R^2: RESIDUAL-GUIDED ATTENTION AND GEOMETRIC POINT REFINEMENT FOR 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in geometric reconstruction have largely followed from feed-forward approaches that jointly estimate cameras, depth, and 3D point maps from images. Such models output these variables through a loosely parallel architecture despite their strong geometric dependencies. We propose R^2, an approach based on residual-guided camera attention and camera- and depth-conditioned point refinement. Starting from a pretrained VGGT representation, a frozen teacher first produces provisional camera parameters and geometry. We then construct a multi-view rigid-consistency residual and use it as a signed geometric prior for camera-query attention. The final camera and depth predictions are combined with the initial point map in a camera- and depth-conditioned refinement stage, which we refer to as geometric point refinement. Experiments show improvements in selected camera and depth metrics, while the proposed point refinement reduces median point error by 12.57% in a matched ablation.

1 INTRODUCTION

Feed-forward reconstruction models have demonstrated that classical SfM workflows can be substituted with learned prediction. DUSt3R introduced direct point-map regression for unconstrained image pairs, MASt3R extended this representation with dense matching, and VGGT further unified camera parameters, depth maps, point maps, and point tracks within a single feed-forward Transformer Wang et al. (2024b); Leroy et al. (2024); Wang et al. (2025a). These efficient models replace much of classical geometric reasoning with learned feature interactions, leaving two relationships largely implicit. First, feature aggregation doesn't verify cross-view consistency of observations against current camera and geometry predictions. It is especially problematic in dynamic scenes when moving or deforming areas contradict the rigid-scene assumption used for camera estimation. Secondly, although being projectively dependent, camera, depth, and 3D point predictions are calculated independently and mostly in parallel using shared features without explicitly refining each other.

Recent works tackled the first problem by distinguishing camera-consistent and dynamic parts within feed-forward reconstruction models. MonST3R adapts DUSt3R to dynamic video through fine-tuning on dynamic data Zhang et al. (2025a). Easi3R observes that DUSt3R attention already contains information about camera and object motion and uses these attention statistics to identify dynamic regions and modify attention at inference time Chen et al. (2025). PAGE-4D extends VGGT with a dynamics-aware aggregator that predicts a learned mask and uses it to reduce the influence of dynamic content on pose estimation while preserving dynamic information for scene reconstruction Zhou et al. (2026). VGGT4D extracts motion cues from VGGT's attention representations for dynamic-scene reconstruction Hu et al. (2025). These methods show that controlling how non-rigid content participates in cross-view reasoning is important for robust reconstruction.

We consider an additional source of camera support based on geometry. With some preliminary estimates of camera parameters, depth, and cross-view feature correspondences, an observation in one view can be matched against the prediction in another view with respect to the existing rigid scene hypothesis. The distance between the actual correspondence and the rigid projection under the current hypothesis is a clear indication of cross-view consistency in terms of geometry. Contrary to the dynamic mask of a semantic kind, this indicator does not imply that a region belongs to any particular object category or that it moves.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Based on this idea, we introduce a *cross-view rigid-consistency residual* to guide camera-specific attention in VGGT. A frozen teacher pass first produces provisional camera parameters, depth, and patch features, from which cross-view feature correspondences are compared with rigid reprojections to construct a signed patch-level camera-support prior. In a second VGGT pass, this prior guides only camera-token queries attending to patch-token keys in global-attention blocks, while frame attention and non-camera query rows remain unchanged. We additionally introduce lightweight camera-query adapters and layer-specific guidance strengths to learn how the camera pathway uses this geometric prior. Global blocks 8–12 are fine-tuned with a smaller learning rate, while DINO, frame-attention blocks, the remaining global blocks, and the original VGGT prediction heads remain frozen.

We address the second issue through *camera- and depth-conditioned point refinement*. Rather than replacing VGGT's parallel prediction heads with a hard geometric cascade, we retain their provisional outputs: camera parameters, depth, and an initial world-space point map P_1. These predictions are then used jointly to refine the 3D point map. This preserves the flexibility of VGGT's original prediction structure while introducing explicit conditioning at the final point-refinement stage. We refer to this camera- and depth-conditioned stage as geometric point refinement.

The two modules operate on different aspects of the reconstruction task. Residual-based camera attention identifies *which* cross-view observations are most relevant to camera estimation, whereas camera- and depth-conditioned point refinement addresses *how* to integrate the resulting geometric estimates for better reconstruction.

