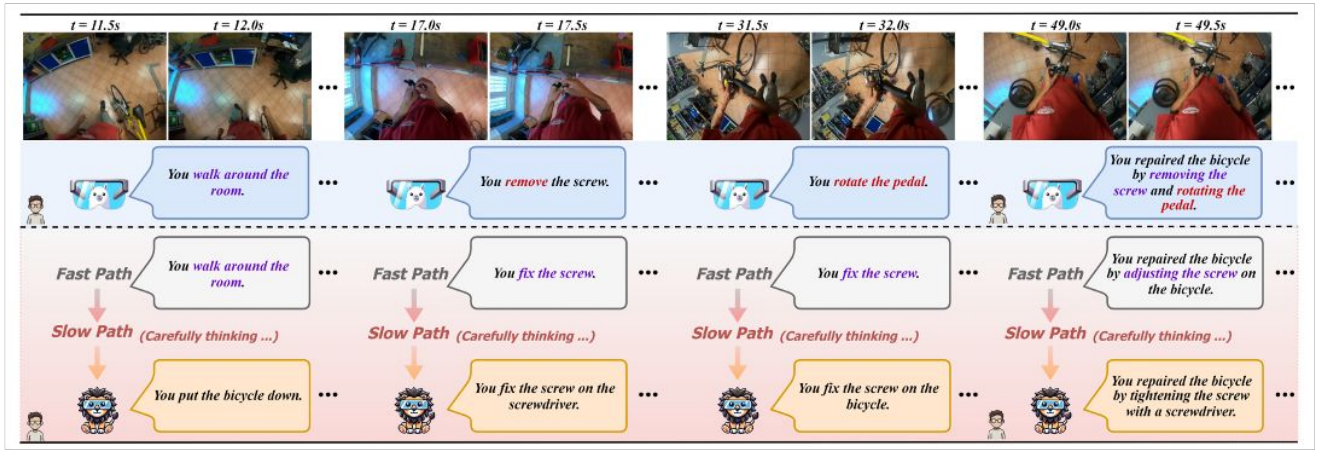


CVPR 2025 の動向・気付き（123/181）

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

- **概要：** 本論文では、特に一人称視点（FPV）動画のコンテキストにおいて、オンライン動画アシスタントの新しい方法である LION-FS（Fast-Slow）を提案する。ファストトラックは高フレームレート処理を使用して迅速な応答を提供し、スロートラックはより詳細な特徴を階層的に処理することにより、より深く詳細な推論を行う。
- **新規性：** もともとアクション認識に使用されていた SlowFastアーキテクチャをオンライン動画アシスタントの領域に適応させ、処理速度と詳細な理解の両方を可能にする。
- **気付き：** SlowFastのコンセプト自体は新しいものではないが、オンライン動画アシスタントへの応用は実用的で有望。このアプローチは他の動画認識タスクにも拡張できる可能性があり、検出をスロートラックに明示的に組み込むことでパフォーマンスをさらに向上させることができる。

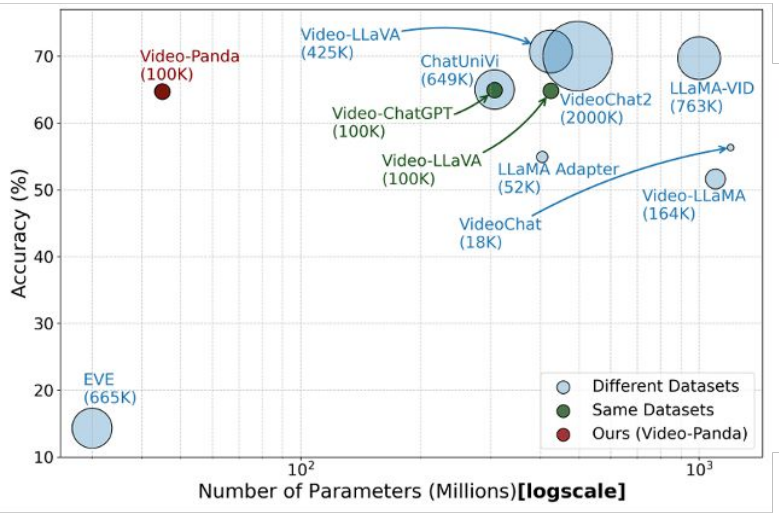


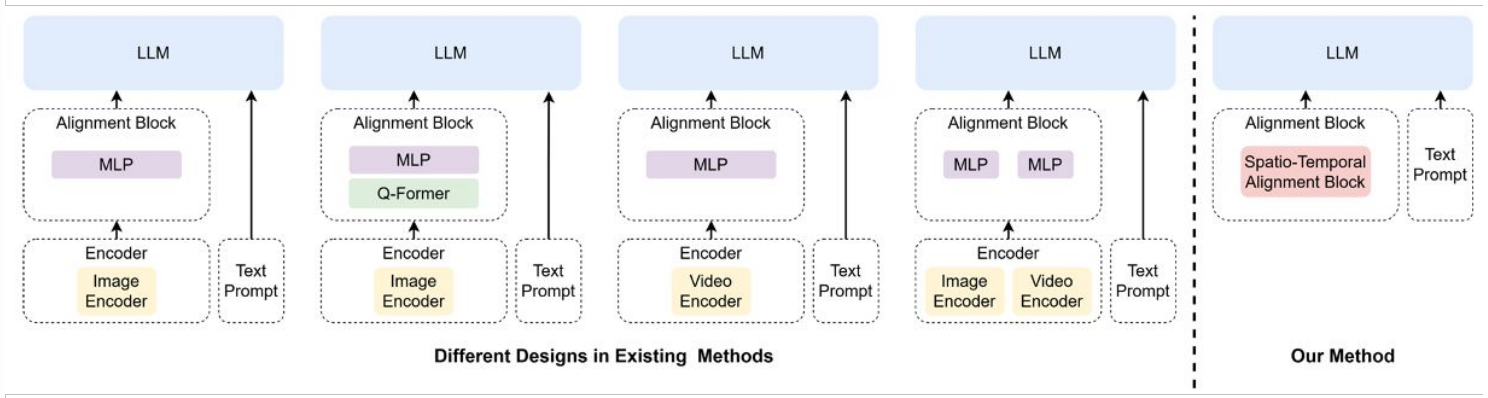


CVPR 2025 の動向・気付き (124/181)

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

- **概要:** 本研究では、エンコーダ不要の動画 LLM アーキテクチャである Video-Panda を紹介する。このモデルは、従来の動画エンコーダーを使用する代わりに、時空間アラインメントブロック (STAB) を使用して、きめ細かな空間的特徴と時間的特徴を動画入力から直接抽出する。Video-PandaはVideo-LlamaやVideo-ChatGPTなどのモデルよりも優れていますが、ビジュアルエンコーディングに使用するパラメーターはわずか4,500万個。
- **新規性:** 通常大型の動画エンコーダモジュールを排除したこと。その結果、軽量でありながら高性能のモデルが得られ、パラメータが大幅に少なくても高い精度を維持できる。
- **気付き:** Video-Pandaは非常にコンパクトだが、それでも複数のベンチマークで競争力のあるパフォーマンスを発揮する。このパフォーマンスを どのように達成するかをより包括的に評価すれば、特に同じようなサイズのモデルと比較すると有効性が確認できる。また、モデルがどのようにスケーリングされ、構成が大きくなるとパフォーマンスがどのように変化するかを調べることも重要。提案されている STAB モジュールは複雑に見えるため、さらに検討する必要がある。

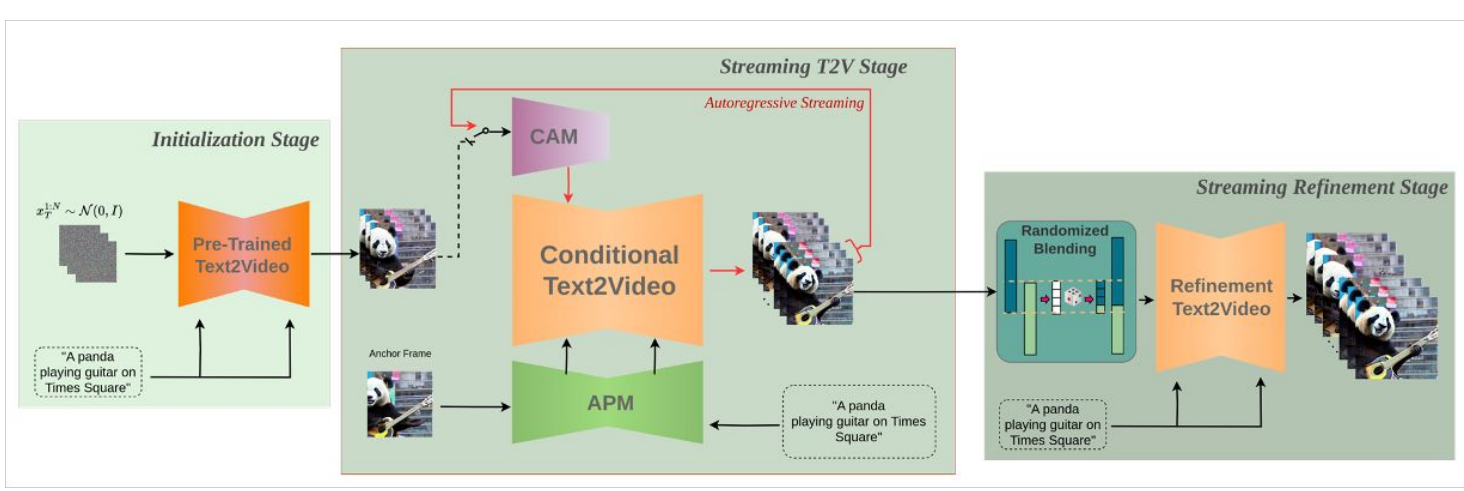


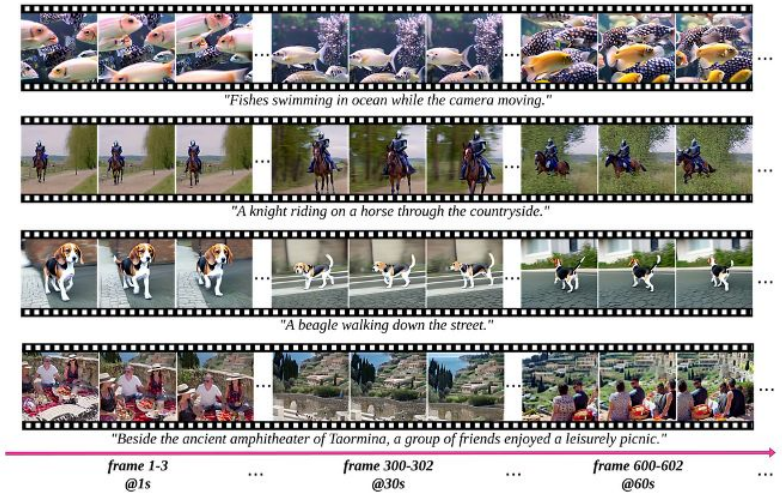


CVPR 2025 の動向・気付き (125/181)

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

- **概要:** 本稿では、StreamingT2Vという3段階のText to Video 生成方式を紹介し、約2分の長さの動画を生成できる。このモデルには、現在のフレームを最近のフレームに合わせるコンディションアテンションモジュールと、長いシーケンスでもグローバルな一貫性を保つための外観保存モジュールが組み込まれている。最後に、長い動画はランダムブレンドを使用して自己回帰的に生成される。このアプローチは、 OpenSora や OpenSoraPlan などの従来の方法よりも大幅に優れている。
- **新規性:** この改善は、既存のベースラインよりも発電品質を大幅に向上させるエンジニアリングの改良とトレーニング戦略によるもの。
- **気付き:** この方法はすでに注目を集めている。カメラ設定 (動き、FPV など) が異なれば、どのような設計上の考慮事項が必要になるのかという興味深い疑問が生じる。また、多段階アプローチが本当に最適かどうかも不明。このようなシステムのトレーニングには、かなりの複雑さが伴う可能性がある。

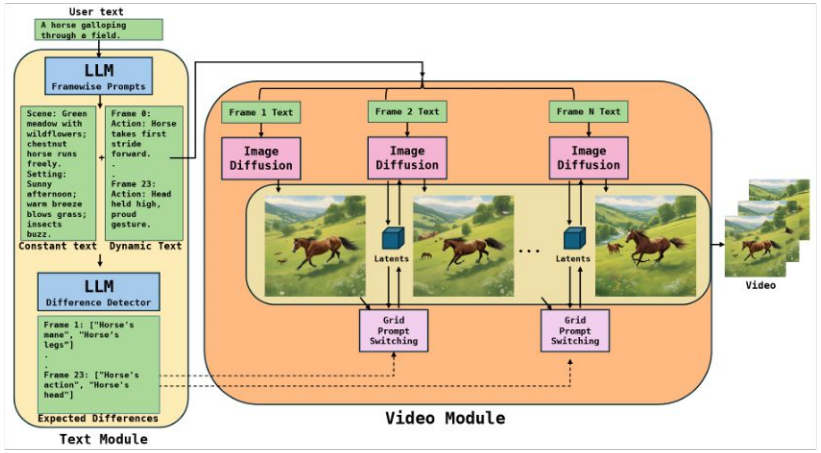


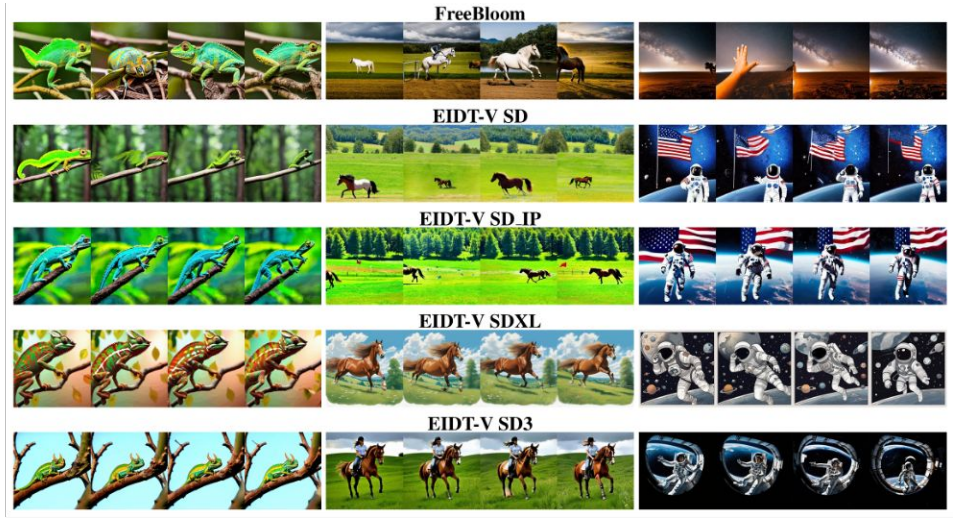


CVPR 2025 の動向・気付き (126/181)

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

- **概要:** 本論文では、既存の画像ベースのテキストから動画への生成モデルを改善するためのトレーニング不要の最適化手法である EIDT-Vを提案する。フレーム単位の記述とフレーム間のモーションプロンプトを取り入れることで、EIDT-V は潜在空間での生成を調整し、フレームごとの品質と時間的コヒーレンスの両方を向上させる。この方法では、モデルを再トレーニングしなくても高いパフォーマンスが得られる。
- **新規性:** テキストから動画への生成のためのトレーニングなしのアプローチは、まだ比較的十分に検討されていない。EIDT-Vは、テキストによるプロンプトによる画像レベルの制御を可能にし、さまざまな用途に使える可能性を秘めた創造的で柔軟なメカニズムを提供する。
- **気付き:** フレームごとに異なるプロンプトを使用することは説得力のあるアイデアであり、時間的な位置合わせにも役立つかもしれない。このアプローチを、画像ベースのフレームワーク以外の動画生成モデルにも拡張できるかどうかを検討するのは興味深い方向性である。

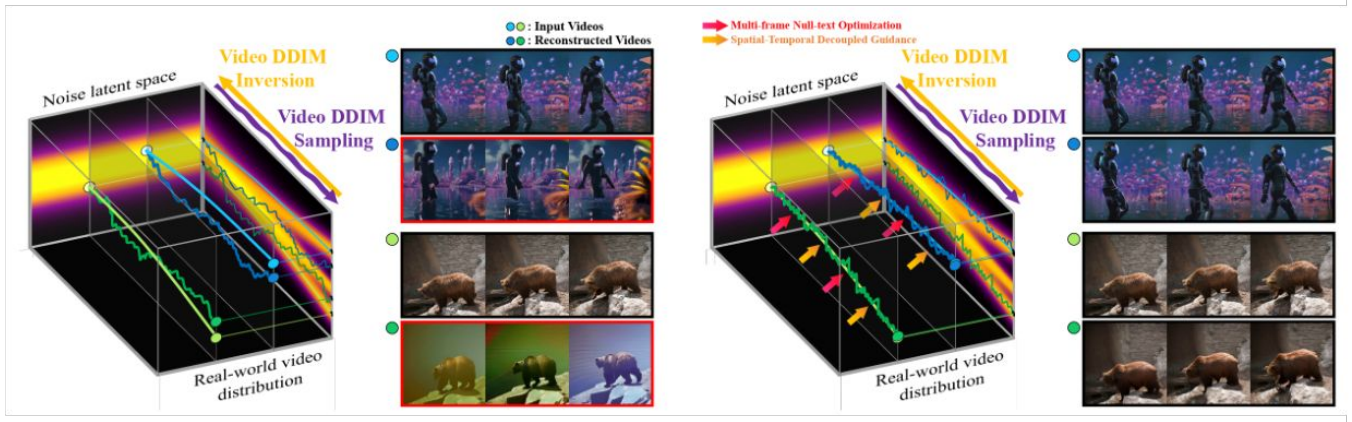


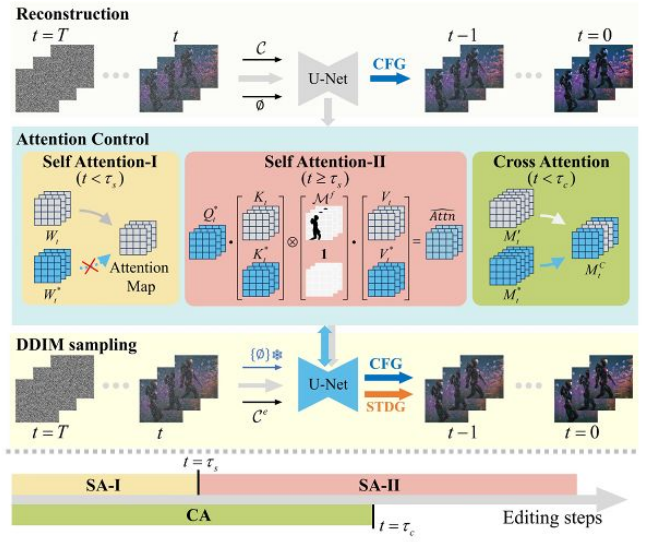


CVPR 2025 の動向・気付き (127/181)

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

- **概要:** 本論文では、text-to-video (T2V) 生成モデルを直接編集する方法である VideoDirector を提案する。このアプローチでは、空間的特徴と時間的特徴について個別のガイダンスを導入し、それぞれの側面を個別に編集しやすくしている。さらに、マルチフレームのヌルテキスト最適化も提案されており、詳細な時間編集が可能になります。この方法は最先端のパフォーマンスを実現する。
- **新規性:** 本論文では、緊密に結合した時空間的特徴や複雑なレイアウトの絡み合いなど、動画編集に対する既存の反転してから編集するアプローチの限界を徹底的に分析している。VideoDirector は、空間編集メカニズムと時間編集メカニズムを明示的に切り離すことでこれらの問題に対処している。
- **気付き:** 従来のT2V世代における時空間結合の課題はよく知られている。本研究は、それを軽減するための明確で構造化されたアプローチを示しており、より制御しやすく正確な動画編集に向けた有望な方向を示している。

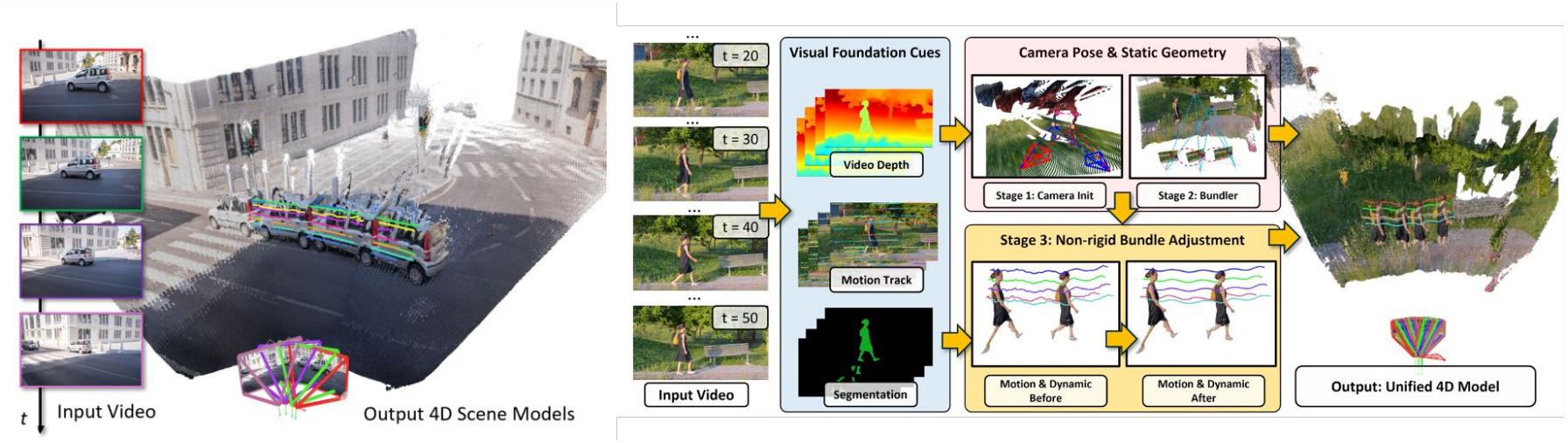




CVPR 2025 の動向・気付き (128/181)

