SAGE: Scalable Adaptation of 3D Geometric Foundation Models via Weak Supervision from Internet Video

Zihui Gao[2,1] Ke Liu[3] Donny Y. Chen[4] Duochao Shi[2] Guosheng Lin[1]
Jia-Wang Bian[1†] Hao Chen[2†] Chunhua Shen[2,5]

[1]Nanyang Technological University [2]Zhejiang University, State Key Lab of CAD & CG
[3]Zhejiang University, School of Software Engineering [4]Monash University [5]Ant Group [†]Corresponding authors







NeurIPS 2026

3D foundation models are starved of 3D labels

~1K scenes

ScanNet: sensor-captured 3D annotations.
Costly, slow to scale, narrow in diversity.

10K+ scenes

RealEstate10K + DL3DV: static-scene Internet video.
Abundant and diverse, but **no geometry, no poses**.

The obvious recipe fails

Fine-tune MV-DUSt3R on pseudo-depth from a video depth model (VDS):

	TUM-RGBD		Matterport3D	
	DAc ↑	CD ↓	DAc ↑	CD ↓
MV-DUSt3R	87.5	2.60	30.3	43.2
+ VDS	84.7	2.68	14.3	82.1

Pseudo-depth is noisy and **multi-view inconsistent**, so zero-shot accuracy gets worse.

We need supervision that is consistent across views.

SAGE: Internet video as a scaling source

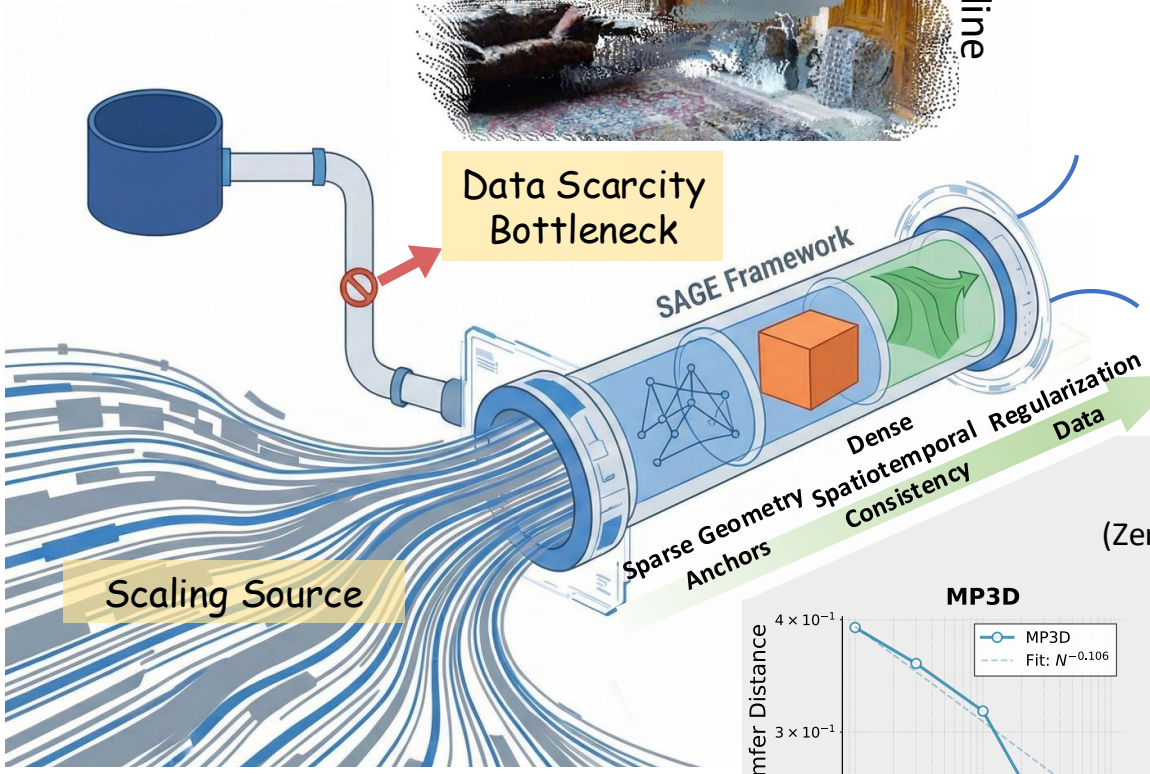
Limited Labeled 3D Data

(e.g., ScanNet, 1K scenes)



