# 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, Singapore [2]Zhejiang University, State Key Lab of CAD & CG
[3] Zhejiang University, School of Software Engineering [4]Monash University [5]Ant Group

## Abstract

Geometric foundation models show promise in 3D reconstruction. However, their progress is severely constrained by the scarcity of diverse, large-scale 3D annotations. While static-scene Internet videos are abundant and diverse, utilizing them as a scaling source for geometric learning is challenging due to the absence of ground-truth geometry and the presence of observational noise. To address this, we propose *SAGE*, a framework for *S*calable *A*daptation of *GE*ometric foundation models from raw video streams. SAGE leverages a hierarchical mining pipeline to transform videos into training trajectories and hybrid supervisions: (1) Informative training trajectory selection; (2) Sparse Geometric Anchoring via SfM point clouds for global structural guidance; and (3) Dense Differentiable Consistency via 3D Gaussian rendering for multi-view constraints. Besides, we also introduce a regularization strategy that replays a small fraction of supervised data to prevent catastrophic forgetting. Applied to two architecturally distinct backbones, SAGE reduces Chamfer Distance on unseen benchmarks (7Scenes, TUM-RGBD, Matterport3D) by 20–42% with MV-DUSt3R and consistently improves VGGT. SAGE establishes Internet video as a viable and scalable adaptation source for geometric foundation models. Code is available at `https://github.com/aim-uofa/SAGE`.

## 1 Introduction

Recent advances in 3D geometric foundation models, such as DUSt3R [35] and MASt3R [19], have reshaped the pipeline of reconstructing 3D geometry from unposed 2D images [37]. These methods move beyond traditional Structure-from-Motion (SfM) [26] pipelines with end-to-end data-driven networks that directly estimate pixel-aligned pointmaps in a unified coordinate system. Further developments like MV-DUSt3R [29], Fast3R [46], and VGGT [32] extend the framework to multi-view inputs, enabling dense 3D reconstruction in a single forward pass. These 3D foundation models achieve impressive results even with extremely sparse inputs or in zero-shot settings, demonstrating strong generalization across unseen scenes. However, *their further scaling is severely bottlenecked by the scarcity of high-quality 3D annotations*. Unlike 2D vision models trained on billions of images, 3D models heavily rely on sensor-captured datasets like ScanNet, which are costly to scale and restricted to static, indoor environments. This reliance on limited supervised data results in a significant distribution shift when applying these models to diverse, in-the-wild scenes.

To overcome this data scalability bottleneck, Internet videos emerge as a promising alternative, offering abundant coverage of diverse real-world static scenes. For example, the RealEstate10K [50] dataset, sourced from YouTube videos, contains over **10,000** unique scenes, far more than 3D

†Corresponding authors: Hao Chen (`haochen.cad@zju.edu.cn`) and Jia-Wang Bian (`jiawang.bian@ntu.edu.sg`).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

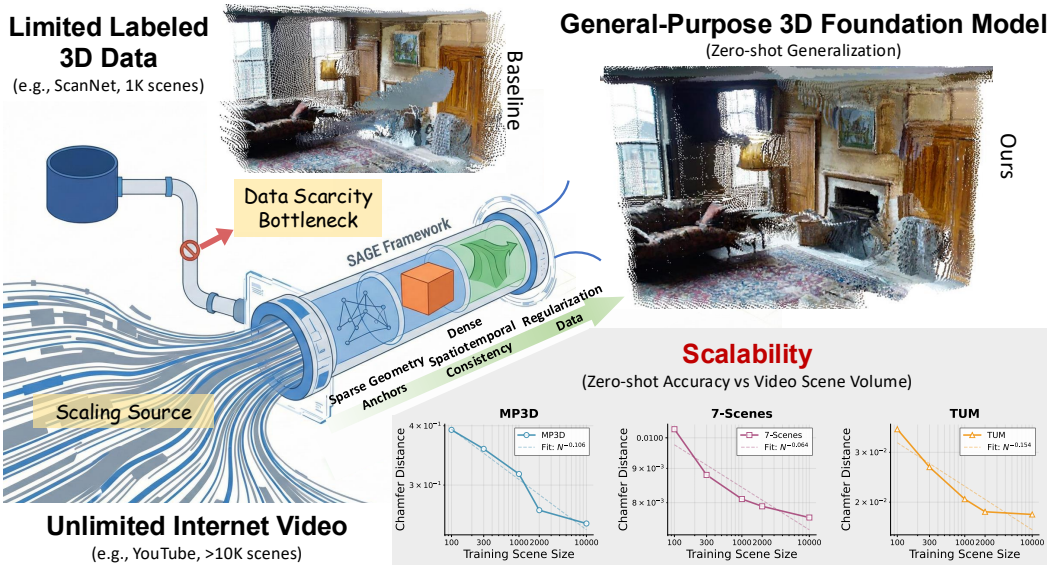



Figure 1: Overview of the SAGE framework for scaling a General-Purpose 3D Foundation Model. To overcome the Data Scarcity Bottleneck inherent in limited labeled 3D datasets, we propose a pipeline that leverages abundant static-scene Internet videos to achieve robust zero-shot generalization on complex unseen scenes. Bottom right: Scalability analysis demonstrates zero-shot reconstruction performance as training data scales from 100 to 10K video scenes, showing consistent improvement with increased data volume across various benchmarks (MP3D, 7-Scenes, and TUM).

datasets like ScanNet [10], which has only around **1,000** depth sensor-annotated scenes. However, *unlocking this potential is non-trivial due to the lack of ground-truth geometry and camera poses.* A straightforward approach, which we refer to as Video Depth Supervision (VDS), fine-tunes models using pseudo-depth labels from video depth estimators [45], but often fails to improve or even degrades performance in our empirical analysis. The main reason is that the predicted video depths contain substantial noise and often lack multi-view consistency, making accurate alignment difficult and ultimately degrading reconstruction quality. Therefore, the core challenge lies in extracting robust supervision signals from noisy, unlabelled video streams without drifting into degenerate solutions.

To tackle this challenge, we propose SAGE, a framework for the *S*calable *A*daptation of *GE*ometric foundation models. Specifically, we develop a spatio-temporal mining pipeline to curate informative training trajectories as well as complementary supervisions from raw videos. We first perform stochastic trajectory sampling to curate frame sequences balancing viewpoint diversity and geometric overlap, providing the data foundation for multi-view learning. Upon these curated trajectories, we utilize sparse point clouds derived from SfM as geometric anchors to provide reliable global structural guidance. To complement these sparse anchors, we leverage differentiable 3D Gaussian Splatting to enforce dense photometric consistency across video frames. Furthermore, to stabilize the adaptation process on diverse video distributions, we introduce a regularization strategy using a small fraction of supervised replay data, effectively mitigating catastrophic forgetting of the original geometric priors.

Extensive experiments demonstrate that SAGE turns Internet video into an effective and scalable supervision source. As shown in Figure 1, zero-shot accuracy improves consistently as the amount of video data scales from 100 to 10K scenes, following a clear log-linear trend without saturation. We validate our approach by adapting pretrained multi-view reconstruction models using over 10,000 video clips from RealEstate10K and DL3DV [21], which is 10 times larger than the standard 3D datasets. Both in-distribution performance and zero-shot generalization capabilities are improved by our SAGE. On unseen benchmarks (7Scenes, TUM-RGBD, Matterport3D), our method reduces the Chamfer Distance of MV-DUSt3R by 20–42% and consistently improves VGGT.

Importantly, SAGE provides an orthogonal scaling axis to recent efforts that improve 3D foundation models through larger architectures and larger curated 3D corpora, such as VGGT [32], $\pi^3$ [40],

and Depth Anything 3 [20]. Rather than training a larger model from scratch with more supervised 3D data, SAGE adapts existing pretrained backbones using cheap and scalable Internet video. To show that this recipe is not tied to a specific architecture, we instantiate SAGE on two representative backbones with different multi-view processing paradigms: MV-DUSt3R, which uses an encoder–decoder architecture with cross-view cross-attention and predicts pointmaps in a selected reference frame; and VGGT, which uses a single transformer with alternating frame-wise and global self-attention, together with multi-task heads for point maps, depth, and camera parameters. Consistent improvements on both backbones suggest that video-based adaptation is complementary to architectural and supervised-data scaling.

Our main contributions are summarized as follows:

- We propose SAGE, a framework for scalable adaptation of pretrained feed-forward geometric foundation models using unlabelled static-scene Internet video, which converts sparse SfM anchors and frozen-renderer multi-view consistency into task-preserving supervision.
- We design a robust hybrid supervision strategy that integrates sparse geometric anchors, dense spatiotemporal consistency, and regularization to effectively learn from noisy video data.
- We instantiate SAGE on two representative backbones, MV-DUSt3R and VGGT, and obtain consistent gains across four benchmarks, establishing Internet video as a viable scaling source for general-purpose 3D foundation models.

## 2 Related work

### 2.1 Geometric Foundation Models & Data Scarcity

Traditional approaches to pose-free multi-view 3D reconstruction include Structure-from-Motion (SfM) [26, 8, 22, 9] and Multi-View Stereo (MVS) [23, 47, 14]. While effective, they rely on complex multi-stage optimization. Recently, the field has shifted towards Geometric Foundation Models such as DUSt3R [35] and VGGSfM [31], which jointly estimate camera parameters and 3D structure in a data-driven manner. Extensions including MV-DUSt3R, Fast3R [46], VGGT [32] and $\pi^3$ [40] further scale this to multi-view inputs.

*However, these models are severely bottlenecked by the scarcity of 3D supervision.* Unlike 2D models trained on web-scale data, they rely on costly annotated datasets like ScanNet, limiting their generalization to in-the-wild distributions. While methods like CUT3R [34] use video sequences, they directly supervise on noisy SfM poses. In contrast, our approach addresses this scalability bottleneck by leveraging unlabelled Internet video. We propose a weakly-supervised adaptation framework that combines sparse geometric anchors with multi-view consistency, enabling foundation models to learn from diverse, real-world data without ground-truth poses. Concurrently, VGGT-$\Omega$ [33] scales foundation-model training to large video corpora using dense annotations from COLMAP bundle adjustment and MVS, together with teacher–student self-training. SAGE is complementary: rather than re-training a foundation model at scale, it adapts released checkpoints through lightweight post-training, using only sparse SfM points and rendering consistency.

### 2.2 Differentiable Rendering as Weak Supervision

Recent works PixelSplat [5], MVSplat [6], and DepthSplat [44] predict 3D Gaussian representations from sparse views for novel view synthesis [36, 38, 39]. Compared with optimization-based methods like 3DGS [17, 12], these feedforward approaches offer better generalization. Extensions such as MVSplat360 [7] and LatentSplat [43] further improve rendering flexibility.

*While these methods target view synthesis, we repurpose the differentiable nature of 3D Gaussians as a dense supervision mechanism.* In our framework, the Gaussian renderer serves as a differentiable loss function that enforces photometric consistency across video frames. This allows us to propagate dense gradients to the underlying geometry in regions undefined by sparse points. Unlike MV-DUSt3R, which trains a Gaussian head for downstream rendering, we utilize this differentiable consistency to robustly adapt the geometric backbone itself.