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

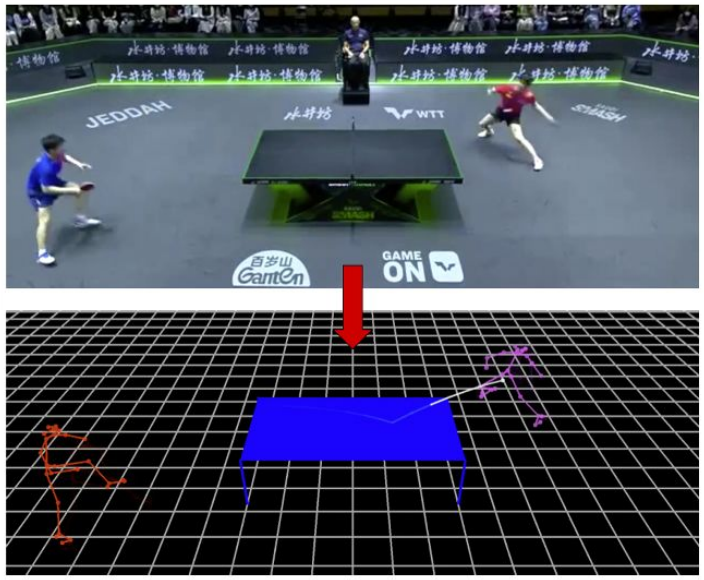
- **概要:** 本論文では、動画から直接4Dシーン(時間の経過に伴う3D幾何)を生成する方法であるUni4Dを提案する。Uni4D は視覚基盤モデルを活用して、動画フレームから奥行き、モーショントラッキング、セグメンテーションを抽出する。これらの出力を使用して、システムはカメラの姿勢とシーンの形状を推定し、続いてバンドル調整して結果を絞り込む。特に、視覚基盤モデルは再トレーニングなしで使用されているが、Uni4D は複数のベンチマークで最高のパフォーマンスを達成。
- **新規性:** 事前にトレーニングされた視覚基盤モデルをVideo-to-4Dタスクに適用すること。これにより、追加のトレーニングや教師なしで高品質の結果を得ることができる。
- **気付き:** この方法は既存のコンポーネントから構築されているが、それらを効果的に組み合わせて最先端のパフォーマンスを実現する。基盤モデルを動的なシーン再構築に統合することは有望かつ実用的。



CVPR 2025 の動向・気付き (129/181)

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

- **概要:** 本研究は、卓球エージェントの開発に 2つの重要な貢献をしている。まず、テーブルとボールのマスク抽出、テーブルの四隅の推定、カメラのキャリブレーション、ポーズの推定、最終的なボールの軌跡の生成など、動画から　3D シーンを再構築するためのパイプラインを提案する。次に、生成モデルに基づく不確実性を認識するコントローラーを導入して、相手の今後のアクションを予測する。プロ卓球の約　50時間の試合のデータセットはオンラインソースから収集されている。
- **新規性:** 予測モデリング（不確実な状況下での対戦相手の行動の予測）が含まれているため、この作業は以前のアプローチと区別される。
- **気付き:** フィジカルダイナミクスを取り入れると、リアリズムがさらに高まる。完全な　4Dシーンの再構築 (ジオメトリ+時間の経過に伴う動き) を実現することは、有望な次のステップとなる。



Algorithm 1 Ball Trajectory Reconstruction

Require: 2D ball positions $\{b_{2D,t}\}$, camera intrinsic parameters $\mathbf{K}$ and extrinsic parameters $\mathbf{R}$, $\mathbf{t}$

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

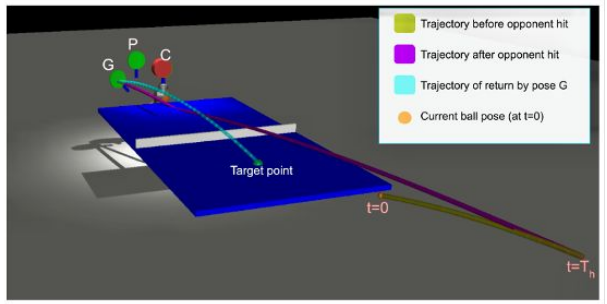
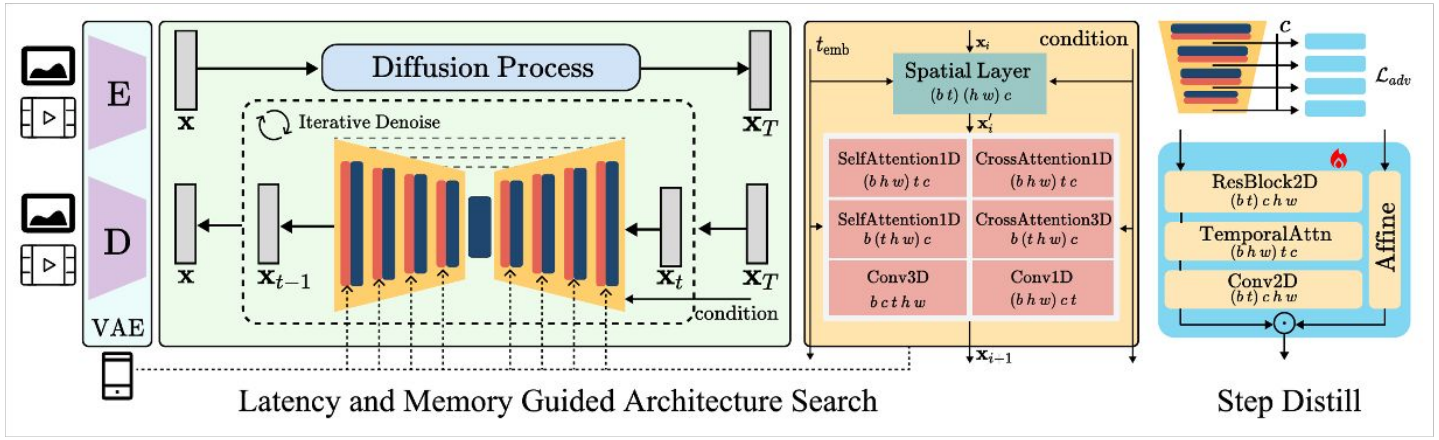


Figure 7. Simulated target poses and ball trajectory.

CVPR 2025 の動向・気付き（130/181）

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

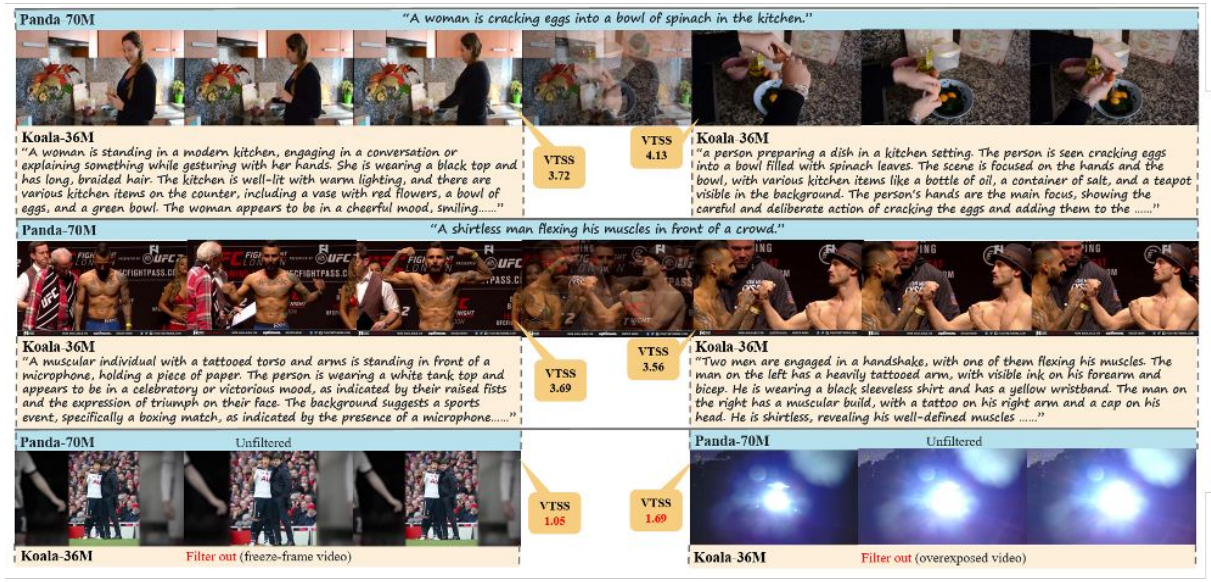
- **概要：** 本論文では、モバイルデバイス（ iPhone 16 Pro Max）で5秒以内に5秒の動画を生成できる、軽量でありながら高性能なテキスト /動画生成モデルであるSnapGen-Vを提案する。この効率化を実現するために、著者らは広範囲にわたる実験を行い、コンパクトでありながら効果的な時間的バックボーンを特定し、効率的な時間的レイヤーを設計。さらに、必要なノイズ除去ステップをわずか　4つに減らす蒸留技術とともに、拡散モデルの新たな敵対的微調整戦略が導入された。
- **新規性：** 軽量アーキテクチャを体系的に探求した点で際立っており、モバイルデバイスで効果的に機能する最初の　text-to-video 生成モデルという可能性を提示。
- **気付き：** モデル効率化は、ますます注目が高まっている研究分野。この研究は、リアルタイムのオンデバイス動画生成への扉を開き、近い将来、モバイルアプリケーションの波への道を開く可能性がある。



CVPR 2025 の動向・気付き (131/181)

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

- **概要**: 本論文では、動画認識と生成の両方を目的として設計された大規模なデータセットである Koala-36Mを提案。各動画クリップ (約 10 秒) には、モデルによって自動的に生成された、平均約 200 語の詳細なテキストによる説明が付けられる。動画品質の評価と向上のための動画トレーニング適合性スコア (VTSS) を提案している。さらに、動画とそれに対応するテキストの間の時間的一貫性を向上させるために、拡散モデルを採用している。
- **新規性**: Koala-36Mは、Panda70Mなどの同様のデータセットと比較して、より詳細で高品質のキャプションを提供する。品質と一貫性のためにVTSSと拡散モデルを使用することで、その有用性がさらに高まる。
- **気付き**: 動画や画像を理解するための高品質できめ細かなデータセットは、ますます重要になっている。重要な問題は、自動生成テキストの品質をどのように保証するか。それでも、現在のモデルで適度に優れたキャプションを大量に生成することは、トレーニングには効果的な戦略。

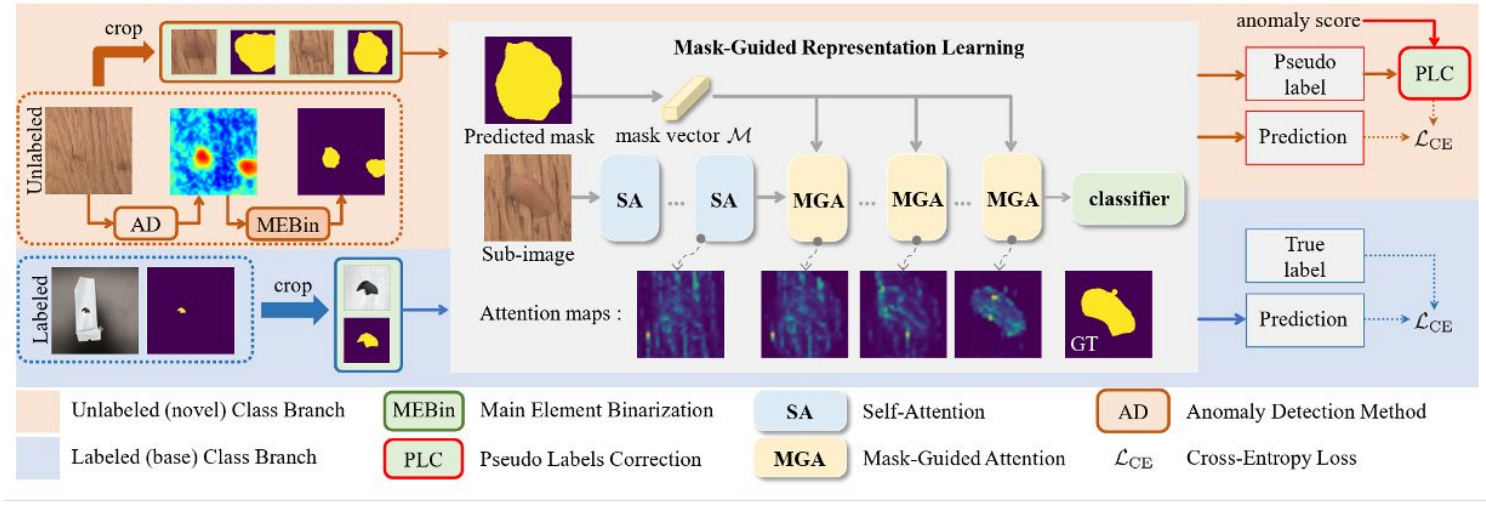


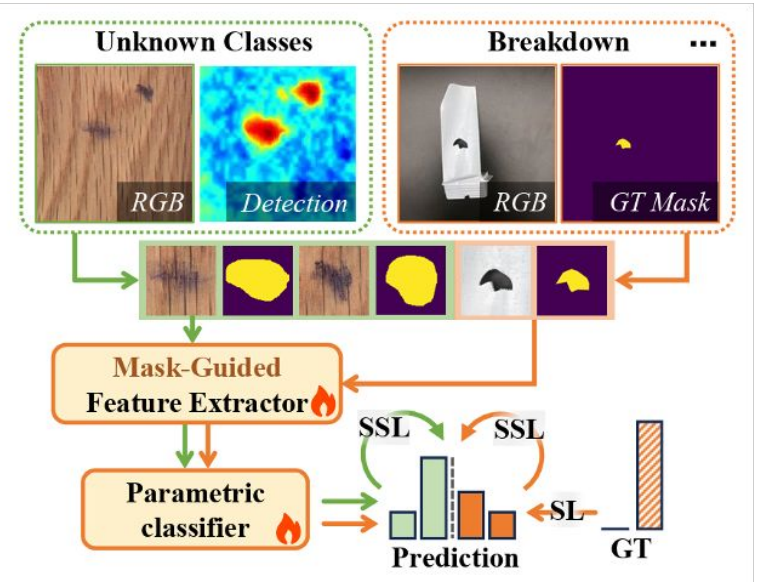
Dataset	#Videos	ATL(words)	TVL(hours)	Text	Filtering	Resolution
LSMDC (Rohrbach et al., 2015)	118K	7.0	158	Manual	Sub-metrics	1080p
DiDeMo (Anne Hendricks et al., 2017)	27K	8.0	87	Manual	Sub-metrics	-
YouCook2 (Zhou et al., 2018)	14K	8.8	176	Manual	Sub-metrics	-
ActivityNet (Caba Heilbron et al., 2015)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

CVPR 2025 の動向・気付き (132/181)

AnomalyNCD: Towards Novel Anomaly Class Discovery in Industrial Scenarios

- **概要:** 本論文では、マルチクラス異常検出の新しい方法である AnomalynCDを提案する。通常、異常領域は画像のごく一部しか占めないため、著者らは異常の中心領域の検出に重点を置いた主要素二値化 (MEBin) モジュールを導入した。この方法には、領域セグメンテーションを通じてセマンティック情報を充実させるためのマスクガイドによる特徴抽出も組み込まれている。このアプローチでは、複数のベンチマークで高いパフォーマンスが得られる。
- **新規性:** イノベーションは主に方法論的設計にある。個々のコンポーネントが完全に新しいわけではないが、それらを統合することで非常に高い精度が得られる。
- **気付き:** マルチモーダル言語モデル(MLLM)を異常検出に適用することは有望な方向であり、検出だけでなく異常タイプの解釈も可能になる可能性がある。また、異常検知とオブジェクトカウントを単一のフレームワークに統合できる可能性もある。

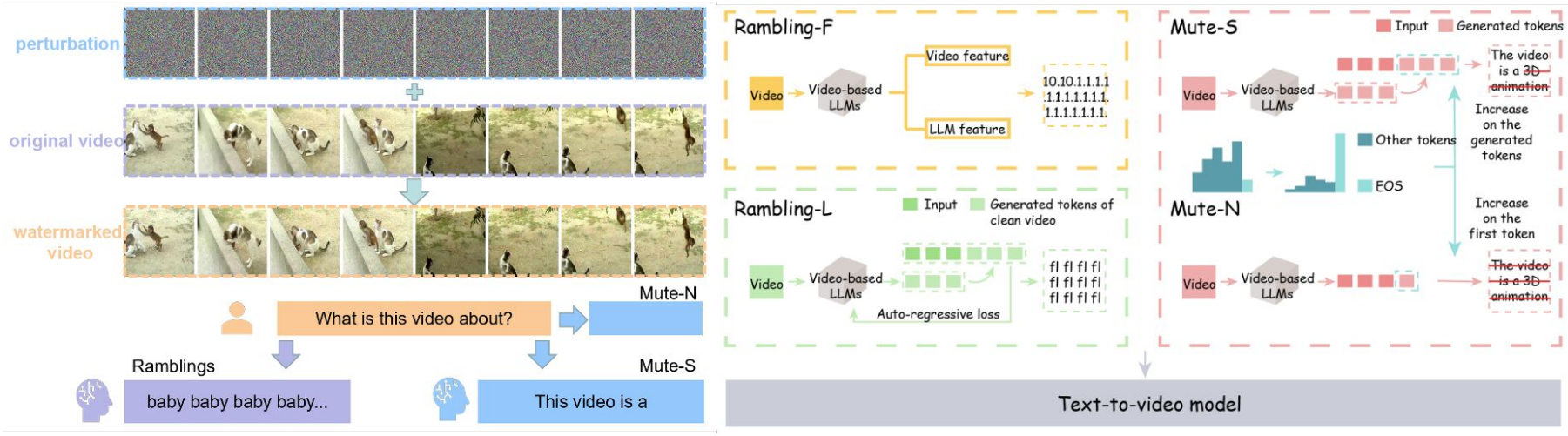




CVPR 2025 の動向・気付き (133/181)

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

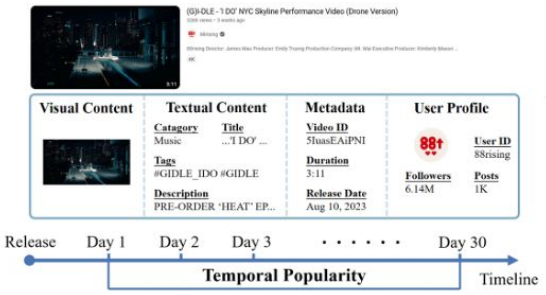
- **概要**: 本論文では、データプライバシ保護を目的として、動画 LLMが動画コンテンツを誤って解釈したり、過剰にアクセスしたりする問題に対処するために、ウォーターマーク方式の 2つの方法(ランブリングとミュート)を提案する。ランブリングス法は、誤解を招くような信号を埋め込むことで動画 LLM を誤った応答に導く。一方、ミュート法はモデルのシーケンス終了 (EOS) 動作を対象として出力品質を低下させることに重点を置いている。どちらの方法も、動画 ChatGPT、動画Llama、動画Vicunaなどのモデルに対して強いパフォーマンスを示す。
- **新規性**: 動画LLMから動画コンテンツのデータプライバシ保護するという新しい重要な研究課題に取り組んでいる。提案されている手法では、マルチモーダルな大規模言語モデルに特有の敵対的な防御策が導入。
- **気付き**: LLMの動画認識機能が急速に向上するにつれて、MLLMを対象としたプライバシ保護および敵対的技術の需要が高まる可能性がある。この研究は、責任あるAI開発における重要な新たな方向性を浮き彫りにしている。



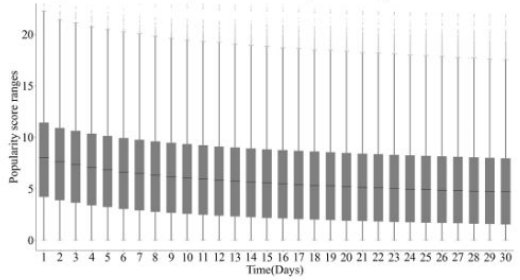
CVPR 2025 の動向・気付き (134/181)

SMTPD: A NewBenchmark for Temporal Prediction of Social Media Popularity

- **概要:** 本論文では、ソーシャルメディア動画の人気を評価するための新しいベンチマークであるSMTPDを、特に人気の時間的ダイナミクスに焦点を当てて提案する。既存のデータセットと比較して、SMTPD は動画の人気が時間の経過とともにどのように変化するかを捉えている。
- **新規性:** この課題自体は斬新で社会的に関連性がある。ソーシャルメディアの人気は、公衆の影響力や経済的価値に大きな影響を及ぼし、機械学習研究にとって重要な分野となっている。
- **気付き:** ソーシャル・ポピュラーな話題は説得力があり、人気の背景にある要因を探ったり、人気を他のコンテンツに移したりするなどの可能性を秘めている。ただし、人気は一定ではなく、実際のユーザーとのやり取りによって変化する可能性があるため、このタスクを学習上の問題として完全に切り分けることができるかどうかは不明。この研究は、ソーシャルメディアのような複雑で現実世界のシステムにMLLMを適用する傾向が広まっていることも反映している。



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



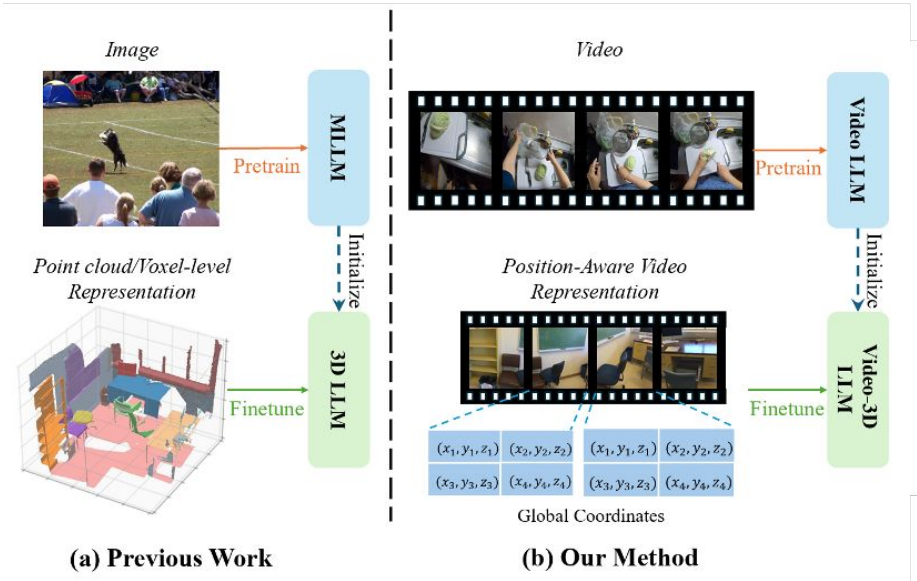
(b) Box plots of popularity scores over time

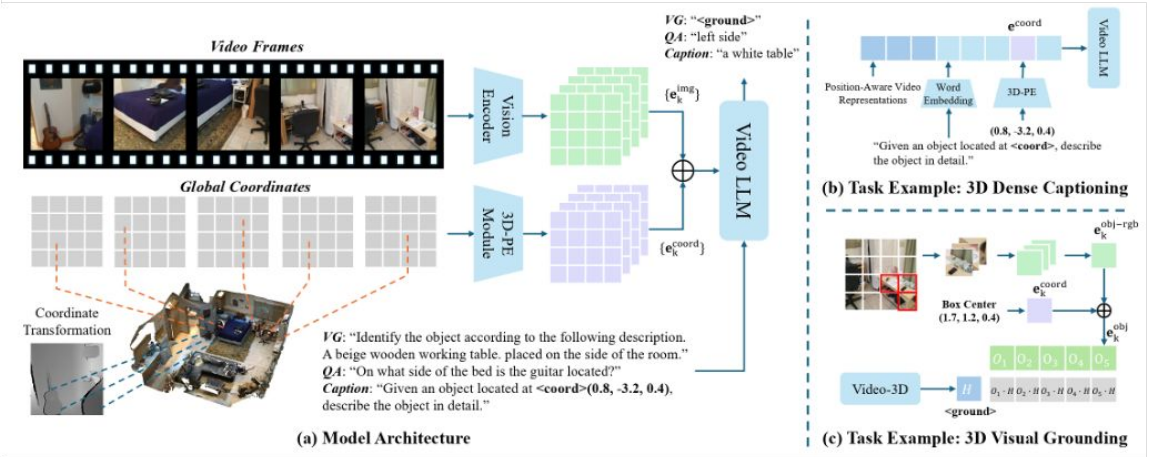
Dataset	Source	Category	Samples	Language	Prediction Type
Mazloom [35]	Instagram	fast food brand	75K	English	single
Sanjo [42]	Cookpad	recipe	150K	Japanese	single
TPIC17 [47]	Flickr	-	680K	English	single
SMPD [48]	Flickr	11 categories	486K	English	single
AMPS [11]	YouTube	shorts	13K	Korean	single
SMTPD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

