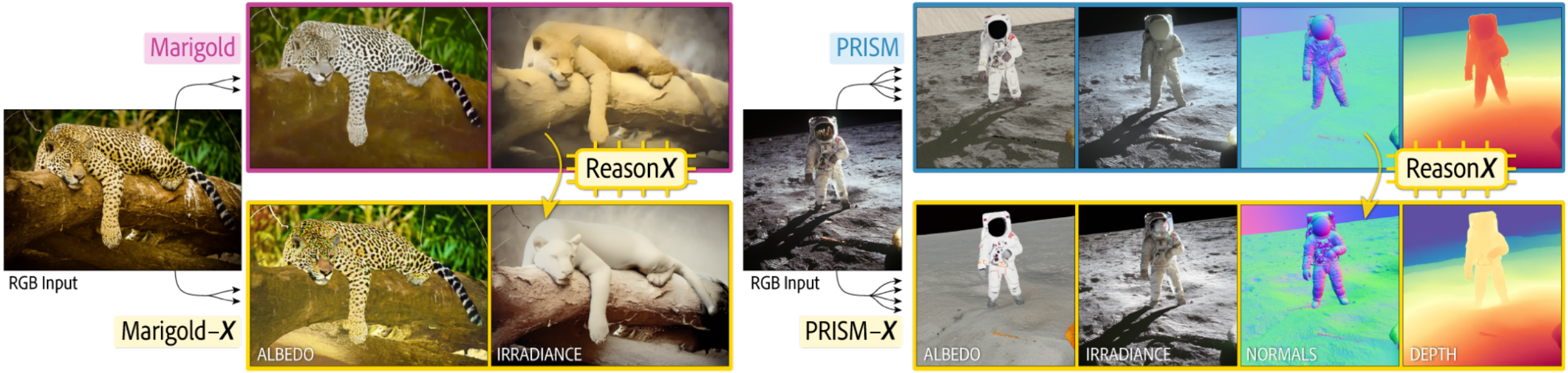


ReasonX: MLLM-Guided Intrinsic Image Decomposition

Alara Dirik[1*] · Tuanfeng Wang[2] · Duygu Ceylan[2] · Stefanos Zafeiriou[1] · Anna Frühstück[2]

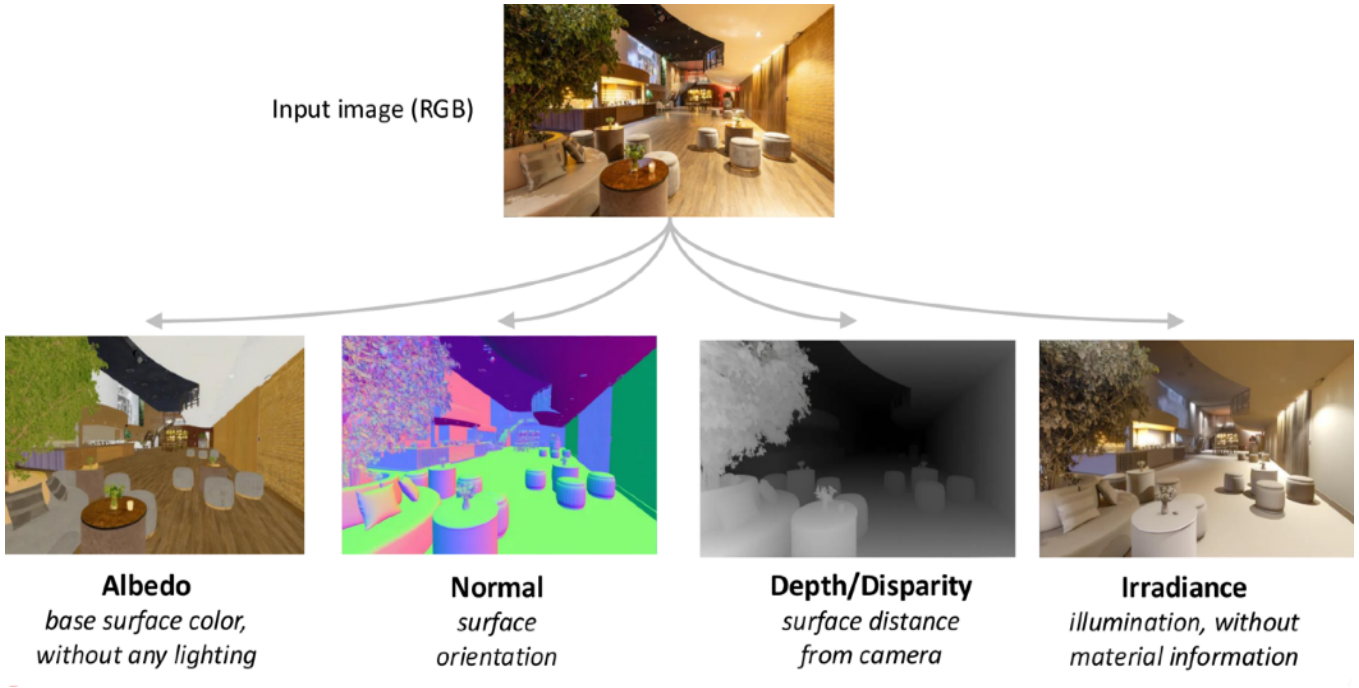
[1]*Imperial College London* · [2]*Adobe Research*
**Work done during internship at Adobe.*