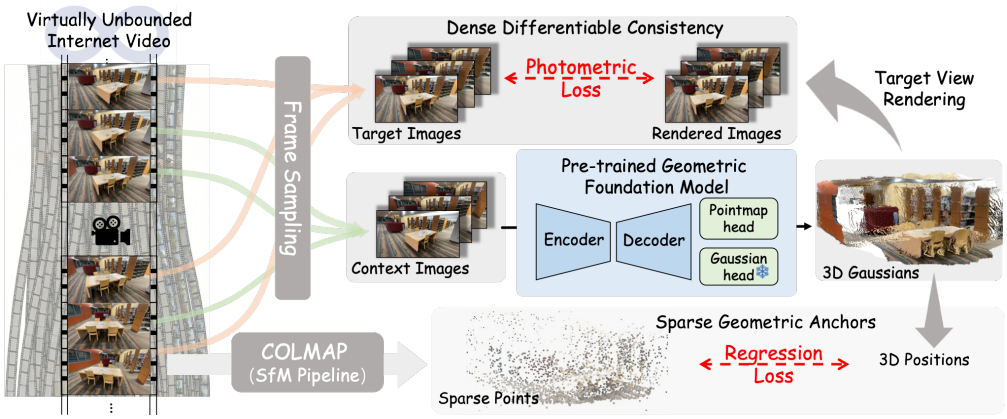



Figure 2: Illustration of SAGE. For each video sequence, we sample context frames as model inputs and designate target frames for novel-view supervision, providing photometric constraints to refine the reconstructed 3D point cloud. Furthermore, we incorporate sparse 3D point clouds that provide consistent geometric constraints, complementing the photometric supervision from the target views.

### 2.3 Scalable Representation Learning from Video

Self-supervised learning on video sequences leverages spatiotemporal consistency as an intrinsic supervision signal. Early works [49] use image-warping losses, while subsequent methods [3, 13] introduce scale-consistent predictions for depth estimation. More recently, CroCo [41, 42] extends masked image modeling to cross-view settings for representation learning. E-RayZer [48] further advances this line by learning geometrically grounded representations through self-supervised 3D Gaussian reconstruction from unlabeled video. NAS3R [15] similarly trains, or self-supervisedly adapts, multi-view geometric models such as MASt3R and VGGT with differentiable Gaussian rendering. These methods primarily target novel-view synthesis and camera estimation, with photometric rendering as the main objective. In contrast, SAGE targets surface-geometry reconstruction evaluated with point-cloud metrics: SfM anchors provide the primary geometric supervision, while rendering through a *frozen* Gaussian head serves only as an auxiliary dense constraint, forcing gradients to refine the backbone's geometry rather than the renderer.

*While most of these methods train task-specific networks from scratch, they share the principle of deriving supervision from video consistency.* Our work extends this principle to the adaptation of pre-trained geometric foundation models. By integrating sparse geometric constraints (bias-reduction) with dense video consistency (variance-reduction), we propose a hybrid learning framework that effectively fine-tunes general-purpose models on noisy, large-scale video data, significantly enhancing their zero-shot generalization.

## 3 Method

We address the problem of scaling up a pre-trained geometric foundation model, $f_\theta$, by adapting it to the domain of unlabelled Internet videos. We formulate this as a weakly-supervised optimization problem, where the goal is to leverage intrinsic spatiotemporal consistency to refine the model without requiring ground-truth 3D annotations.

### 3.1 Preliminary & Problem Formulation

We formulate the training of geometric foundation models as a scalable optimization problem, transitioning from a data-scarce supervised regime to a data-abundant weakly-supervised regime.

#### 3.1.1 Task Instantiation

We focus on *unconstrained sparse-view 3D reconstruction*.

**Definition 3.1** (pose-free sparse-view 3D reconstruction)**.** Let $f_\theta : \mathcal{I}^N \to \mathcal{X}$ denote a geometric foundation model parameterized by $\theta$. Given a set of unposed RGB images $I \in \mathcal{I}^N$, the model

directly predicts a dense 3D representation $\mathcal{X}$ (specifically, pixel-aligned 3D pointmaps) without requiring ground-truth camera poses.

Our objective is to leverage the abundance of static-scene Internet video to push the generalization limits of $f_\theta$.

#### 3.1.2 Bottleneck & Scaling Source

**Supervised Regime (Bottleneck)** Existing models are typically initialized on a supervised dataset $\mathcal{D}_{sup} = \{(I, Y)\}$, where $Y$ represents expensive GT 3D annotations (e.g., depth, poses). While $\mathcal{D}_{sup}$ offers high-fidelity supervision, it is fundamentally non-scalable due to the prohibitive cost of 3D data acquisition. Consequently, $|\mathcal{D}_{sup}|$ remains small, restricting the model's exposure to a narrow distribution of scenes and limiting its zero-shot generalization capabilities to in-the-wild scenarios.

**Weakly-Supervised Regime (Scaling Source)** To address this bottleneck, we introduce a large collection of static-scene Internet videos, denoted as $\mathcal{D}_{video} = \{V_j\}$. In contrast to $\mathcal{D}_{sup}$, this dataset is scalable ($|\mathcal{D}_{video}| \gg |\mathcal{D}_{sup}|$) but *unlabelled*: it lacks explicit 3D geometry ($Y = \emptyset$) and camera poses. However, it contains rich intrinsic supervision encoded in physical constraints, specifically spatiotemporal consistency.

#### 3.1.3 Optimization Objective

We aim to learn an optimal parameter set $\theta^*$ that absorbs geometric knowledge from massive video data while preserving the structural priors learned from the supervised regime. We formulate it as a constrained optimization problem:

$$\min_{\theta} \underbrace{\mathbb{E}_{V\sim\mathcal{D}_{video}}[\mathcal{L}_{consistency}(f_\theta(V))]}_{\text{Scaling via Weak Supervision}} + \gamma \cdot \underbrace{\mathbb{E}_{(I,Y)\sim\mathcal{D}_{sup}}[\mathcal{L}_{sup}(f_\theta(I), Y)]}_{\text{Foundation Regularization}}, \tag{1}$$

where $\mathcal{L}_{consistency}$ represents a hybrid loss function designed to distill robust geometric signals from noisy video consistency, and the second term serves as a regularization constraint to prevent catastrophic forgetting.

### 3.2 Scaling with Spatio-Temporal Weak Supervision

A fundamental requirement for scaling 3D foundation models is the **ability to harvest informative training trajectories and latent supervision from massive, unconstrained video streams**. Unlike structured 3D datasets, raw videos are inherently redundant and exhibit inconsistent viewpoint transitions. To address this, we first introduce a video data processing pipeline in App. B.1, followed by a spatio-temporal mining pipeline that extracts informative training trajectories from raw videos, as shown in Figure 2. First, we curate training trajectories by balancing viewpoint diversity and geometric overlap to ensure representative scene coverage. For these trajectories, we derive two levels of weak supervision: (1) Sparse Geometric Anchors from COLMAP point clouds to provide global structural bias; (2) Dense Differentiable Consistency via 3D Gaussian rendering to propagate gradients through video spatiotemporal consistency. This synergy allows our model to effectively optimize Eq. 1, providing essential supervision for learning robust 3D geometry from diverse real-world scenes in a scalable and stable manner.

#### 3.2.1 Spatiotemporal Sampling from Videos

Unlike prior works [35, 29] that rely on GT depth to determine image overlap, our approach tackles the *scaling* challenge where only raw video streams are available. We exploit the inherent spatiotemporal continuity of videos as a geometric overlap proxy to mine multi-view training samples.

Given a video sequence, we construct training trajectories $\mathcal{V}$ by balancing viewpoint overlap (ensuring stable supervision) and baseline diversity (promoting geometric learning). While uniform temporal sampling with interval $\Delta t$: $\mathcal{V}_{\text{uniform}} = \{I_{t_0+i\Delta t} \mid i = 0, 1, \ldots, n\}$ maintains consistent visual transitions, it constrains viewpoint diversity and introduce motion pattern bias.

We instead employ *stochastic temporal perturbations* by sampling offsets $\delta_i \sim \mathcal{U}(-\epsilon, \epsilon)$: $\mathcal{V}_{\text{perturbed}} = \{I_{t_0+i\Delta t+\delta_i} \mid i = 0, 1, \ldots, n\}$, which preserves temporal structure for geometric consistency while

maximizing camera pose diversity. This curation establishes the data foundation for scalable learning, from which we subsequently derive hierarchical supervision signals (Sec. 3.2.2, 3.2.3).

### 3.2.2 Sparse Geometric Anchors

While dense pseudo-depth maps can be easily generated by off-the-shelf models, they often suffer from multi-view inconsistency and scale drift, which can destabilize the scaling process. In contrast, sparse point clouds derived from SfM provide globally consistent geometric anchors, offering a reliable structural bias. Specifically, the sparse point cloud is projected onto each image to generate corresponding sparse depth maps, using the camera intrinsics $\mathbf{K}$ and extrinsics $[\mathbf{R}\,|\,\mathbf{t}]$ from COLMAP. Each 3D point $\mathbf{P} = (x, y, z)$ is projected to image coordinates $(u, v, d)$ via $\begin{bmatrix} u & v & 1 \end{bmatrix}^\top \sim \mathbf{K}[\mathbf{R}\,|\,\mathbf{t}]\begin{bmatrix} x & y & z & 1 \end{bmatrix}^\top$, where $d$ denotes the projected depth. Then following DUSt3R and MV-DUSt3R, the confidence-aware 3D loss is defined as:

$$\mathcal{L}_{\text{anchor}} = \sum_{v\in\{1,\ldots,N\}} \sum_{i\in\mathcal{D}^v} C_i^{v,1}\ell_{\text{regr}}(v,i) - \alpha \log C_i^{v,1}, \ell_{\text{regr}}(v,i) = \left\| \frac{1}{z} X_i^{v,1} - \frac{1}{\bar{z}} \bar{X}_i^{v,1} \right\|, \quad (2)$$

where $z$, $\bar{z}$ are scale normalization factors and $C_i^{v,1}$ is the predicted confidence score for pixel $i$. The term $X_i^{v,1}$ represents the predicted 3D coordinate in the reference coordinate. This term anchors the predicted geometry to the global SfM structure, mitigating potential geometric drift.

### 3.2.3 Dense Differentiable Consistency

Sparse anchors alone leave the majority of the surface undefined. To propagate geometric supervision to dense regions, we derive the Spatiotemporal Consistency from video. While prior works optimize 3D Gaussians as a scene representation for view synthesis, *we repurpose the splatting pipeline solely as a differentiable supervision module*, keeping all Gaussian parameters frozen. The Gaussian centers are the backbone's predicted 3D points, while opacity, scale, rotation, and color are regressed by a Gaussian head. For MV-DUSt3R we use its pretrained Gaussian head; for point-map backbones without one (*e.g.*, VGGT), we attach a lightweight Gaussian head, train it for 2–3 epochs, and then freeze it during adaptation.

Given a sampled video clip of $M$ temporally ordered frames $\mathcal{V}_{\text{clip}} = \{I_{t_1}, I_{t_2}, \ldots, I_{t_M}\}$, we divide them into *context frames* as inputs and *target frames* as supervision for novel view synthesis. Target views are unseen frames from the same video, used to enforce multi-view consistency during training. To ensure effective supervision, we random sample intermediate frames $t_j \in \{t_2, t_3, \ldots, t_{M-1}\}$ as target views, avoiding extrapolated viewpoints that may introduce artifacts in Gaussian splatting reconstructions. $\hat{I}_k$ is rendered from differentiable Gaussian Splatting from the predicted point cloud. The loss function is:

$$\mathcal{L}_{\text{ren}} = \sum_{I_k\in\{I_{con},I_{tgt}\}} \gamma \cdot \text{LPIPS}(I_k, \hat{I}_k) + \left\| I_k - \hat{I}_k \right\|_2, \quad (3)$$

where $I_{con}$ and $I_{tgt}$ are context and target frames. This loss enforces image consistency through novel view synthesis of 3D Gaussians, thereby constraining the Gaussian centers and improving the accuracy of the predicted 3D point cloud.