Our main contributions are:

1. We introduce a *cross-view rigid-consistency residual* that compares observed feature correspondences with rigid reprojections induced by provisional VGGT camera and depth predictions, providing an explicit geometry-derived measure of camera-supporting evidence.
2. We introduce a residual-guided camera-attention mechanism that injects this geometric prior specifically into camera-token-to-patch-token interactions in VGGT global attention, together with lightweight camera-query adapters.
3. We present camera- and depth-conditioned point refinement, which uses predicted cameras, depth, and initial VGGT points to refine the final 3D geometry. In a matched ablation, this component substantially improves point reconstruction while producing only small changes in camera and depth metrics.
4. We evaluate the individual and combined components on dynamic-scene camera-pose, depth, and point-reconstruction benchmarks, including ablations of camera guidance and point refinement.

2 Related Work

2.1 Feed-Forward and Stateful 3D Reconstruction

Cross-view pretraining and dense point-map prediction have advanced learned reconstruction, with CroCo v2 learning transferable geometric features, DUSt3R regressing pairwise geometry, and MASt3R adding local descriptors for matching and localization Weinzaepfel et al. (2023); Wang et al. (2024b); Leroy et al. (2024). VGGT and Fast3R extend feed-forward reconstruction to multiple views, with VGGT jointly predicting cameras, depth, point maps, and tracks Wang et al. (2025a); Yang et al. (2025). Subsequent designs explore spatial memory, recurrent state, camera-conditioned prediction, and scene-token aggregation Wang & Agapito (2025); Wang et al. (2025b); Zhang et al. (2025b); Wang et al. (2026).

R^2 retains VGGT's parallel prediction heads and subsequently refines their point-map output using provisional camera and depth estimates, differing from FLARE's camera-conditioned prediction cascade. Analysis of emergent outlier-view rejection further suggests that pretrained VGGT representations encode input reliability Han et al. (2026); our guidance instead operates on patch evidence for camera estimation.

2.2 Dynamic-Scene Geometry and Motion Disentanglement

Independent scene motion violates the rigid-scene assumption, complicating camera estimation. Dynamic-aware SLAM and odometry address this through moving-object detection, trajectory-based motion segmentation, or long-term tracking with uncertainty estimation Bescos et al. (2018); Zhao et al. (2022); Chen et al. (2024). MonST3R, Align3R, and Easi3R adapt learned reconstruction to dynamic video through dynamic-data fine-tuning, depth-assisted optimization, and inference-time attention modification, respectively Zhang et al. (2025a); Lu et al. (2025a); Chen et al. (2025). Video diffusion priors support temporally coherent reconstruction Jiang et al. (2025); Xu et al. (2025), while GeoMotion segments motion using latent geometric representations He et al. (2026).

Among VGGT-based methods, PAGE-4D suppresses dynamic information for camera estimation while emphasizing it for geometry, whereas VGGT4D extracts and refines attention-derived dynamic masks without fine-tuning Zhou et al. (2026); Hu et al. (2025). R^2 instead compares feature correspondences with rigid reprojections from provisional camera and depth estimates, using their discrepancy as a soft prior for camera-token-to-patch-token attention. This measures geometric agreement rather than semantic motion: occlusion, correspondence errors, and inaccurate provisional geometry can also produce large residuals.

Forests combine moving foliage, stable trunks, thin structures, and occlusion. Applications in crop reconstruction, plant phenotyping, and forest fuel estimation highlight the utility of image-based vegetation modeling Zhai et al. (2026); Jia et al. (2026); Wu et al. (2026). VidFin further couples UAV trajectory and tree-distance estimation for monocular video-based forest inventory Lu et al. (2025b). These applications motivate investigating motion-aware camera estimation and coherent reconstruction in leaf-on forests.

2.3 Geometric Reasoning and Reconstruction Refinement

Classical reconstruction relies on explicit geometric estimation and bundle adjustment Schönberger & Frahm (2016); Triggs et al. (2000). Learned multi-view stereo incorporates camera geometry through differentiable warping, learned PatchMatch, and attention-based aggregation Yao et al. (2018); Wang et al. (2021); Cao et al. (2024). Joint camera and scene estimation is also central to differentiable geometric networks and learned SLAM systems Tang & Tan (2019); Teed & Deng (2020; 2021); Czarnowski et al. (2020). Related systems couple learned correspondences with geometric optimization through sparse patch tracking, loop closure, differentiable SfM, or dynamic-video reconstruction Teed et al. (2023); Lipson et al. (2024); Wang et al. (2024a); Li et al. (2025).