Adobe Imperial College London

CVPR JUNE 3-7, 2026 DENVER COLORADO

The Problem

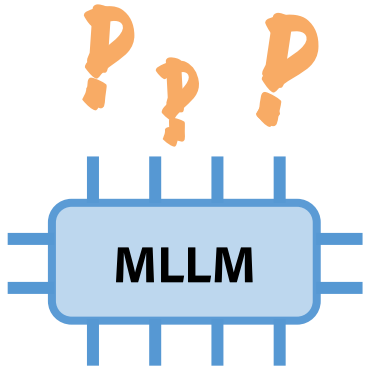
- Intrinsic decomposition needs paired data with GT intrinsics.
- SOTA models fail on in-the-wild images.
- Real image intrinsic labels are prohibitively expensive to collect.



Intrinsic decomposition separates RGB into isolated components: albedo, normals, depth, irradiance, and more...

Key Idea

Judge Questions	GT Judgments
ALBEDO DO THE RED AND GREEN POINTS HAVE THE SAME BASE COLOR?	ALBEDO YES, THE RED AND GREEN POINTS HAVE THE SAME BASE COLOR.
IRRADIANCE WHICH POINT IS MORE ILLUMINATED, RED OR GREEN?	IRRADIANCE THE RED POINT IS MORE ILLUMINATED.
NORMALS WHICH POINT HAS A SURFACE MORE FACING TOWARDS THE CAMERA, RED OR GREEN?	NORMALS THE RED POINT IS MORE FACING TOWARDS THE CAMERA.
DEPTH WHICH POINT APPEARS TO BE CLOSER TO THE CAMERA - RED OR GREEN?	DEPTH THE GREEN POINT IS CLOSER.



MLLMs are unreliable at absolute estimation

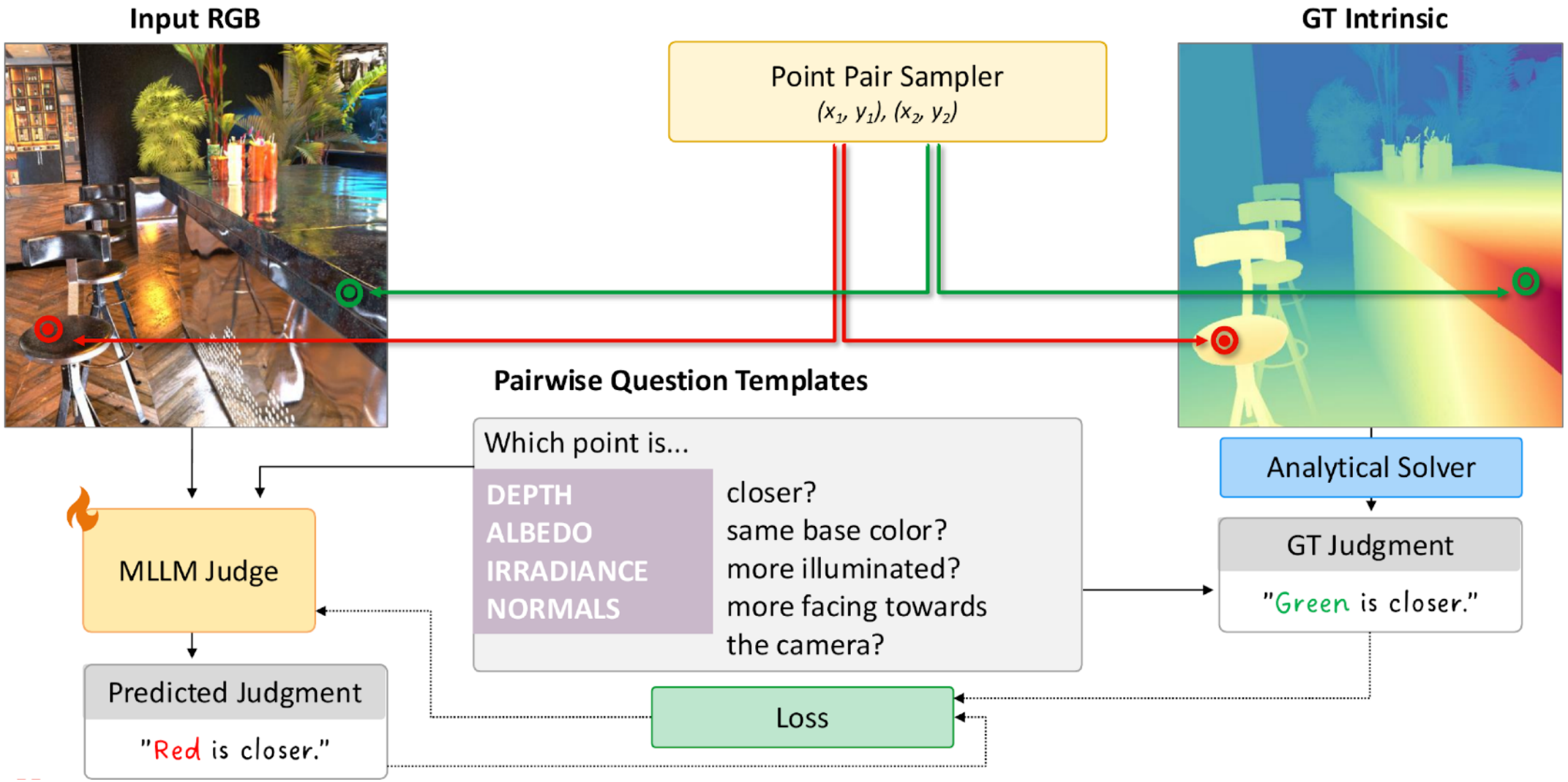
"How far is this point from the camera?"