## 3.3 Regularization against Catastrophic Forgetting

A critical challenge in adapting foundation models to noisy video data is distribution drift, where the model overfits to the video domain and loses its original general-purpose capabilities. To explicitly mitigate this, we implement the regularization term in Eq. 1 by mixing a small set of supervised replay data $\mathcal{D}_{sup}$ from the original pre-training distribution.

We define the sampling ratio between video data and supervised replay data as $\eta$. Empirically, we find that a small fraction ($\eta \approx 3\%$) is sufficient to stabilize the optimization landscape. This regularization constrains the model parameters to remain in the valid manifold of plausible 3D geometries, preventing catastrophic forgetting while allowing the model to absorb new diversity from Internet videos.

# 4 Experiments

We use the pretrained MV-DUSt3R [29] and VGGT [32] as our baseline, originally trained on 3D ground-truth annotated datasets, and we fine-tune them with Internet video data to enhance its 3D reconstruction performance. Video preprocessing is a one-time, cacheable sparse-SfM pass ($\sim$15 s per clip; over 13K clips in 5–6 hours on four RTX 4090 GPUs), and adaptation completes in under two days on the same hardware. For more details about dataset, Implementation details and Evaluation metrics, please refer to App. B.2.

Table 1: **8-view reconstruction results** on four benchmarks at $224 \times 224$ resolution. Group (A) includes feed-forward foundation models trained from scratch with full 3D supervision, while Groups (B,C) adapt pretrained backbones using only Internet video, with no additional depth or pose annotations. SAGE consistently improves both MV-DUSt3R and VGGT, and matches or surpasses the strongest 3D-supervised models on the zero-shot 7Scenes and TUM-RGBD benchmarks. **VDS** denotes video-depth supervision using pseudo depth from Depth Any Video. **SAGE ($n$K)** uses $n$K video clips for adaptation. Subscripts in green/red show $\Delta$ over the corresponding pretrained backbone. Underline: best within Group (A); **bold**: best within Groups (B,C).

| Method | In-distribution | | Zero-shot Generalization | | | | | |
|---|---|---|---|---|---|---|---|---|
| | ScanNet | | 7Scenes | | TUM-RGBD | | Matterport3D | |
| | DAc@0.2 ↑ | CD ↓ | DAc@0.2 ↑ | CD ↓ | DAc@0.2 ↑ | CD ↓ | DAc@0.5 ↑ | CD ↓ |
| *(A) Foundation models trained from scratch with full 3D supervision (depth / poses)* | | | | | | | | |
| Spann3R [30] | 59.0 | 13.48 | 90.0 | 1.48 | 35.7 | 7.06 | 15.8 | 69.7 |
| DUSt3R [35] | 88.0 | 4.50 | 89.5 | 1.50 | 71.4 | 4.65 | 45.4 | 44.1 |
| MapAnything [16] | 79.5 | 3.05 | 88.1 | 1.56 | 82.0 | 2.96 | 51.8 | 24.0 |
| $\pi^3$ [40] | 97.0 | 0.95 | 95.0 | 0.80 | 92.7 | 1.90 | 57.4 | 22.9 |
| Depth Anything 3 [20] | 89.8 | 2.45 | 91.5 | 0.84 | 85.7 | 1.69 | 67.3 | 17.1 |
| *(B) MV-DUSt3R + video-based adaptation (no new 3D annotations)* | | | | | | | | |
| MV-DUSt3R [29] *(pretrained backbone)* | 88.2 | 1.79 | 89.6 | 0.95 | 87.5 | 2.60 | 30.3 | 43.2 |
| + VDS *(prior recipe)* | 90.0 +1.8 | 1.78 -0.01 | 92.5 +2.9 | 0.89 -0.06 | 84.7 -2.8 | 2.68 +0.08 | 14.3 -16.0 | 82.1 +38.9 |
| + SAGE (1K) | 90.0 +1.8 | **1.61** -0.18 | 96.0 +6.4 | 0.79 -0.16 | 91.7 +4.2 | 2.05 -0.55 | 47.4 +17.1 | 31.7 -11.5 |
| **+ SAGE (10K) (Ours)** | **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 |
| *(C) VGGT + video-based adaptation (no new 3D annotations)* | | | | | | | | |
| VGGT [32] *(pretrained backbone)* | 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 |

## 4.1 Sparse 3D Reconstruction from Multi Inputs

We evaluate SAGE on ScanNet as the in-distribution benchmark, and on 7Scenes, TUM-RGBD, and MP3D to assess zero-shot generalization. All methods are compared under the standard 8-view setting at $224 \times 224$ resolution, with results using 4–24 input views reported in App. C.1. Table 1 compares SAGE with representative pose-free 3D foundation models, including $\pi^3$ [40], Depth Anything 3 [20], and MapAnything [16]. These foundation models in Group (A) are typically trained from scratch with large-scale, carefully curated 3D supervision, higher-capacity architectures, and/or more involved training recipes. In contrast, SAGE in Groups (B,C) adapts pretrained backbones using only unlabelled videos, making the comparison a stringent test of whether video-based adaptation can provide effective 3D generalization.

With the proposed hybrid supervision, SAGE consistently improves both MV-DUSt3R and VGGT across all four benchmarks. For MV-DUSt3R, SAGE (10K) reduces CD by 6%, 20%, 29%, and 42% on ScanNet, 7Scenes, TUM-RGBD, and MP3D, respectively, and further brings a substantial DAc gain of +17.7 on MP3D. For VGGT, SAGE also improves both DAc and CD on every dataset, showing that our adaptation strategy is effective across different pretrained 3D backbones.

Notably, on the zero-shot benchmarks 7Scenes and TUM-RGBD, MV-DUSt3R+SAGE matches or surpasses the strongest fully 3D-supervised foundation models on both DAc and CD, achieving

Table 2: **Ablation study on hybrid supervision and training settings.** We evaluate the contribution of each supervision component and the impact of different training strategies.

| | ScanNet | | 7Scenes | | TUM | | MP3D | |
|---|---|---|---|---|---|---|---|---|
| Method | DAc↑ | CD↓ | DAc↑ | CD↓ | DAc↑ | CD↓ | DAc↑ | CD↓ |
| Baseline (Foundation) | 88.2 | 1.79 | 89.6 | 0.95 | 87.5 | 2.60 | 30.3 | 43.2 |
| **Ours (Full)** | **91.2** | **1.69** | **98.0** | **0.76** | **93.1** | **1.85** | **48.0** | **24.9** |
| *Component Ablation:* | | | | | | | | |
| w/o Sparse Anchors ($\mathcal{L}_{\text{anchor}}$) | 77.0 | 4.33 | 88.5 | 1.43 | 69.4 | 5.20 | 18.9 | 33.6 |
| w/o Dense Consistency ($\mathcal{L}_{\text{ren}}$) | 89.2 | 1.89 | 93.5 | 0.79 | 86.1 | 2.26 | 49.2 | 26.0 |
| w/o Supervised Replay | 76.0 | 3.32 | 88.5 | 1.25 | 80.6 | 2.14 | 45.4 | 32.5 |
| w/o Frozen GS Params | 91.2 | 1.70 | 93.5 | 0.72 | 91.7 | 2.09 | 46.7 | 23.7 |
| *Strategy Variants:* | | | | | | | | |
| w/ Linear Sampling (vs. Jitter) | 91.0 | 1.88 | 95.0 | 0.84 | 90.2 | 2.06 | 49.0 | 25.1 |
| w/ 8 Target Views (vs. 2) | 90.2 | 1.73 | 96.5 | 0.78 | 90.3 | 1.84 | 49.7 | 26.0 |

98.0/0.76 on 7Scenes and 93.1/1.85 on TUM-RGBD. On ScanNet and MP3D, models trained with substantially larger curated 3D corpora still retain advantages in some metrics. This suggests that video-based adaptation is not a replacement for large-scale 3D pretraining, but rather an orthogonal and potentially composable direction for improving pose-free 3D foundation models.

We also compare with VDS, a representative video-based baseline that fine-tunes MV-DUSt3R using pseudo depth from Depth Any Video [45]. While VDS slightly improves in-distribution performance, it degrades zero-shot generalization on TUM-RGBD and MP3D, indicating that noisy and temporally inconsistent pseudo-depth can harm geometric representations. This supports the need for SAGE's SfM-anchored and differentiable consistency design. Qualitative comparisons are provided in App. E.

### 4.2 Ablations

**Effect of Hybrid Supervision** We evaluate the contribution of each component in our hybrid training strategy by removing one supervision at a time. Specifically, we independently ablate the Sparse Anchors, the Dense Consistency, and the small amount of supervised replay data used to stabilize fine-tuning. The results in Table 2 show that the full combination achieves the best reconstruction quality, while excluding any single component leads to a noticeable performance drop. We further examine whether to freeze the pretrained Gaussian representation module during fine-tuning. Results indicate that freezing the GS module leads to more stable optimization and better adaptation of the reconstruction network, whereas updating GS jointly with the network tends to degrade performance.

**Number of Rendered Views** We analyze the effect of varying the number of rendered novel views per training sample, as shown in Table 2. By default, we supervise the model with two target views to balance training efficiency and supervision quality. Increasing the number of rendered views (e.g., w/ 8 Target Views) provides stronger geometric constraints but also exposes the model to more unseen regions across views, which may introduce noise into the supervision signal.

**Frame Sampling Strategy** We ablate the effect of our perturbed frame sampling strategy against linear sampling. As shown in Table 2, compared with video frame sampling without perturbation (w/ Linear Sampling), our perturbed sampling consistently yields higher accuracy and more stable performance across benchmarks. These improvements validate its effectiveness, and we adopt it as the default strategy for constructing multi-view training sequences. We further compare against content-aware context and target selection based on rotation-compensated flow and view overlap (App. C.3). Temporal jitter remains a strong accuracy–simplicity trade-off on RE10K/DL3DV, whereas content-aware selection is most beneficial on wide-baseline MP3D (CD 24.9→23.4), suggesting its value for less regular camera motion.

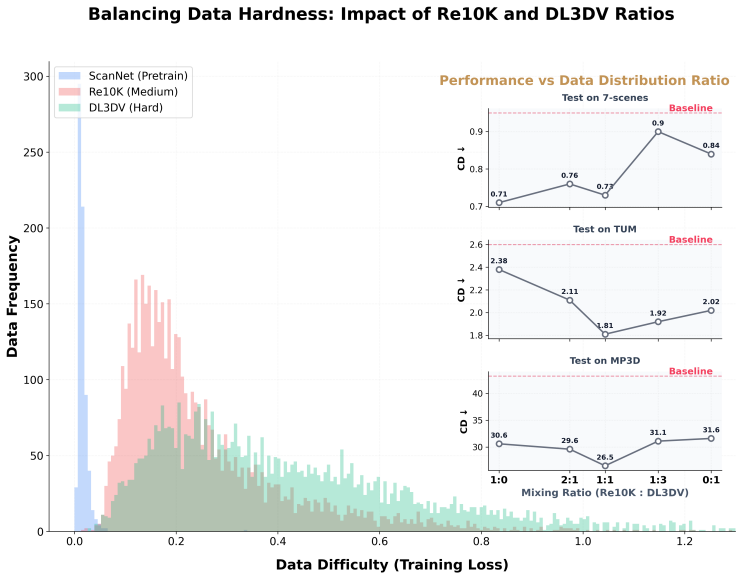



Figure 3: Empirical analysis of training data difficulty and its impact on model generalization. (Left) Training distribution across three datasets. (Right) Generalization performance (CD ↓) on three test sets under various mixing ratios of Re10K to DL3DV samples. Subplots indicate that while pure medium-hardness data (1:0) benefits simple scenes, a balanced mixture (1:1) yields the most robust generalization on large-scale environments like MP3D. The dashed red lines denote the pre-trained baseline performance.