CVPR 2025 の動向・気付き (135/181)

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

- **概要:** 本稿では、動画を用いた3D認識が可能な新しいマルチモーダル大言語モデルであるVideo-3D LLMを提案する。この手法では、3D理解を動的な動画分析の一形態として扱い、動画表現と3D表現を一致させる。このモデルは、複数の 3D 認識ベンチマークで高いパフォーマンスを発揮する。
- **新規性:** 動画と3Dモダリティの橋渡しにより、MLLMが明示的な3D入力を必要とせずに3D構造を認識できるようにすることにある。
- **気付き:** 動画だけから3D情報を認識することは非常に実用的。このアプローチにより、専用の 3D 入力が利用できないことが多い現実世界のシナリオでも、3D 認識が容易になる。これにより使いやすさが向上し、空間理解におけるMLLMの新たな応用の可能性が開かれた。

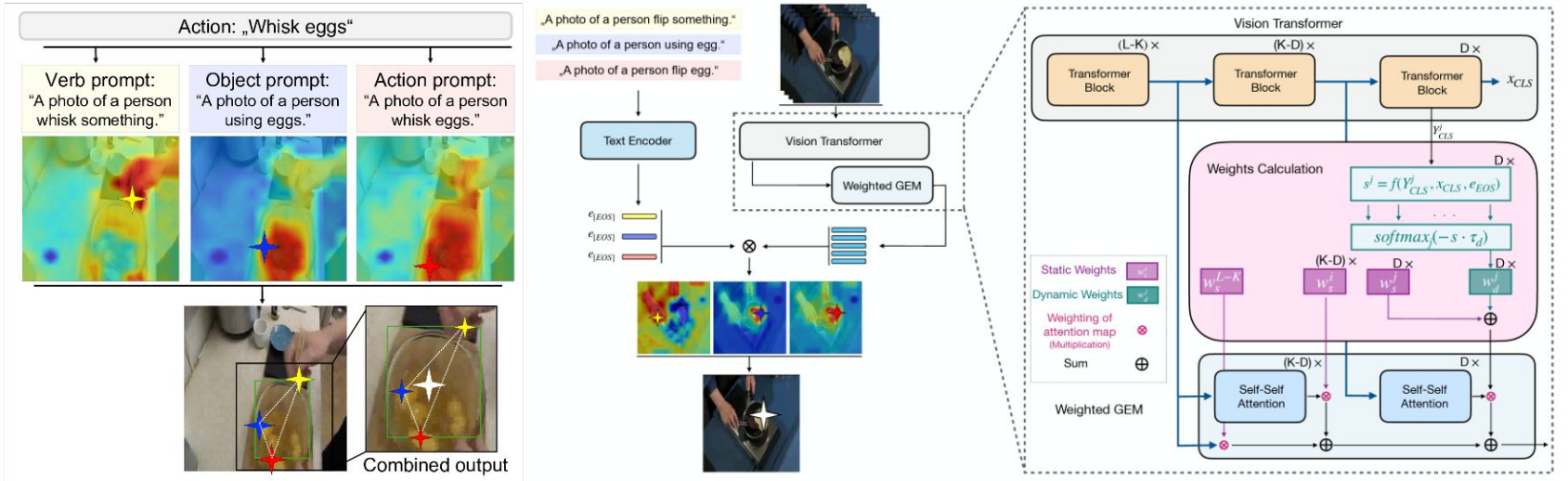




CVPR 2025 の動向・気付き (136/181)

VideoGEM: Training-free Action Grounding in Videos

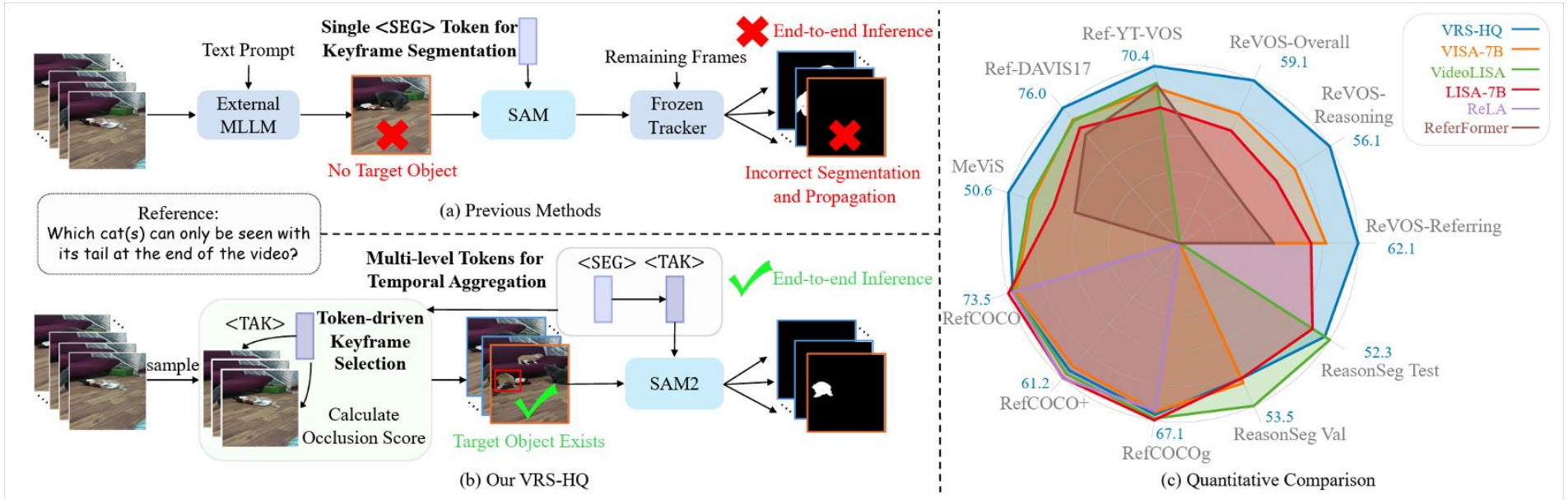
- **概要:** 本論文では、追加のモデルトレーニングなしで動画内のアクションとオブジェクトをローカライズする、トレーニング不要の最初の動画グラウンディング方法である VideoGem を紹介する。画像ベースのグラウンディング手法 GEM を動画にも応用している。この手法では、さまざまなアテンションレイヤーがアクションやオブジェクトに対応しているという観察に基づいて、レイヤーの重みが動的に調整される。また、アクションラベルからアクション、動詞、オブジェクトのプロンプトを生成し、これらのプロンプトを使用して注意を誘導し、対応するアテンションマップを抽出する。
- **新規性:** これは、アテンションレイヤーのダイナミクスとプロンプトベースのガイダンスを活用して関連コンテンツをローカライズし、トレーニングなしで動画のグラウンディングを実現する最初のアプローチである。
- **気付き:** VideoGem のトレーニング不要の性質は、特にスケーラブルな環境やリソースが少ない環境では非常に魅力的。プロンプトに柔軟に対応できるため、きめ細かくカスタマイズ可能なグラウンディングが可能になり、動画ベースのデータセットの自動構築が可能になり、幅広い用途が広がる。



CVPR 2025 の動向・気付き (137/181)

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

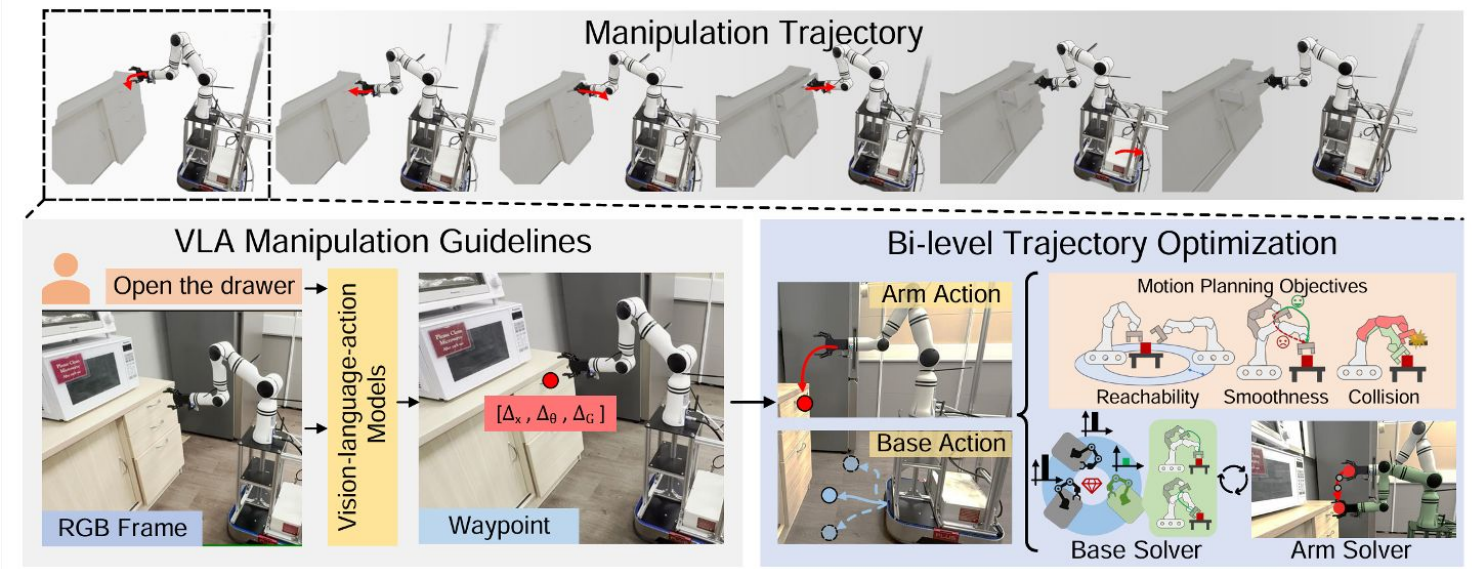
- **概要:** 本論文では、動画推論セグメンテーションの新しい方法であるVRS-HQを提案する。単一の [SEG] トークンを使用して動画フレーム全体でターゲットオブジェクトを検出する従来の方法とは異なり、VRS-HQ では時間的ローカリゼーションの重要性が強調されている。ターゲットオブジェクトを含むフレームを識別するテンポラルアテンショントークン([TAK]) が導入されている。また、この方法では、類似性に基づく融合メカニズムとフレーム選択メカニズムを統合して、時間的推論を強化している。
- **新規性:** 主な革新点は、[TAK] 時間トークンの導入である。これにより、オブジェクトのより正確な時間的グラウンディングが可能になる。この方法はシンプルだが、複数のデータセットやタスクにわたって高いパフォーマンスを発揮する。
- **気付き:** [TAK] トークンは、他の動画推論タスクで広く使用される可能性を示しており、同様の時間トークンをオーディオやその他のシーケンシャルモダリティに適用できるかどうかという興味深い疑問が生じる。このアプローチは実用的で一般化できる。



CVPR 2025 の動向・気付き (138/181)

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

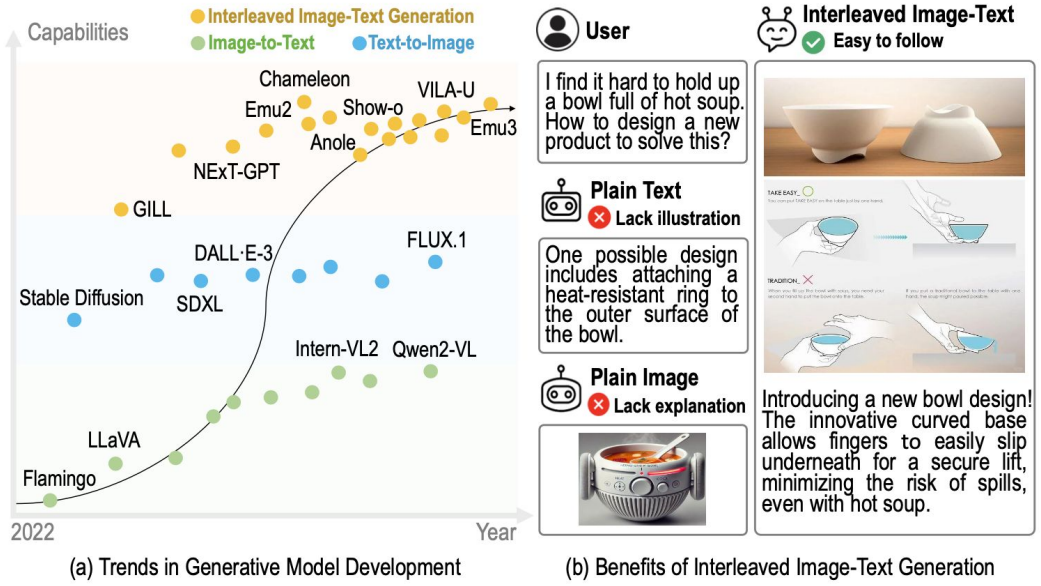
- **概要：** 本稿では、大規模なVison-Language-Action（VLA）モデルをロボット工学のモバイルマニピュレーションタスクに移行するためのフレームワークを提示する。提案手法では、VLA 出力からウェイポイント表現を予測する。これは、さまざまな操作タスクで広く一般化されている。また、腕とベースアクションの両方にモーションプランニングが組み込まれているため、現実世界での動きと操作の精度が向上する。このシステムは、最小限の追加トレーニングでモバイルマニピュレーションを実現する。
- **新規性：** 主なイノベーションは、ウェイポイントを効果的かつ一般化可能な中間表現として使用して、VLAモデルを現実世界のモバイルマニピュレーションに移行することである。
- **気付き：** MLLMをロボット工学に適用することへの関心が高まっている。この研究は、ウェイポイントやシーングラフなど、現実世界のタスクにはどのようなシーン表現が最も効果的かという重要な疑問を投げかけている。構造化された表現を統合することで、ロボットプランニングの精度と一般化がさらに向上する可能性がある。



CVPR 2025 の動向・気付き (139/181)

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

- **概要:** 本論文では、マルチモーダル大規模言語モデル（MLLM）におけるインターリーブされた画像テキスト生成を評価するために設計された包括的なベンチマークであるOpenINGを紹介する。OpenING には、旅行案内、デザイン、ブレーンストーミングなどの 56 の現実世界のタスクを対象とした 5,400 個の人間が注釈を付けたインスタンスが含まれてる。信頼できる評価を支援するために、著者らはまた、カスタム・データ・パイプラインを介してトレーニングされた新しい自動判定モデルIntJudgedも提案している。このモデルでは、GPTベースの評価者を上回り、人間の評価との一致率が 82.42% に達している。現在のインターリーブ型世代モデルはまだ不十分であることが実験によって示されており、この論文は、この分野における将来の進歩を促進するための知見を与える。
- **新規性:** インターリーブジェネレーションについては、まだあまり議論されておらず、新規性があると言える。
- **気付き:** 詳細生成のためのデータセットはまだ不十分なので重要である。インターリーブされたマルチモーダルデータを使用してモデルを事前トレーニングすると、モデルのパフォーマンスが向上する可能性がある。実際のユーザー入力では、データにさまざまなデータ型が組み込まれている可能性があるため、どのような種類の入力データが重要で、扱いが少なく、根本的に難しいかを慎重に検討することが重要である。

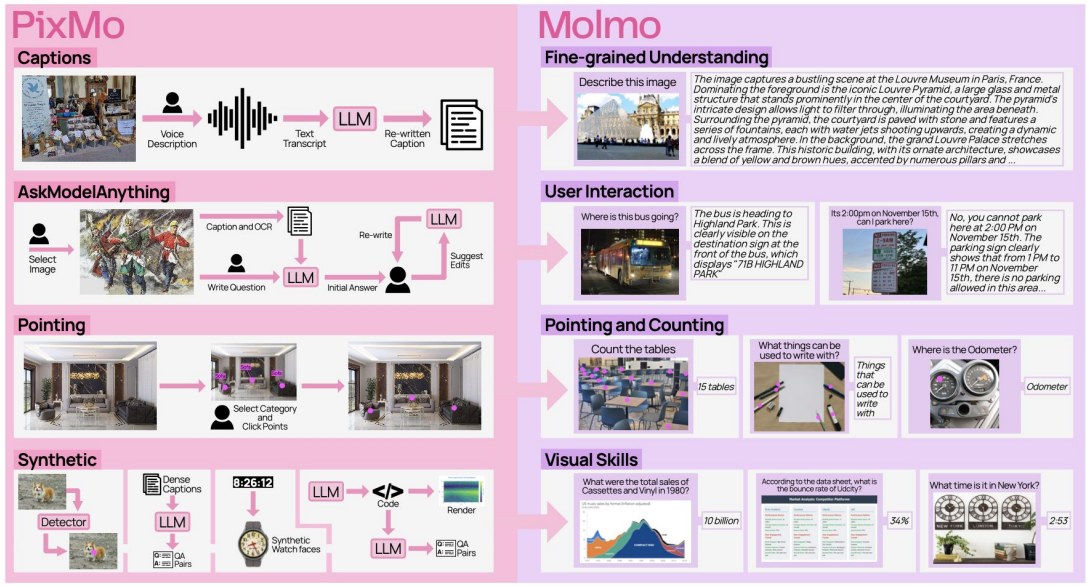


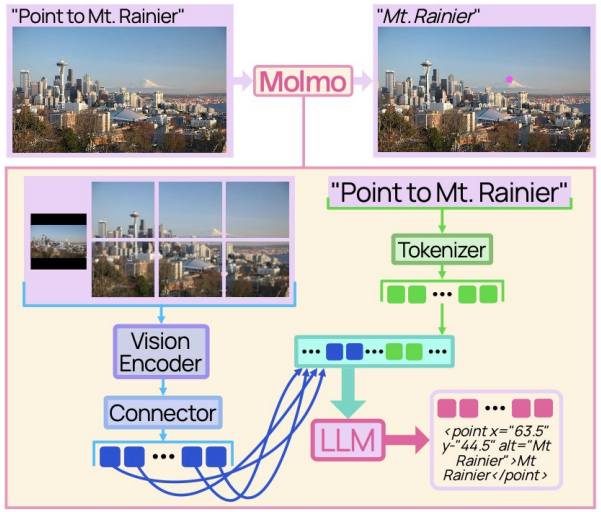
(a) Trends in Generative Model Development　(b) Benefits of Interleaved Image-Text Generation

CVPR 2025 の動向・気付き (140/181)

Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models (Best Paper Honorable Mention)

- **概要:** 本論文では、新しく収集された PixMOと呼ばれるデータセットスイートを使用して完全にゼロからトレーニングされたオープンウェイトのビジョン言語モデルのファミリーである Molmoを紹介する。独自の VLMに頼ることなく、Molmoはオープンモデルの中でも最先端のパフォーマンスを実現し、Claude 3.5 SonnetやGemini 1.5 Proなどのいくつかの主要なプロプライエタリシステムよりも優れたパフォーマンスを実現している。Molmoには、詳細なキャプション、優れたグラウンディング能力、詳細な知識ベースの推論が組み込まれている。
- **新規性:** 第一にデータとモデルがオープンである。Molmoはオープンソースだが、クローズドソースの商用モデルと比較しても最先端の結果が得られている。
- **気付き:** 詳細な認識とグラウンディング能力は非常に重要であり、 Molmoを含む最近のいくつかの作品ではパフォーマンスが向上している。

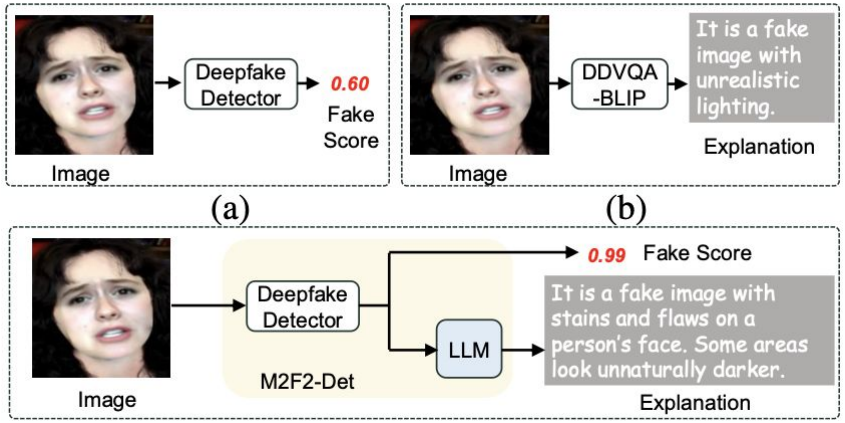


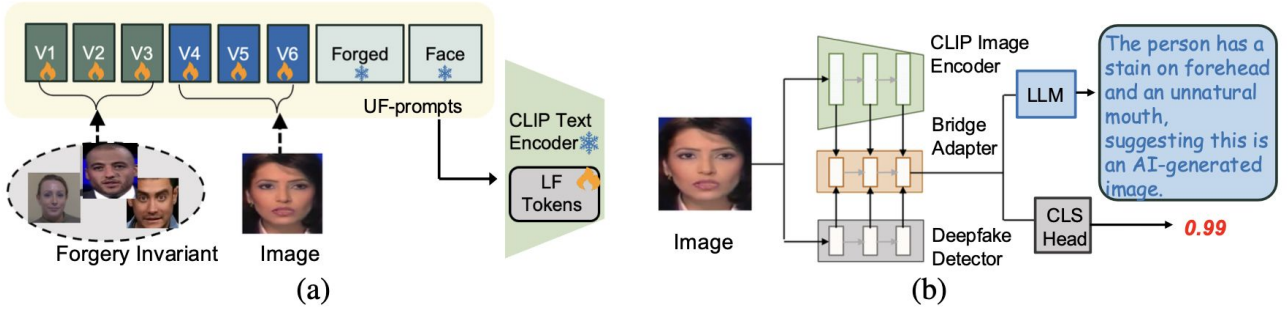


CVPR 2025 の動向・気付き (141/181)

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

- **概要:** 本稿では、マルチモーダルディープフェイク検出モデルであるM2F2-Detを提案する。これは、画像を本物か偽物かに分類するだけでなく、その判断について詳細な自然言語による説明を提供し、解釈可能性を高める。このモデルでは、画像エンコーダーを LLM ベースの説明ジェネレーターに合わせるためにブリッジアダプターを導入し、偽造プロンプトラーニングとレイヤー単位の LF トークンを使用して、微妙な顔の操作をより正確にキャプチャする。M2F2-Det は多段階トレーニングを通じて、目に見えない偽造品への強力な汎用化を実現し、検出精度と説明品質の両面で新たな基準を打ち立てる。
- **新規性:** スコアと解釈の同時生成タスク(左の画像)。また、このモデルは複雑だが、新規性がある (右図)。
- **気付き:** 顔、身振り、視線は社会的インタラクションの認識において重要だが、データ面ではあまりうまく扱われていない。スコアの代わりに、偽造された顔の理由を説明するために領域を接地する機能の方が面白いかもしれない。別の問題として、このモデルが複雑であることが挙げられる。

