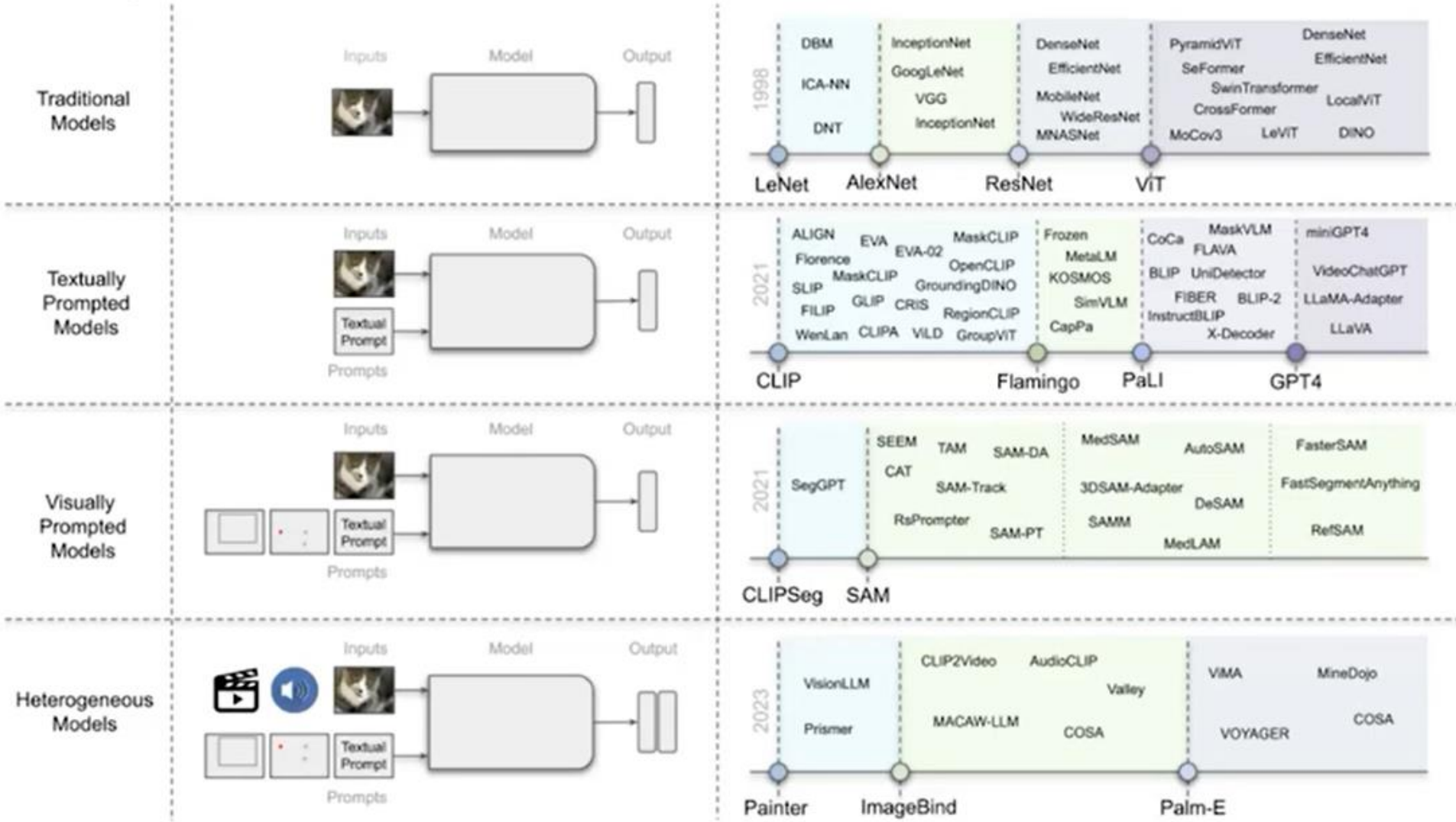
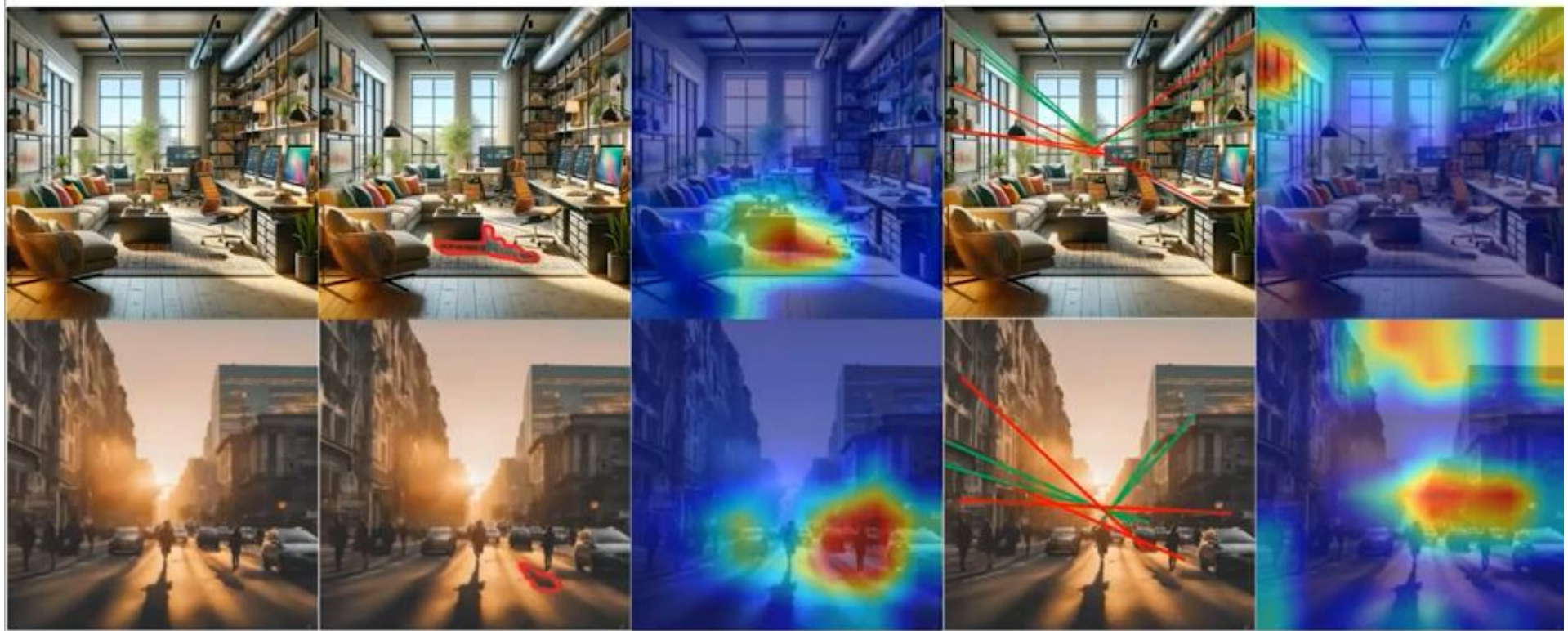
Computer Vision

Evolution of foundational model in computer vision



Awais et al., Foundational Models Defining a New Era in Vision: A Survey and Outlook arXiv preprint arXiv:2307.13721, 2023

Hallucination



Generated Image | Shadow Errors | Detected Shadow Errors | Vanishing Point Errors | Detected Perspective Errors

Sarkar et al., Shadows Don't Lie and Lines Can't Bend! Generative Models don't know Projective Geometry...for now, CVPR 2024

Bias and fairness

- Biases, stereotypes, and prejudices related to race, underrepresented groups, minority cultures, and gender can make the models output biased predictions or exhibit skewed behavior

Figure 8: **Bias of CLIP.** We zero-shot classify a random COCO image with a male-looking and a female-looking individual, also adding a circle over each of them. We score the following sentences to the images: *This is an image of a* {*woman, man, missing person, suspected murderer* }. The apparent gender resolution is correct, but the circled images tend to be scored higher as missing or murderers. Blur added for privacy.