# 5 Discussion

## 5.1 Beyond Scale: How Does the Data Difficulty Spectrum Shape Generalization?

We analyze the training data distribution in terms of sample difficulty, measured by the per-sample training loss of the pretrained model. As shown in the left panel of Figure 3, the Re10K dataset predominantly contains medium-difficulty samples, while DL3DV offers a higher proportion of challenging samples with elevated training losses. This observation aligns with our intuition, as Re10K features more structured indoor scenes, whereas DL3DV encompasses complex real-world environments with diverse lighting conditions and cluttered backgrounds. As shown in the right panel of Figure 3, we conduct a sensitivity analysis by varying the mixing ratio between medium-difficulty samples (Re10K) and high-difficulty samples (DL3DV) to investigate how the difficulty composition of training data influences cross-dataset generalization. The pre-trained baseline (red dashed line) consistently exhibits the highest error across all benchmarks, underscoring the necessity of fine-tuning on diverse data. For relatively structured and simple scenes, such as 7Scenes, the model achieves its peak performance with a higher proportion of medium-difficulty data (*e.g.*, the 1:0 ratio). In contrast, for large-scale and complex environments like MP3D and TUM, the inclusion of high-difficulty samples is crucial for pushing the performance boundary. As shown in the right panels of Figure 3, the Chamfer Distance (CD) significantly drops as the DL3DV ratio increases, reaching its optimum at the balanced 1:1 ratio. Beyond this point, however, we observe diminishing returns or slight performance fluctuations. These results suggest that a balanced curriculum of medium and high difficulty is essential for robust generalization.

## 5.2 Regularization Strength & Prevention of Forgetting

As formulated in Eq. 1, the supervised replay data acts as a regularization term. A key question is: *what is the minimal regularization strength required to prevent catastrophic forgetting?* To answer this, we vary the ratio between weakly-supervised video data $\mathcal{D}_{video}$ and supervised replay data $\mathcal{D}_{sup}$. As shown in Table 3, a $30:1$ ratio achieves the optimal balance. Without sufficient regularization (*e.g.*, ratio $\to \infty$), the model drifts from the metric scale derived from ScanNet. Conversely, even a very sparse signal (100:1) significantly stabilizes the optimization, confirming that the foundation model requires only lightweight anchoring to maintain its priors while scaling up.

Table 3: **Effect of Video-to-Supervised Data Ratio.** Evaluation on four datasets with different video-to-supervised sample ratios. A 30:1 ratio is adopted as the default, balancing reconstruction quality with reliance on pretrained supervised data.

| | ScanNet | | 7Scenes | | TUM | | MP3D | |
|---|---|---|---|---|---|---|---|---|
| Ratio | DAc↑ | CD↓ | DAc↑ | CD↓ | DAc↑ | CD↓ | DAc↑ | CD↓ |
| 10:1 | 90.0 | 1.77 | 94.5 | 0.75 | 91.6 | 1.84 | 44.9 | 24.55 |
| 20:1 | 91.2 | 1.62 | 98.0 | 0.76 | 89.0 | 2.08 | 47.2 | 24.90 |
| 30:1 | 91.2 | 1.69 | 98.0 | 0.76 | 93.1 | 1.85 | 48.0 | 24.90 |
| 100:1 | 90.0 | 1.84 | 94.5 | 0.87 | 93.0 | 1.89 | 43.4 | 23.73 |

Table 4: **Impact of Fine-tuning with Videos on Pose Estimation.** Comparison between the baseline and our method on relative rotation error (RRE), relative translation error (RTE), and mean average error (mAE). Lower is better.

| | 7Scenes | | | TUM RGBD | | | MP3D | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | RRE | RTE | mAE | RRE | RTE | mAE | RRE | RTE | mAE |
| MV-DUSt3R | 0.3 | 3.2 | 13.8 | 2.7 | 16.0 | 27.0 | 56.9 | 55.9 | 67.3 |
| Ours (SAGE) | 0.2 | 2.9 | 13.3 | 2.0 | 15.7 | 26.4 | 41.2 | 42.5 | 55.9 |

Table 5: **Reconstruction on UASOL (out-of-distribution, outdoor dataset).** SAGE outperforms MV-DUSt3R and the VDS variant, which suffers from inconsistent and noisy multi-frame depth predictions.

| **Method** | **DAc-0.5 ↑** | **CD ↓** |
|---|---|---|
| MV-DUSt3R | 3.3 | 27.8 |
| MV-DUSt3R + VDS | 26.0 | 31.8 |
| Ours (SAGE) | **27.8** | **18.8** |

### 5.3 Capability Gains from Video-based Adaptation

**Pose Estimation** We observe that the proposed training scheme with video data not only improves the quality of the reconstruction, but also improves the accuracy of the model pose estimation. Following MV-DUSt3R [29], we estimate camera poses using a standard RANSAC [11] and PnP [18] pipeline, where 2D–3D correspondences are formed between image pixels and the predicted pointmaps generated by the network. In Table 4, our SAGE yields consistently lower rotation and translation errors across multiple datasets, demonstrating that video-based fine-tuning leads to more geometrically reliable outputs even without access to ground-truth annotations of video.

**Evaluation on OOD Outdoor Scenes** Although the training videos primarily consist of indoor scenes, they also include a portion of outdoor scenes. As shown in Table 5, SAGE generalizes surprisingly well to the out-of-distribution outdoor UASOL dataset [2], achieving a DAc@0.5 of 27.8 and outperforming the baseline MV-DUSt3R [29]. This demonstrates the advantage of leveraging diverse video data, where incidental exposure to outdoor environments contributes to improved robustness beyond the training distribution.

## 6 Conclusion

We introduce **SAGE**, a framework that unlocks the scalability of geometric foundation models by exploiting the intrinsic consistency of Internet video. By formulating adaptation as a regularized weakly-supervised optimization, SAGE balances structural bias (from sparse anchors) and variance reduction (from dense differentiable consistency) to robustly learn from noisy data. Our experiments show log-linear scaling laws in geometric performance, reducing zero-shot Chamfer Distance by 20–42% on MV-DUSt3R with consistent gains on VGGT. This work establishes a scalable paradigm for evolving general-purpose 3D models using abundant static-scene video, turning the data scarcity bottleneck into a data abundance opportunity.

## Acknowledgments and Disclosure of Funding

This work was in part supported by National Natural Science Foundation of China (No. 62576315), National Research Foundation, Singapore under its National Research Foundation Fellowship in Artificial Intelligence (NRF Award NRFFIAI1-2024-0010), NTU FA-SUG, Ningbo Global Innovation Center, Zhejiang University (Grant No. 701G26003), and Hangzhou National AI Application Pilot Base (Embodied Intelligence). The authors declare no competing interests.

## References


[1] Aanæs, H., Jensen, R. R., Vogiatzis, G., Tola, E., and Dahl, A. B. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016.

[2] Bauer, Z., Gomez-Donoso, F., Cruz, E., Orts-Escolano, S., and Cazorla, M. UASOL, a large-scale high-resolution outdoor stereo dataset. *Scientific Data*, 6(1):162, 2019. doi: 10.1038/s41597-019-0168-5.

[3] Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.-M., and Reid, I. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32, 2019.

[4] Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. Matterport3D: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pp. 667–676, 2017. doi: 10.1109/3DV.2017.00081.

[5] Charatan, D., Li, S. L., Tagliasacchi, A., and Sitzmann, V. pixelSplat: 3D Gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

[6] Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.-J., and Cai, J. MVSplat: Efficient 3D Gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pp. 370–386. Springer, 2024.

[7] Chen, Y., Zheng, C., Xu, H., Zhuang, B., Vedaldi, A., Cham, T.-J., and Cai, J. MVSplat360: Feed-forward 360 scene synthesis from sparse views. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107064–107086. Curran Associates, Inc., 2024. doi: 10.52202/079017-3399.

[8] Crandall, D. J., Owens, A., Snavely, N., and Huttenlocher, D. P. SfM with MRFs: Discrete-continuous optimization for large-scale structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2841–2853, 2013. doi: 10.1109/TPAMI.2012.218.

[9] Cui, H., Gao, X., Shen, S., and Hu, Z. HSfM: Hybrid structure-from-motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1212–1221, 2017.

[10] Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

[11] Fischler, M. A. and Bolles, R. C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[12] Gao, Z., Bian, J.-W., Lin, G., Chen, H., and Shen, C. SurfaceSplat: Connecting surface reconstruction and Gaussian splatting. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10, 2025. doi: 10.1109/ICCV51701.2025.02649.

[13] Godard, C., Mac Aodha, O., Firman, M., and Brostow, G. J. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3828–3838, 2019.

[14] Goesele, M., Curless, B., and Seitz, S. M. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 2402–2409. IEEE, 2006.

[15] Huang, R., Luo, W., Mao, Y., and Mikolajczyk, K. From none to all: Self-supervised 3D reconstruction via novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37358–37369, June 2026.

[16] Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S. R., Richardt, C., Ramanan, D., Scherer, S., and Kontschieder, P. MapAnything: Universal feed-forward metric 3D reconstruction. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509, 2026. doi: 10.1109/3DV69130.2026.00054.

[17] Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, July 2023. doi: 10.1145/3592433.

[18] Lepetit, V., Moreno-Noguer, F., and Fua, P. EPnP: An accurate $O(n)$ solution to the PnP problem. *International journal of computer vision*, 81:155–166, 2009.

[19] Leroy, V., Cabon, Y., and Revaud, J. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, pp. 71–91. Springer, 2024.

[20] Lin, H., Chen, S., Liew, J. H., Chen, D. Y., Li, Z., Zhao, Y., Peng, S., Guo, H., Zhou, X., Shi, G., Feng, J., and Kang, B. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, volume 2026, pp. 141261–141285, 2026.

[21] Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

[22] Schonberger, J. L. and Frahm, J.-M. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016.

[23] Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pp. 519–528. IEEE, 2006.

[24] Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., and Fitzgibbon, A. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2930–2937, 2013.

[25] Smart, B., Zheng, C., Laina, I., and Prisacariu, V. A. Splatt3R: Zero-shot Gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024.

[26] Snavely, N., Seitz, S. M., and Szeliski, R. Photo tourism: exploring photo collections in 3D. In *ACM SIGGRAPH 2006 Papers*, SIGGRAPH '06, pp. 835–846. Association for Computing Machinery, 2006. doi: 10.1145/1179352.1141964.

[27] Sturm, J., Engelhard, N., Endres, F., Burgard, W., and Cremers, D. A benchmark for the evaluation of RGB-D SLAM systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580. IEEE, 2012.

[28] Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., and Batra, D. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[29] Tang, Z., Fan, Y., Wang, D., Xu, H., Ranjan, R., Schwing, A., and Yan, Z. MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5283–5293, 2025. doi: 10.1109/CVPR52734.2025.00498.