MASt3R-SLAM and GLUEMAP integrate learned reconstruction priors into tracking, mapping, and global optimization pipelines Murai et al. (2025); Pan et al. (2026). R^2 uses explicit geometry within intermediate camera attention and subsequently predicts a bounded point-map correction conditioned on provisional camera, depth, and point outputs, without introducing an iterative bundle-adjustment solver. Its distinction lies in where and how geometric evidence guides reconstruction.

3 Method

3.1 Overview

Our method augments VGGT with two complementary geometric mechanisms: (i) a cross-view rigid-consistency prior that guides camera-token attention, and (ii) a camera- and depth-conditioned point refinement module that improves the final 3D point prediction.

As shown in Fig. 1, a frozen VGGT teacher first produces camera parameters, depth, and patch features. These predictions are used to compute a cross-view rigid-consistency residual, which is converted into a patch-level camera-support prior. A student VGGT then processes the same input. The prior modifies only camera-token queries attending to patch-token keys in selected global-attention blocks. Based on the placement ablation in Sec. A.4, we apply guidance to global blocks 8–12.

The student produces provisional camera, depth, and point predictions through the original VGGT heads. We subsequently refine the initial point map using the predicted camera and depth, introducing an explicit dependency among the otherwise parallel geometric outputs.

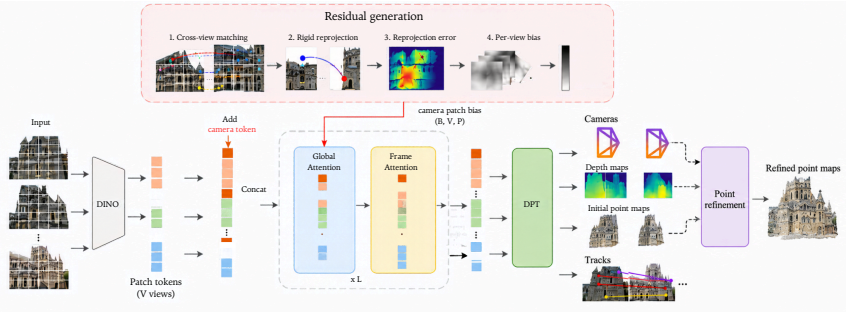

Figure 1: Overview of our geometry-guided VGGT. A frozen teacher produces a cross-view rigid-consistency prior that guides camera-query attention in student global blocks 8–12. The signed prior denotes rigid-consistent and rigid-inconsistent evidence, not semantic static and dynamic classes. The resulting provisional camera, depth, and point predictions are then jointly used to refine the final point map.

3.2 Cross-View Rigid-Consistency Residual

Let $p_{s,i}$ denote patch i in source view s, and let $m_{s\to t,i}$ denote its observed correspondence in target view t, obtained from frozen VGGT patch features. Using the teacher-predicted depth D_s, intrinsics K_s, and extrinsics E_s, we first unproject the source patch into 3D,

$$X_{s,i} = \pi^{-1}\left(p_{s,i}, D_s(p_{s,i}); K_s, E_s\right). \tag{1}$$

Under a rigid-scene hypothesis, its predicted target location is

$$\hat{p}_{s\to t,i} = \pi\left(X_{s,i}; K_t, E_t\right). \tag{2}$$

We define the rigid-consistency residual as

$$r_{s\to t,i} = \left\| m_{s\to t,i} - \hat{p}_{s\to t,i} \right\|_2. \tag{3}$$

Correspondences are filtered using feature similarity, cycle consistency, visibility, positive depth, and image bounds. Valid residuals across target views are aggregated by their median and normalized by patch scale and correspondence uncertainty,

$$z_{s,i} = \frac{r_{s,i}}{\sqrt{d_{\text{patch}}^2 + u_{s,i}^2}}. \tag{4}$$

Because residual magnitudes vary across scenes, we rank $z_{s,i}$ within each view and denote its percentile by $q_{s,i} \in [0,1]$. We convert this rank into a signed camera-support prior,

$$b(q) = \begin{cases} 1 - \dfrac{q}{0.75}, & q \le 0.75, \\ -\dfrac{q - 0.75}{0.25}, & q > 0.75. \end{cases} \tag{5}$$

Thus, relatively rigid-consistent patches receive positive values, strongly inconsistent patches receive negative values, and unsupported patches receive zero. The resulting prior $B \in \mathbb{R}^{V\times P}$ is used only for camera attention.