but excel at relative comparisons.

"Which point is closer?"

"Do these points share the same base color?"

Stage A — Train the MLLM Judge



We fine-tune InternVL2.5-4B on synthetic RGB + point pairs. Ground truth derived analytically from GT intrinsics.

Judge Performance

Tested on held-out synthetic test set

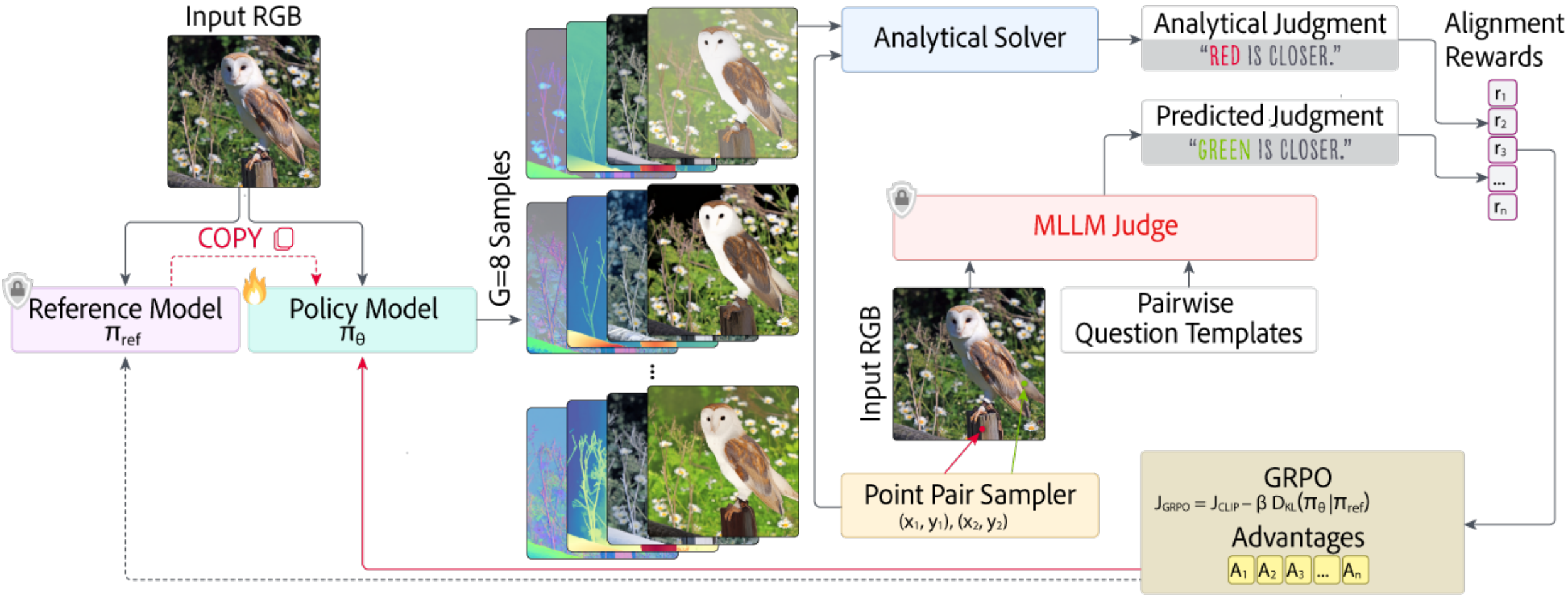
Modality	Depth	Normal	Albedo	Irradiance
Accuracy	0.962	0.935	0.894	0.876
F1 Score	0.962	0.933	0.889	0.878

Reliable enough to drive a reward signal without GT intrinsics on real images.