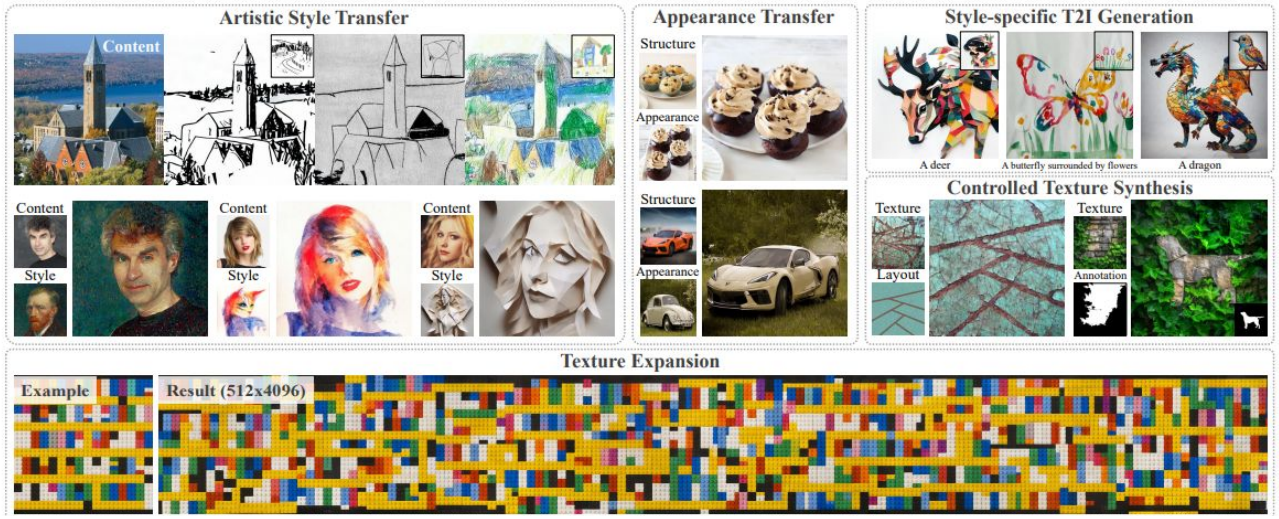


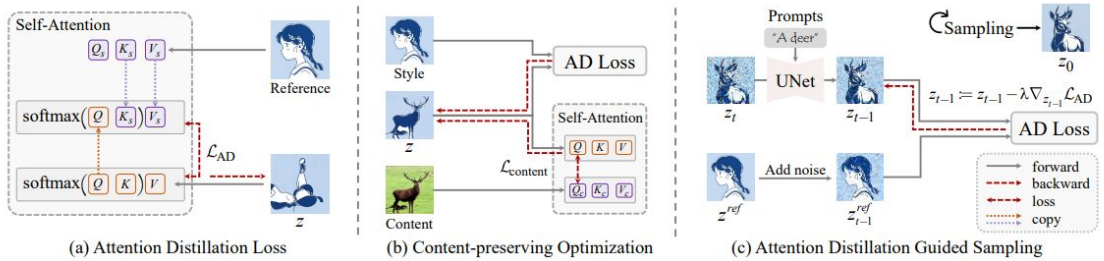


CVPR 2025 の動向・気付き (142/181)

Attention Distillation: A Unified Approach to Visual Characteristics Transfer

- **概要:** 拡散モデルの自己注意(self-attention)機能には、画像のスタイルや意味構造が自然に学習されている。この attention 情報を活用し、参考画像からスタイルやテクスチャなどの視覚的特性を別の生成画像に転写する統一的な手法を提案した。
- **新規性:** 技術としてはAttention Distillation Lossを理想的スタイル付与後の注意マップと、生成中の画像の注意マップとの差を誤差として定義した点が優れている。結果的に、Latent 空間での逆伝播による画像最適化により視覚特性を転写できた。また、従来の分類器ガイダンスを強化し、attention distillation loss を拡散サンプリング過程に統合することで生成速度を速めた。
- **気付き:** スタイル・質感の転移能力は素晴らしいが、「構図」や「内容」など高度な意味内容の転写はまだ難しい。しかし普遍的な技術であるため今後発展が期待され、画像を特定のスタイルにそろえるなどの現実のニーズに適応できるため実用性が高い。



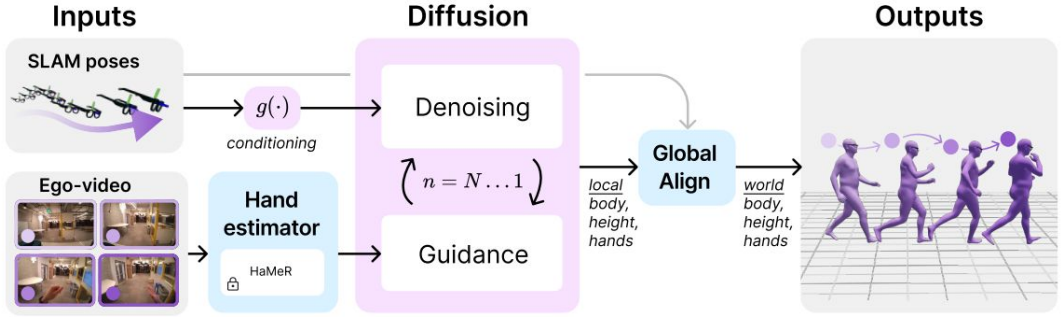


https://github.com/xugao97/AttentionDistillation

CVPR 2025 の動向・気付き (143/181)

Estimating Body and Hand Motion in an Ego-sensed World

- **概要:** ヘッドマウント型デバイスの姿勢(6DoF)データと画像データだけを用い、その場でワールド地図を構築しながら、ワールド座標系における3D身体ポーズ・身長・手の動きを統合的に推定する仕組み。
- **新規性:** 拡散モデル(Diffusion-based motion prior)に頭部動作を条件付ける形で身体の動作生成を行う。絶対座標ではなく、頭部基準で正規化された相対空間での推定を行うことで、空間的・時間的なばらつきを抑え、推定精度を大幅に向上。
- **気付き:** 姿勢情報とヘッドマウントカメラ画像だけで推定できる点は大きな利点。身長の推定は地面が平坦であることを仮定することで可能となっているが、この仮定ゆえに傾斜や階段といった非平面環境では対応が難しいという制限もある。



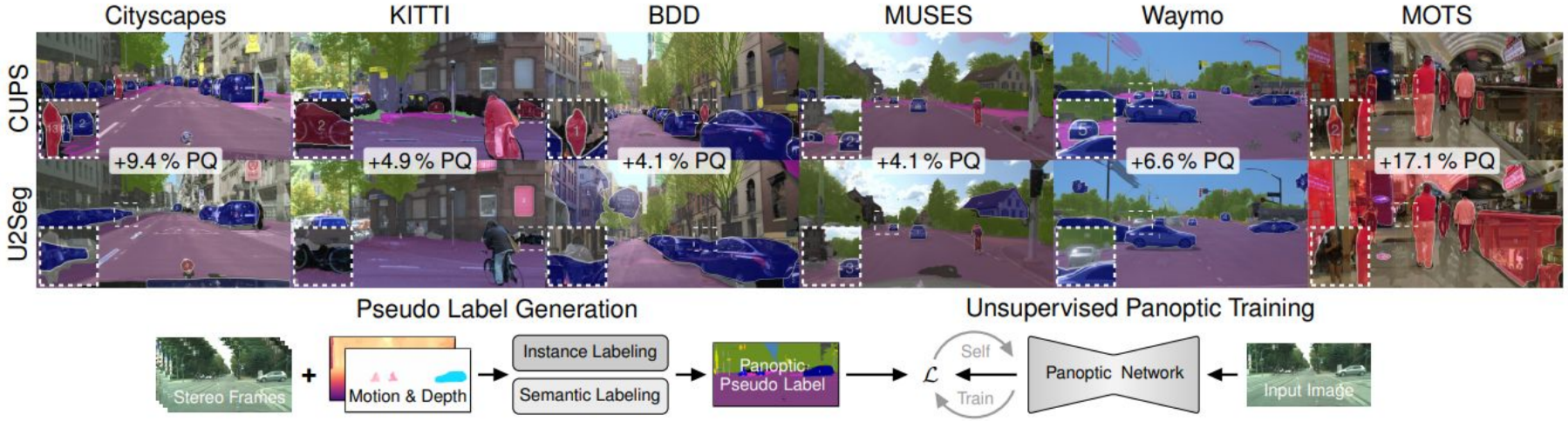
完成度高い動画！

https://egoallo.github.io/

CVPR 2025 の動向・気付き (144/181)

Scene-Centric Unsupervised Panoptic Segmentation

- **概要:** 完全ラベルなしの状態でも、シーン全体を対象に panoptic segmentation を実現する新手法。
- **新規性:** ステレオ映像と自己教師あり手法（SMURFなど）を用い、SF2SE3で動きのある物体を3次元モーションに基づいてクラスタリング。DINOやDepth-Gによる自己教師あり特徴を使い、ステレオ深度を加味して高解像度セマンティックマスクを生成し、モーションから「thing」、セマンティックから「stuff」を識別し、高精度なパンオプティック擬似ラベルを構築。擬似ラベルで訓練したネットワークに、copy-paste augmentationやモーメンタムモデルを用いた自己訓練（self-training）を追加し性能向上。
- **気付き:** 複雑なシーンでも有効な疑似ラベル生成を行う手法として非常に有望。手動ラベルの前の荒い検出としても使えるので非常に実用的。

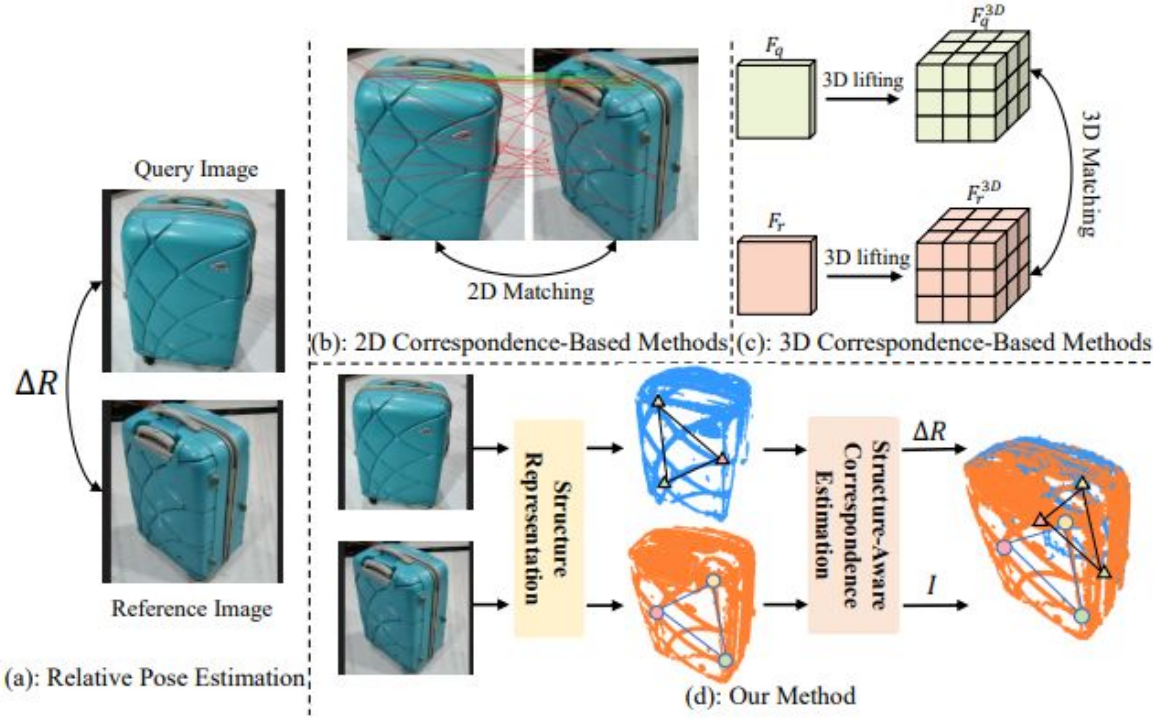


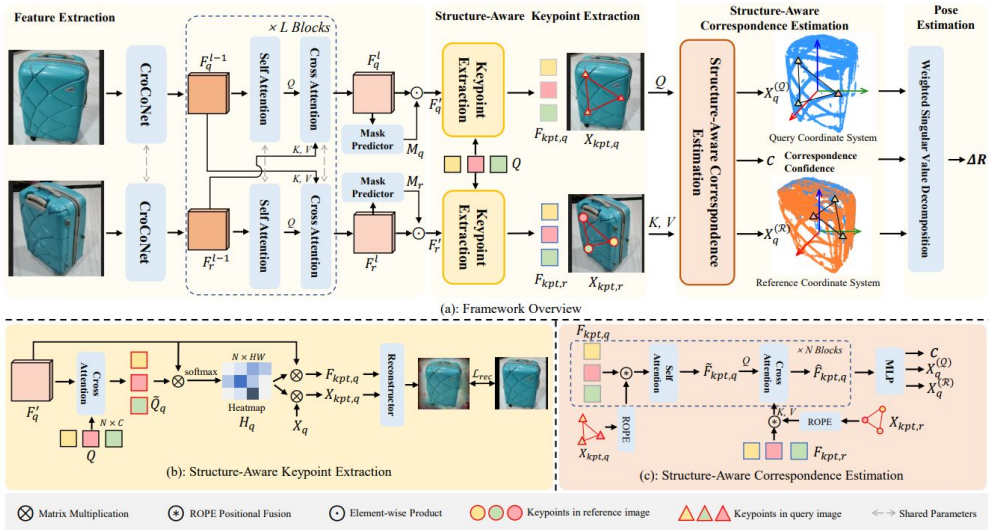
https://visinf.github.io/cups/

CVPR 2025 の動向・気付き (145/181)

Structure-Aware Correspondence Learning for Relative Pose Estimation

- **概要**: 未知の物体に対して参照画像とクエリ画像との間の3D回転・並進関係を推定。
- **新規性**: 物体の構造を代表するキーポイントを画像上に自動抽出することで、見た目が異なる物体でも形構造を把握。画像間の関係性をグラフ構造でとらえることで、明示的な特徴マッチングを行わずに3D・3D対応点を直接予測。
- **気付き**: 物体構造をグラフ構造として理解し、対応点を直接推定する点が優れている。しかしクエリ画像によって元画像の構造推定が変わるため、絶対的な評価法ではない。逆に言うとある画像を用いて特定の画像をとらえなおす手法として使えるため実用性が高い。



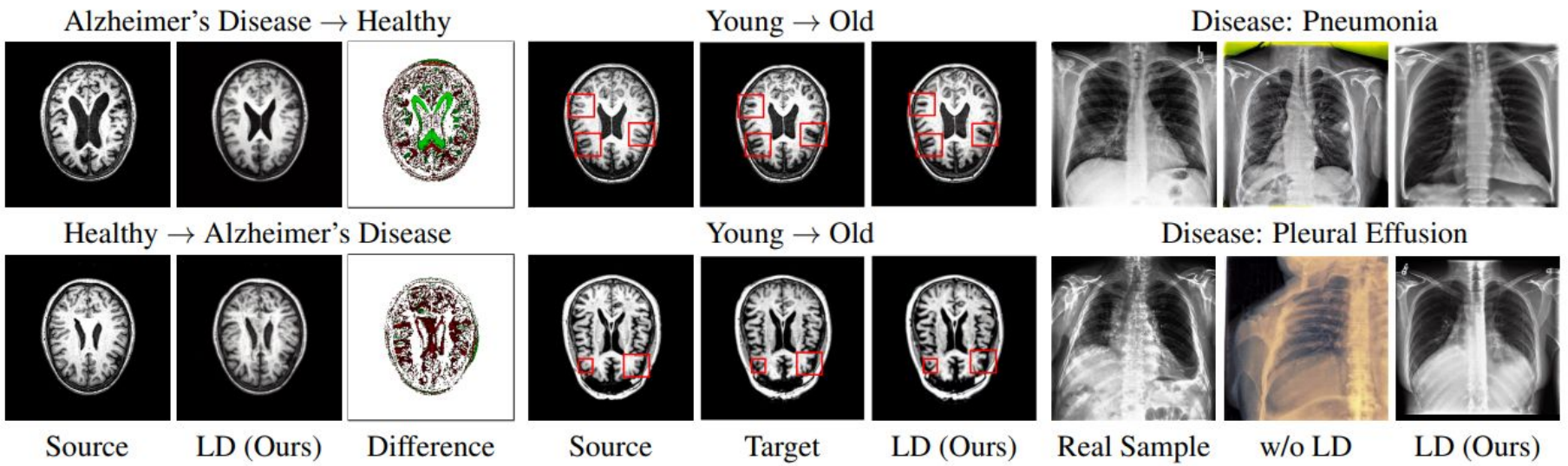


https://cyhhzo02.github.io/SAC-Pose/

CVPR 2025 の動向・気付き（146/181）

Latent Drifting in Diffusion Models for Counterfactual Medical Image Synthesis

- **概要:** 医療画像の生成は、医療画像が少ないという問題があるため困難だった。少ない枚数からでも画像が生成できる新手法「Latent Drifting（LD）」を提案。
- **新規性:** ノイズの剥落・再構成過程に「δ」というパラメータを導入し、潜在分布をドメイン間（自然⇄医療）で徐々にずらす。「ある属性を変えた場合にどう変化するか」を反事実（counterfactual）として数学的に定式化した最小最大化問題として扱う。
- **気付き:** 自然画像→医療画像への分布ギャップを埋めるため、新しい潜在ドリフト制御手法を提案しただけでなく、因果推論の反事実生成を利用した点が面白い。少ない枚数の医療画像データからも良質な画像生成が可能になる。

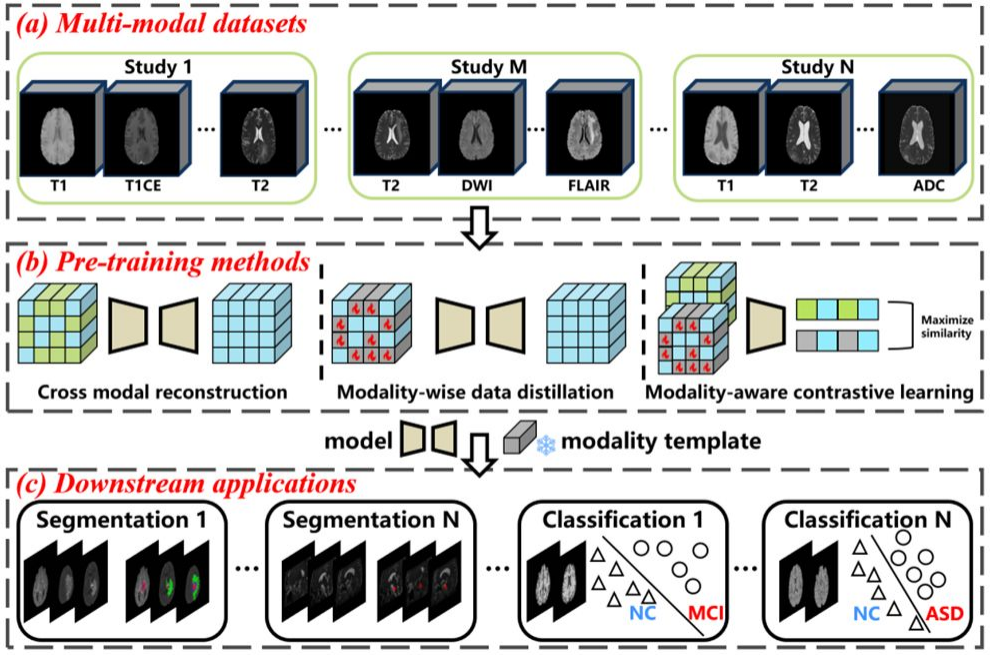


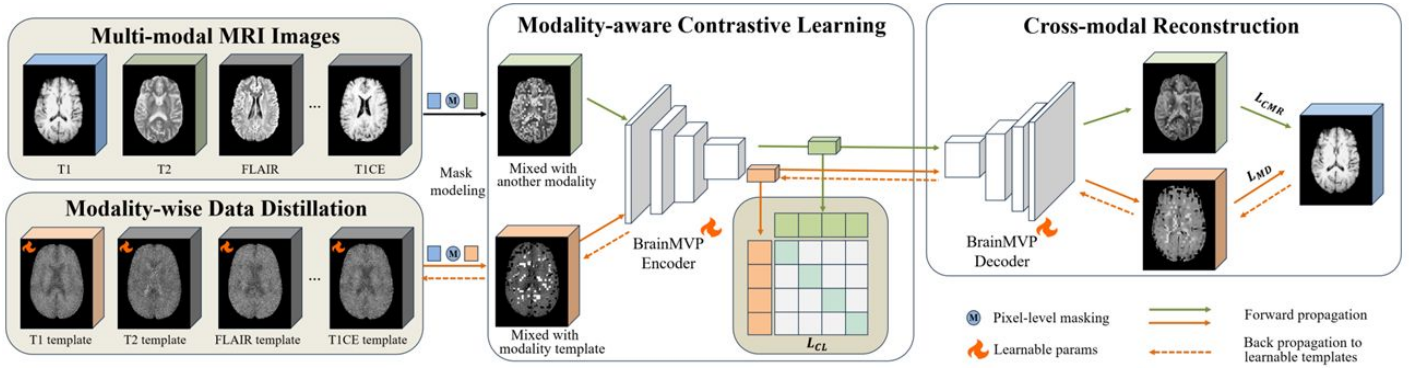
https://latentdrifting.github.io/

CVPR 2025 の動向・気付き (147/181)

Multi-modal Vision Pre-training for Medical Image Analysis

- **概要**: MRIの特徴は同一断面から複数のモダリティーを撮像可能。そのためモダリティ間に相関がある。MRI画像の自己教師あり学習のためには複数のモダリティーを含むデータセットが必要である。
- **新規性**: モダリティ間の相関を含めて学習するために、16,022件(240万枚以上)、8種類のMRIモダリティーデータセットを構築クロスモダリティ表現と相関の学習を促進するための3つの代理タスク(クロスモーダル画像再構成、モダリティ認識型対照学習、モダリティテンプレート蒸留)でマルチモーダル画像事前学習を実施し、下流タスクで良好な結果。
- **気付き**: これだけの規模のデータセットを構築した点が素晴らしい。データセットだけでも医療画像研究に大きく貢献。