3.3 Residual-Guided Camera Attention

VGGT alternates frame attention within each view and global attention across views. Since our residual is defined through cross-view geometry, we inject it only into global attention and modify only camera-token query rows.

For guided global layer ℓ, let Q^{ℓ}_{cam} be the original camera query and X^{ℓ}_{cam} its input representation. We introduce a lightweight bottleneck adapter,

$$\tilde{Q}^{\ell}_{\mathrm{cam}} = Q^{\ell}_{\mathrm{cam}} + A^{\ell}_{Q}(X^{\ell}_{\mathrm{cam}}), \tag{6}$$

where A^{ℓ}_{Q} is a $1024 \to 64 \to 1024$ MLP. The geometry-guided camera logits are

$$\tilde{L}^{\ell}_{\mathrm{cam},j} = \frac{\tilde{Q}^{\ell}_{\mathrm{cam}}(K^{\ell}_{j})^{\top}}{\sqrt{d}} + B_j, \qquad j \in \mathcal{P}, \tag{7}$$

where $\mathcal{P}$ contains patch-token keys. Camera and register keys receive no geometric bias.

Let O^{ℓ}_{geo} be the resulting camera-attention output and $O^{\ell}_{\mathrm{vanilla}}$ the original VGGT output. We combine them through a learned bounded strength,

$$O^{\ell}_{\mathrm{cam}} = O^{\ell}_{\mathrm{vanilla}} + \alpha_{\ell}\left(O^{\ell}_{\mathrm{geo}} - O^{\ell}_{\mathrm{vanilla}}\right), \qquad \alpha_{\ell} = \tanh(\theta_{\ell}). \tag{8}$$

The parameters θ_{ℓ} are initialized to zero, so the modification initially reproduces vanilla VGGT. Patch-query, register-query, and frame attention remain unchanged.

We apply this mechanism to global blocks 8–12, as determined by the layer-placement ablation. These blocks are fine-tuned at a lower learning rate together with the newly introduced Q adapters and guidance parameters. DINO, all frame-attention blocks, global blocks outside 8–12, and the original prediction heads remain frozen.

3.4 Camera- and Depth-Conditioned Point Refinement

VGGT produces provisional camera parameters C, depth D, and a world-space point map P_1 through parallel prediction heads. Rather than replacing this parallel prediction structure, we introduce explicit geometric conditioning only at the final point-refinement stage.

We construct a patch-resolution conditioning tensor

$$H_{\mathrm{ref}} = \left[\Phi_{\mathrm{cam}}(C), \log D, P_1\right], \tag{9}$$

where $\Phi_{\mathrm{cam}}(C)$ is a 16-channel spatial encoding of the predicted camera parameters. Together with log depth and the three coordinates of P_1, this gives a 20-channel refinement input.

A lightweight convolutional network predicts a residual correction,

$$\Delta P = f_{\mathrm{ref}}(H_{\mathrm{ref}}), \tag{10}$$

using two 3×3 convolutional layers with 64 channels followed by a 1×1 output projection. The final point prediction is

$$P_{\mathrm{final}} = P_1 + 0.1 \tanh(\Delta P). \tag{11}$$

This bounded residual formulation preserves the pretrained point estimate while allowing camera and depth information to explicitly refine the final 3D geometry.

3.5 Training Objective

We jointly supervise camera pose, depth, and the refined point map,

$$\mathcal{L} = \mathcal{L}_{\mathrm{cam}} + \mathcal{L}_{\mathrm{depth}} + \mathcal{L}_{\mathrm{point}}(P_{\mathrm{final}}, P^{\star}). \tag{12}$$

The trainable parameters are the original global blocks 8–12, their camera-query adapters and guidance parameters, and the point-refinement network. We use a lower learning rate for the pretrained global blocks than for the newly introduced modules.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

All variants of our method are initialized from pretrained VGGT and trained using the same multi-dataset mixture, consisting of PointOdyssey, DynamicReplica, Kubric MOVi-E, Spring, CO3D, and Waymo. Unless otherwise specified, training uses ten input views resized to a width of 518 pixels. All architectural comparisons use identical data sampling, preprocessing, and loss configurations.