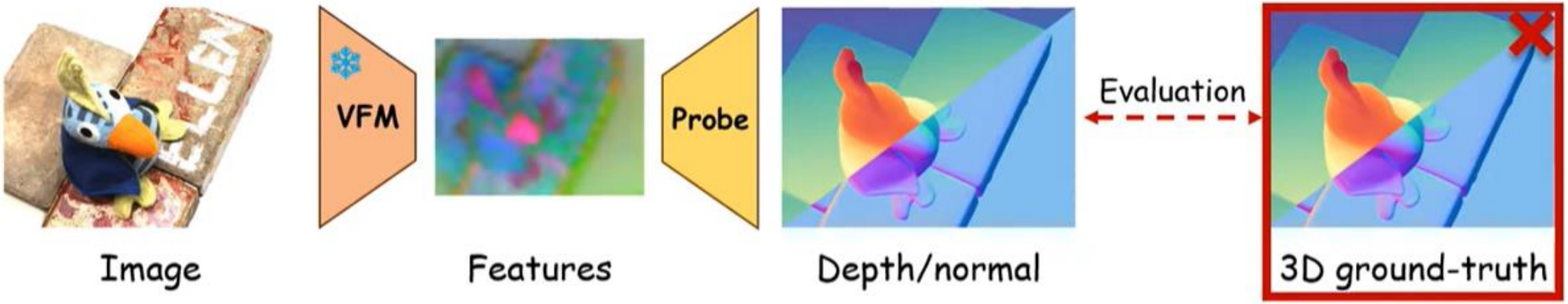
Shtedritski et al., What does CLIP know about a red circle? Visual prompt engineering for VLMs. ICCV 2023

Visual Foundation Models

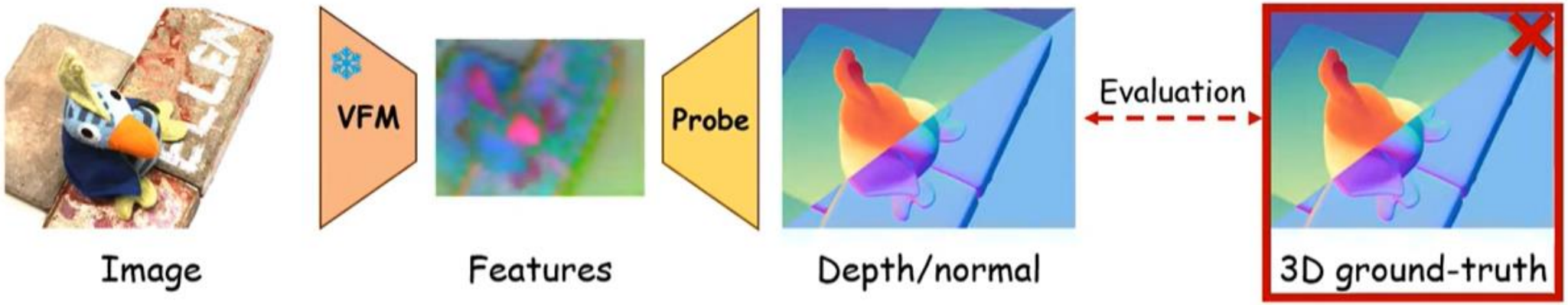
How well do they understand the **3D world**?

3D Probing!