Baseline

Data Scarcity Bottleneck



Unlimited Internet Video

(e.g., YouTube, >10K scenes)

General-Purpose 3D Foundation Model

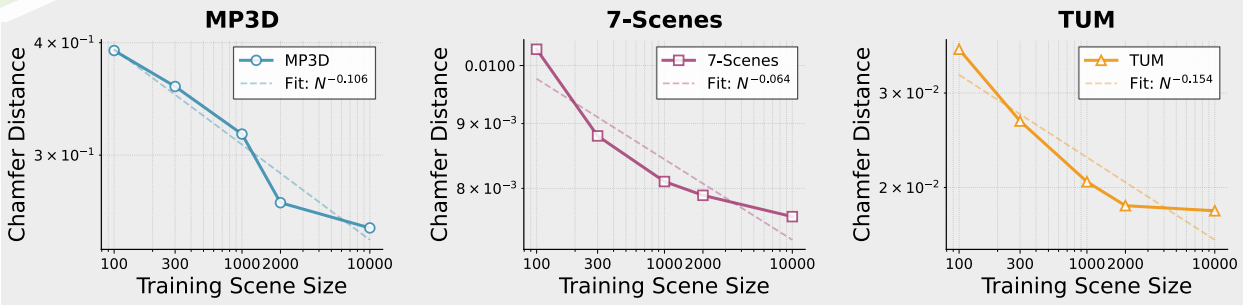
(Zero-shot Generalization)