We initialize from the pretrained VGGT checkpoint and train the final model for 18,600 optimization steps. We keep most of the pretrained backbone frozen and fine-tune global-attention blocks 8–12 together with the newly introduced modules, while the remaining pretrained VGGT parameters are kept fixed.

4.2 EVALUATION PROTOCOL

Our primary evaluation uses ten endpoint-sampled frames per sequence. We evaluate up to ten sequences per dataset; TUM contains eight evaluated sequences. For camera estimation, we report Absolute Trajectory Error (ATE), relative translation and rotation errors (RPE-T and RPE-R), pairwise rotation error, pairwise translation-direction error, angular AUC and AUC@30. Complete 3°, 5°, 10°, 15°, and 30° breakdown is provided in Appendix A.2.

Depth is evaluated using the released PAGE4D evaluation code. Following their protocol, predicted depth is aligned to ground truth using a single median-based scale factor computed separately for each sequence before calculating the depth metrics. We report Abs Rel, RMSE, and $\delta < 1.25$ in the results section. Complete depth metrics are provided in Appendix A.3.

4.3 ABLATIONS

VGGT observes that 3D points obtained by unprojecting its predicted depth using the predicted camera parameters can be more accurate than the points predicted directly by its point-map head Wang et al. (2025a).

Motivated by this observation, we evaluate whether camera and depth predictions can also be used to refine the directly predicted point map. We further study the placement of geometric guidance within the pretrained VGGT global-attention stack and separate the contribution of the geometric residual from that of the additional trainable parameters. Additional analyses are provided in Appendix A.

5 RESULTS

5.1 DYNAMIC-SCENE CAMERA ESTIMATION

Table 1 summarizes the matched camera evaluation. The complete angular AUC breakdown is provided in Appendix A.2.

On Sintel, our model improves ATE relative to VGGT, but has slightly higher RPE-T and RPE-R and lower angular accuracy. PAGE4D achieves a lower ATE, whereas VGGT-Omega provides the strongest overall camera performance. On TUM, our model improves ATE, RPE-R, and pairwise rotation error over VGGT, achieving the lowest RPE-R and rotation error among the matched methods, but its RPE-T is slightly higher. VGGT-Omega obtains the best ATE, RPE-T, translation-direction error, and AUC@30. On Bonn, our model achieves the lowest RPE-T, PAGE4D performs best in RPE-R and rotation error, and VGGT-Omega achieves the best ATE, translation-direction error, and AUC@30. On DyCheck, VGGT-Omega is strongest overall and VGGT also outperforms our model, although our method improves over PAGE4D in ATE, RPE-T, RPE-R, translation-direction error, and AUC@30. Overall, the proposed camera refinement improves selected aspects of the pretrained VGGT solution, with the clearest gains observed on TUM.

Table 1: Matched camera evaluation. All methods use ten endpoint-sampled frames per sequence. Lower is better for error metrics and higher is better for AUC. Bold indicates the best result among the four matched models.

Dataset	Method	ATE ↓	RPE-T ↓	RPE-R ↓	Rot. ↓	Trans. ↓	AUC@30 ↑
Sintel	VGGT	0.200536	0.135324	1.322223	3.235	17.943	0.6011
Sintel	PAGE4D	0.168138	0.164221	1.607194	3.741	22.605	0.5144
Sintel	VGGT-Omega	**0.080760**	**0.079508**	**0.780726**	**2.058**	**10.067**	**0.7514**
Sintel	Ours	0.188077	0.137474	1.392261	3.244	18.589	0.5640
TUM	VGGT	0.010892	0.015797	0.461763	0.541	7.438	0.7883
TUM	PAGE4D	0.016190	0.022170	0.635481	0.775	14.064	0.6001
TUM	VGGT-Omega	**0.009690**	**0.014111**	0.436785	0.542	**6.831**	**0.8150**
TUM	Ours	0.010705	0.016129	**0.425038**	**0.470**	7.838	0.7806
Bonn	VGGT	0.020677	0.585083	34.981841	41.214	46.404	0.0362
Bonn	PAGE4D	0.038977	0.584361	**34.591028**	**40.720**	46.636	0.0317
Bonn	VGGT-Omega	**0.012773**	0.585643	34.969557	41.186	**45.937**	**0.0367**
Bonn	Ours	0.033702	**0.581412**	35.157318	41.403	46.256	0.0364
DyCheck	VGGT	0.012745	0.019087	1.068782	1.213	6.560	0.8184
DyCheck	PAGE4D	0.016324	0.024192	1.558721	1.890	9.190	0.7321
DyCheck	VGGT-Omega	**0.004917**	**0.007722**	**0.439855**	**0.538**	**2.462**	**0.9340**
DyCheck	Ours	0.014923	0.019932	1.546136	1.898	7.645	0.7756