[30] Wang, H. and Agapito, L. 3D reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89, 2025. doi: 10.1109/3DV66043.2025.00013.

[31] Wang, J., Karaev, N., Rupprecht, C., and Novotny, D. VGGSfM: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21686–21697, 2024.

[32] Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., and Novotny, D. VGGT: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

[33] Wang, J., Chen, M., Zhang, S., Karaev, N., Schönberger, J., Labatut, P., Bojanowski, P., Novotny, D., Vedaldi, A., and Rupprecht, C. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21486–21499, June 2026.

[34] Wang, Q., Zhang, Y., Holynski, A., Efros, A. A., and Kanazawa, A. Continuous 3D perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025. doi: 10.1109/CVPR52734.2025.00983.

[35] Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., and Revaud, J. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

[36] Wang, W., Chen, D. Y., Zhang, Z., Shi, D., Liu, A., and Zhuang, B. ZPressor: Bottleneck-aware compression for scalable feed-forward 3DGS. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113407–113436, 2025.

[37] Wang, W., Cao, Q., Gao, S., Chen, D. Y., Xu, H., Bian, W., Peng, S., Cham, T.-J., Zheng, C., Geiger, A., Cai, J., Bian, J.-W., and Zhuang, B. Feed-forward 3D scene modeling: A problem-driven perspective. *arXiv preprint arXiv:2604.14025*, 2026.

[38] Wang, W., Chen, Y., Zhang, Z., Liu, H., Wang, H., Feng, Z., Qin, W., Chen, F., Bian, J.-W., Zhu, Z., Chen, D. Y., and Zhuang, B. VolSplat: Rethinking feed-forward 3D Gaussian splatting with voxel-aligned prediction. In *European Conference on Computer Vision (ECCV)*, pp. 1–19, 2026.

[39] Wang, W., Li, Z., Shi, J., Zhang, Z., Ye, B., Pollefeys, M., Chen, D. Y., and Zhuang, B. TriSplat: Simulation-ready feed-forward 3D scene reconstruction. *arXiv preprint arXiv:2605.26115*, 2026.

[40] Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., and He, T. $\pi^3$: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026.

[41] Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., and Revaud, J. CroCo: Self-supervised pre-training for 3D vision tasks by cross-view completion. *Advances in Neural Information Processing Systems*, 35: 3502–3516, 2022.

[42] Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., and Revaud, J. CroCo v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17969–17980, 2023.

[43] Wewer, C., Raj, K., Ilg, E., Schiele, B., and Lenssen, J. E. latentSplat: Autoencoding variational Gaussians for fast generalizable 3D reconstruction. In *European Conference on Computer Vision*, pp. 456–473. Springer, 2024.

[44] Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., and Pollefeys, M. DepthSplat: Connecting Gaussian splatting and depth. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16453–16463, 2025. doi: 10.1109/CVPR52734.2025.01534.

[45] Yang, H., Huang, D., Yin, W., Shen, C., Liu, H., He, X., Lin, B., Ouyang, W., and He, T. Depth Any Video with scalable synthetic data. In *International Conference on Learning Representations*, volume 2025, pp. 97335–97349, 2025.

[46] Yang, J., Sax, A., Liang, K. J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., and Feiszli, M. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935, 2025. doi: 10.1109/CVPR52734.2025.02042.

[47] Yao, Y., Luo, Z., Li, S., Fang, T., and Quan, L. MVSNet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 767–783, 2018.

[48] Zhao, Q., Tan, H., Wang, Q., Bi, S., Zhang, K., Sunkavalli, K., Tulsiani, S., and Jiang, H. E-RayZer: Self-supervised 3D reconstruction as spatial visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7525–7535, June 2026.

[49] Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1851–1858, 2017.

[50] Zhou, T., Tucker, R., Flynn, J., Fyffe, G., and Snavely, N. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4), July 2018. doi: 10.1145/3197517.3201323.

# A Preliminary

## A.1 Common Paradigm of 3D Geometric Foundation Models

Recent 3D geometric foundation models (3D GFMs), represented by DUSt3R [35] and VGGT [32], follow a largely unified pipeline that performs end-to-end geometric reasoning from RGB images, as illustrated in Figure 4. Given one or multiple input views, these models first extract visual tokens with a shared image encoder, then perform cross-view feature interaction to capture multi-view geometric consistency, and finally regress geometry in pixel-aligned forms such as 3D point maps, depth maps, and camera parameters.

We instantiate this paradigm with two representative backbones that differ in their cross-view interaction and output parameterization. MV-DUSt3R extends DUSt3R to multiple views using an encoder–decoder architecture with cross-view cross-attention, and predicts point maps in a selected reference frame. VGGT, in contrast, uses a single transformer with alternating frame-wise and global self-attention, together with multi-task heads for point maps, depth, and camera parameters. These complementary designs allow us to evaluate whether our adaptation strategy transfers across both reference-frame point-map reconstruction and unified multi-view transformer prediction. Despite such architectural progress, existing 3D GFMs still rely heavily on large-scale, strongly supervised 3D datasets for pretraining, which constrains their scalability and motivates alternative adaptation strategies using weaker but more scalable supervision sources.

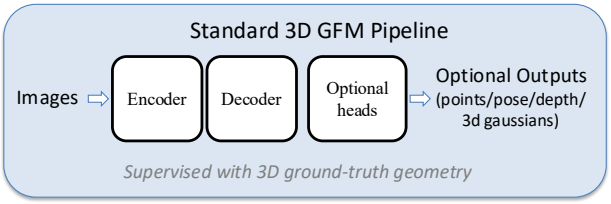



Figure 4: **Standard 3D GFM Training and Inference Framework.** A generic 3D GFM consists of an image encoder, a decoder for cross-view feature interaction, and optional output heads that regress various geometric representations. These models are commonly trained with explicit 3D supervision.

# B Experimental Details

## B.1 Video Data Processing Pipeline

This section details the video data processing pipeline used to construct the training set for weakly-supervised adaptation of 3D geometric foundation models. An overview of the workflow is illustrated in Fig. 5.

**1) Video Acquisition.** We collect raw videos from large-scale Internet video datasets, RealEstate10K and DL3DV contain minimal dynamic content, making them well-suited for fine-tuning pretrained 3D reconstruction models. These videos provide rich multi-view observations while avoiding reliance on curated 3D ground-truth geometry. **2) Video Filtering.** To ensure sufficient geometric content and temporal coverage for reliable multi-view reconstruction, we further filter out low-quality video sequences, those are too short, exhibit large viewpoint discontinuities, or contain dynamic objects, as they may cause COLMAP reconstruction failures or produce excessively noisy point clouds: (1) minimum duration of 10 seconds, and (2) minimum spatial resolution of 480p. Videos that do not meet these requirements are discarded. This filtering step removes low-quality or short clips that are unlikely to yield stable camera poses or consistent multi-view correspondences. **3) Structure-from-Motion via COLMAP.** For each retained video, we run a standard Structure-from-Motion (SfM) pipeline using COLMAP to recover sparse 3D geometry and camera parameters. Specifically, we first extract up to 16K local features per frame using GPU acceleration, which improves robustness under challenging conditions such as motion blur, illumination changes, and textureless regions. We then perform exhaustive feature matching across frames to

maximize reliable correspondences and support wide-baseline reconstruction. Finally, COLMAP performs triangulation and bundle adjustment to estimate sparse 3D point clouds, camera extrinsic parameters, and camera intrinsic parameters. Only reconstructions that successfully converge are retained for subsequent training. **4) Output Representation.** The final output of the pipeline consists of sparse 3D points with associated camera intrinsics and extrinsics, which serve as weak geometric supervision for model adaptation. Notably, no dense depth maps or ground-truth meshes are used at any stage of training. This design allows the proposed method to scale to large and diverse Internet video collections while maintaining compatibility with standard SfM tools.

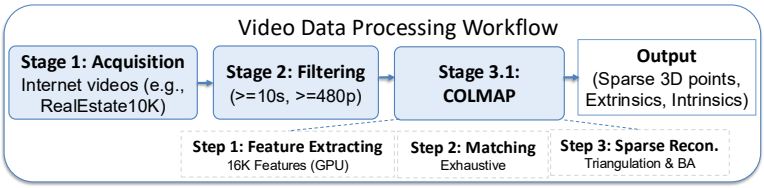



Figure 5: Video Data Processing Workflow for Weak Geometric Supervision. Raw Internet videos are filtered based on duration and resolution and processed using a COLMAP-based structure-from-motion pipeline. Feature extraction, exhaustive matching, and sparse reconstruction are performed to obtain sparse 3D points and camera parameters, which serve as weak geometric supervision for scalable model adaptation.

**Pipeline statistics at scale.** To quantify the practical scalability of our preprocessing pipeline, we report the following statistics measured on the RE10K subset used for training. The high success rate benefits from RE10K's upstream pre-filtering: clips with motion blur or aggressive viewpoint discontinuity are largely already excluded by the dataset curators. Crucially, COLMAP failures produce *missing* data rather than *erroneous* data: clips that do not converge are simply discarded, so no noisy supervision is injected into training. The end-to-end preprocessing for our $10\text{K}+$ scene corpus is a modest one-time cost that runs in parallel on commodity GPUs.

Table 6: **COLMAP preprocessing statistics on the RealEstate10K subset.** The pipeline scales to thousands of clips with negligible failure rate; preprocessing for the entire $10\text{K}+$ scene set is a one-time, modest cost.

| Statistic | Value |
|---|---|
| RE10K clips processed | $>$13K (50–200 frames per clip) |
| COLMAP failures | 8 clips ($<0.1\%$) |
| Average processing time | $\sim$15 sec / clip (170 frames), GPU-accelerated |
| Sparse points / frame | 300–2000 (mean $\sim$1500) |
| Total preprocessing time | $\sim$5–6 h for $>$13K clips (4$\times$ RTX 4090, 2 workers) |
| Reconstruction stage | sparse SfM only (no dense MVS) |

### B.2 Experimental Setup

**Datasets** We use videos from RealEstate10K [50] and DL3DV [21] as the primary training sources. These datasets contain a large number of diverse scenes of common real-life environments. We preprocess the dataset by filtering out short videos. In total, over 10,000 video clips are utilized for training, resulting in a 10-fold increase in scene diversity compared to standard 3D datasets. To assess the reconstruction performance, we evaluate all models on four benchmarks. ScanNet [10] is divided into training and test sets; all methods are trained on the training split and tested on the official test split. In addition, we include 7Scenes [24], TUM RGB-D [27], and Matterport3D [4] as zero-shot datasets, which are not seen during training by any of the methods. These datasets provide a more rigorous assessment of generalization ability to unseen and diverse real-world scenes.

**Evaluation metrics** We evaluate reconstruction quality using two main metrics: Chamfer Distance (CD) [1] and Distance Accuracy @0.2 (DAc), which measures the proportion of pixels whose

normalized distance to the ground-truth pointmap is less than or equal to 0.2. CD serves as a global geometric error metric but is sensitive to point cloud density and outliers, and can be affected by prediction sparsity or noise. In contrast, DAc directly quantifies the proportion of well-aligned points and more reliably reflects overall structural consistency, making it particularly suitable for evaluating dense and accurate reconstructions. In more challenging settings, such as the MP3D [4] dataset characterized by low viewpoint overlap and textureless surfaces, we use Distance Accuracy @0.5 (DAc@0.5) as a more appropriate metric. It offers a more reliable assessment of reconstruction performance under such conditions.