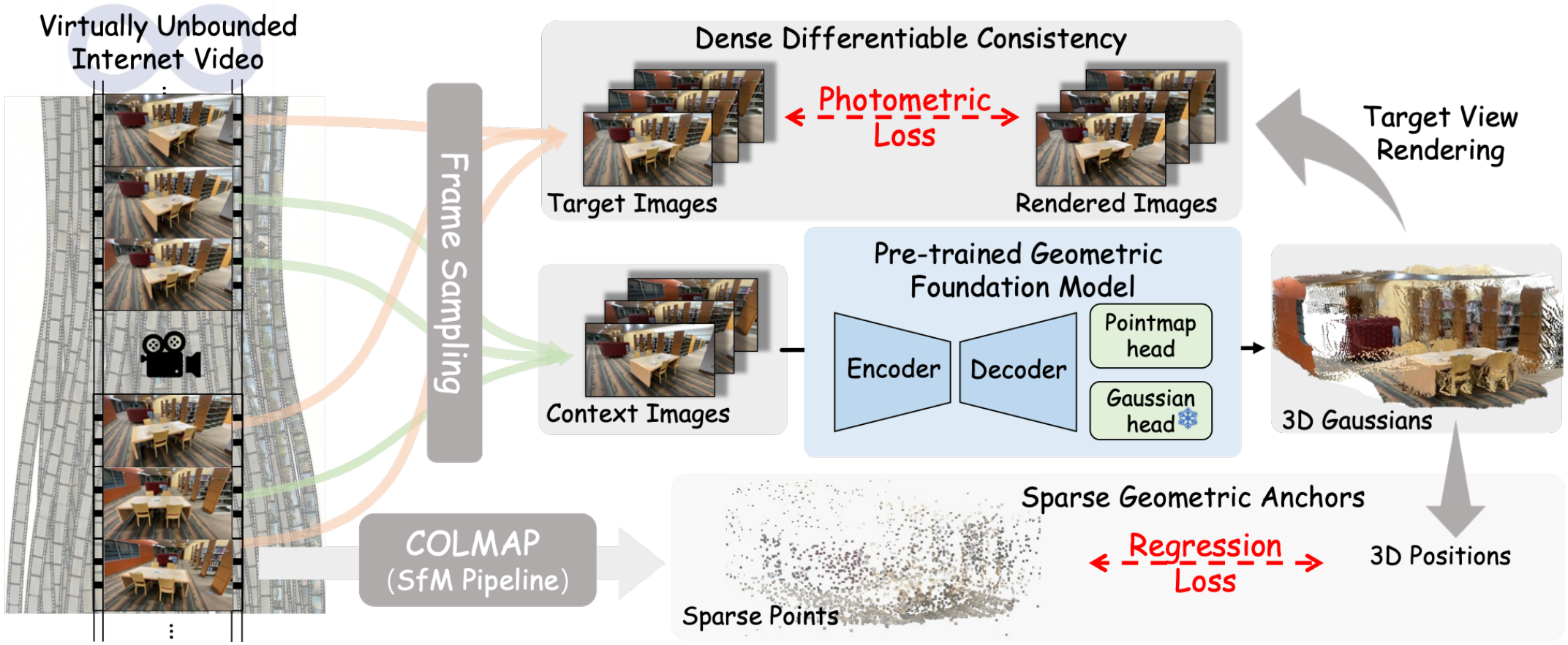
Ours

Scalability

(Zero-shot Accuracy vs Video Scene Volume)



Hybrid weak supervision from raw video



$$\min_{\theta} \underbrace{\mathbb{E}_{V\sim\mathcal{D}_{video}}\left[\mathcal{L}_{\text{anchor}}+\mathcal{L}_{\text{ren}}\right]}_{\text{scaling via weak supervision}} + \gamma\cdot\underbrace{\mathbb{E}_{(I,Y)\sim\mathcal{D}_{sup}}\left[\mathcal{L}_{sup}(f_\theta(I),Y)\right]}_{\text{foundation regularization}}$$

Three signals, each one necessary

1 Sparse geometric anchors

COLMAP SfM points, projected to sparse depth, with a confidence-aware regression loss. Globally consistent structure.

Remove it:

TUM CD 1.85 → **5.20**
MP3D DAc 48.0 → **18.9**

2 Dense differentiable consistency

Render held-out video frames through frozen 3D Gaussians and compare photometrically. Supervises the pixels the anchors miss.

Remove it:

TUM CD 1.85 → **2.26**
7Scenes DAc 98.0 → **93.5**

3 Supervised replay

Mix in about 3% labelled data from pretraining (30 : 1). Keeps the model from drifting and forgetting.

Remove it:

ScanNet CD 1.69 → **3.32**
7Scenes DAc 98.0 → **88.5**

Training clips: 8 context frames sampled with temporal jitter, 2 interpolated target frames. Preprocessing is one sparse-SfM pass, about 15 s per clip.

Consistent gains on two different backbones

	ScanNet		7Scenes		TUM-RGBD		Matterport3D	
	in-distribution				zero-shot			
	DAc@0.2 ↑	CD ↓	DAc@0.2 ↑	CD ↓	DAc@0.2 ↑	CD ↓	DAc@0.5 ↑	CD ↓
MV-DUSt3R *(pretrained)*	88.2	1.79	89.6	0.95	87.5	2.60	30.3	43.2
+ SAGE (10K)	**91.2** +3.0	**1.69** -0.10	**98.0** +8.4	**0.76** -0.19	**93.1** +5.6	**1.85** -0.75	**48.0** +17.7	**24.9** -18.3
VGGT *(pretrained)*	87.0	3.30	88.5	1.13	80.4	4.24	53.1	20.7
+ SAGE (10K)	**89.5** +2.5	**2.01** -1.29	**92.0** +3.5	**0.79** -0.34	**84.3** +3.9	**2.84** -1.40	**55.4** +2.3	**18.3** -2.4

20–42% lower CD

zero-shot, on MV-DUSt3R (7Scenes, TUM-RGBD, Matterport3D)