Table 2: Matched depth evaluation. One median scale is estimated independently for each sequence. Lower is better for Abs Rel and RMSE and higher is better for $\delta < 1.25$. Bold indicates the best result among the four matched models.

Dataset	Method	Abs Rel ↓	RMSE ↓	$\delta < 1.25$ ↑
Sintel	VGGT	0.303145	6.832123	0.574288
Sintel	PAGE4D	**0.266954**	**4.287417**	**0.681670**
Sintel	VGGT-Omega	0.287431	5.257597	0.603519
Sintel	Ours	0.274007	5.922952	0.591045
TUM	VGGT	0.090255	0.643425	0.936373
TUM	PAGE4D	**0.046739**	**0.424289**	**0.971684**
TUM	VGGT-Omega	0.159915	1.005817	0.925290
TUM	Ours	0.076969	0.680385	0.930448
Bonn	VGGT	0.037611	0.172370	0.983350
Bonn	PAGE4D	**0.029027**	**0.125713**	**0.991630**
Bonn	VGGT-Omega	0.053029	0.205739	0.971284
Bonn	Ours	0.036982	0.146042	0.989262
DyCheck	VGGT	0.095184	0.332549	0.880865
DyCheck	PAGE4D	0.077773	0.288543	0.924029
DyCheck	VGGT-Omega	**0.071551**	**0.254568**	**0.931606**
DyCheck	Ours	0.100759	0.366818	0.882383

5.2 Depth Estimation

Table 2 summarizes the corresponding depth evaluation. The complete metric breakdown is reported in Appendix A.3.

On Sintel, our model improves all three summarized depth metrics relative to VGGT, reducing Abs Rel from 0.3031 to 0.2740 and RMSE from 6.8321 to 5.9230 while increasing $\delta < 1.25$ from 0.5743 to 0.5910. On TUM, it improves Abs Rel from 0.0903 to 0.0770, although RMSE is slightly higher and $\delta < 1.25$ slightly lower. On Bonn, it improves Abs Rel from 0.0376 to 0.0370, reduces RMSE from 0.1724 to 0.1460, and increases $\delta < 1.25$ from 0.9834 to 0.9893. On DyCheck, our model slightly improves $\delta < 1.25$ from 0.8809 to 0.8824 but is weaker in Abs Rel and RMSE, while PAGE4D and VGGT-Omega perform better overall. These results show that the proposed architecture improves selected depth metrics over VGGT on several datasets.

Table 3: Ablation of camera- and depth-conditioned point refinement at matched 20K training step.

Metric	No refinement	Refinement	Improvement
Point median ↓	0.19600	**0.17136**	12.57%
Point mean ↓	0.19729	**0.17467**	11.47%
Point P90 ↓	0.29793	**0.26818**	9.99%
Depth Abs Rel ↓	0.06610	**0.06604**	0.09%
Rotation error ↓	15.841	**15.769**	0.45%
ATE ↓	0.13698	**0.13609**	0.65%

5.3 Camera- and Depth-Conditioned Point Refinement

We isolate the contribution of the proposed point refinement by comparing the camera- and depth-conditioned point refinement module with a variant that predicts the point map without this refinement stage under identical training settings. Unlike the full-model training described in Section 4.1, this ablation keeps global blocks 8–12 and all other backbone parameters frozen; the prediction heads and refinement module are initialized from scratch.

At the matched 20K step, conditioning the point refinement on camera and depth reduces median point error from 0.19600 to 0.17136, corresponding to a 12.57% reduction. Mean point error decreases by 11.47% and P90 error by 9.99%. In contrast, camera and depth quantities change modestly: depth Abs Rel improves by 0.09%, rotation error by 0.45%, and ATE by 0.65%.