### B.3 Implementation details

**Video Frame Sampling** When constructing training inputs, we follow the uniform sampling strategy described in the main text. For RealEstate10K [50], we set the temporal interval $\Delta t = 20$ based on its typical recording frame rate, while for DL3DV [21], we use $\Delta t = 15$ due to its faster motion and denser frame sampling. For the perturbed sampling variant, we apply a random offset of up to $\pm 5$ frames around each sampled timestamp to increase viewpoint diversity.

**Test Dataset Construction** Our test dataset construction follows the protocol proposed in CroCo [41] and MV-DUSt3R [29], but with key modifications. Unlike MV-DUSt3R, we do not use synthetic images rendered via Habitat-Sim [28], as such rendered data often suffers from limited photorealism and domain discrepancies compared to real-world images. Instead, for all test datasets, we begin by randomly selecting the first image in each test sequence. Subsequent views are sampled based on their geometric overlap with the previously selected views, which is computed using ground-truth depth. A predefined threshold is applied to determine whether a candidate frame provides sufficient overlap to be included in the test sequence.

**Training Process** When we finetune the MV-DUSt3R, the resolution of input images is $224 \times 224$. Each training sample consists of 8 input views, with either 2 or 8 novel views rendered for supervision. During training, we mix the video data from RealEstate10K and DL3DV. Additionally, we include approximately 3% (about 400) pretrained samples, resulting in a 30:1 ratio between video data and pretrained supervision data. To preserve the pretrained 3D Gaussian representation, we freeze the gaussian head and only optimize the reconstruction network. The model is trained for 5 epochs. In each epoch, we sample 15,000 training trajectories from each dataset. The entire training process is highly efficient, requiring less than two days on 4 RTX 4090 GPUs using only low-cost video data. As for the VGGT finetuning, we follow the original VGGT resolution setting and the model is trained for 15 epochs. Since VGGT has no Gaussian head, we attach a lightweight Gaussian head that regresses opacity, scale, rotation, and color, with the Gaussian centers given by VGGT's predicted 3D points; the head is trained for 2–3 epochs and then frozen during adaptation.

## C Additional Experimental Results

### C.1 Multi-view Reconstruction on ScanNet

In the main paper, we report 3D point cloud reconstruction results using 8 input views. Here, we provide additional results under varying numbers of input views. Table 7 shows the performance on the ScanNet test set across different view counts. As shown in Table 7, our method achieves the best overall reconstruction performance across most settings. However, as the number of input views increases, single forward inference becomes more challenging. We also report more results on MP3D [4] with varying number of inputs in Table 8. For multi-view reconstruction, Spann3R [30] and DUSt3R [35] are excluded due to high memory usage and optimization overhead, respectively. Only single forward inference methods are compared.

### C.2 Effect of Viewpoint Interpolation for Novel View Supervision

As discussed in the main paper, when constructing novel view supervision, we observe that interpolated target views (views that lie spatially between the input images) consistently yield better results than extrapolated ones. This is because extrapolated views tend to fall outside the effective coverage of the input images, making view consistency supervision more susceptible to ambiguity and error, which is also analyzed in Splatt3r [25]. In contrast, interpolated views benefit from overlapping

Table 7: **Multi-view 3D reconstruction performance on ScanNet.** Comparison across five ScanNet scenes with varying number of images (4 to 24). Each column reports Distance Accuracy (DAc↑) and Chamfer Distance (CD↓).

| Method | 4 | | 12 | | 16 | | 20 | | 24 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DAc | CD | DAc | CD | DAc | CD | DAc | CD | DAc | CD |
| VGGT | 72.0 | 5.19 | 58.0 | 4.83 | 50.0 | 4.91 | 45.0 | 5.56 | 36.0 | 5.62 |
| MV-DUSt3R | 88.0 | **1.67** | **86.0** | 2.63 | 82.0 | 2.43 | 80.0 | 2.95 | **73.0** | 2.49 |
| Ours | **90.0** | 1.71 | 84.0 | **2.31** | **82.0** | **1.98** | **82.0** | **2.41** | 72.0 | **2.34** |

Table 8: **Multi-view 3D performance on MP3D.** Evaluation across different number of images. Each column reports Distance Accuracy (DAc@0.5↑) and Chamfer Distance (CD↓). All methods are evaluated without global alignment.

| Method | 4 | | 12 | | 16 | | 20 | | 24 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | DAc | CD | DAc | CD | DAc | CD | DAc | CD | DAc | CD |
| VGGT | 44.1 | 40.65 | 24.6 | 46.28 | 19.5 | 57.45 | 13.3 | 57.56 | 13.3 | 78.27 |
| MV-DUSt3R | 42.9 | 42.08 | 23.0 | 49.28 | 24.0 | 62.29 | 11.7 | 77.89 | 7.1 | 92.19 |
| SAGE (Ours) | **62.2** | **19.28** | **35.2** | **30.41** | **30.1** | **34.73** | **18.9** | **39.69** | **14.3** | **52.61** |

content on both sides, leading to more stable and effective training signals. In our experiments, we provide a comparative analysis of both strategies.

We ablate the effect of target view selection for novel-view supervision. As shown in Table 9, interpolated views consistently yield better reconstruction performance compared to extrapolated ones under the same training setup. Note that this experiment is conducted under a slightly different configuration from the main paper and serves primarily for comparative analysis.

Table 9: **Comparison of novel-view supervision strategies.** Performance across four datasets using extrapolated vs. interpolated target views. Interpolated targets consistently yield better reconstruction results.

| Method | ScanNet | | 7Scenes | | TUM | | MP3D | |
|---|---|---|---|---|---|---|---|---|
| | DAc | CD | DAc | CD | DAc | CD | DAc | CD |
| Ours (extra) | 88.0 | 2.31 | 100.0 | 0.74 | 87.5 | 2.53 | 30.6 | 39.91 |
| Ours (inter) | **90.0** | **1.88** | **100.0** | **0.48** | **91.7** | **1.78** | **46.4** | **21.80** |

### C.3 Content-Aware Frame Sampling

Our default strategy selects context frames with stochastically perturbed temporal intervals and randomly samples intermediate frames as targets, using timestamps as a proxy for viewpoint change. Since a fixed temporal interval can correspond to very different baselines (*e.g.*, a slow translational pan versus a fast rotation), we compare it against content-aware alternatives that use geometric cues to drive sampling:

- **Context selection (A1).** We use rotation-compensated optical flow to estimate translational parallax and a coverage score to measure view overlap, selecting frames that balance geometric diversity with reliable shared content.
- **Target selection (B1).** We enforce an overlap ratio between $0.6$ and $0.9$ between each target view and the selected context views.

A0 and B0 denote our default temporal-jitter context sampling and random intermediate target sampling, respectively.

As shown in Table 10, flow-based context selection (A1B0) yields mixed changes on ScanNet, 7Scenes, and TUM, but clearer gains on MP3D (DAc/CD from 48.0/24.9 to 49.1/24.6). This suggests that the temporal heuristic is adequate for RE10K and DL3DV, whose camera trajectories are relatively regular, whereas content-aware selection is more beneficial under larger viewpoint variation. Overlap-aware target curation alone (A0B1) does not improve performance; its benefit emerges only when combined with flow-based context selection (A1B1), which yields the best MP3D result of 49.5/23.4. This indicates that moderate-overlap targets are most effective when the context views already provide sufficient parallax. Overall, temporal jitter offers a strong accuracy–simplicity tradeoff on the datasets considered, while content-aware sampling is expected to become increasingly valuable as adaptation expands to more diverse videos with less regular camera motion.

Table 10: **Temporal vs. content-aware frame sampling.** A0/B0: default temporal-jitter context sampling and random intermediate targets. A1: flow-based context selection. B1: overlap-curated targets.

| Method | ScanNet | | 7Scenes | | TUM | | MP3D | |
|---|---|---|---|---|---|---|---|---|
| | DAc↑ | CD↓ | DAc↑ | CD↓ | DAc↑ | CD↓ | DAc↑ | CD↓ |
| A0B0: temporal jitter (Ours) | 91.2 | 1.69 | **98.0** | **0.76** | 93.1 | **1.85** | 48.0 | 24.9 |
| A1B0: flow-based context | **91.4** | **1.63** | 95.5 | 0.83 | **93.4** | 1.89 | 49.1 | 24.6 |
| A0B1: curated targets | 90.6 | 1.70 | 96.4 | 0.89 | 92.3 | 1.86 | 48.0 | 25.1 |
| A1B1: flow-based context + curated targets | 90.4 | **1.63** | 97.9 | 0.78 | 91.2 | 1.86 | **49.5** | **23.4** |

### C.4 Pose Estimation: Comparison with State-of-the-Art Foundation Models

In addition to the MV-DUSt3R vs. SAGE comparison reported in the main paper (Table 4), we further compare against three recent feed-forward foundation models that natively predict (or implicitly encode) camera parameters: $\pi^3$ [40], Depth Anything 3 (DA3) [20], and VGGT [32]. Following the same RANSAC + PnP readout protocol as the main paper, we report the relative rotation error (RRE), relative translation error (RTE), and mean average error (mAE) on 7Scenes in Table 11.

Table 11: **Cross-method pose estimation on 7Scenes.** SAGE delivers competitive or superior pose accuracy over $\pi^3$, DA3, and VGGT, despite using only video-based weak supervision. Notably, VGGT+SAGE surpasses $\pi^3$ on all three pose metrics.

| **Method** | **RRE ↓** | **RTE ↓** | **mAE ↓** |
|---|---|---|---|
| $\pi^3$ [40] | 0.3 | 3.2 | 15.4 |
| Depth Anything 3 [20] | 0.1 | 2.6 | 13.1 |
| VGGT [32] | 0.4 | 2.7 | 12.9 |
| **VGGT + SAGE** | **0.1** | **2.1** | **12.1** |

This further demonstrates that the SAGE recipe transfers consistently across architecturally distinct backbones (CroCo-based MV-DUSt3R and ViT-based VGGT), and that video-based adaptation can match or exceed pose-estimation quality of from-scratch foundation models trained with substantially larger 3D-supervised corpora.

### C.5 Robustness to Anchor Quality and Density

A natural concern about SAGE's reliance on COLMAP-derived sparse anchors is brittleness: what happens when SfM points are sparse, or when they contain noise? Since our anchor loss is confidence-weighted, an adversarial concern is that the model could trivially down-weight all SfM points and thereby render the supervision inert. To rule out this failure mode, we conduct two controlled-perturbation experiments on the converged SAGE (MV-DUSt3R, 10K) checkpoint.

**Density: subsampling SfM anchors.** We randomly retain a fraction of the COLMAP sparse points per training trajectory, holding all other settings fixed. Table 12 reports Chamfer Distance (↓) on the four reconstruction benchmarks under retentions of $100\%$ (default), $75\%$, $50\%$, and $25\%$.

Table 12: **Anchor density ablation: Chamfer Distance (↓) under random subsampling of SfM points.** SAGE exhibits graceful degradation: even at $25\%$ retention, performance is largely preserved on three of four datasets, indicating that the dense differentiable consistency loss provides geometric supervision largely independent of anchor density.

| Retention | ScanNet | 7Scenes | TUM | MP3D |
|---|---|---|---|---|
| 100% (default) | **1.69** | **0.76** | **1.85** | **24.9** |
| 75% | 1.93 | 1.13 | 1.89 | 29.9 |
| 50% | 1.98 | 1.31 | 1.89 | 28.5 |
| 25% | 1.79 | 1.12 | 1.94 | 26.6 |

**Quality: injecting Gaussian noise on SfM anchors.** We perturb each retained 3D anchor point by isotropic Gaussian noise with standard deviation $\sigma = r \cdot \mathrm{median_depth}$, where $r$ is the noise ratio. Table 13 reports Chamfer Distance (↓) at $r \in \{5\%, 20\%, 50\%\}$.