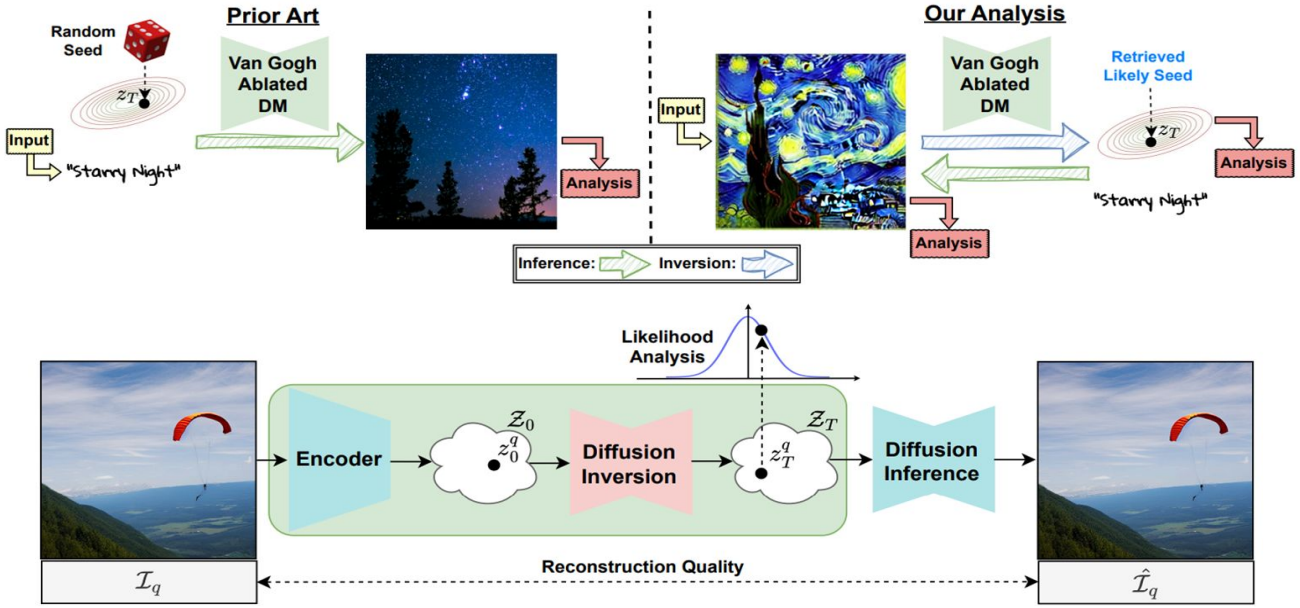


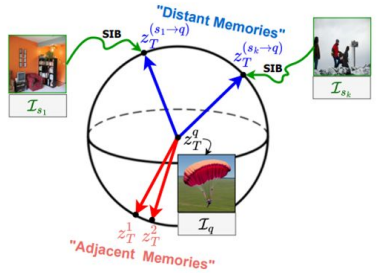
https://github.com/shaohao011/BrainMVP

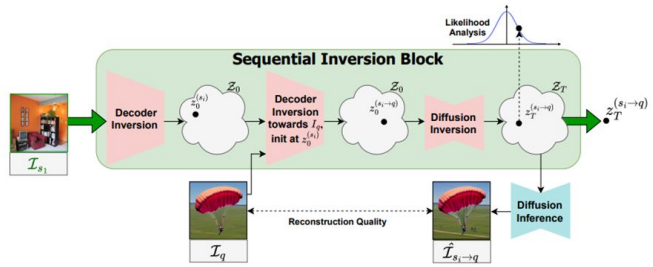
CVPR 2025 の動向・気付き (148/181)

Memories of Forgotten Concepts

- **概要:** 本研究は、AIモデル内に消去された概念情報が残存し、復活可能であることを明らかにした。特定の「潜在的な種」と「逆変換手法」を用いることで、忘却された概念の高品質な画像を生成できることを示した。
- **新規性:** 本研究は、AIから概念情報を完全に消去することが困難であることを唯一無二に示した。消去されたデータが再構築可能であることを示すことで、現在の概念消去技術における潜在的な脆弱性を浮き彫りにした。
- **気付き:** 概念が離散的な単位ではなく、重なり合い分散した情報として潜在空間に存在することは驚き。これは、絵画から特定の色を削除するように、わずかなピクセル調整だけで「真の消去」を実現する処理が極めて困難であることを示している。





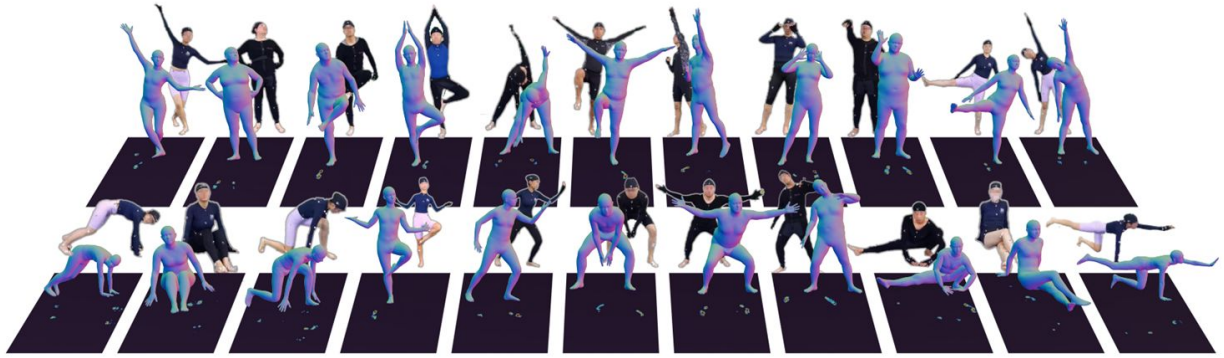


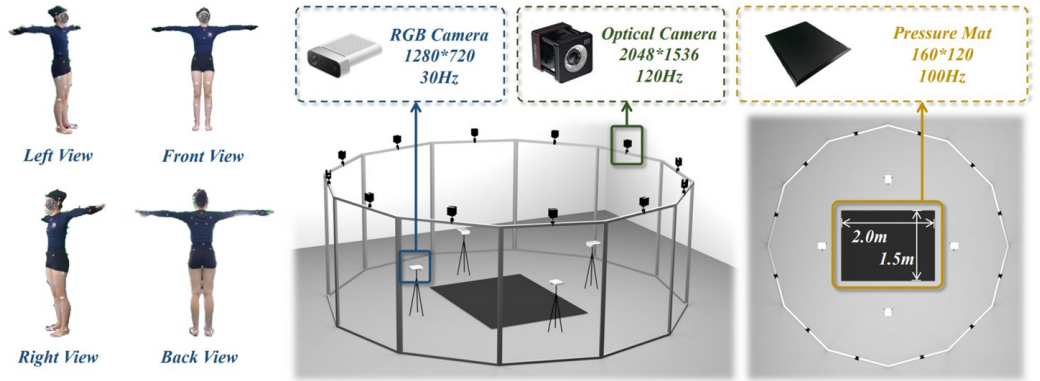
https://matanr.github.io/Memories_of_Forgotten_Concepts/

CVPR 2025 の動向・気付き (149/181)

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

- **概要**: 本論文では、圧力センサー、RGBセンサー、光学センサーを統合した大規模な人体動作キャプチャデータセット「MotionPRO」を提案する。圧力信号の必要性と有効性を、姿勢推定と軌跡推定の精度向上、および現実的なバーチャルヒューマンの駆動に示している。
- **新規性**: 本研究の核心的な新規性は、マルチモーダルデータセットを構築し、人体動作キャプチャにおける圧力情報の重要な役割を厳密に探索したことにある。これにより、動作理解と生成におけるAIモデルの物理的現実感と精度が向上する。
- **気付き**: 圧力センサーデータを同時に取り込むことで、特に次のような場合に非常に実用的になる。ジャンプなどの動的な動きの時系列分析。同時取得された圧力センサーデータは、特にジャンプのような動的動作のタイムシリーズ分析において、実践的な有用性を提供する。視覚的手がかりを超えた重要な物理的洞察を提供し、このような複雑な動作中の地面との相互作用や力分布のより正確で堅牢な理解を可能にする。



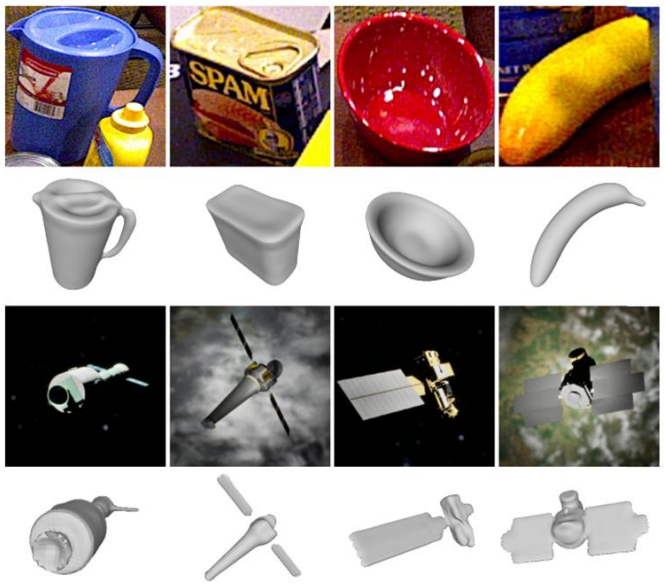


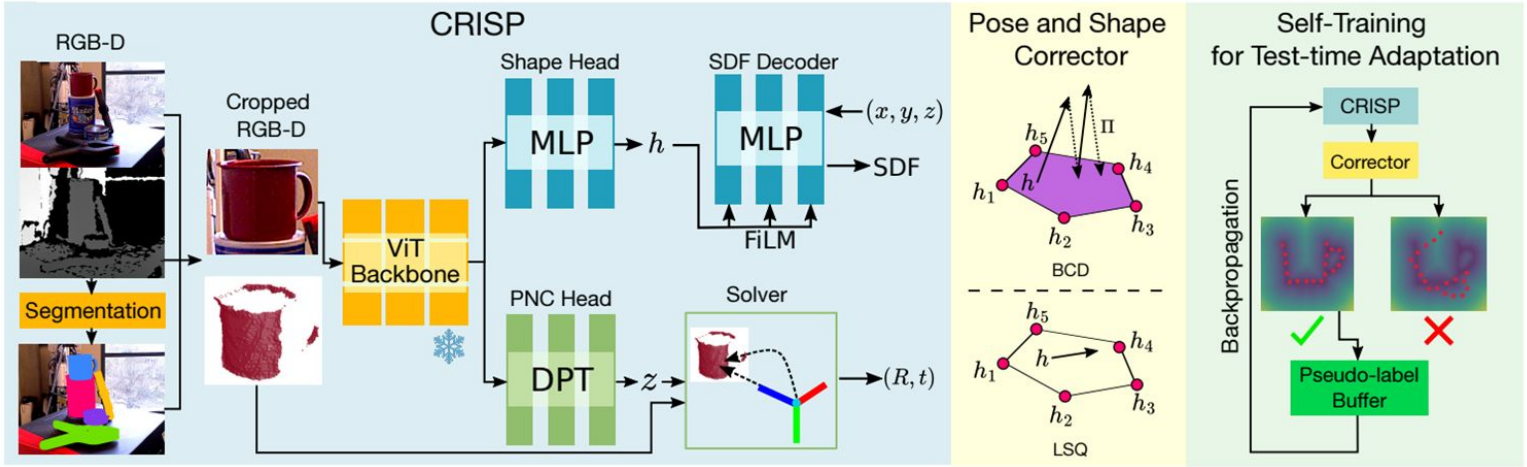
https://nju-cite-mocaphumanoid.github.io/MotionPRO/

CVPR 2025 の動向・気付き (150/181)

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

- **概要:** CRISPは、単一のRGB-D画像から6次元物体姿勢と形状を推定する新しい手法。未見の物体への適応性と、テスト時自己学習による大規模ドメインギャップの橋渡しに優れる。このシステムは、ロボティクスや拡張現実などの実世界応用をターゲットに設計されている。
- **新規性:** CRISPの主な新規性は、物体のカテゴリに依存しない姿勢と形状推定アプローチにあり、事前知識を必要としない。また、効果的なテスト時適応メカニズムを導入し、未知の物体や環境に対する強い汎化性能を実現している。
- **気付き:** CRISPが物体のカテゴリーを事前に知らなくても物体の姿勢と形状推定を実行できる点は印象的。このカテゴリー独立性は大きな進歩であり、未知の物体が存在する現実世界のシナリオにおいて、極めて汎用性の高い手法となる。



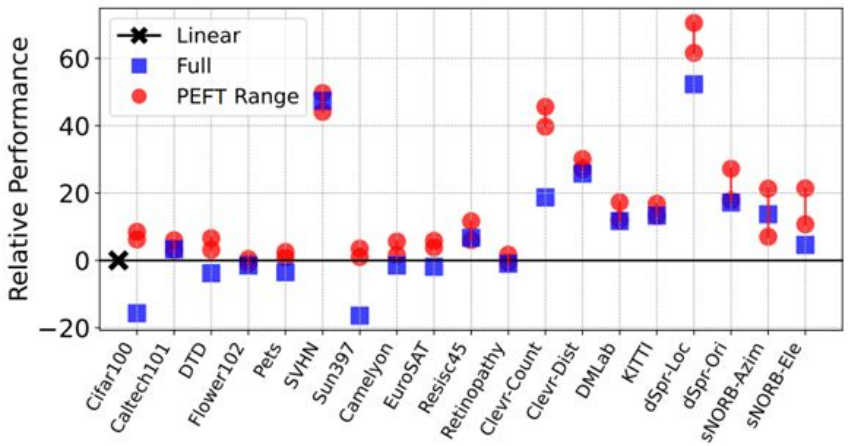


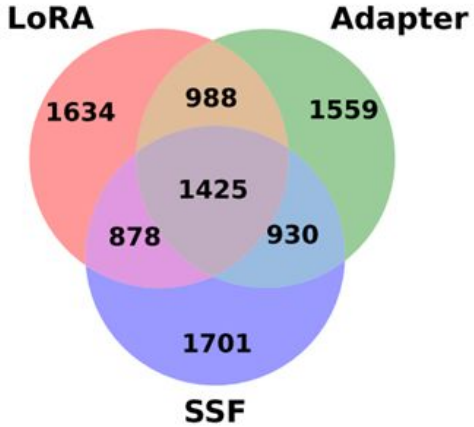
https://github.com/MIT-SPARK/CRISP?tab=readme-ov-file

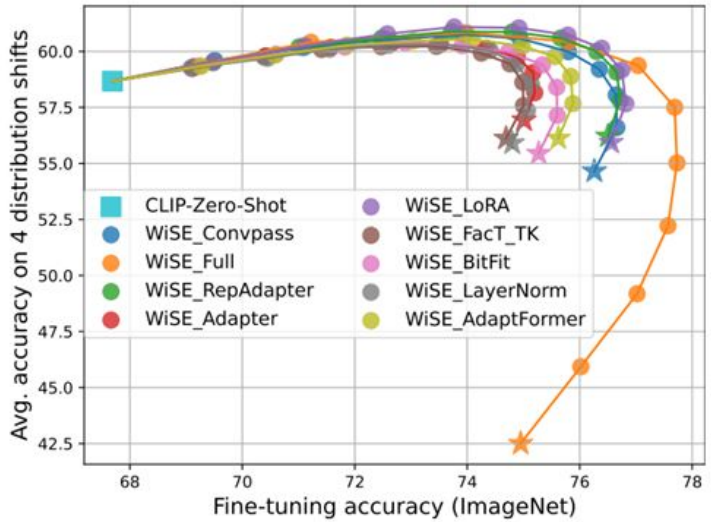
CVPR 2025 の動向・気付き (151/181)

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

- **概要:** 本論文は、視覚認識におけるパラメーター効率型微調整（PEFT）手法の研究を統合し、適切な微調整を施すことで、多様なPEFTアプローチが類似の精度を達成することを示している。
- **新規性:** 類似の性能を示すにもかかわらず、異なるPEFT手法は異なる予測とエラーを生成し、多様な帰納的バイアスが存在することを示唆している。
- **気付き:** 各PEFT手法が独自の機能的バイアスを有することが驚きで、単一の「万能解決策」は存在せず、手法固有の適用が不可欠であることを示している。







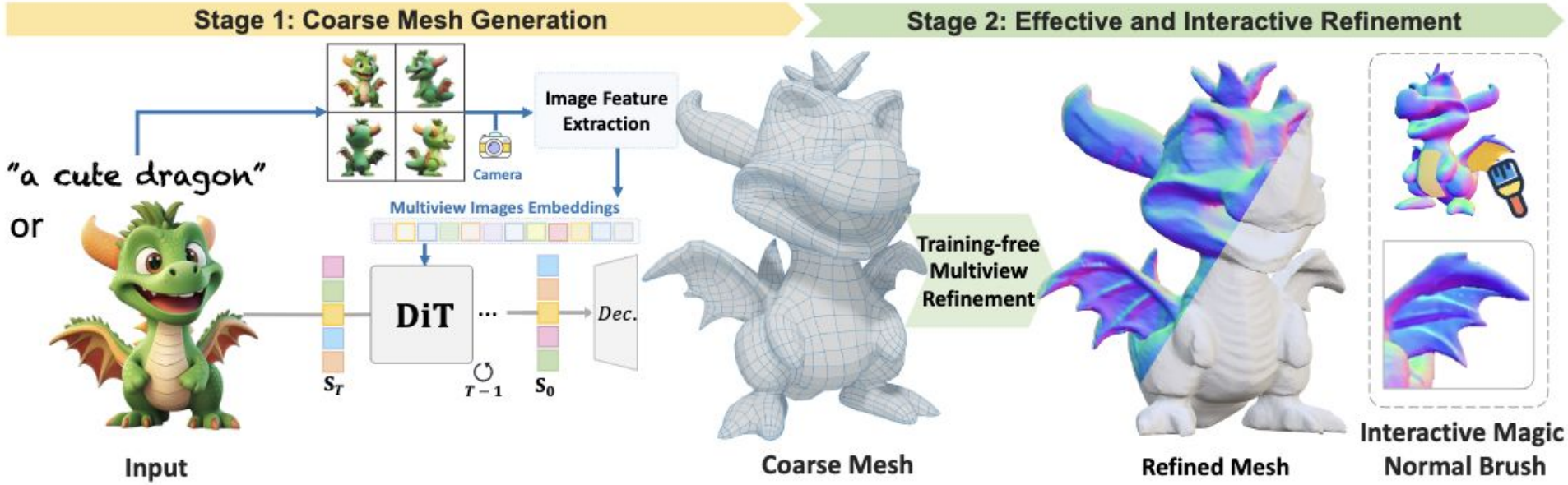
(a) Accuracy gain vs. linear probing on VTAB-1K (19 tasks) (b) Prediction overlaps (5K most confident) (c) Target distribution vs. distribution shifts

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

CVPR 2025 の動向・気付き (152/181)

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

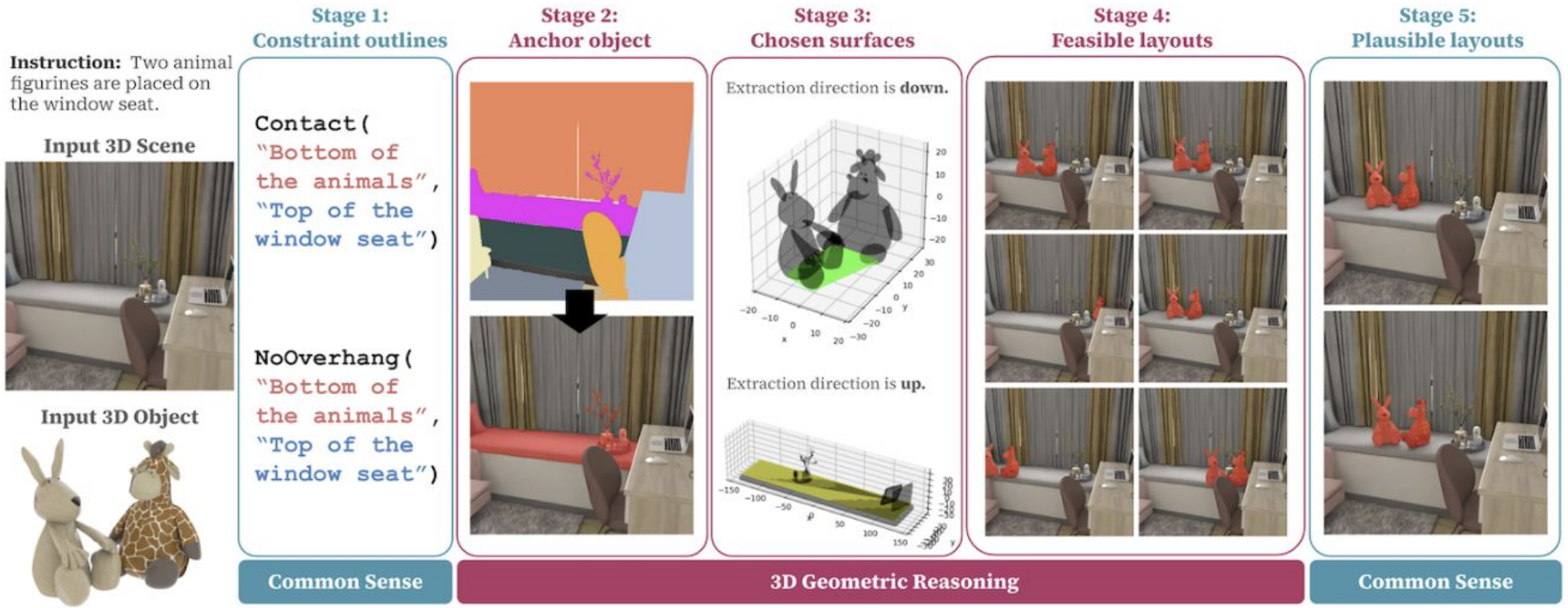
- **概要**: 高品質の3D形状を短時間で作成でき、インタラクティブな形状編集を可能にする新しい3D生成システムを開発。
- **新規性**: 高速、高品質、編集可能な3Dデータの生成モデルを提案する。
- **気付き**: これは3Dアセットを自動生成できる最先端の方法。LLMベースの3Dシーン生成と組み合わせれば、3Dデータ不足の問題を一気に解決できる可能性がある。



CVPR 2025 の動向・気付き (153/181)

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

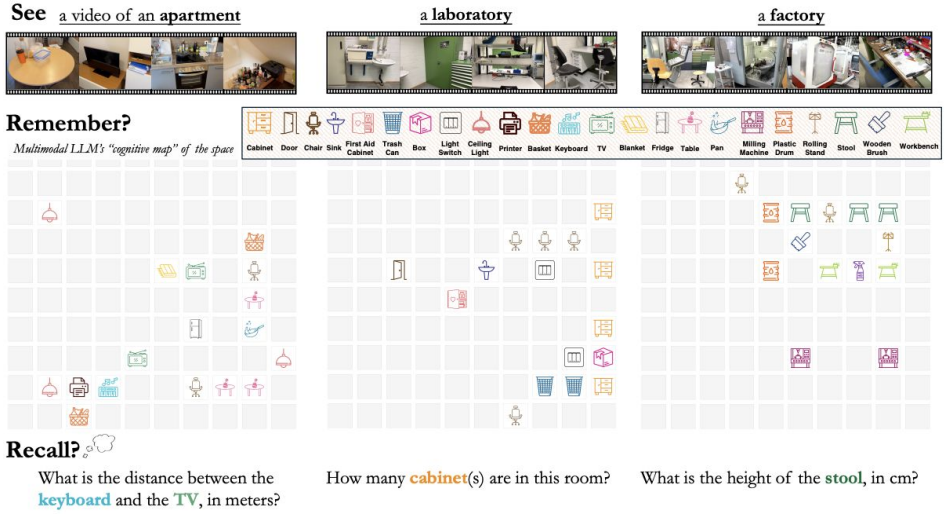
- ❏ **概要**: オブジェクト配置タスクに MLLM を最適に活用する方法を調査し、”FirePlace”を提案した。
- ❏ **新規性**: オプションを複数の段階で提示すると、MLLMの推論精度が向上する。