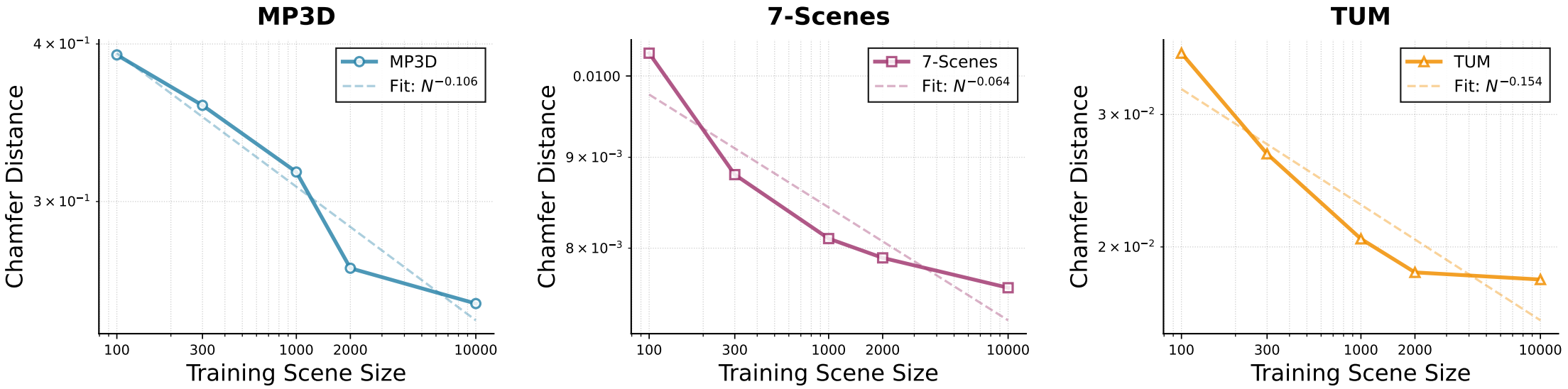
- No new 3D annotations
- Under 2 days on 4× RTX 4090
- Encoder–decoder and single-transformer backbones both improve

Competitive with fully 3D-supervised models

Method	ScanNet		7Scenes		TUM-RGBD		Matterport3D	
	DAc@0.2↑	CD↓	DAc@0.2↑	CD↓	DAc@0.2↑	CD↓	DAc@0.5↑	CD↓
Trained from scratch with full 3D supervision								
Spann3R	59.0	13.48	90.0	1.48	35.7	7.06	15.8	69.7
DUSt3R	88.0	4.50	89.5	1.50	71.4	4.65	45.4	44.1
MapAnything	79.5	3.05	88.1	1.56	82.0	2.96	51.8	24.0
π^3	**97.0**	**0.95**	95.0	0.80	92.7	1.90	57.4	22.9
Depth Anything 3	89.8	2.45	91.5	0.84	85.7	**1.69**	**67.3**	**17.1**
Adapted with Internet video only								
MV-DUSt3R + SAGE (10K)	91.2	1.69	**98.0**	**0.76**	**93.1**	1.85	48.0	24.9
VGGT + SAGE (10K)	89.5	2.01	92.0	0.79	84.3	2.84	55.4	18.3

Zero-shot: best DAc on 7Scenes and TUM-RGBD, best CD on 7Scenes. Larger curated 3D corpora still lead on ScanNet and Matterport3D: video adaptation is complementary, not a replacement.