Failure modes: mirrors, intricate textures, dim lighting, point pairs too close in value.

Stage B — Intrinsic-GRPO

Fine-tune the base model on real images without GT intrinsics.



Quantitative Highlights

PRISM → PRISM-X, Marigold → Marigold-X, no GT intrinsics during fine-tuning

WHDR error of albedo estimates

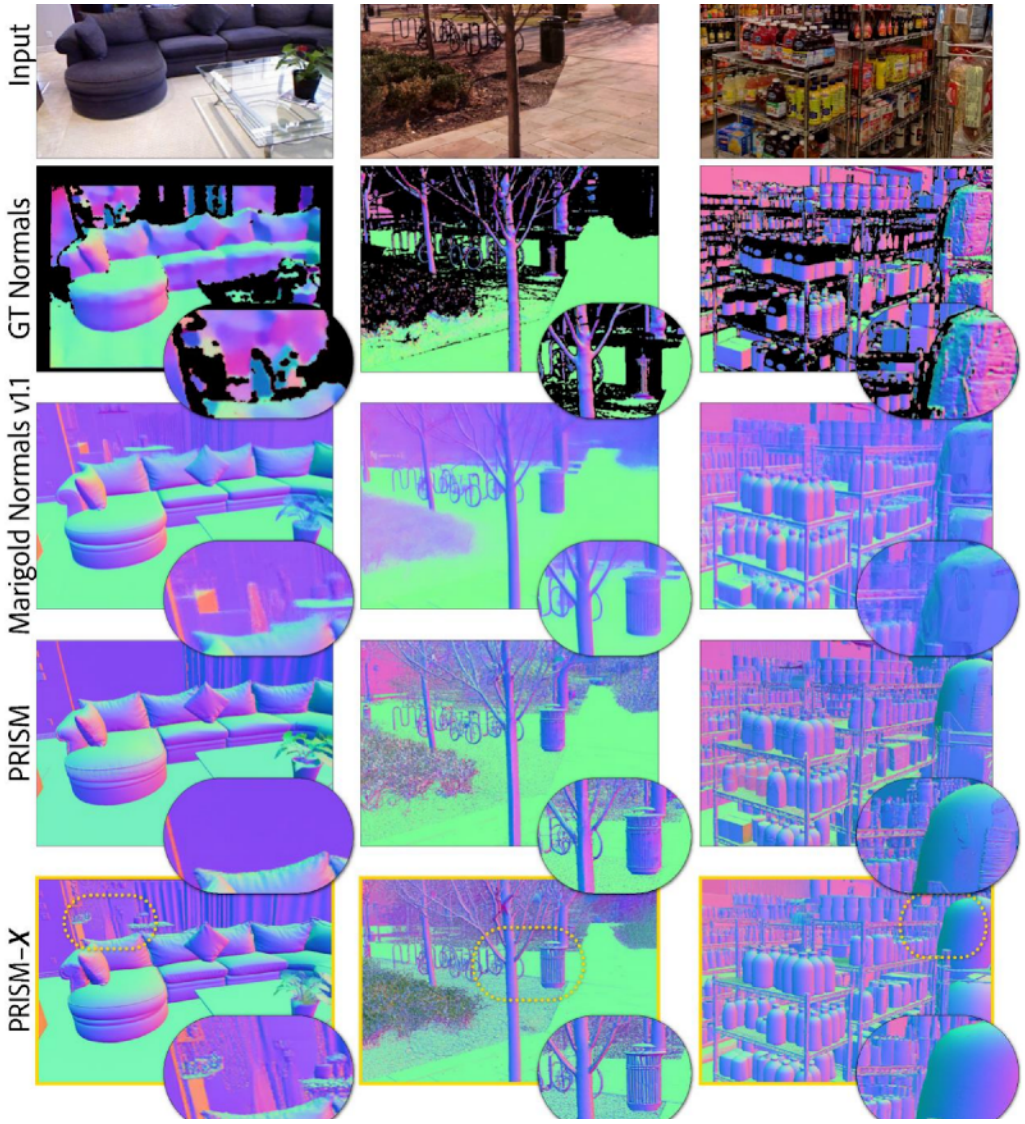
Method	WHDR 10% ↓	WHDR 20% ↓
CRefNet [27]	12.8	10.8
Zhu et al. [53]	34.7	24.1
Careaga and Aksoy [5]	24.8	19.2
Kocsis et al. [19]	26.1	20.7
RGB↔X [49]	23.6	21.1
Careaga and Aksoy [6]	16.8	15.6
OmniGen2 [40]	17.8	14.3
Marigold (base variant) [18]	16.7	14.8
Marigold-**X**	**15.2**	14.0
Improvement	*+9.0%*	*+5.4%*
PRISM [12]	17.2	15.9
PRISM-**X**	**12.9**	**11.9**
Improvement	*+25.0%*	*+25.2%*