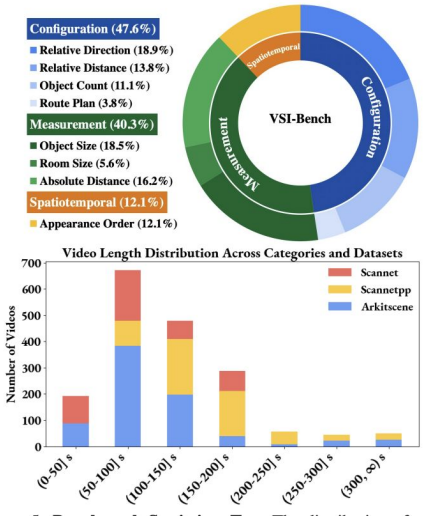


CVPR 2025 の動向・気付き (154/181)

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

- **概要**：MLLMは動画から生まれた新しい空間知能を示すが、推論には手間取る。そこで明示的なコグニティブマップが役に立つ。
- **新規性**: MLLMが動画から人間のような空間理解を深めることができるかどうかを調査し、視覚空間推論における主な制限を特定した。
- **気付き**:
  - 本研究では、現実世界の動画から得た多様な空間課題のベンチマークであるVSI-Benchを導入し、言語的自己説明と視覚的認知マップの両方を分析してMLLMを評価した。
  - データセットは、ScanNet、ScanNet++、ARKitScenesの動画と3Dアノテーションデータを統合し、自動アノテーションとヒューマンインザループリファインメントを使用して8つの空間タスクにわたって5,000を超えるQAペアを生成することによって構築されている。

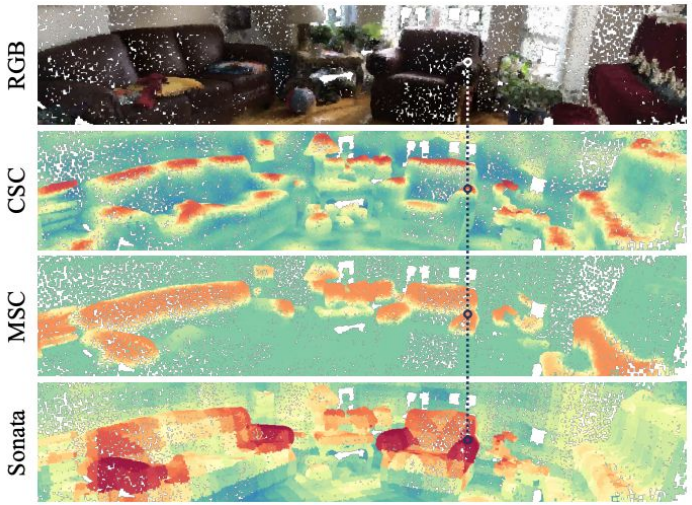


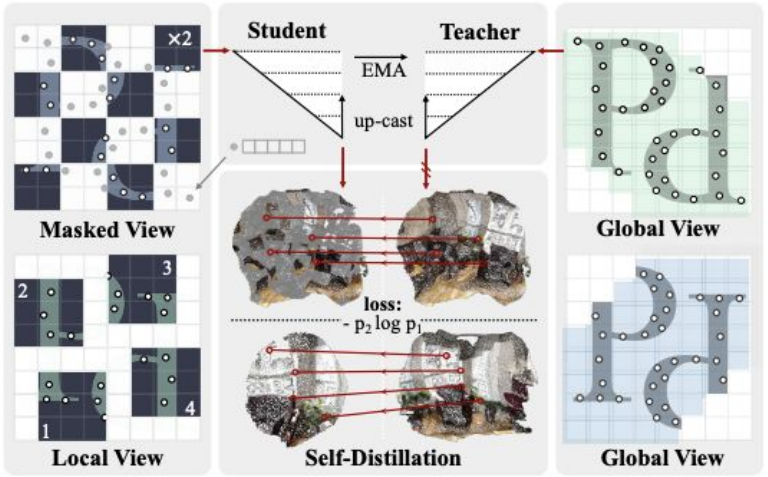


CVPR 2025 の動向・気付き (155/181)

Sonata: Self-supervised learning of reliable point representations

- **概要**: Sonataは、空間的な手がかりを抽象化して、入力された特徴への依存を強化することで、点群の自己教師付き学習における「幾何学的ショートカット」の問題を軽減し、線形プロービングや転送タスクに優れた、信頼性が高く一般化可能な3D表現を実現する。
- **新規性**:
  - マクロフレームワーク: ポイントセルフディスティレーション
  - ツービュー戦略: 拡張ポイントクラウドビュー (クロップ、ジッター、マスク) を生成
  - 機能の整列: EMA教師と生徒の機能を組み合わせて以下の機能を一致させる
    - ローカル/マスクビュー (Student)
    - グローバルビュー (Teacher)
  - Self-distillationはContrastive/Generative Learningに取って代わる
  - 損失: シンクホーンクラスタリング+KoLEO正則化によるロバストアライメント

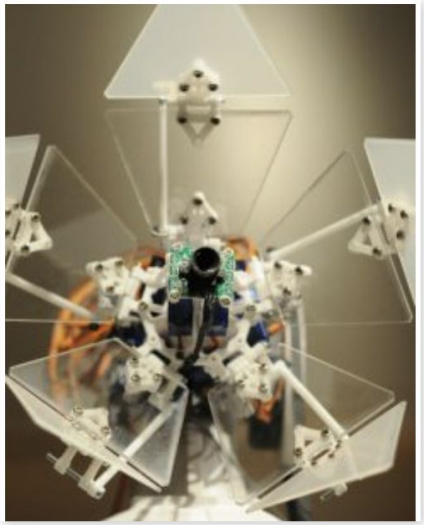


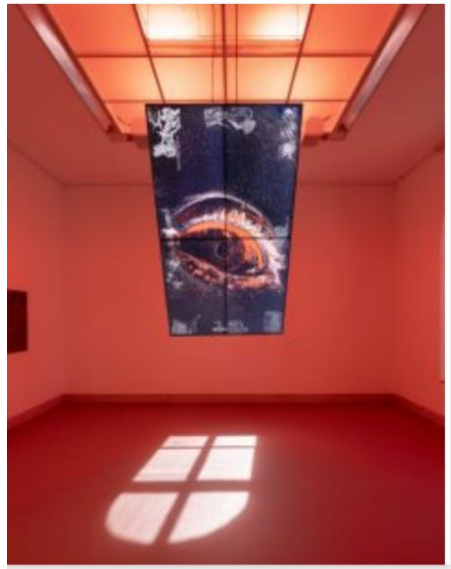


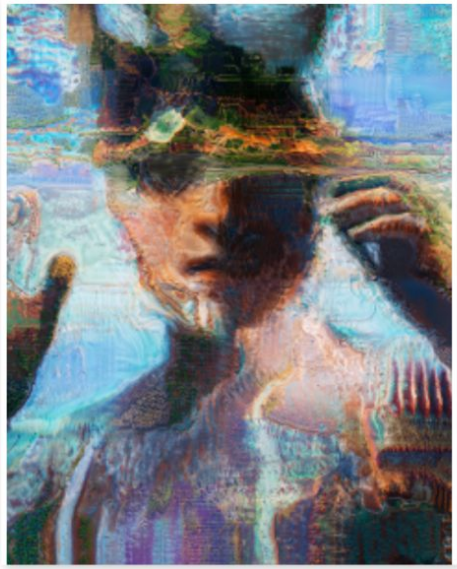
CVPR 2025 の動向・気付き (156/181)

CVPR AI arts

- ホールA2にはAIアートインスタレーションが展示された。最も印象的だったのは “The Flower”という作品である。この作品は人間に反応する 視聴者の顔と花びらを追跡することで感情、または意図が伝わり、静かに反応しているかのような動きを見せる。



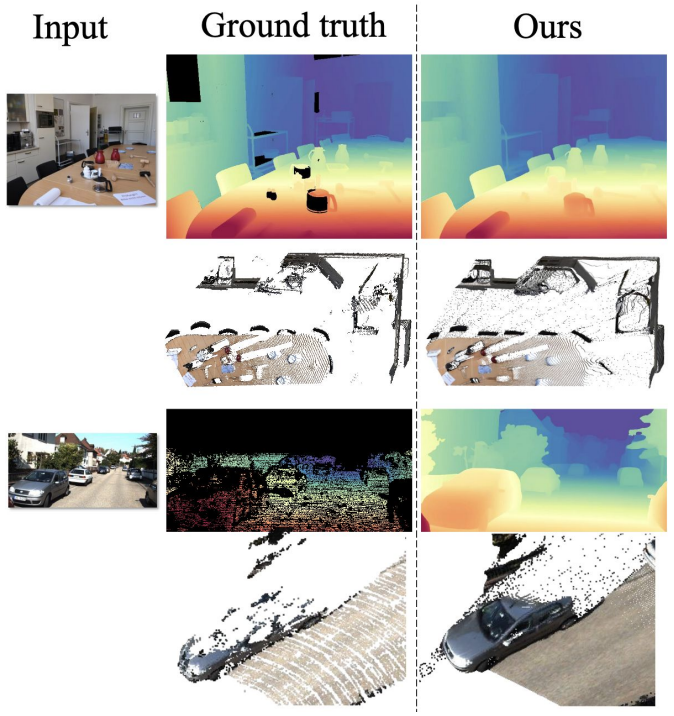




CVPR 2025 の動向・気付き (157/181)

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

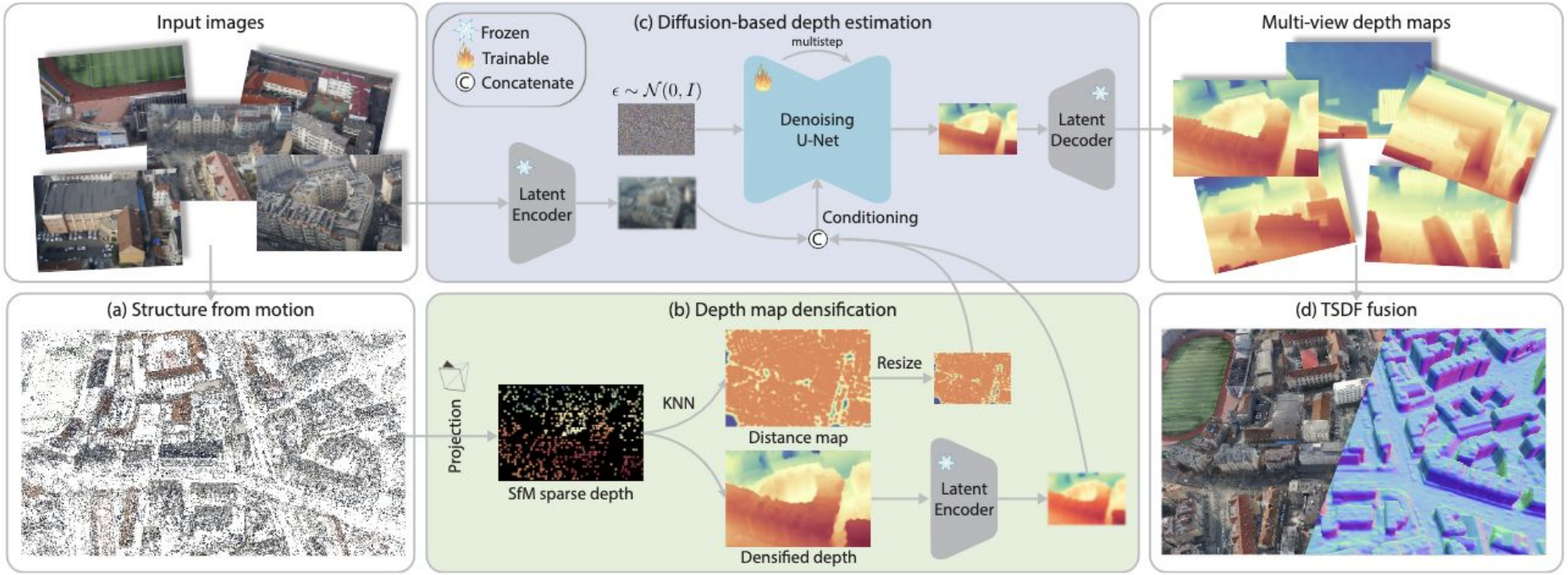
- **概要**：本論文は、単一の入力画像から 3D点群マップを生成するアーキテクチャを提案している。
- **新規性**：この研究では、スケールや変換の影響を受けない「アフィン不変」のポイントマップが導入されている。これにより、焦点距離のあいまいさがなくなり、より安定した学習が可能となり、カメラパラメータを必要とせずに単一の画像から直接　3D再構成が可能になった。
- **気付き**：これを認識タスクにどのように適用できるかに特に興味を惹かれた。具体的には、 1枚の画像から3Dシーンを再構築する際に、自動的に注釈を生成する方法はないかと考えた。たとえば、入力画像にすでにセマンティック・スーパービジョンやラベルが付いている場合、シーンの再構築プロセス中にそれらのアノテーションを伝播できるかも。



CVPR 2025 の動向・気付き (158/181)

Multi-view Reconstruction via SfM-guided Monocular Depth Estimation

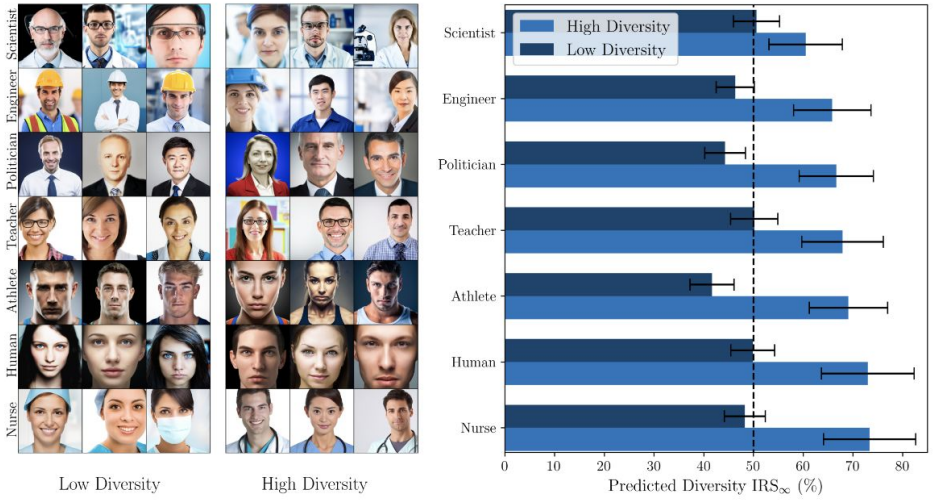
- **概要**: 本研究は、SfMからの事前情報を拡散ベースの深度推定に組み込む新しいアプローチを提案する。これにより、各視点の高精度でマルチビューの一貫した深度予測が可能になる。
- **新規性**: SfM点群を条件とする拡散モデルを使用して、マッチングを必要とせずにマルチビュー画像からスケールに一貫した深度を予測し、これを使用して高精度の3D再構成を実現する。



CVPR 2025 の動向・気付き (159/181)

Image Generation Diversity Issues and How to Tame Them

- **概要**:本研究は、画像生成タスクの多様性を評価するための新しい測定方法(IRS)を提案している。この指標を用いると、従来の拡散モデルは、条件を付けたとしても、トレーニングデータに存在する多様性の最大 77% しかカバーできないことが分かった。また、無条件モデルのパフォーマンスはさらに悪くなる。この洞察に動機付けられて、著者らは、画質を犠牲にすることなく多様性を高めることを目指して、生成のための疑似ラベル付けアプローチを提案している。
- **新規性**:この研究はダイバーシティの新しい指標を定義している。従来のモデルは、条件付き生成を適用しても多様性が低いままであることが示されている。

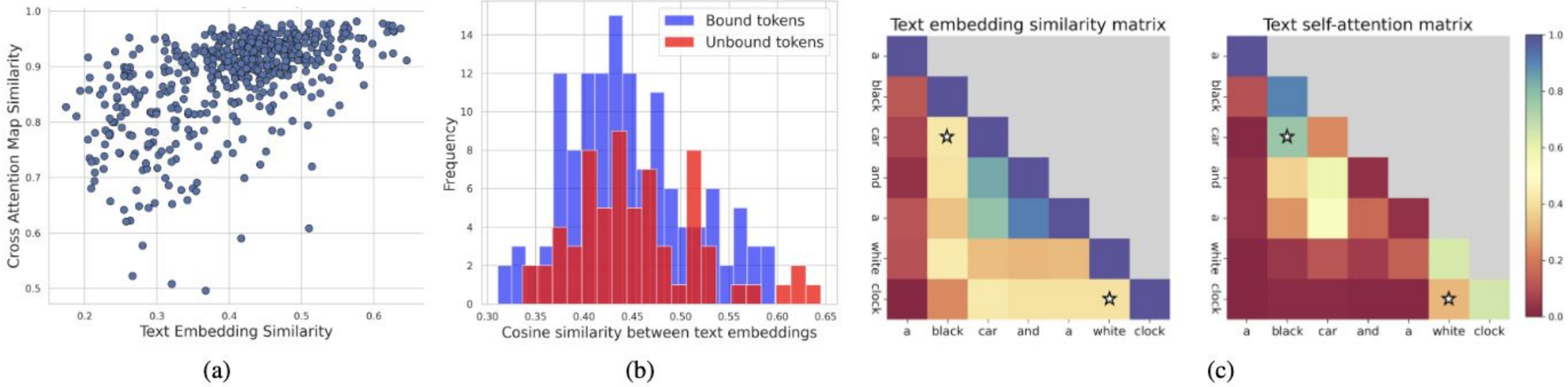


Model	Image Resolution	FID ↓	Prec. ↑	Rec. ↑	Dens. ↑	Cov. ↑	Vendi ↑	$IRS_{\infty,a}$ ↑ (Ours)
Pixel diffusion								
ADM-256 [12]	256	6.01 (30.30)	0.82 (0.57)	0.62 (0.73)	1.08 (0.41)	0.91 (0.40)	70.94 (36.18)	0.44 (0.20)
Transformer								
DiT-XL/2-256 [36]	256	22.15 (**8.72**)	0.94 (0.69)	0.34 (**0.76**)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

CVPR 2025 の動向・気付き (160/181)

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

- **概要**: 本研究では、既存の text-to-image モデルが入力テキストに忠実に追従できない問題を調査し、解決策を提案した。この問題は、クロスアテンションマップでは正しくない領域が強調される傾向があることが原因であることが判明した。この調査では、テキスト埋め込みでは、類似度の高いトークンの方がクロスアテンションマップでの値が高い傾向にあることが明らかになった。
- **新規性**: テキストエンコーダーのセルフアテンションはテキストの構造を捉えるため、この研究では、潜在変数の探索を通じてクロスアテンションがセルフアテンションマップを模倣するように促すことで、忠実度を向上させている。



CVPR 2025 の動向・気付き (161/181)

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

- **概要**: オープンワールド動画の奥行き推定は、カメラの動きやコンテンツのレイアウトが多様であるため困難だった。Depth Anything-V2 のような既存の手法を動画に適用した場合、時間的な一貫性が欠けている。この問題を解決するために、本論文では、事前にトレーニングされた画像から動画への拡散モデルを活用して、時間的に一貫した深度推定を行う方法を提案する。
- **新規性**: 多様な動画を処理するために、モデルはコンテンツの多様性については実際の動画を使用し、正確な深度監視には合成動画を使用してトレーニングしている。さらに、正確なコンディショニングには、クロスアテンションだけでなく、VAEの潜在空間でのコンディショニングも活用している。さらに、このモデルは、時間的レイヤーと空間レイヤーのどちらかを選択的に微調整するトレーニング戦略を採用しているため、ロングフレームの動画を効率的に処理できる。

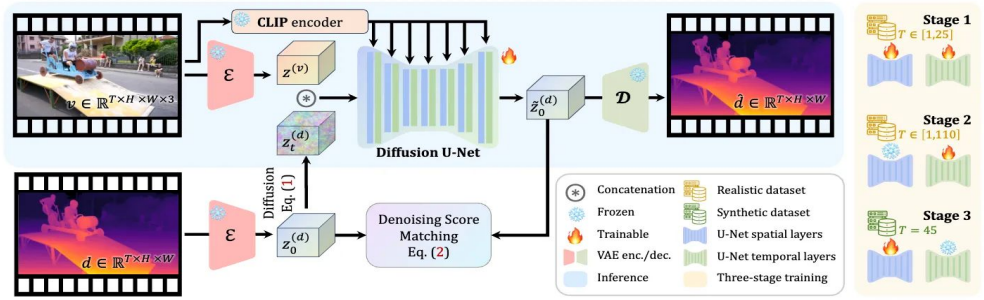


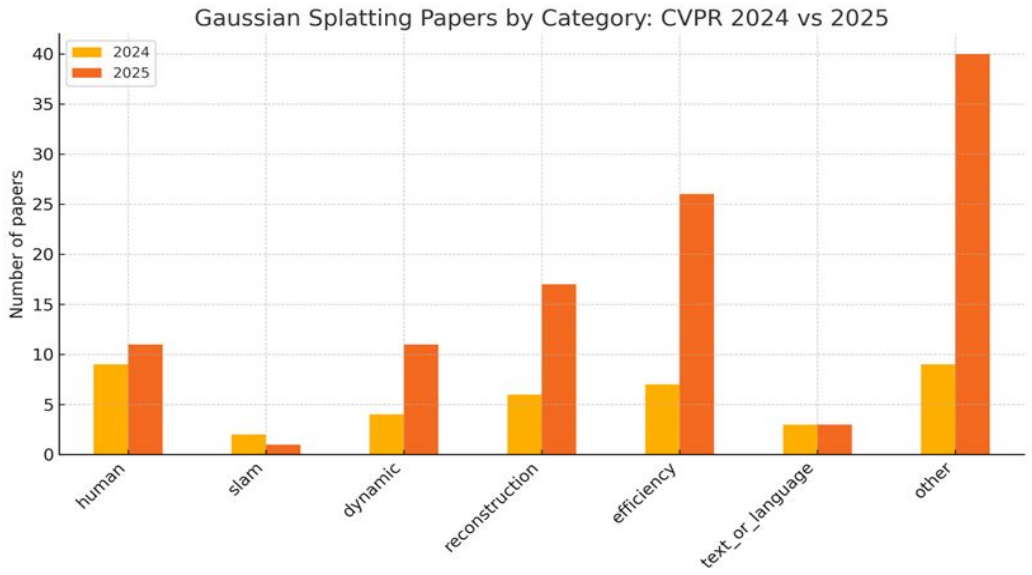
Figure 2. Overview of our *DepthCrafter*. It is a conditional diffusion model that models the distribution $p(\mathbf{d} \mid \mathbf{v})$ over the depth sequence $\mathbf{d}$ conditioned on the input video $\mathbf{v}$. We train the model in three stages, where the spatial or temporal layers of the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

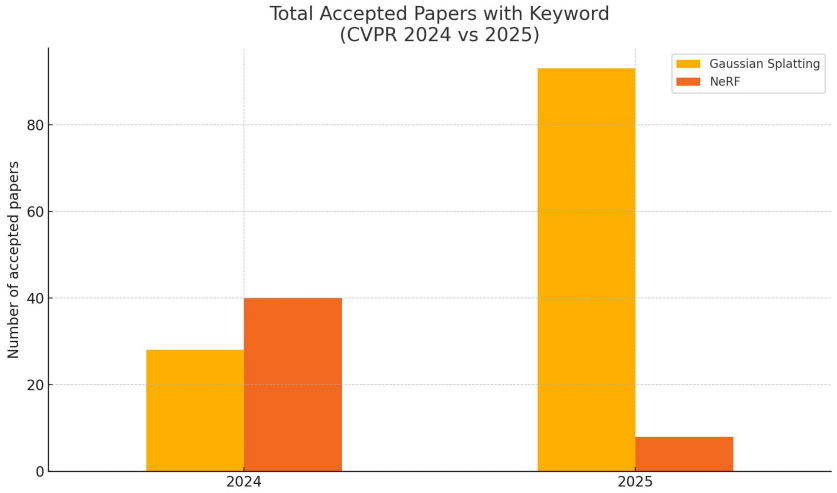
Figure 1. We innovate DepthCrafter, a novel video depth estimation approach, that can generate temporally consistent long depth sequences with fine-grained details for open-world videos, without requiring additional information such as camera poses or optical flow.