More video keeps helping



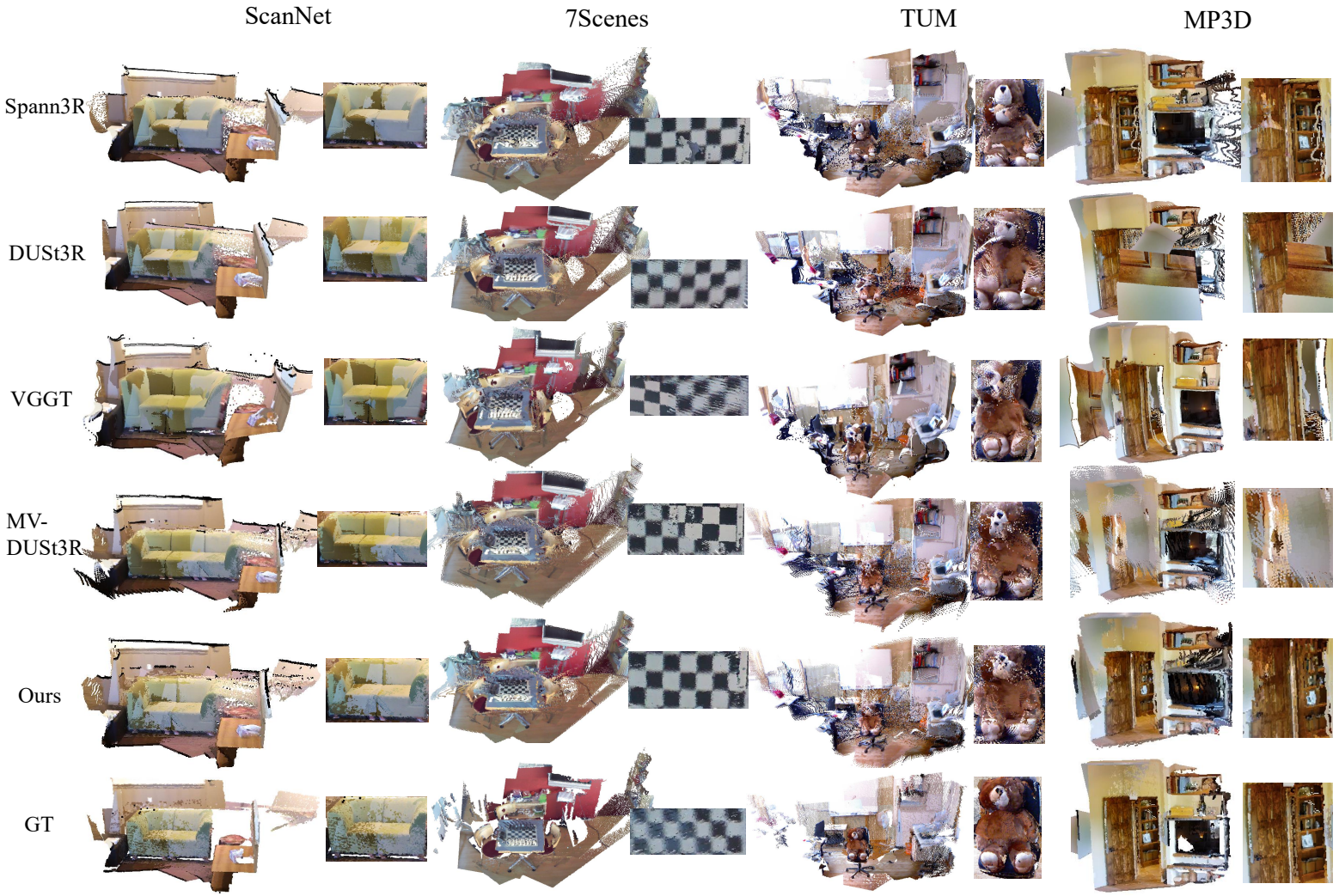
Log-linear gains, no saturation

Zero-shot Chamfer Distance keeps falling as training video grows from 100 to 10K scenes.

Diversity beats density

At a fixed training budget, more unique scenes help more than more frames per scene.

Cleaner, more complete geometry



8 unposed input views.

Ours = MV-DUSt3R + SAGE.

Compared with the MV-DUSt3R row above it: more complete and more coherent geometry, closer to the ground truth in the zoomed crops.

Takeaways

- **Internet video is a viable, scalable adaptation source** for 3D geometric foundation models.
- **Sparse SfM anchors + dense rendering consistency + light replay** turn noisy, unlabelled video into stable supervision.
- **Gains scale log-linearly** from 100 to 10K scenes, and are orthogonal to scaling architectures and 3D-supervised data.

Code
github.com/aim-uofa/SAGE

Thank you!