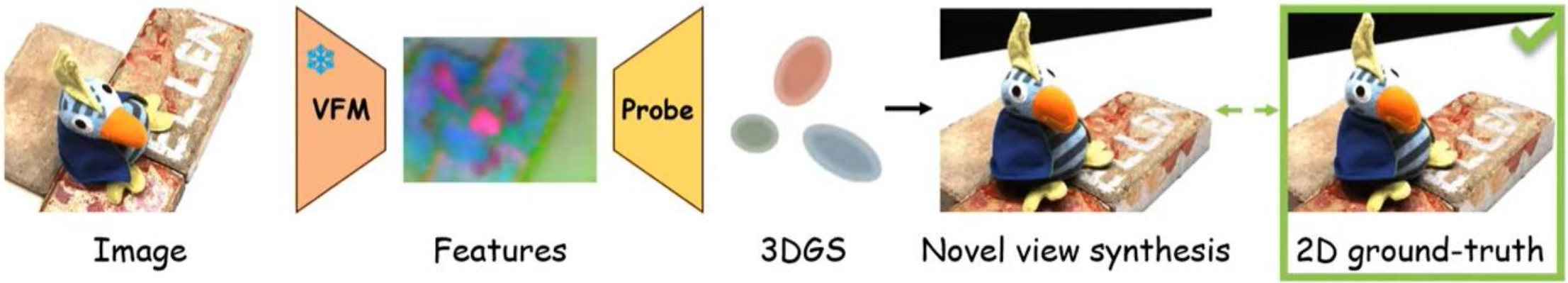
3D Probing


VFM
Probe
Evaluation
Image
Features
Depth/normal
3D ground-truth

3D Probing

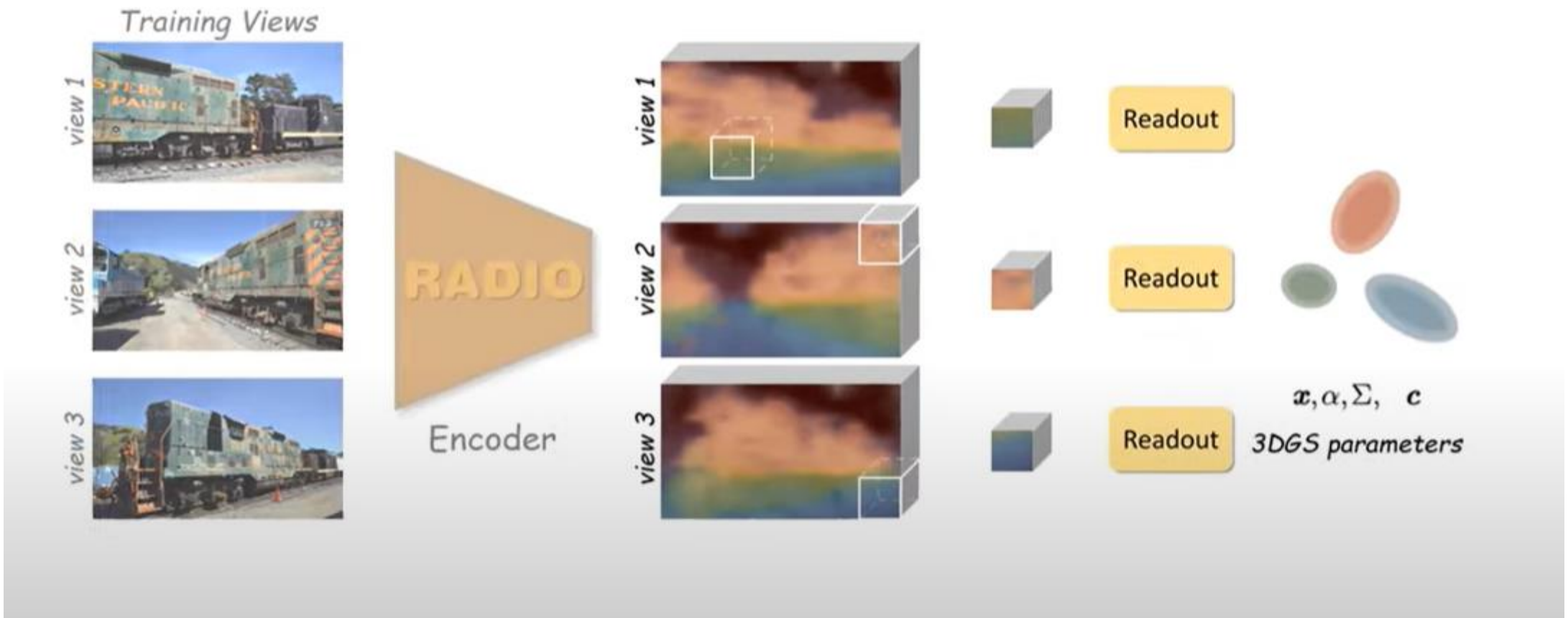


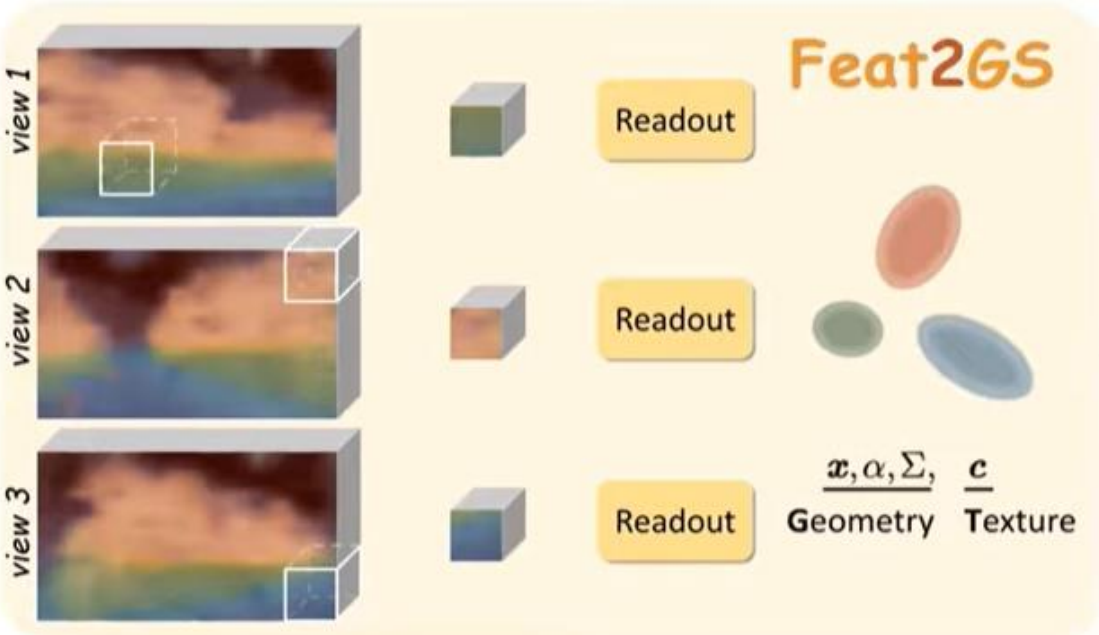
New Benchmark



Feat2GS as VFM Probe

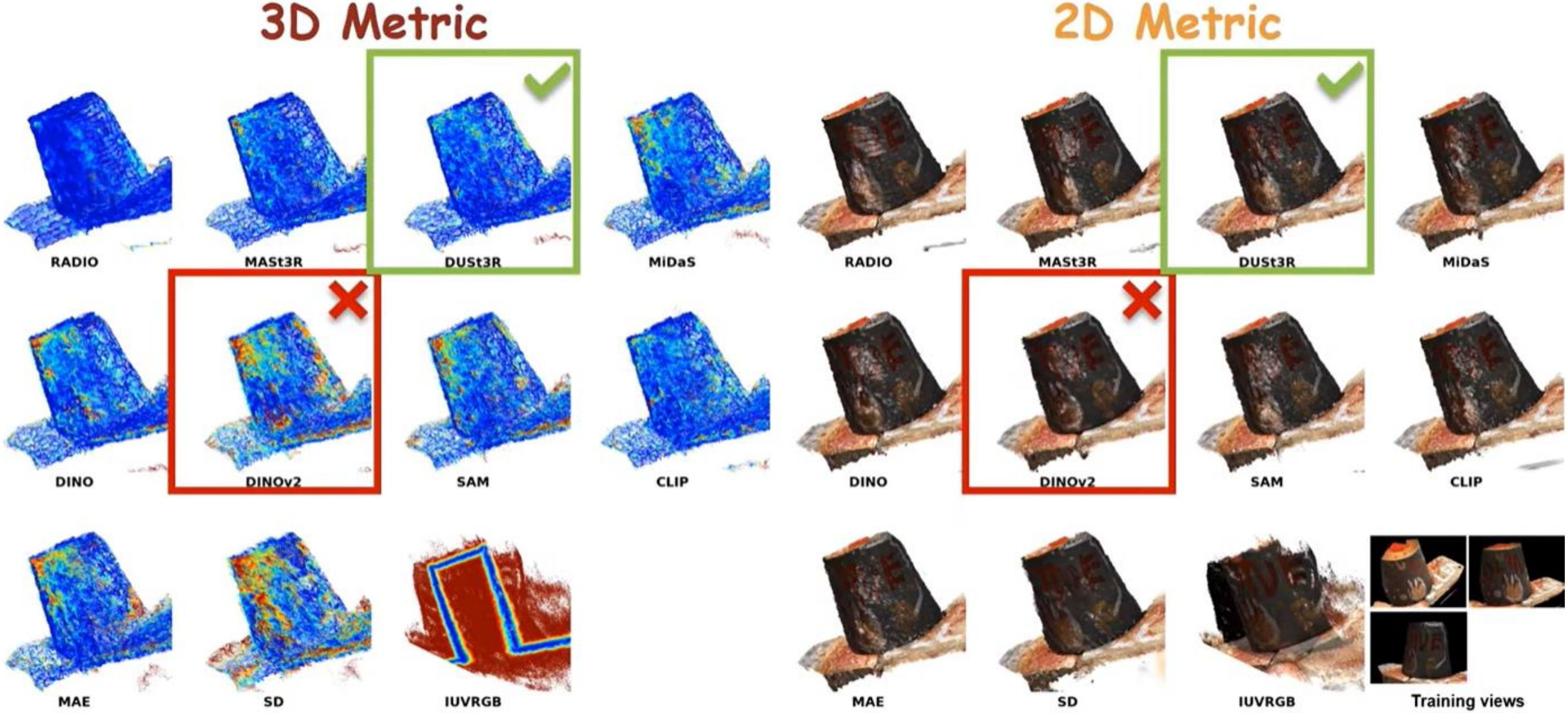
Novel View Synthesis as a proxy task


Training Views
view 1
view 2
view 3
RADIO
Encoder
view 1
view 2
view 3
Readout
Readout
Readout
$x, \alpha, \Sigma, \quad c$
3DGS parameters



Splatting

Novel View Synthesis


3D Metric
2D Metric
RADIO
MASt3R
DUSt3R
MiDaS
DINO
DINOv2
SAM
CLIP
MAE
SD
IUVRGB
Training views
Pointcloud Error Map
Novel View Synthesis