CVPR 2025 の動向・気付き (162/181)

Trends of Gaussian Splatting

- 2024年のタイトルに「Gaussian Splatting（GS）」を含む論文は28件だったが、2025年には93件（約3.3倍の増加）。
  - 3Dシーンの再構築がCVPR 2025の重点分野の1つであったため、Gaussian Splatting は注目されていた。
- CVPR 2024との比較
  - Efficiency/Sparsity: 26 論文 (最も増加が著しい論文)
  - シーンの再構築とダイナミックなシーン ≈ 3倍
  - ヒューマンアバターのシェアは 32% から 12% に減少
- 3D再構築においてGSがニッチからデフォルトにシフト
  - 主要な3Dシーン再構築方法としてNeRFに取って代わった。

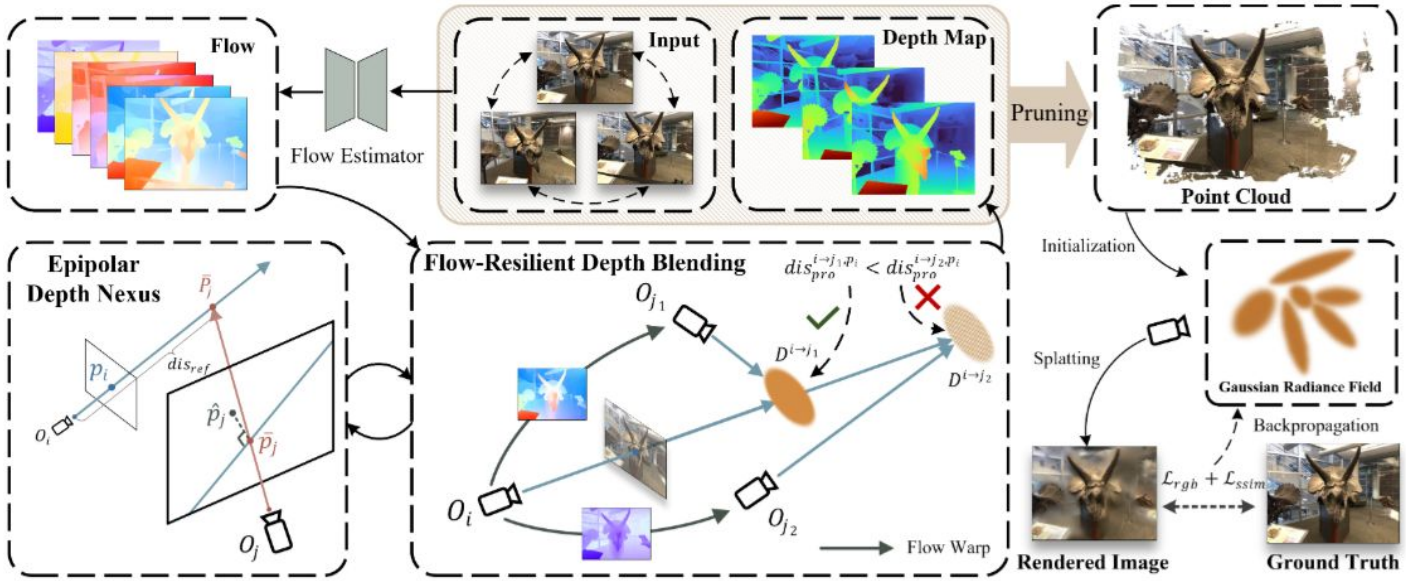


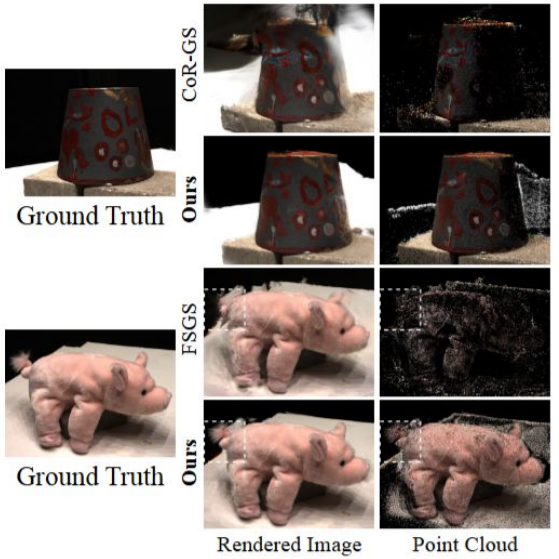


CVPR 2025 の動向・気付き (163/181)

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

- **概要**：エピポーラ深度優先順位を3D Gaussian Splatting に組み込んで、スパースビューの新しいビュー合成を改善。
- **新規性**：オプティカルフローを介してエピポーラ深度優先順位を持つ高密度ガウス分布を初期化する点群高密度化戦略が導入。フローレジリエントデプスブレンディングとフローフィルターデプスプルーニングを採用して、フローエラーを抑制し、まばらなビューでも正確な深度を生成。
- **気付き**：NexusGSはエピポーラ形状を巧みに利用してスパースビューの制限を克服し、大幅な正則化を必要とせずにロバストな深度初期化を実現している。その効果的な3段階のプロセスは、テクスチャのない領域では難しいオプティカルフローに依存。



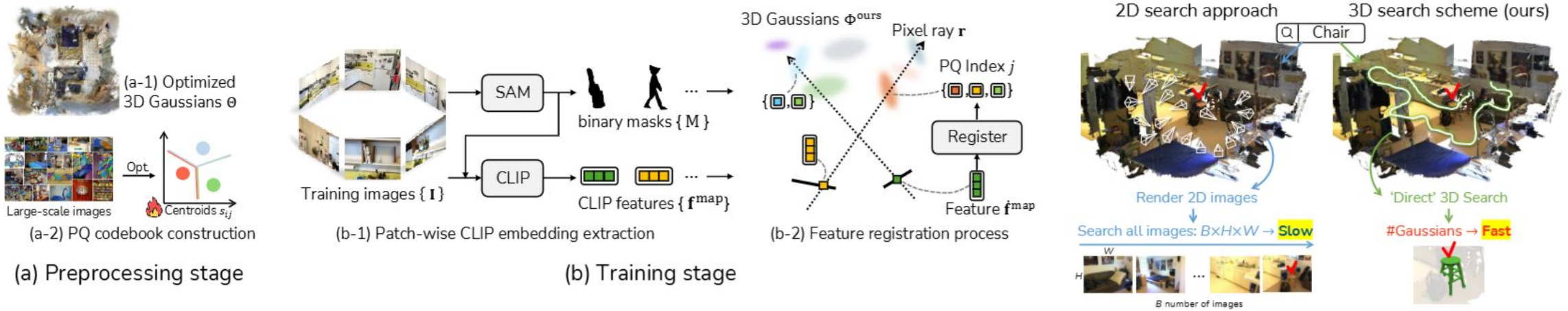


[Y.Zhengら、CVPR 2025。] [リンク]

CVPR 2025 の動向・気付き (164/181)

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

- **概要**: Dr. SplatはCLIP埋め込みを3Dガウシアンに直接登録し、レンダリングをバイパスしてオープンボキャブラリーの3Dシーンを効率的に理解。
- **新規性**: CLIP埋め込みを主要な3Dガウス分布に割り当て、事前学習済みの製品量子化を統合して、シーンごとの最適化を行わずにコンパクトな埋め込み表現を実現する直接特徴登録手法。
- **気付き**: レンダリング段階をなくすというアイデアは魅力的。これにより、フィーチャの歪みが軽減され、3D クエリが高速化される。事前学習済みの PQ は、メモリと精度のバランスを効果的にする。

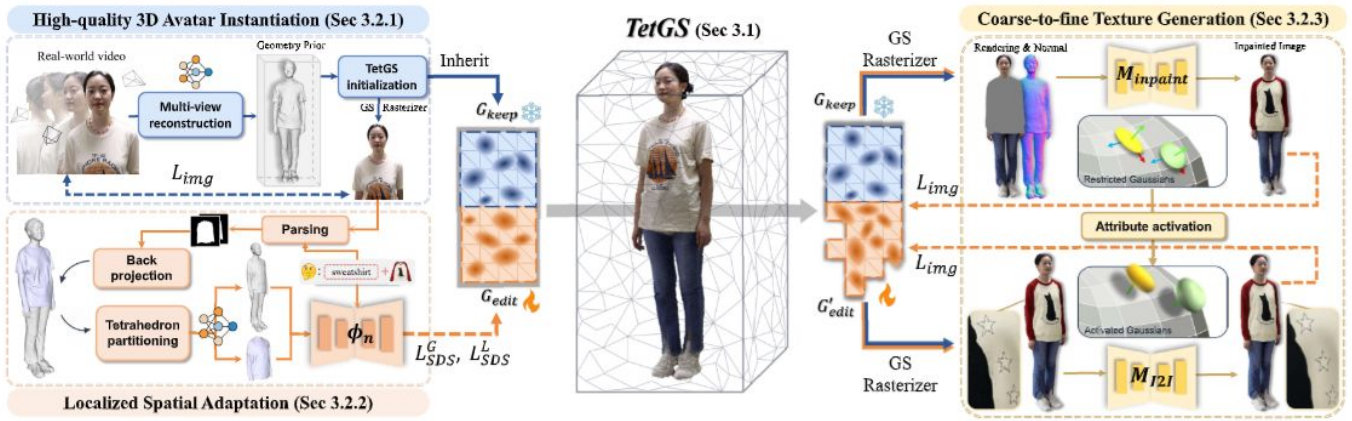


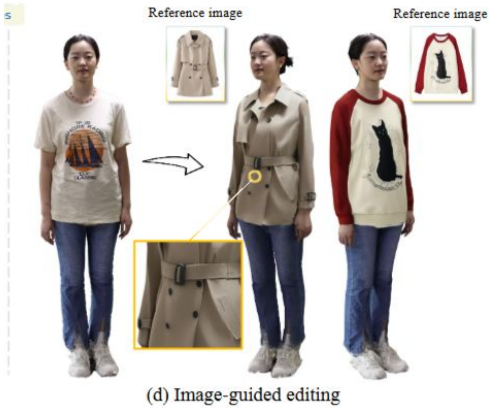
[K・ジュンソン他、CVPR 2025。] [リンク]

CVPR 2025 の動向・気付き (165/181)

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

- **概要**: ハイブリッドTETGSベースのパイプラインは、テキスト/画像ガイド付き編集により、単眼動画から編集可能なフォトリアリスティックな3Dアバターを生成。
- **新規性**: 本研究では、Gaussianカーネルを四面体格子に埋め込み、分離した空間適応と外観学習を実現するTETGを提案。これにより、局所的な正確なジオメトリ編集と、粗～微細かつ数ショットの監視によるフォトリアリスティックなテクスチャ生成が可能。
- **気付き**: 構造化された四面体グリッドを統合することにより、3D Gaussian Splatting編集における不安定性の問題に取り組んでいる。このアプローチにより、明確な幾何学的制御と忠実度の高い結果が得られる。計算負荷は大きいものの、この論文の段階的なパイプラインとハイブリッド表現は、実用的で使いやすい3Dアバターの作成における大きな進歩を示している。



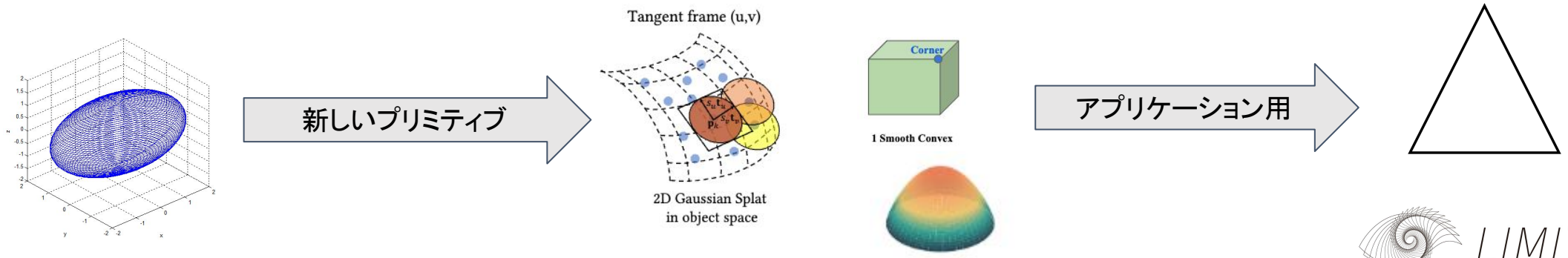


[H. Liu et al., CVPR 2025.] [リンク]

CVPR 2025 の動向・気付き（166/181）

From Photorealism to Geometric Integrity

- **3DGSの弱点: 幾何学的に不正確なサーフェス/ぼやけたエッジ/マルチビューの不一致。**
  - 主な研究トレンドは、単なる視覚的品質の向上から、これらの幾何学的欠陥の修正へと決定的にシフトした。
- **「スプラット」はもはや単なる3D blob（ガウシアン）ではない。**
  - 幾何学的な精度を実現するために、新しいプリミティブが次々と登場。
    - **2D サーフェルスプラッティング** 表面の一貫性を保つため
    - **3D 凸型スプラッティング** 鋭いエッジのオブジェクト用
    - **変形可能なベータスプラッティング** より少ないパラメータでより高い忠実度を実現
- **グラフィックパイプラインとの統合**
  - 高度なプリミティブが導入されたが、それでもカスタム表現どまり。ゲームやVFXのような多くのアプリケーションにとって究極の目標は、標準的な三角メッシュを生成すること（直接ポリゴンを生成すること）。
    - **Triangle Splatting（トライアングル・スプラッティング）** は、この分野の潜在的な収束点を示しており、基本的なグラフィックスプリミティブを直接最適化可能に。



CVPR 2025 の動向・気付き (167/181)

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

- **概要**: 離岸流検出のための大規模な動画インスタンスセグメンテーションベンチマーク (RipVIS) を作成(開発に約3年)。

- **気付き**: 日本にはすでに離岸流を予測するための優れたシステムがあるが、それはクローズなシステムであるため、著者らは日本の共同研究者を見つけたいと考えている。RipVISを基にしたコンペティションがICCV 2025で開催される予定。

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

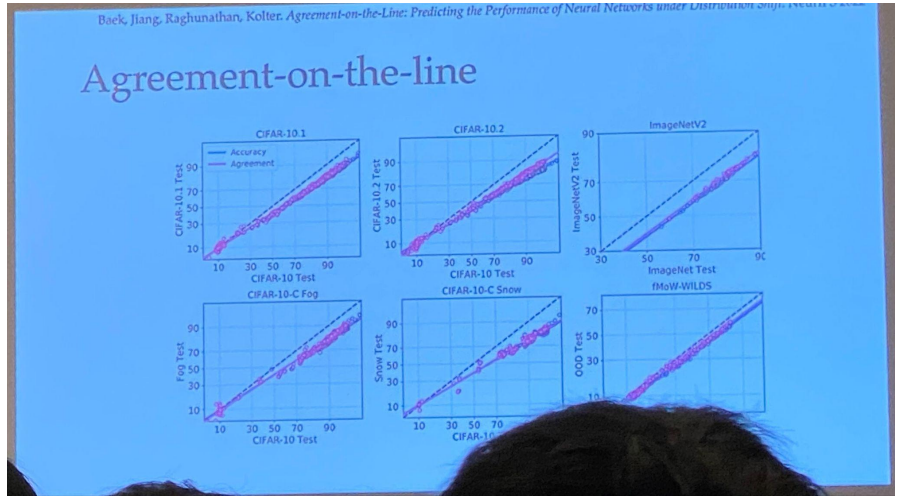
CVPR 2025 の動向・気付き（168/181）

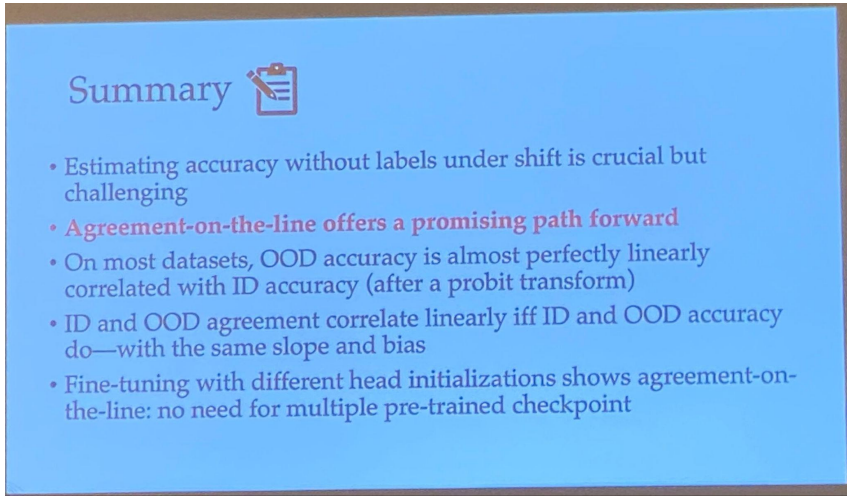
Workshop: Domain Generalization: Evolution, Breakthroughs and Future Horizon

- Speaker: Aditi Raghunathan

  Title: Predicting the Performance of Foundation Models Under Distribution Shift

  - OODの精度はIDの精度とほぼ直線的に相関している。
  - IDとOODの一致度も ID と ODD の精度がそうである限り、線形相関関係が成り立つ。
    - この経験的現象により、**ラベルなしデータを用いて分布シフト下での基盤モデル（FM）の性能を推定**　することが可能。（論文）
  - しかし、複数のFMで合意を計算することは、計算上不可能。
    - **リニアヘッドのランダム初期化**　で代用（複数のチェックポイントは不要）。





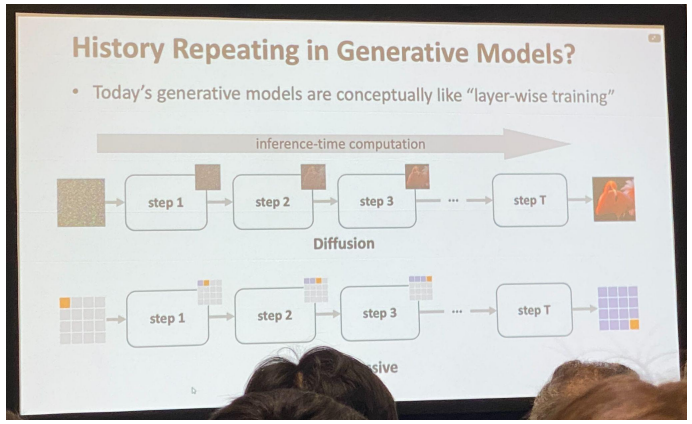
CVPR 2025 の動向・気付き (169/181)

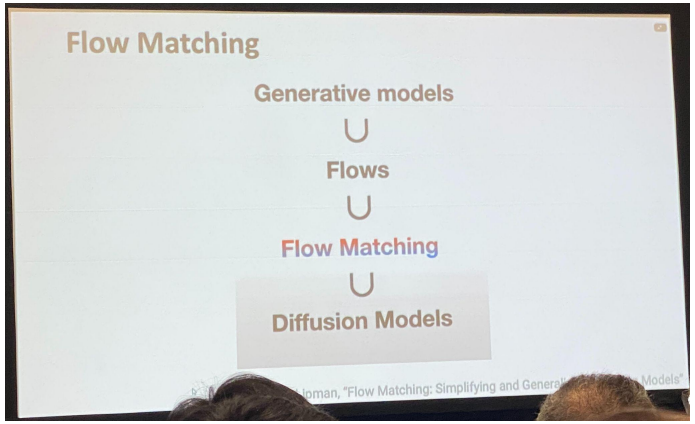
Workshop: Visual Generative Modeling: What's After Diffusion?