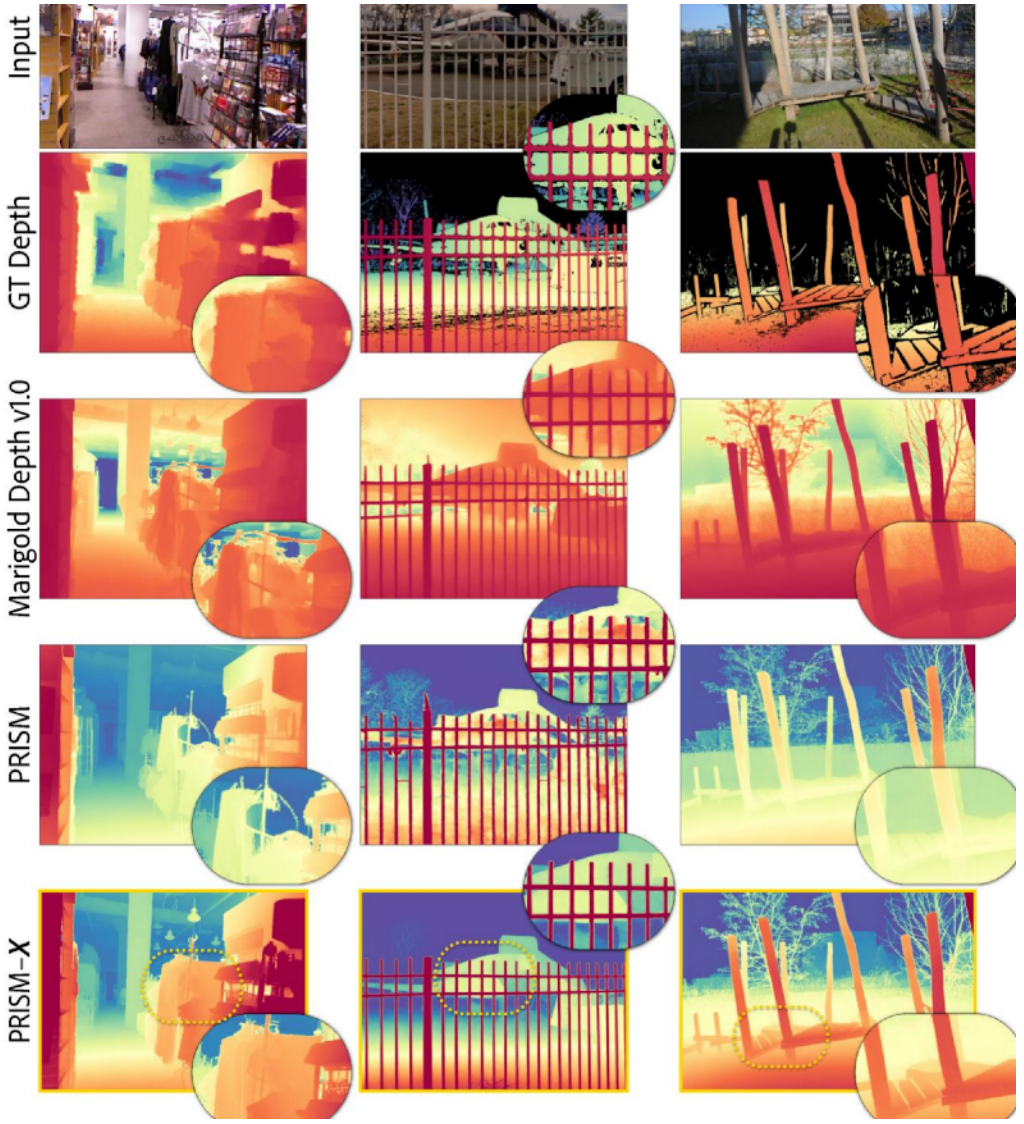
Zero-shot relative depth estimation

Method	NYUv2		ETH3D	
	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
Omnidata [13]	0.074	0.945	0.166	0.778
Depth Anything V2 [46]	0.045	0.979	0.131	0.865
MiDaS [30]	0.111	0.885	0.184	0.752
DPT [31]	0.098	0.903	0.078	0.946
HDN [51]	0.069	0.948	0.121	0.833
OmniGen2 [40]	0.111	0.886	0.123	0.897
Marigold Depth `v1.0` [17]	0.055	0.964	0.065	0.960
PRISM [12]	0.061	0.922	0.142	0.836
PRISM-**X**	**0.053**	0.958	0.077	0.950
Improvement	*+13.1%*	*+3.9%*	*+45.8%*	*+13.6%*