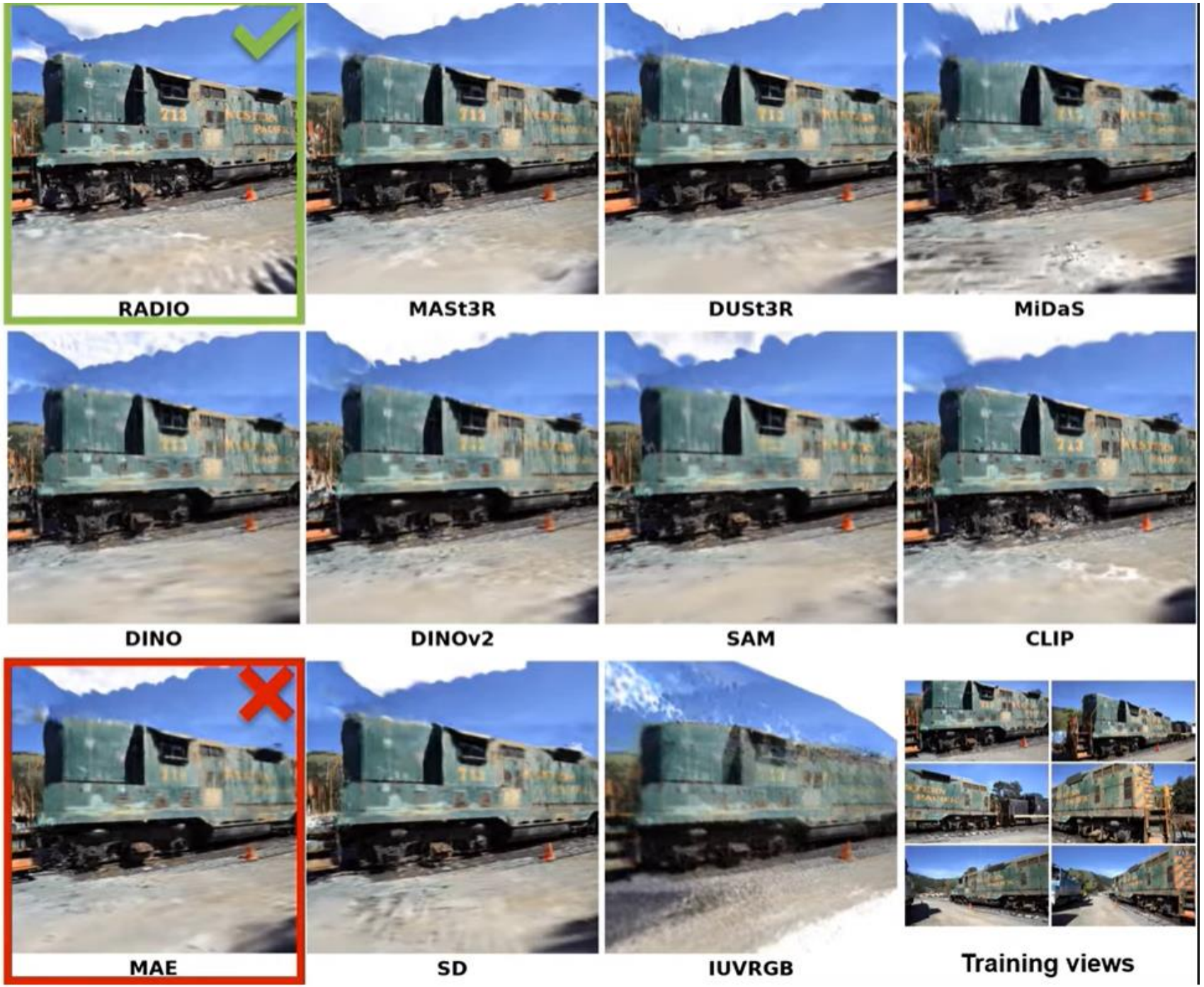
Geometry Probing

Geometry Probing

Geometry

x, α, Σ

Readout

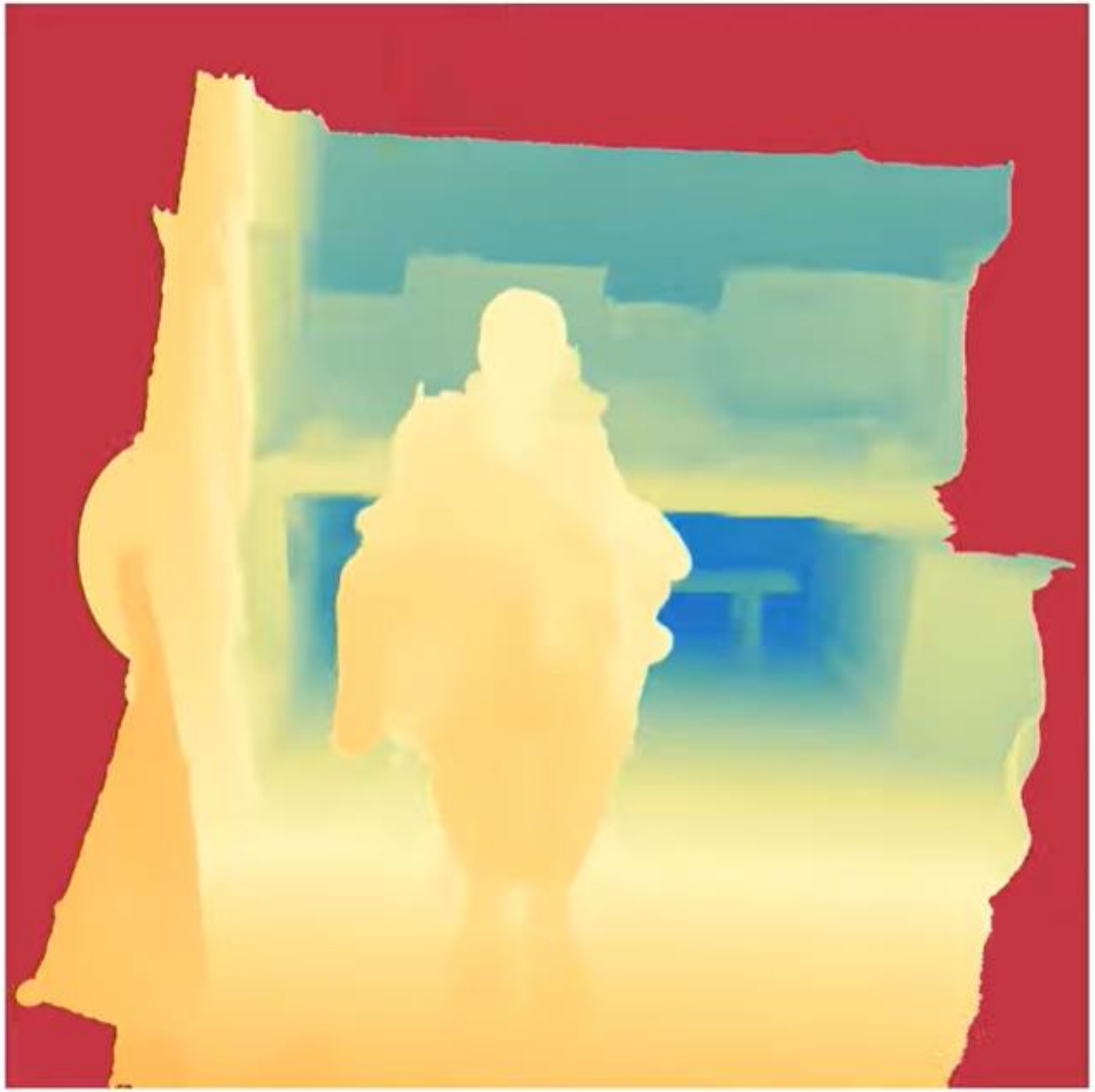


RADIO-Geometry

RADIO-All

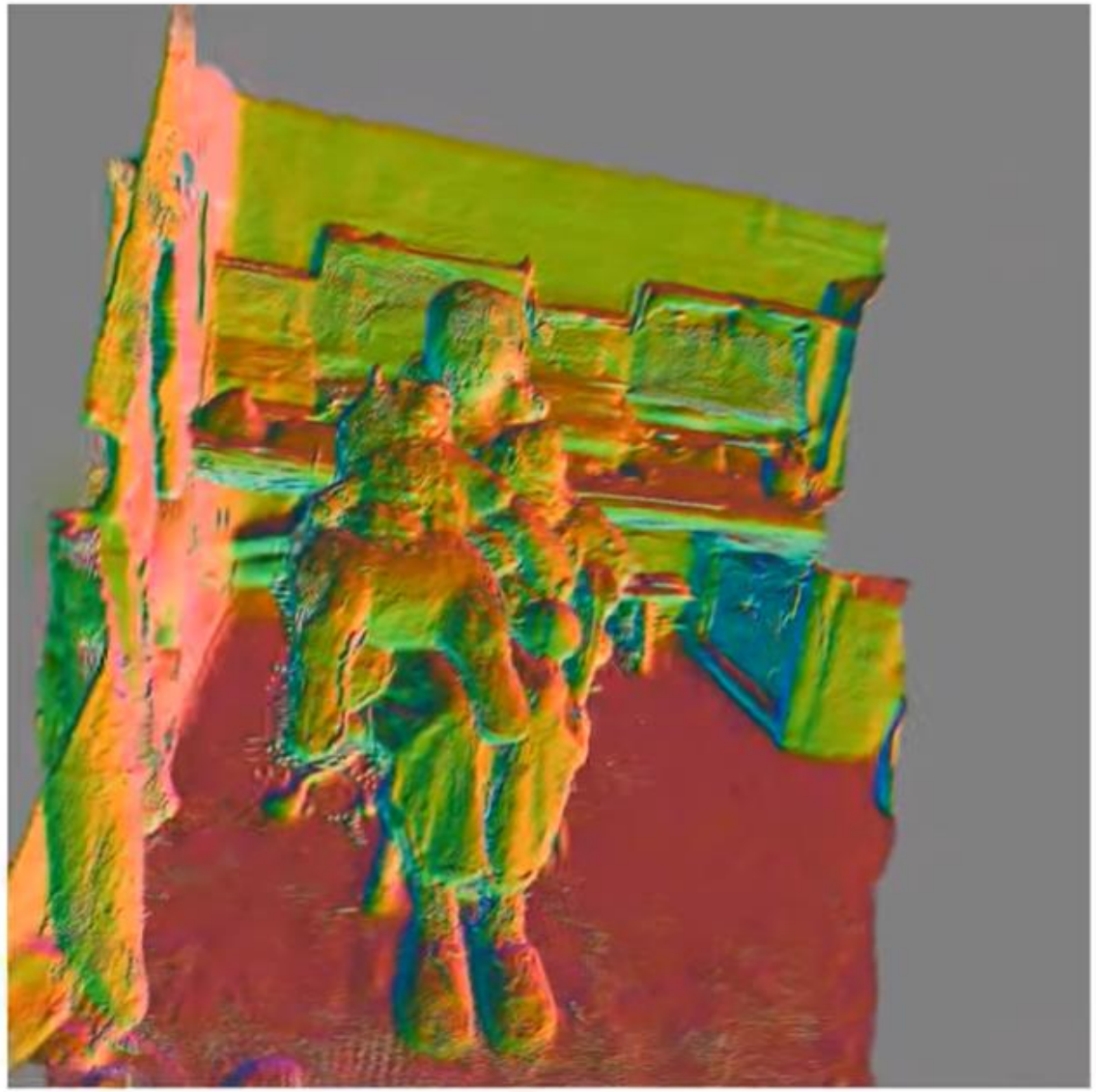
Feat2GS-Geometry

InstantSplat

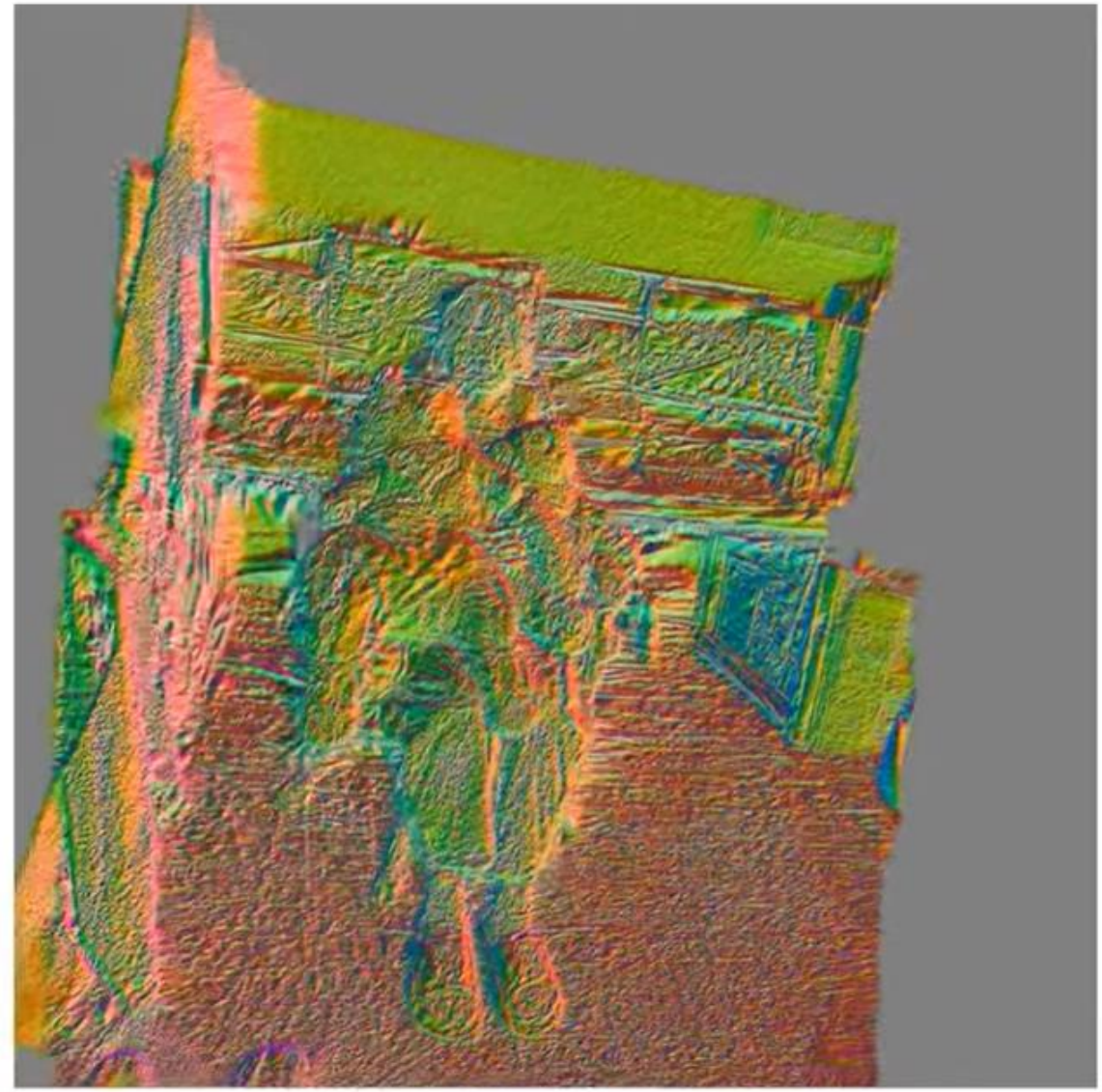


Feat2GS-Geometry

InstantSplat



Feat2GS-Geometry



InstantSplat

Feat2GS

Probing Visual Foundation Models with Gaussian Splatting





Yue Chen[1] Xingyu Chen[1] Anpei Chen[1,3] Gerard Pons-Moll[3,4] Yuliang Xiu[1,2]

[1] Westlake University
[2] Max Planck Institute for Intelligent Systems
[3] University of Tübingen, Tübingen AI Center
[4] Max Planck Institute for Informatics