Table 13: **Anchor quality ablation: Chamfer Distance (↓) under additive Gaussian noise on SfM anchor points.** Performance is essentially unchanged at $\sigma$=$5\%$ and degrades only mildly at $\sigma$=$20\%$. Only an aggressive $50\%$ perturbation causes severe drops, well beyond the noise levels expected from converged COLMAP reconstructions in practice.

| Noise $\sigma$ | ScanNet | 7Scenes | TUM | MP3D |
|---|---|---|---|---|
| 5% | 1.92 | 0.96 | 1.85 | 27.2 |
| 20% | 2.23 | 1.44 | 2.84 | 45.5 |
| 50% | 3.15 | 5.03 | 7.28 | 68.3 |

**Takeaway.** Both experiments demonstrate *graceful degradation*: SAGE remains stable under aggressive sparsification (down to $25\%$ of anchors) and under realistic anchor noise ($\sigma$=$5\%$). This robustness stems from the complementary dense differentiable consistency loss (Sec. 3.2.3), which provides geometric supervision orthogonal to anchor quality. The results also empirically rule out the failure mode where the confidence-weighted loss collapses all SfM contributions to zero: if it did, anchor perturbations would have no effect on performance, but Tables 12 and 13 clearly show graded sensitivity to anchor quality.

### C.6 Novel View Synthesis Quality Comparison

We evaluate the novel view synthesis performance of our method by rendering images from the predicted 3D Gaussians. Among the test datasets, we choose ScanNet [10] for this evaluation due to its high-quality imagery. In contrast, 7Scenes [24] and TUM-RGBD [27] often suffer from motion blur, and MP3D [4] lacks sufficient novel views with appropriate overlap, making them less suitable for consistent evaluation. To ensure fairness, we allow training of the Gaussian head in our method, similar to the baseline. As shown in Table 14, our method achieves slightly better perceptual quality than MV-DUSt3R in terms of PSNR, SSIM, and LPIPS. Although novel view synthesis is not our primary target, the improved rendering results further validate the quality of the reconstructed geometry.

### C.7 The Impact of Scene Diversity vs. Sample Density

To investigate whether the generalization capability stems from the total volume of training samples or the diversity of the scenes, we conduct a controlled scaling experiment. We fix the total number of training iterations and the total sample budget to ensure a fair comparison. We then vary the number of unique training video scenes, ranging from 100 to 2,000. In this setup, as the number of scenes increases, the number of samples drawn from each individual video decreases proportionally.

Table 14: **Novel view synthesis quality on the ScanNet dataset.** Comparison between MV-DUSt3R and our method (SAGE) using three standard metrics: PSNR (↑), SSIM (↑), and LPIPS (↓). Higher PSNR/SSIM and lower LPIPS indicate better perceptual quality.

| Method | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| MV-DUSt3R [29] | 20.72 | 0.66 | 0.188 |
| SAGE (Ours) | **20.98** | **0.67** | **0.184** |

We evaluate the zero-shot generalization performance on the 7-Scenes and TUM datasets using the Chamfer Distance (CD) as the primary metric.

The results, illustrated in Figure 6, reveal a clear power-law relationship between training scene diversity and generalization error. On both the 7-Scenes and TUM datasets, we observe a consistent and significant reduction in Chamfer Distance as the number of training scenes increases. These findings underscore the necessity of leveraging large-scale, diverse internet video data. The diminishing returns of over-sampling from the same scene suggest that the model benefits more from exposure to novel geometric contexts and environmental variations, which are abundant in internet-scale datasets.

Crucially, since the total number of training iterations is kept constant, this performance gain is achieved by training on more diverse scenes with fewer samples per scene, rather than seeing more frames from a limited set of environments. This provides empirical evidence that scene-level diversity (the "breadth" of data) is a more critical factor for robust generalization than the temporal sampling density within a single video (the "depth" of data).

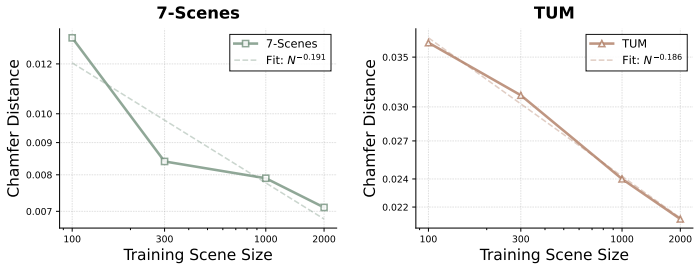



Figure 6: Generalization performance under a fixed number of training iterations, as the diversity of training scenes increases from 100 to 2K video scenes, demonstrating consistent performance gains from broader scene coverage.

## D Limitation and Impact

**Limitation and Further Discussion.** Our method builds on a pretrained sparse 3D reconstruction model for static scenes and we primarily leverage Internet video sequences that capture largely static environments. Although we filter out clips with excessive motion to ensure reliable sparse 3D supervision, this strategy cannot fully exploit the vast scale and diversity of Internet video data, where dynamic content is prevalent. Extending our framework to pretrained models capable of handling both static and dynamic scenes represents a highly promising direction, with the potential to unlock the full value of large-scale Internet videos, enhance model generalization across diverse environments, and alleviate the longstanding scarcity and limited diversity of existing 3D training datasets.

**Impact Statement** This work presents a method for adapting 3D geometric foundation models using weak supervision from internet videos. The proposed approach has the potential to significantly enhance the capabilities of 3D vision systems, enabling them to better understand and interpret com-

plex real-world environments. This advancement could lead to improved applications in various fields, including robotics, augmented reality, and autonomous driving.

# E Qualitative Results

## E.1 Qualitative 3D Reconstruction Results

Qualitative reconstruction comparisons against MV-DUSt3R and VGGT in Fig 7 show that SAGE yields more geometrically consistent reconstructions on its respective backbones.

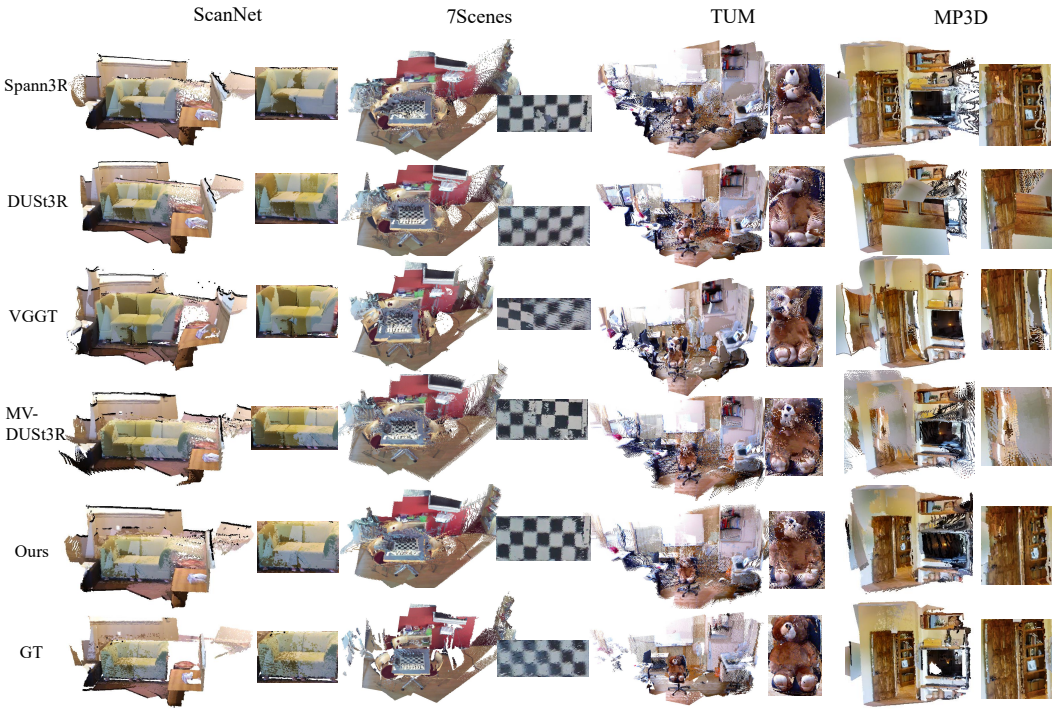



Figure 7: Qualitative comparison of reconstructed point clouds across different methods, alongside ground-truth geometry. Zoomed-in views of representative regions are shown on the right. (Note: The 7Scenes ground-truth may appear slightly misaligned due to minor pose inaccuracies in the dataset.)

## E.2 Additional Qualitative 3D Reconstruction Results

We provide additional qualitative reconstruction results in Figure 8, and further examples on more challenging scenes, low overlap or limited texture, in Figure 9.

## E.3 Qualitative Rendering Results

We visualize the rendering quality of the inferred 3D representation in Figure 10 and Figure 11. Given 8 input images as model input, the network performs a forward pass to infer a 3D Gaussian representation, which is then used to render images from both seen viewpoints and novel viewpoints. Specifically, the reconstructed representation is rendered to reproduce the original 8 input views, as well as 2 additional novel views for evaluating generalization to unseen viewpoints. Due to occlusions and limited field-of-view coverage in the input images, the novel-view renderings are naturally restricted to regions observed by the input views. Despite this limitation, the rendered images exhibit high visual fidelity, sharp geometric structures, and strong consistency with the ground-truth images, demonstrating that the learned 3D Gaussian representation effectively captures the underlying scene geometry and appearance. These qualitative results indicate that the proposed model can recon-




Figure 8: Qualitative results of reconstructed point clouds with 8 input views across diverse scenes, shown alongside ground-truth geometry. All examples are drawn from unseen scenes and demonstrate the generalization ability of the model.

struct accurate 3D scene representations from sparse multi-view inputs and generate high-quality renderings across viewpoints.