Qualitative — Depth & Normals

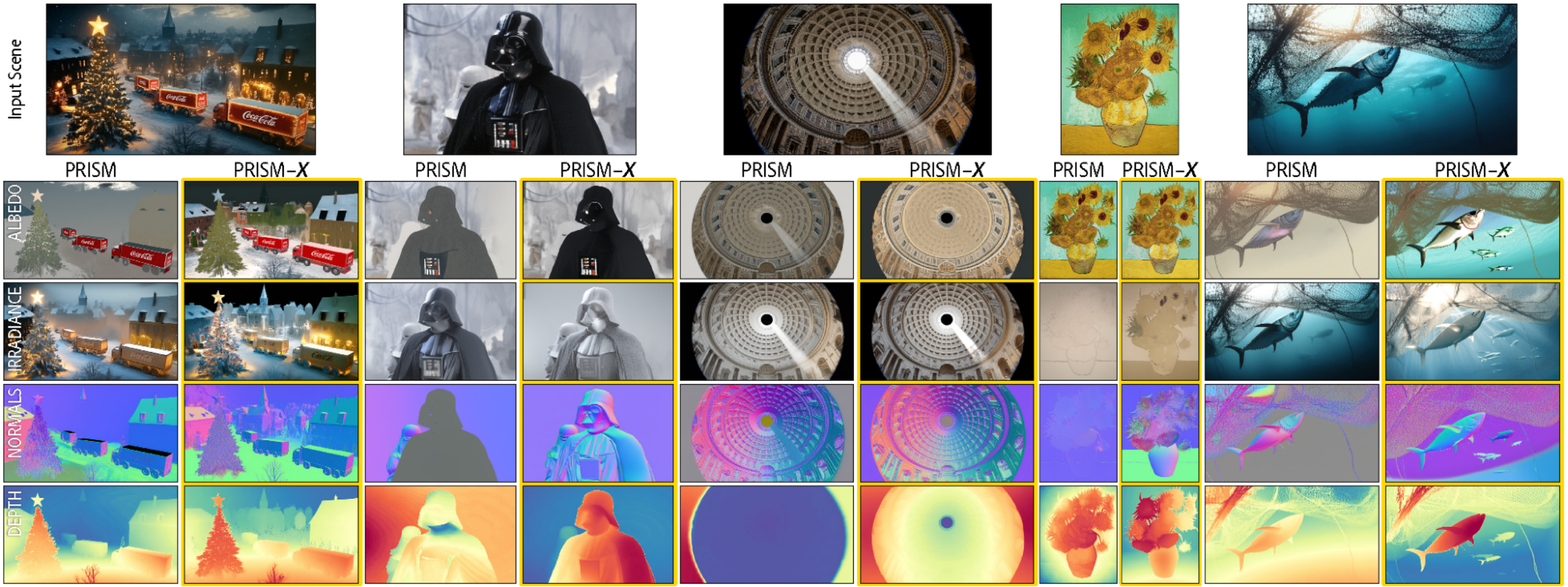
Zero-shot on NYUv2, DIODE, ETH3D — never seen during training