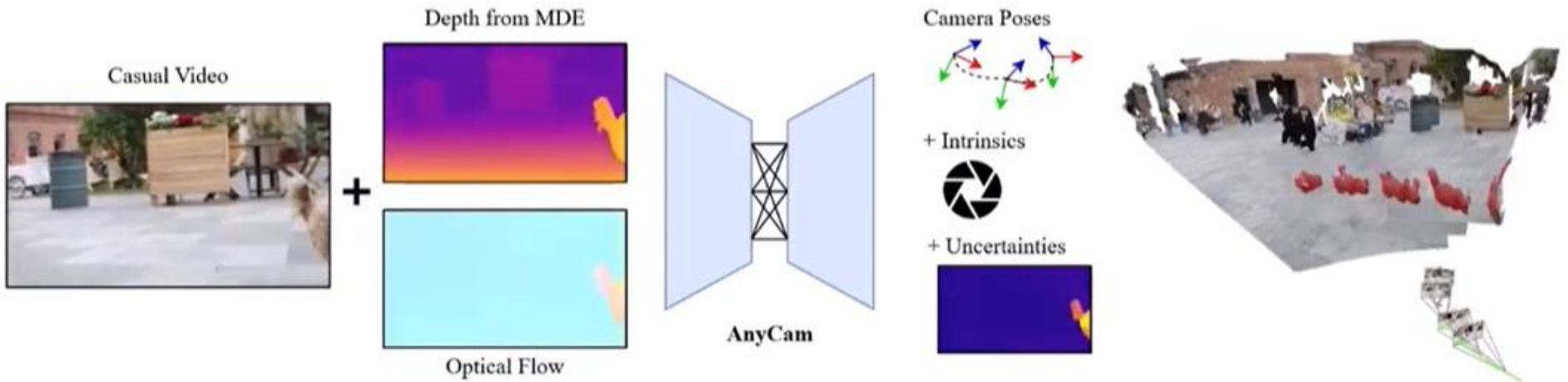
WESTLAKE UNIVERSITY

Tübingen AI Center
tuebingen.ai

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

AnyCam: Learning to Recover Camera Poses and Intrinsics from Casual Videos

fwmb.github.io/anycam



Felix Wimbauer[1,2,3] Weirong Chen[1,2,3] Dominik Muhle[1,2] Christian Rupprecht[3] Daniel Cremers[1,2]

[1]Technical University of Munich [2]MCML [3]University of Oxford

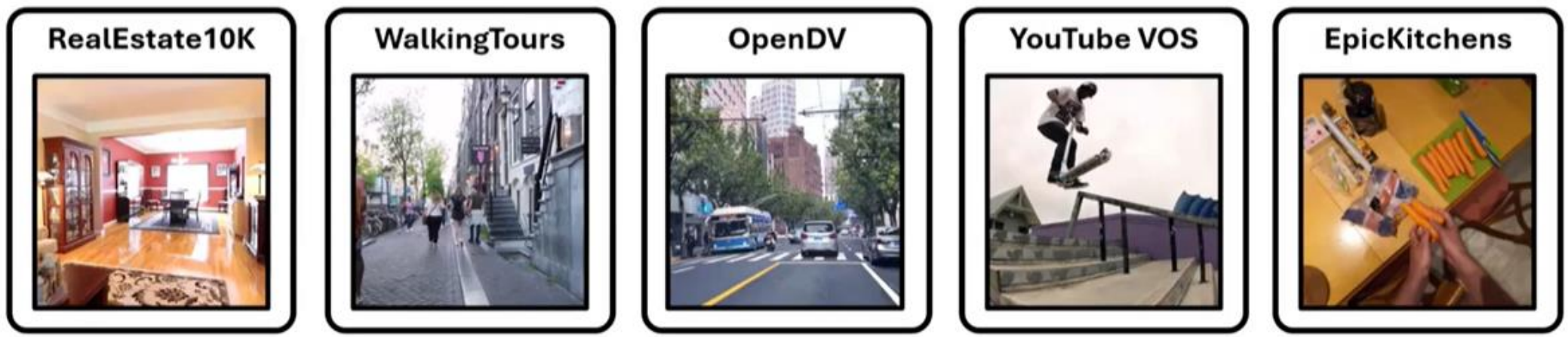
AnyCam is a learning-based method that processes a **casual input video** and computes **camera poses** and **intrinsics**.

Unlike other methods, ***AnyCam*** is

not trained in a **supervised way**,

but through **self-supervision on casual videos**.

AnyCam Training Datasets



Comparing to Existing Works

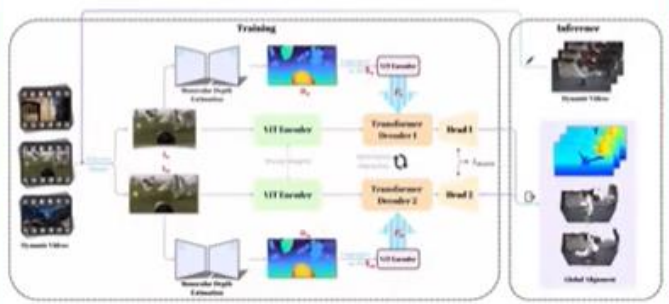
Monst3r (ICLR 2025)

We reconstruct dynamic point cloud & camera poses

MegaSaM (CVPR 2025)

Cut3r (CVPR 2025)

Align3r (CVPR 2025)



And many more

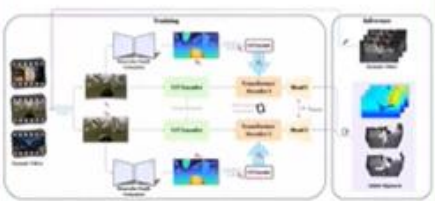
Comparing to Existing Works

Monst3r (ICLR 2025)

MegaSaM (CVPR 2025)

Cut3r (CVPR 2025)

Align3r (CVPR 2025)