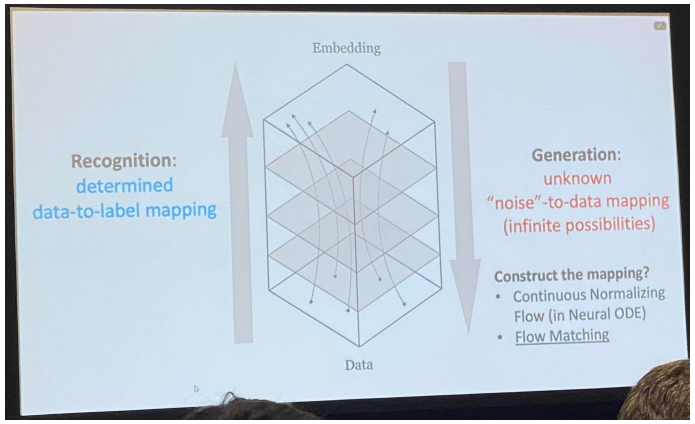
- Speaker: Kaiming He

  Title: Towards End-to-End Generative Modeling

  - AlexNet導入前 レコグニションでは、レイヤーワイズトレーニングが一般的なソリューションだった。
  - 今日のジェネレーティブモデルは、概念的にはまだ「レイヤーワイズトレーニング」のようなもの。
    - **ジェネレーティブモデルを端から端まで作ることができれば**　改善の余地がかなりあるかもしれない 。
  - Flow Matching手法は有望な方向。
    - 関連：Mean Flows for One-step Generative Modeling



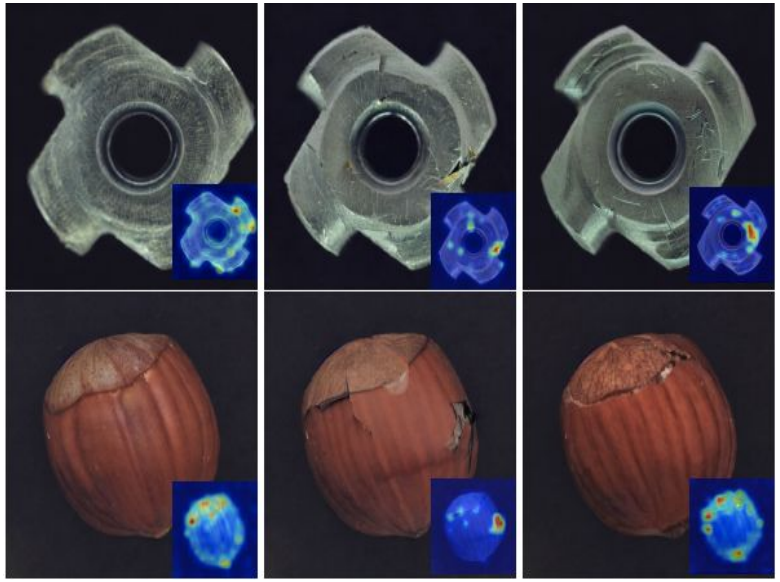




CVPR 2025 の動向・気付き (170/181)

Unseen Visual Anomaly Generation

- 概要:
  - **テスト時に与えられた単一の正常画像と事前学習済みの　Stable Diffusionモデルのみを使い** 、多様かつ現実的な「未知の」異常画像を生成する新しいフレームワーク「 Anomaly Anything (AnomalyAny)」を提案。
- 新規性:
  - クロスアテンションに基づく最適化と迅速な改良を独自に活用して、追加のトレーニングなしで異常発生を誘導。
- 気付き:
  - トレーニングなしのアプローチは優雅だが、実際のワンショット異常の例を取り入れることで生成精度をさらに高める。



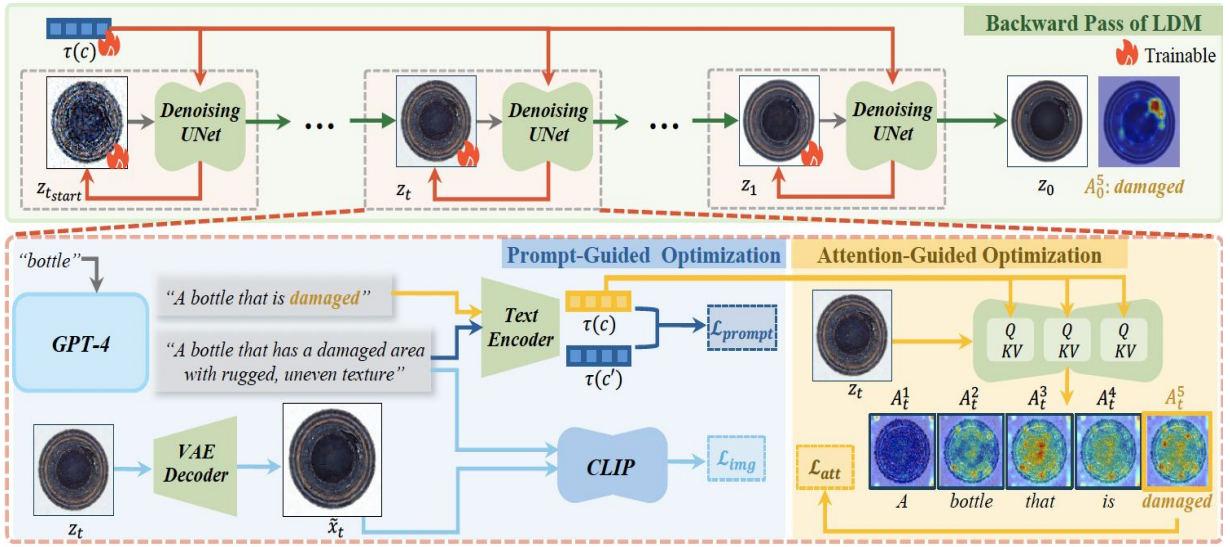


Figure 2. **Illustration of AnomalyAny** with details of the attention-guided & prompt-guided optimization process at time step t.

CVPR 2025 の動向・気付き (171/181)

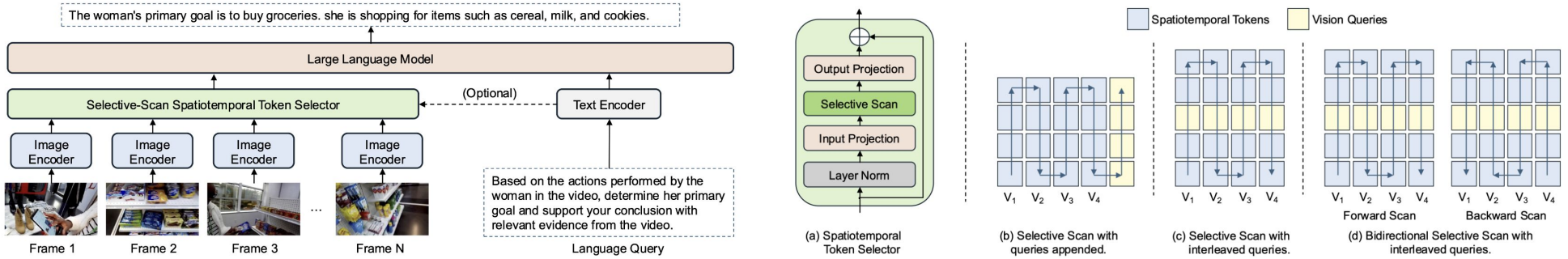
Workshop on Video Large Language Models

- 評価ベンチマーク
  - 幻覚と不作為 (アーガス)
  - 構成上の推論 (ベロシティ)
  - ロードイベント (ロードソーシャル)
  - 時間的理解 (ロスト・イン・タイム)
- 長い動画の効率的な処理
  - マルチフレームフュージョン (フィラ・ビデオ)
  - 状態空間モデル (ビンバ)
  - フレームサンプリング (モーメントサンプリング)
- 動画グラウンディング
  - 監視が弱い (コスパル、STPro)
  - 低コスト (NumPro)
  - ピクセルグラウンディング (PG-動画ラバ)
- その他
  - 現在のトレーニングのボトルネック
  - ゼロショットアクションのローカリゼーション

CVPR 2025 の動向・気付き (172/181)

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

- 概要
  - VMLLMによる動画理解は、長い動画を扱う場合にかなり困難に直面。膨大な数の時空間トークンの処理の中心となるセルフアテンションメカニズムの二次計算コストが非常に高いことが問題。
- 新規性
  - BIMBAは、Mambaの効率的なセレクティブスキャンを使用して高次元動画から重要な情報を抽出し、コンパクトで情報豊富なトークンシーケンスに縮小する。
- 気付き
  - 「バニラ」(線形、Qフォーマー) メソッドでは、GPUメモリの問題がすぐに発生し、プーリングが長期的な依存関係をキャプチャするのに苦労する。そのため、時間情報を効率的にモデル化できる SSM ベースの手法は、動画の理解においてさらに進歩すると予想される。



CVPR 2025 の動向・気付き (173/181)

Watermark

- 生成モデル用
  - Robust Watermark against Fine-Tuning
  - Box-Free Watermark Removal
  - Black-Box Forgery Attacks on Semantic Watermarks for Diffusion Models
- 局所的なウォーターマーク/部分的な盗難防止
  - Robust Watermarking Scheme Against Partial Image Theft
  - Manipulation Localization
- 3D Gaussian Splatting用
  - GuardSplat, 3D-GSW
- その他
  - Robust Message Steganography
  - Open-source Dataset Copyright

CVPR 2025 の動向・気付き (174/181)

Video LLM

- フレーム/トークンの選択と圧縮
  - Dynamic Compression of Tokens
  - Adaptive Keyframe Sampling, Flexible Frame Selection, M-LLM Based Video Frame Selection
- 時間的理解
  - Sequential Knowledge Transfer
  - Consistency of Video Large Language Models in Temporal Comprehension
  - Fine-Grained Compositional and Temporal Alignment
- 時空間グラウンディング
  - VideoRefer Suite, LLaVA-ST, VideoGLaMM
- その他
  - Mitigating Action-Scene Hallucination
  - Real-time interactive streaming

CVPR 2025 の動向・気付き (175/181)

Hand-Object Interaction

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

CVPR 2025 の動向・気付き (176/181)

Workshop: 4D vision Modeling the Dynamic World

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

2. 3D リギング、4D オブジェクト組み込み関数の生成
   CVPR Poster Category-Agnostic Neural Object Rigging
   Birth and Death of a Rose
   Anymate: A Dataset and Baselines for Learning 3D Object Rigging

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

CVPR 2025 の動向・気付き (177/181)

Insight : 4D vision Modeling the Dynamic World

・Gaussian-Splattingは4D分野でも強力 4D LangSplat

- 4D LangSplatは、マルチモーダルLLMによって生成された、テキストキャプションを、4D gaussian splatting表現に埋め込むことで、オープンボキャブラリーかつ時間変化を意識したセマンティック・クエリを可能に。

・オブジェクト固有の時間変化の生成

- The rose of Deathは2D Diffusionから蒸留した4Dの事前知識によって、時間的オブジェクトの固有表現を学習。
  Birth and Death of a Rose

CVPR 2025 の動向・気付き (178/181)

VideoGigaGAN: Towards Detail-rich Video Super-Resolution

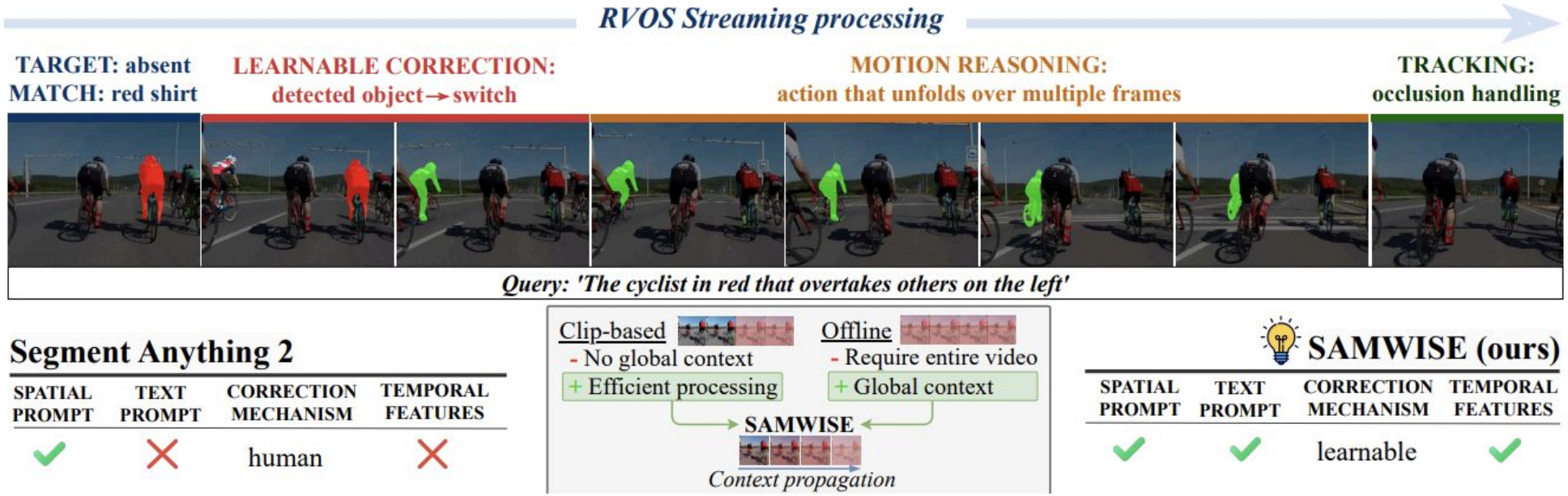
- 概要：提案されたモデル（動画GaN）は、出力フレーム全体で高品質と時間的一貫性の両方を実現し、既存のモデルよりも高い詳細度を実現。
- 新規性：VideoGigaganは、エイリアシングなどのアーティファクトを大幅に軽減しながら、高周波数のディテールと時間的な一貫性の両方を備えた動画結果を生成。
- 気付き：フレーム間の一貫性と品質の両方を実現するという問題 は将来の課題。

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

CVPR 2025 の動向・気付き (179/181)

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

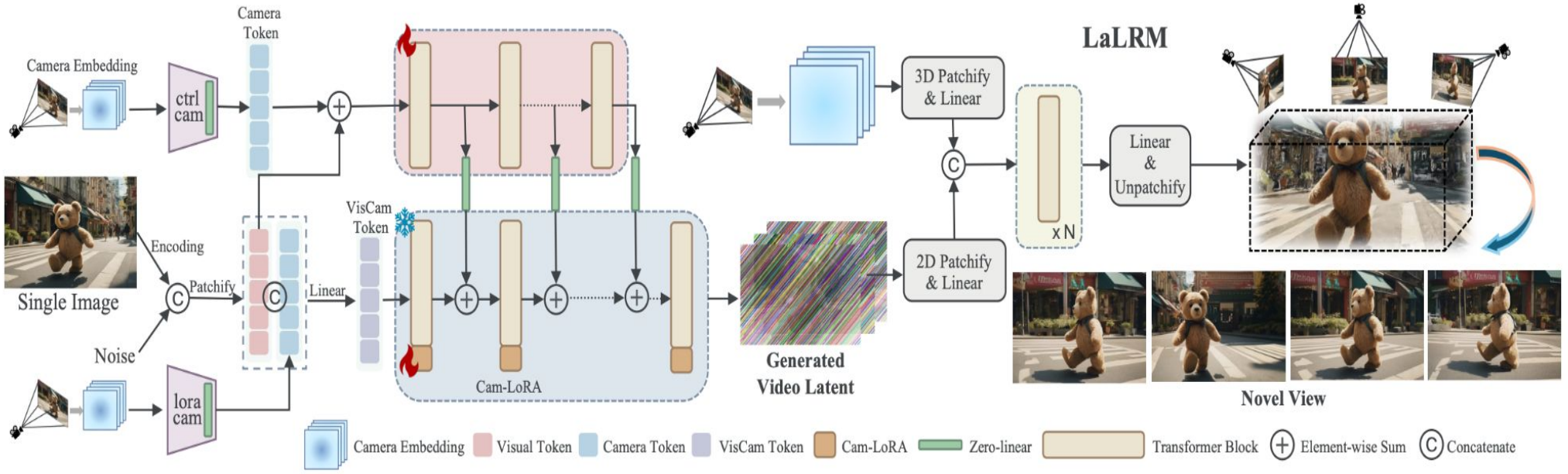
- **概要:** SAMWISEは自然言語理解と時系列モデリングをFine-tuningなしで統合することによりテキスト駆動型の動画セグメンテーションにSAM2を活用。動的オブジェクトのパフォーマンスを向上させるためにSAM2固有のトラッキングバイアスを調整する新しいモジュールも導入。
- **新規性:** SAM2のトラッキング制限を克服しながら、基盤モデルを再学習することなく、テキスト主導の動画セグメンテーションを効率的に実現。
- **気付き:** 昨年と同様、SAMやSAM2をベースにした派生研究が数多く行われている。Fine-tuningを行わずにテキストプロンプトを統合できたことが特に印象的。



CVPR25: Meta Insights into Trends and Tendencies (180/181)

Wonderland: Navigating 3D Scenes from a Single Image

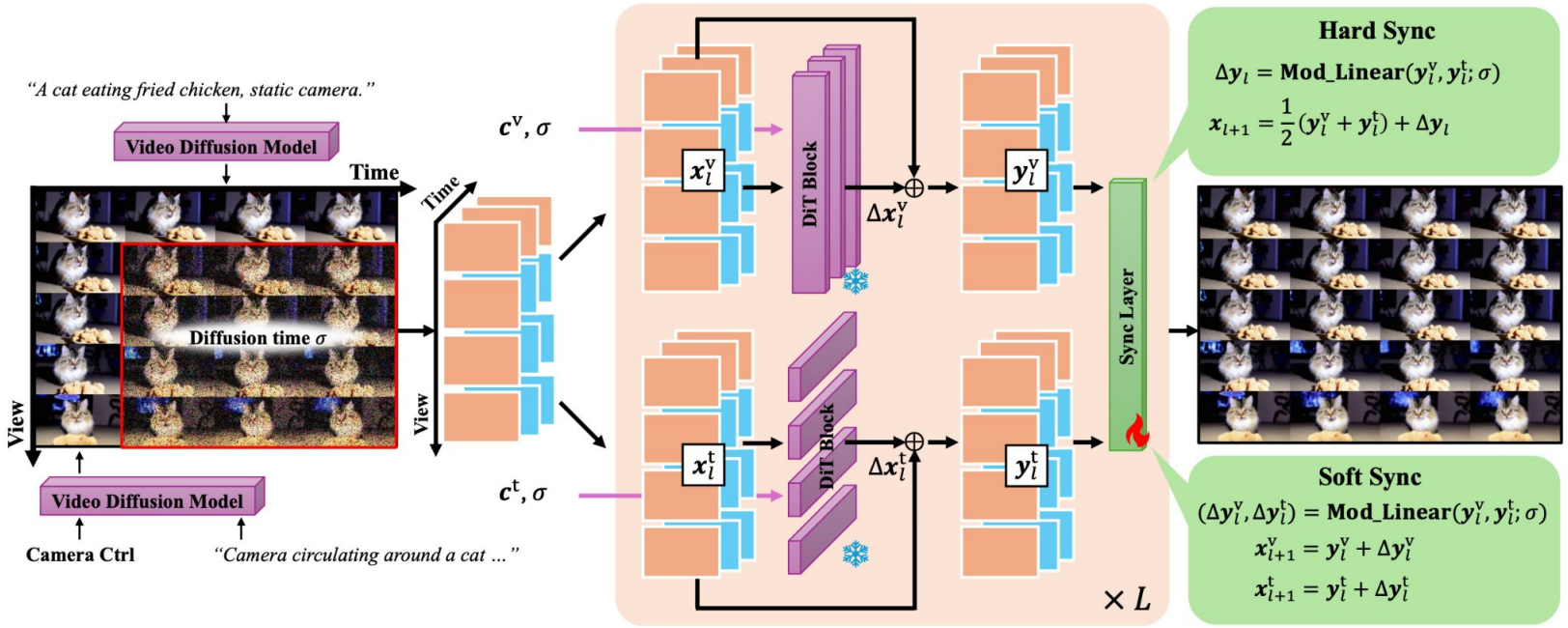
- 概要：画像＋カメラの軌跡から3Dシーンへ。カメラ制御可能な動画拡散モデルを再利用し、動画潜在空間からガウシアンをデコード（動画VAEデコーダを潜在LRMモデルと入れ替える）
- 新規性：デュアルカメラ制御モジュール- ローラ＋制御ネットの両方で、モデルがカメラの軌跡に厳密に従うように制御。潜在LRM - 凍結された動画拡散モデルの潜在空間から直接ガウシアンをデコード。
- 気付き：この方法は、動画を生成し、既製の3D再構成モデルを使用してGSを予測する（2ステップ）という単純なアプローチを簡素化し、動画VAEデコーダを再構成モデルに直接置き換えることで、これを1ステップのアプローチに抽出。

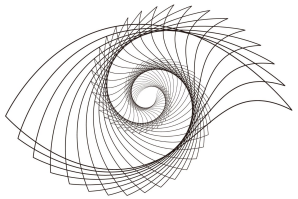


CVPR25: Meta Insights into Trends and Tendencies (181/181)

4Real-Video: Learning Generalizable Photo-Realistic 4D Video Diffusion

- **概要:** 4次元動画は、2次元行列（Y軸がカメラ、X軸が時間）として表現できる。2つの動画が与えられたとして、1つ目の動画は行（カメラが静止・時間が動的に変化）に対応し、もう1つの動画は、列（時間が静止・カメラが動的に変化）。このデータ構造をもとに、残りの行列を生成する拡散モデルを構築。
- **新規性:** 時間固定と動的カメラモジュールを導入。2ストリーム拡散アーキテクチャを採用。一方のストリームは列に対して視点更新を行い、もう一方のストリームは行に対して時間更新を行う。各拡散トランスフォーマー層の後に同期処理が実行。





LIMIT.Lab

A collaboration hub

for building multimodal AI models under limited resources

Our Members

Hirokatsu Kataoka
AIST / Oxford VGG

Yoshihiro Fukuhara
AIST

Rintaro Yanagi
AIST

Ryousuke Yamada
AIST

Daichi Otsuka
AIST

Partha Das
UvA

Nakamasa Inoue
Science Tokyo

Go Irie
TUS

Rio Yokota
Science Tokyo

Ikuro Sato
Science Tokyo

Rei Kawakami
Science Tokyo

Christian Rupprecht
Oxford VGG

Iro Laina
Oxford VGG

Yuki M. Asano
UTN FunAI Lab

Elliott (Shangzhe) Wu
Cambridge

Daniel Schofield
Oxford VGG

Limited Resources, Unlimited Impact with Multimodal AI Models

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

Left: Core members / Right: Our mission

LIMIT.Workshop

Home Call for Papers Program Organizers Contact

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

ICCV 2025 Workshop

October, 2025 Honolulu, Hawaii

Submit Paper Check Program

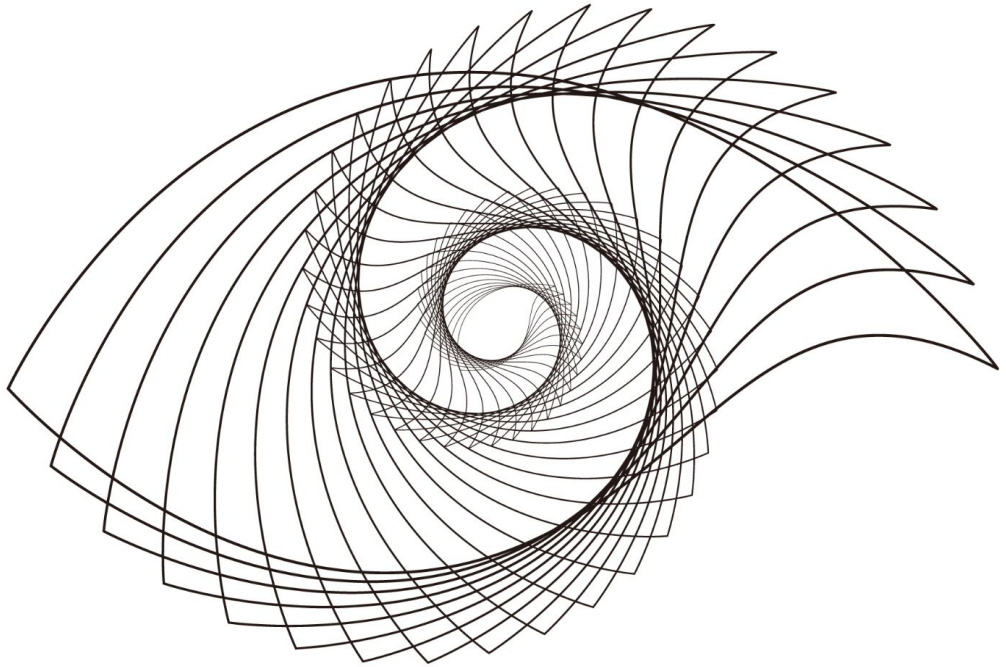
Please submit your paper to the ICCV25 LIMIT Workshop!

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

About LIMIT Workshop

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

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



LIMIT.LAB

https://limitlab.xyz/

Join us! -> Slack invitation [Link]