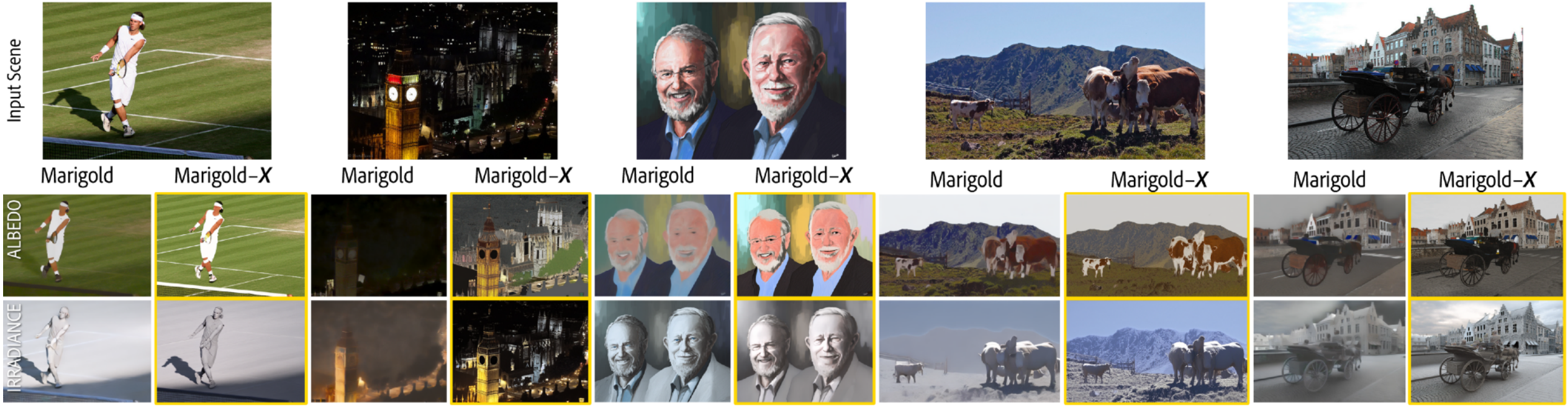
Out-of-Domain Generalization — PRISM / PRISM-*X*

ReasonX significantly improves generalization on challenging scenes.



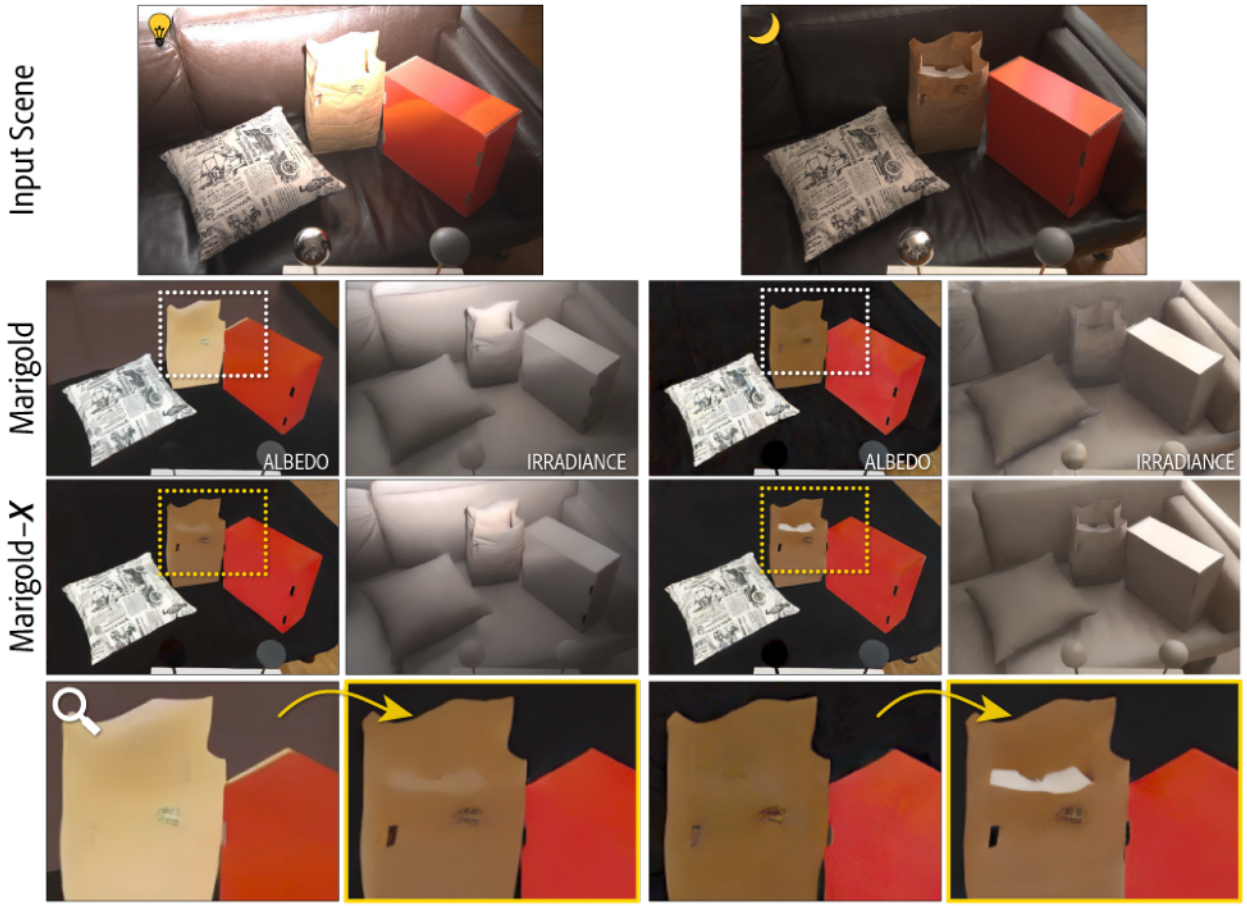
Out-of-Domain Generalization — Marigold / Marigold-*X*