Impressive results

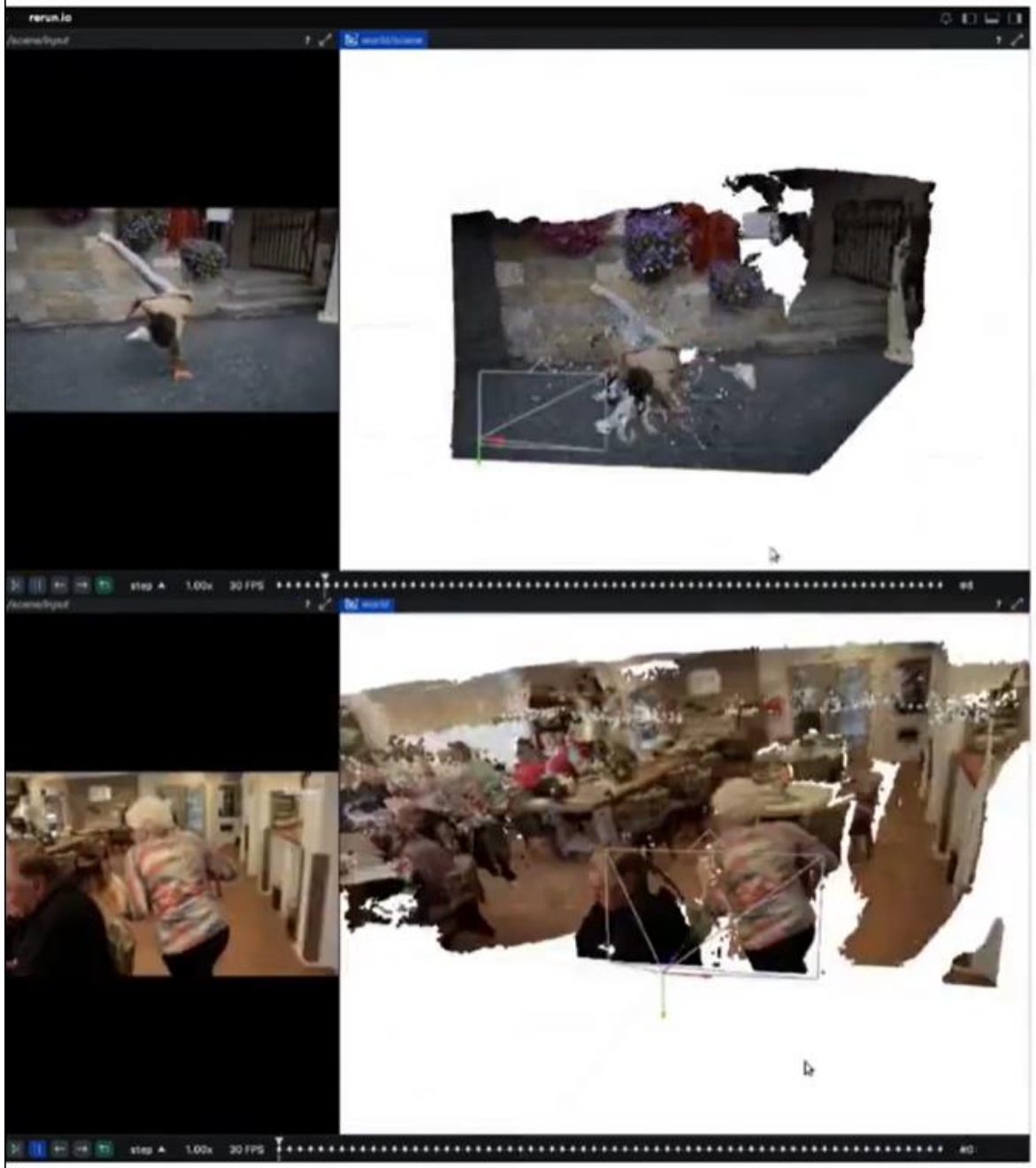
but

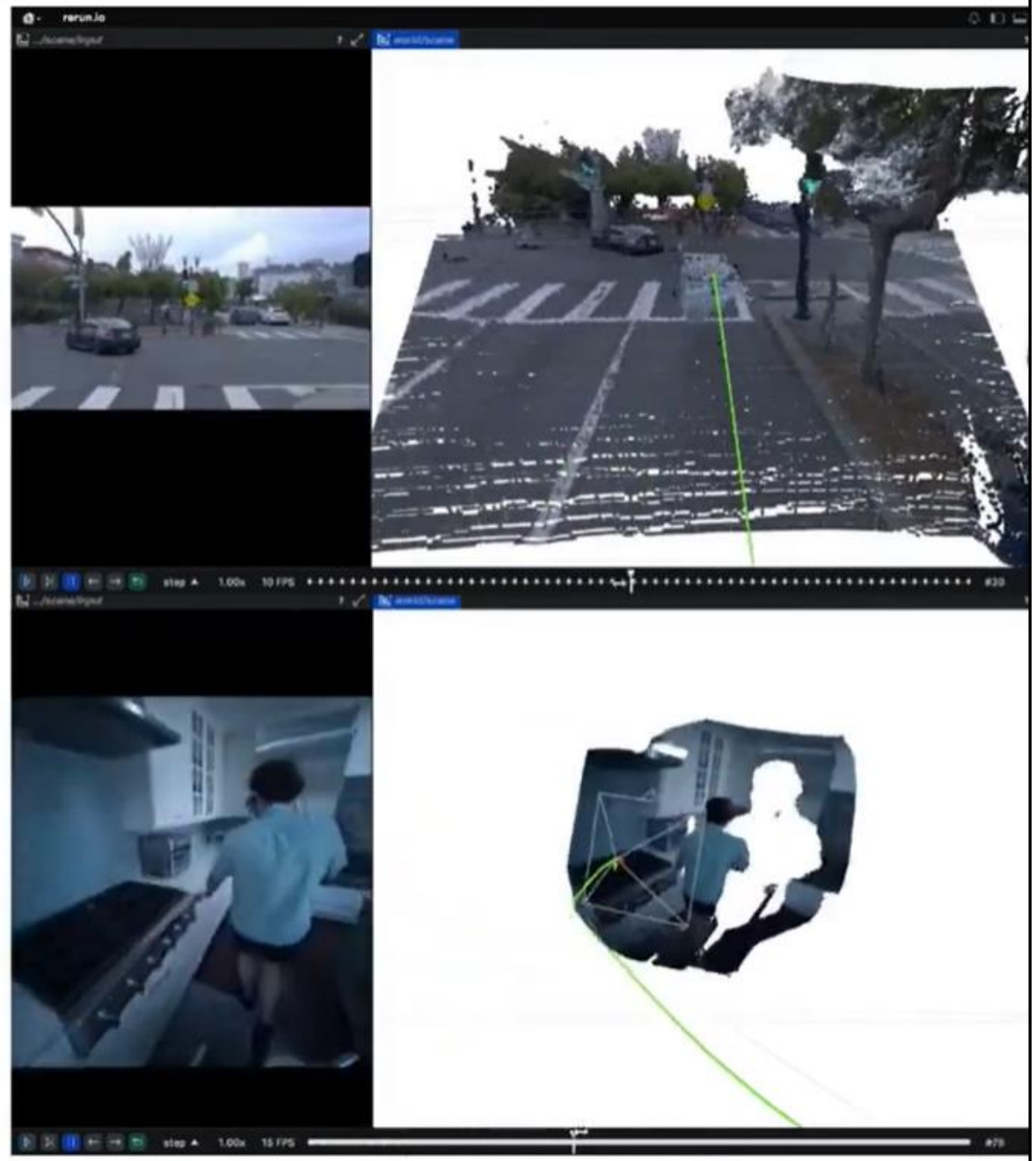
Supervised training

⚠ Expensive data collection ⚠ Dataset biases

⚠ Limited datasets ⚠ Synthetic-to-real gap

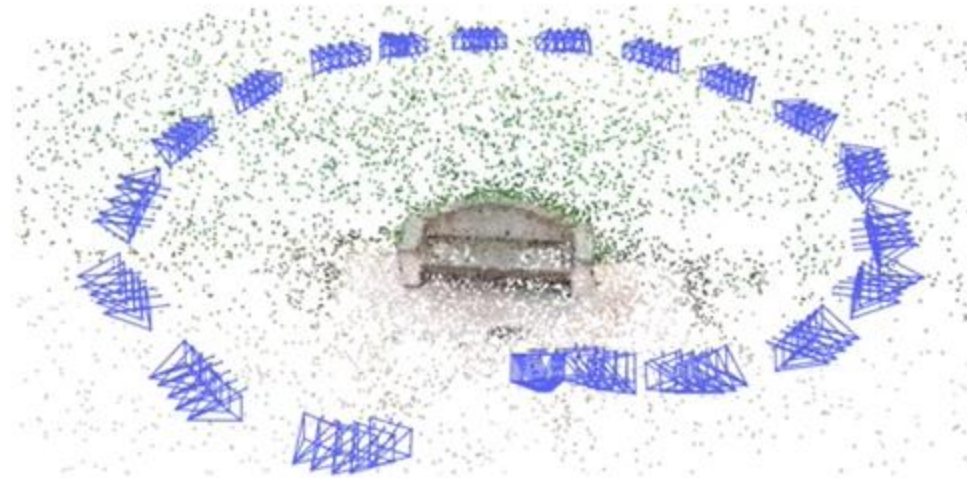
AnyCam *solves these issues*


rerun.io
/scene/input
world/scene
step
1.00x
30 FPS

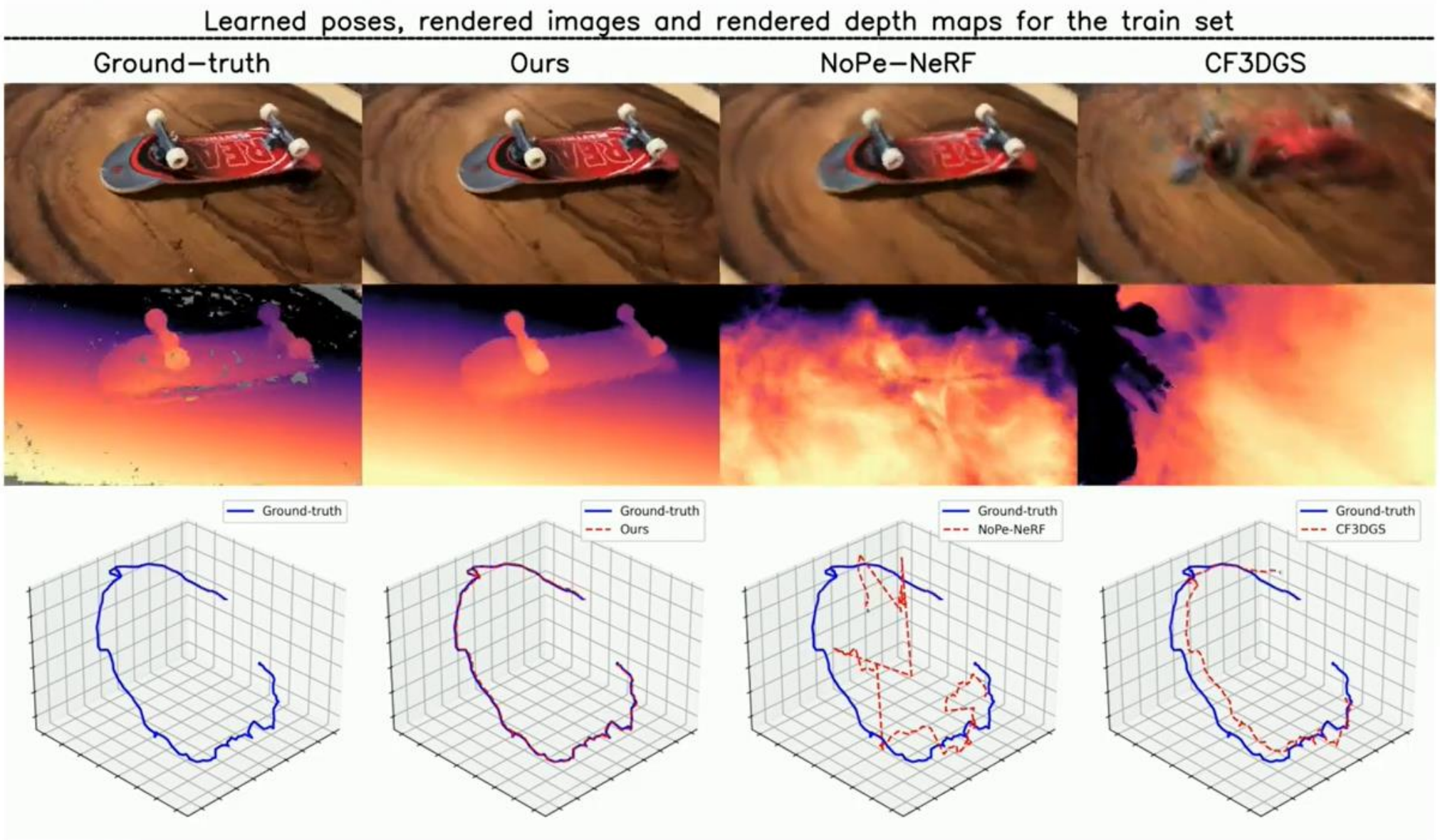

rerun.io
.../scene/input
world/scene
step
1.00x
10 FPS

Joint optimization of NeRF and camera poses from a monocular video

Input: Monocular video



Output: Scene geometry and camera poses.
The scene geometry is represented via NeRF.



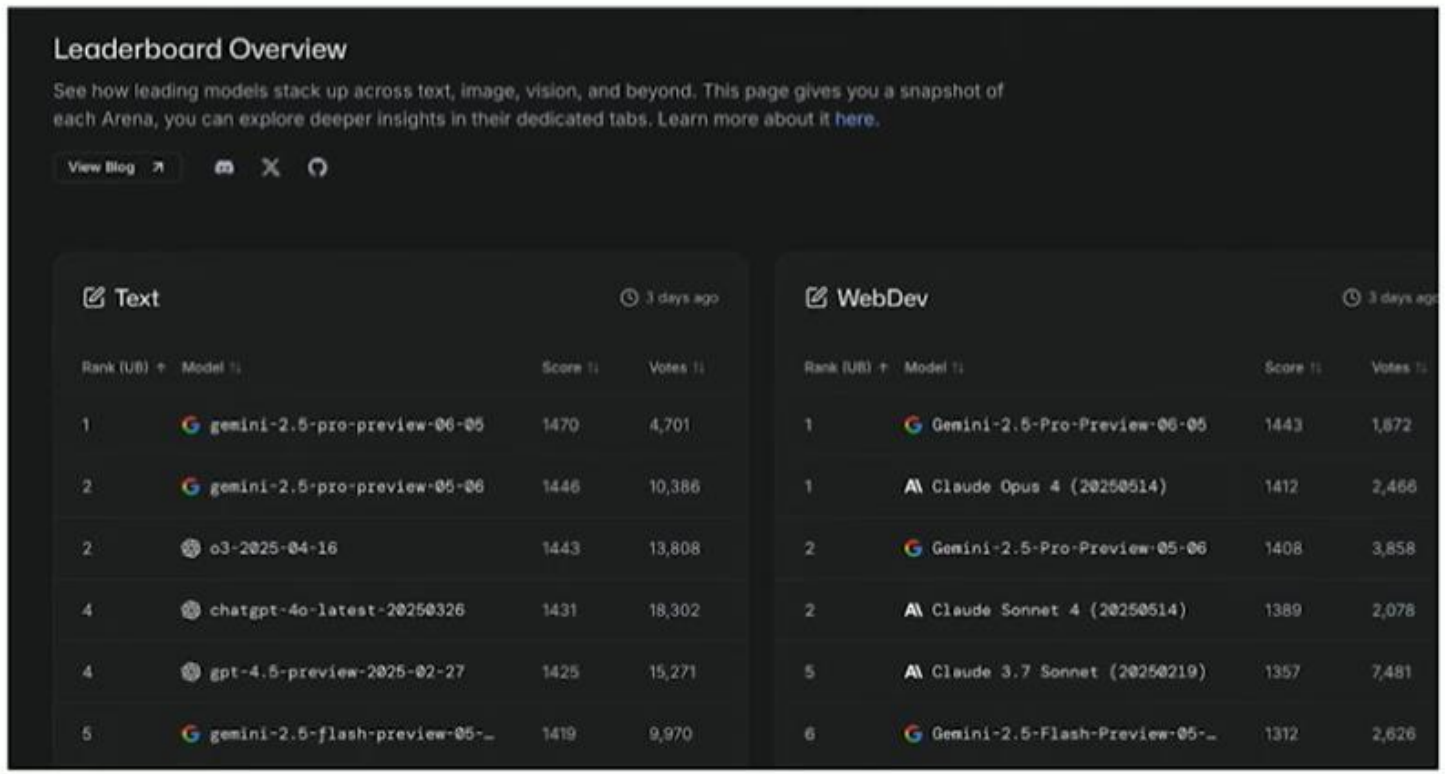
[CVPR 2025] Joint Opt. of NeRFs and Continuous Camera Motion from a Monocular Video (Nguyen et al.)

Arena Style

- Voters perform side-by-side comparisons of (typically) two anonymized models

https://lmarena.ai/leaderboard

- Challenges
  - Zero-sum, dynamic
  - Difficult to control for prompts, distributions, difficulty
  - Live traffic can be noisy over time
  - Relies on expensive human resources



Fixed Test Sets

- Annotate and collect examples with labelled ground truth

- Challenges
  - Poor Annotation quality
  - Training Data leakage
  - Overfitting to the test set
  - Biases
  - Creating difficult examples
  - Metrics for open-ended evaluation

"When a measure becomes a target, it ceases to be a good measure" – Goodhart's Law.

Unbiasing through Textual Descriptions: Mitigating Representation Bias in Video Benchmarks

CVPR 2025

Nina Shvetsova
University of Tuebingen /
MPII Saarbrucken /
Goethe University Frankfurt

Arsha Nagrani
University of Oxford

Bernt Schiele
MPII Saarbrucken

Hilde Kuehne
University of Tuebingen /
MIT-IBM Watson AI Lab /
Goethe University Frankfurt

Christian Rupprecht
University of Oxford

References

- [CVPR 2025] Feat2GS: Probing Visual Foundation Models with Gaussian Splatting (https://www.youtube.com/watch?v=4fT5lzcAJqo)
- [CVPR 2025] AnyCam: Learning to Recover Camera Poses and Intrinsics from Casual Videos
- [CVPR 2025] Joint Opt. of NeRFs and Continuous Camera Motion from a Monocular Video (Nguyen et al.)