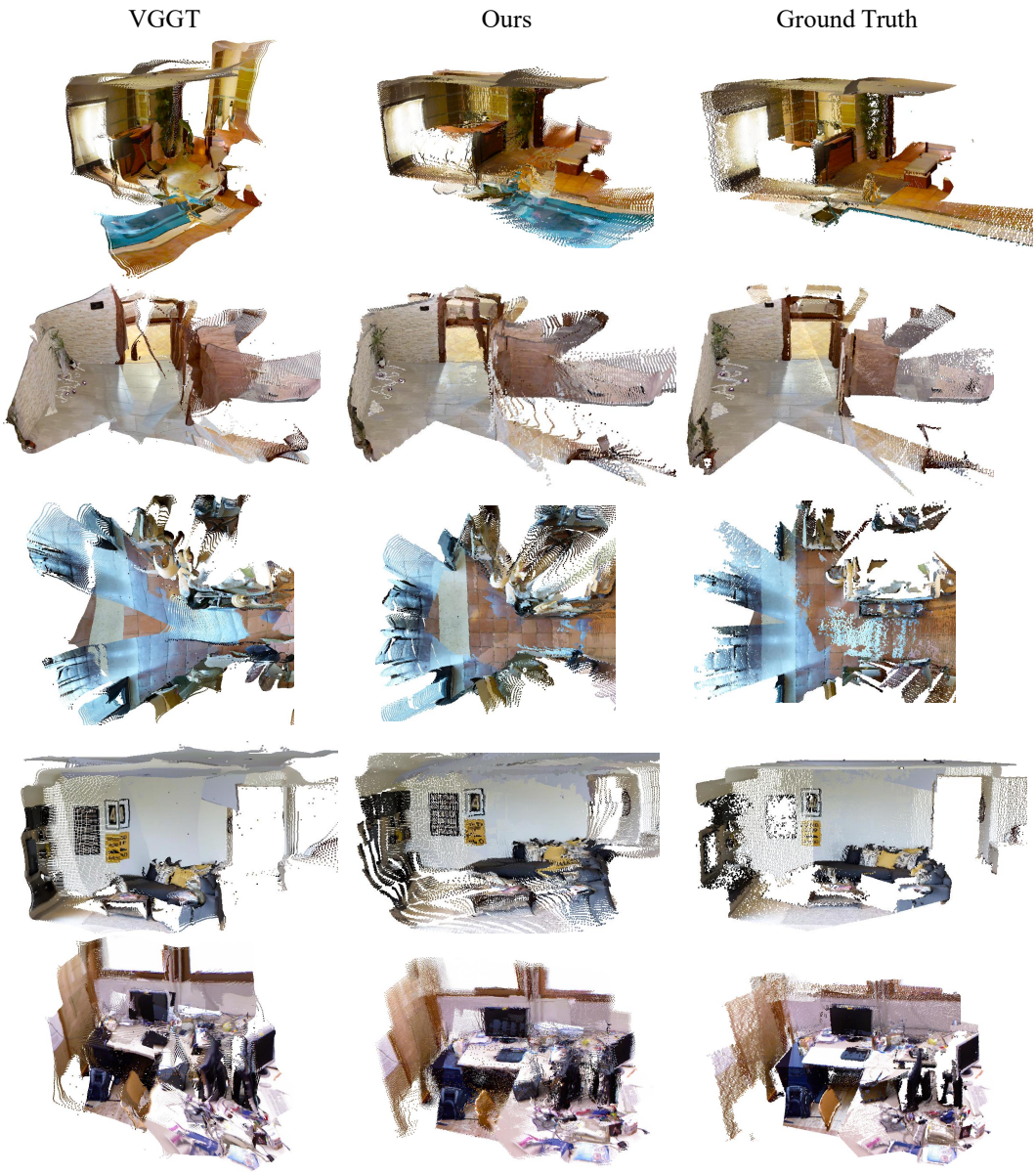



Figure 9: Qualitative comparisons results of reconstructed point clouds views on challenging scenes, shown alongside ground-truth geometry. These examples feature low image overlap, texture-less surfaces (e.g., walls or floors), or inherently difficult matching conditions, and are drawn from previously unseen test scenes.

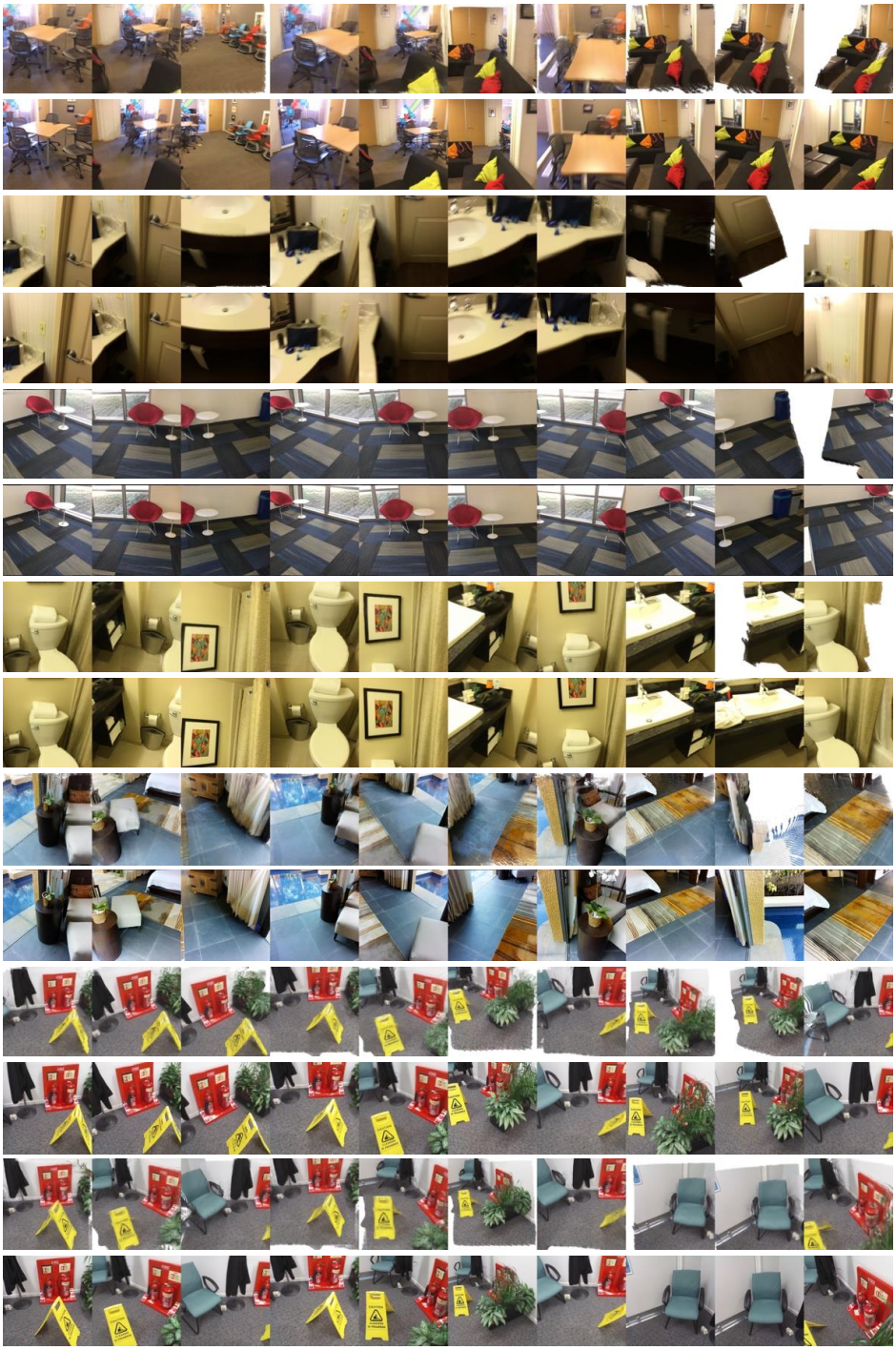


Figure 10: Qualitative visualization of rendered images and ground-truth views.

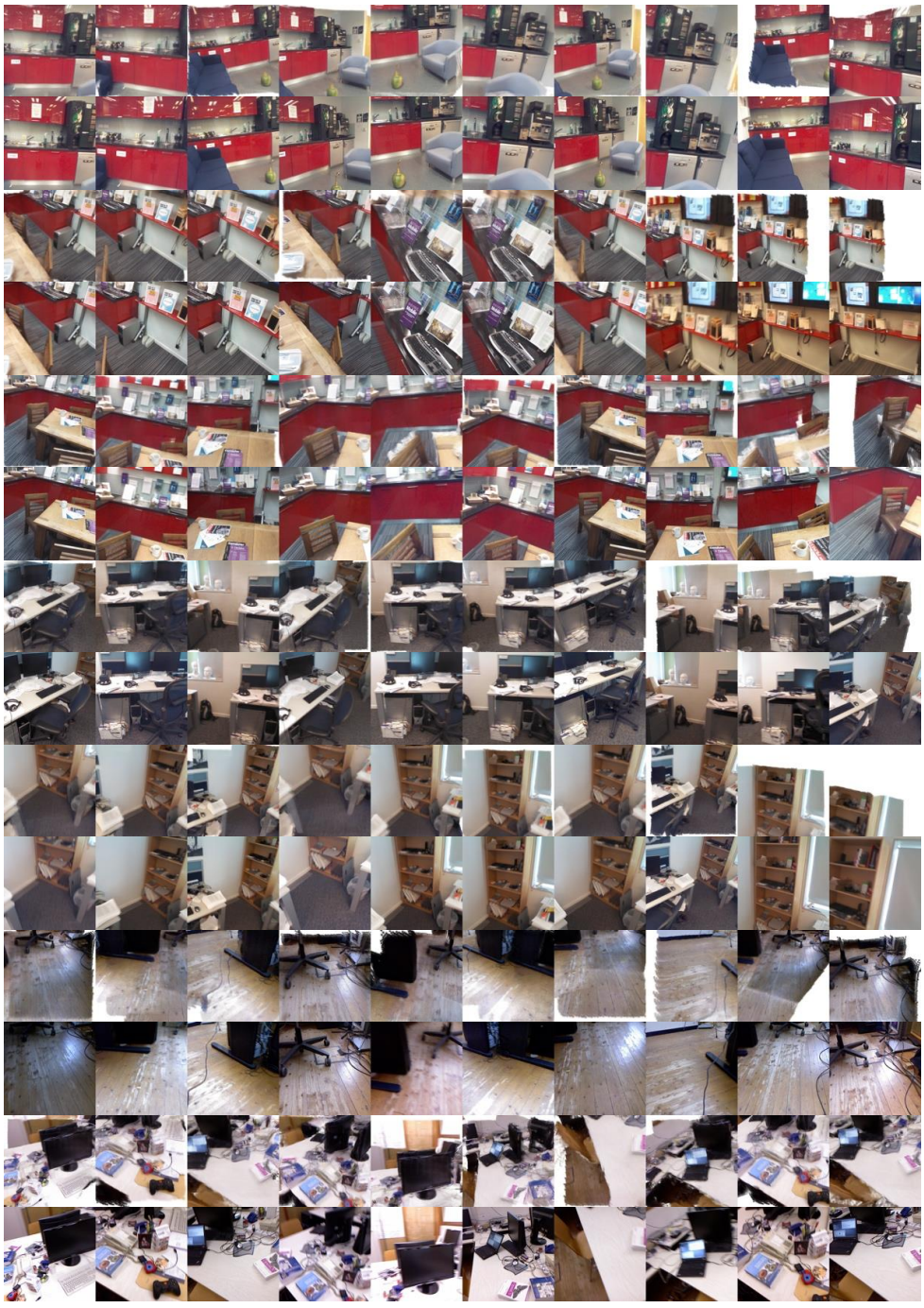


Figure 11: Qualitative visualization of rendered images and ground-truth views. 8 images are provided as model input, from which a 3D Gaussian representation is inferred and rendered to reconstruct the original 8 views as well as 2 novel viewpoints. Due to occlusions, novel-view renderings are limited to regions covered by the input views. The rendered images exhibit high visual fidelity and consistency with ground truth.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state our claims clearly in the abstract and introduction. The claims are supported by various experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in App D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not have any theoretical result

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the complete experimental setup and data in the App. B.2, ensuring the reproducibility of the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is publicly available at `https://github.com/aim-uofa/SAGE`. The training videos come from the public RealEstate10K and DL3DV datasets, and the data processing pipeline and training details are described in App. B.1 and App. B.2.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We discussed the experimental setup and hyperparameter selection in detail in the appendix to ensure the reproducibility of the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We did not report error bars or confidence intervals in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided the amount of compute required in experimental details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not foresee any such issue in our work

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed the potential social impacts in App. D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited the original papers that produced the code packages and datasets used in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the details in the App. B.1.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: This is not applicable to our research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.