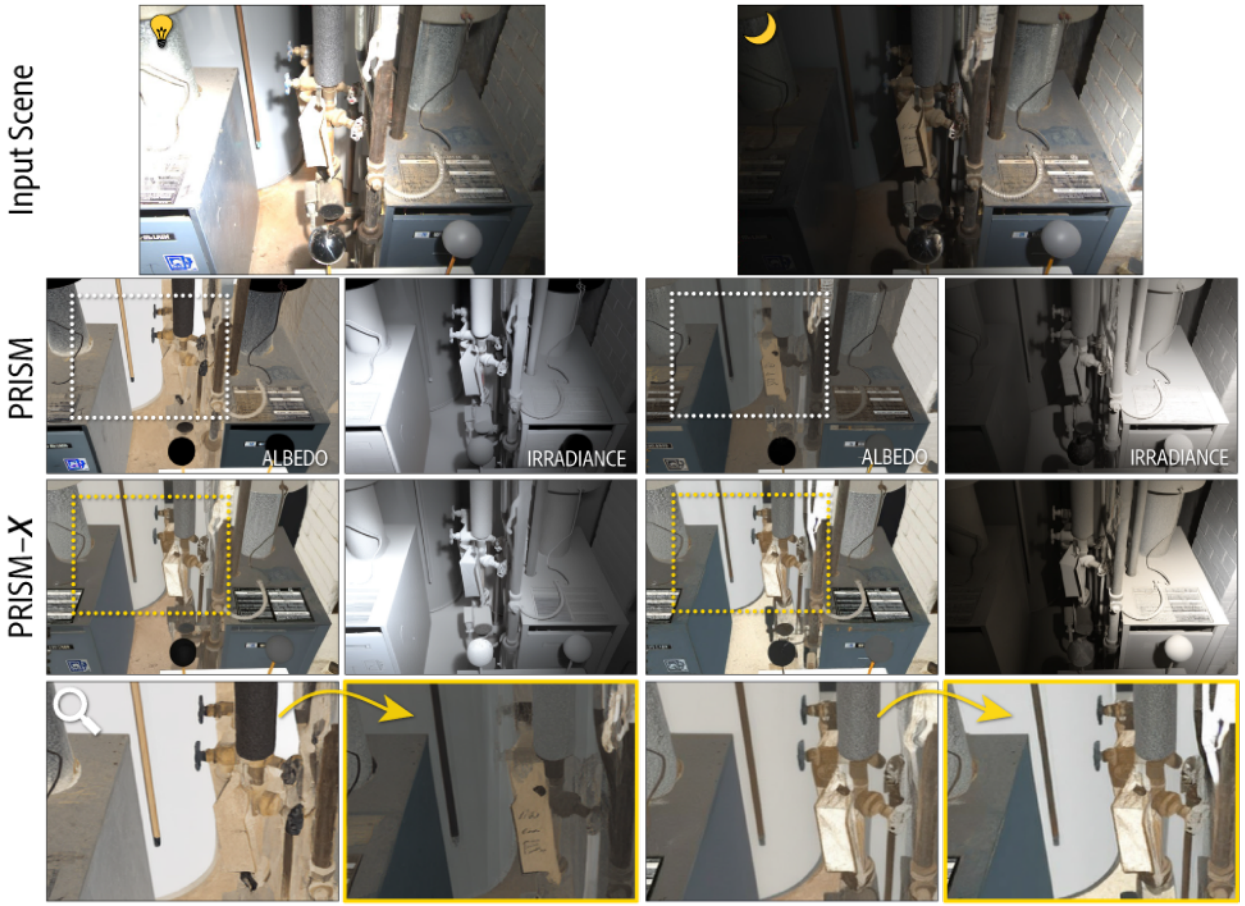
ReasonX significantly improves generalization on challenging scenes.



Decomposition Consistency

Can predictions for the same scene stay stable under different lighting?





Tested on MIT Multi-Illumination scenes.

Takeaways

1. Relative MLLM judgments are a scalable, GT-free supervision signal for intrinsic decomposition.

2. Intrinsic-GRPO fine-tunes intrinsic predictors on real images with no paired labels.

3. Consistent gains across 4 modalities and 2 base models → bridges synthetic-to-real gap.

Project Page

Code

Thank you!

arXiv: *arxiv.org/abs/2512.04222*

For questions: *a.dirik22@imperial.ac.uk*