The larger improvement in point reconstruction suggests that the camera and depth predictions provide complementary geometric information that is useful for refining the final 3D points. This indicates that explicitly conditioning point refinement on these intermediate geometric estimates can recover information that is not fully captured by the direct point prediction alone.

5.4 Qualitative Results

Comparison in a leaf-on forest. Figure 2 compares R^2, VGGT, PAGE-4D, and VGGT-Ω using the same ten frames uniformly sampled from the first 30 seconds of a forest video. The scene combines moving foliage, stable trunks, thin structures, and substantial occlusion. The bottom row shows depth predictions for the frame at $t = 16.6\,\mathrm{s}$.

All methods delineate the foreground trunk and ground. VGGT produces smoother vegetation depth transitions, while PAGE-4D resolves smaller structures clearly. R^2 shows sharper boundaries around several slender trunks and vegetation, closely resembling VGGT-Ω in these areas.

R^2 and VGGT-Ω display more extended point-cloud structures than VGGT and PAGE-4D, although R^2 retains scattered points and elongated peripheral geometry. These observations suggest clearer depth delineation, but do not establish greater reconstruction completeness or accuracy.

6 Discussion

From our findings, we conclude that explicitly using geometric conditioning can be an effective addition to the learned representations from feed-forward reconstruction models, but its effectiveness depends on the part of the reconstruction pipeline that is being improved. The cross-view rigid-consistency prior is a geometry-based input for refining camera estimates through comparison of observed cross-view correspondences with rigid reprojections induced by the provisional camera and depth estimates. In contrast to other methods based mainly on learned dynamic masks and motion, the residual represents the disagreement between the current rigid reconstruction hypothesis and the observation.

Camera results demonstrate the improvement of selected metrics using this intervention with respect to the pretrained VGGT representation, especially on the TUM dataset. Still, the improvements are not consistent across datasets, as on the DyCheck dataset, the pretrained VGGT proves to be better in several metrics, whereas VGGT-Ω and VGGT4D give better camera estimates overall. From that, we conclude that geometry-guided attention is not always beneficial for the camera prediction task.

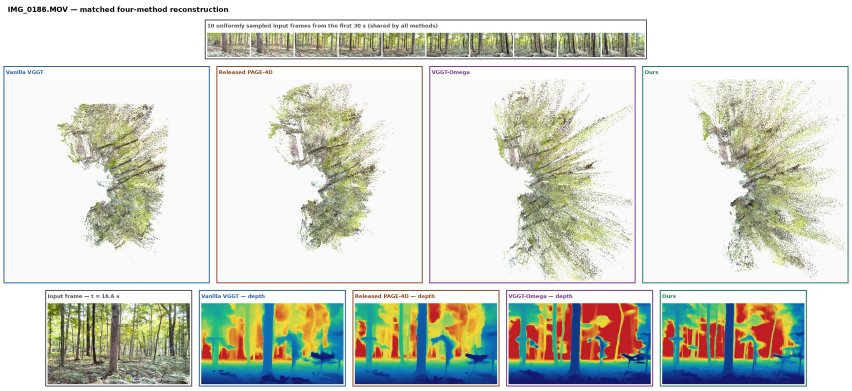

Figure 2: **Qualitative comparison in a leaf-on forest.** All methods share ten frames uniformly sampled from the video's first 30 seconds (top). Middle: point clouds from VGGT, PAGE-4D, VGGT-Ω, and R^2 (step 15,200), left to right. Bottom: input at $t = 16.6$ s and corresponding depth maps in the same order. R^2 shows clearer boundaries around several slender trunks than VGGT and PAGE-4D, but retains dispersed peripheral points.

Errors in depth, correspondence estimation, occlusion handling, or the initial camera estimate can all affect the residual and therefore the guidance signal.

Point refinement experiments demonstrate more consistent improvements. Conditioning the point refinement on camera and depth improves median point error by 12.57% at the matched 20K training step, whereas changes in camera and depth metrics are relatively small. This demonstrates that the predictions of camera and depth carry complementary information useful for refining the direct point-map prediction even when the intermediate predictions are not improving substantially. Instead of cascading VGGT's independent prediction heads into a strict geometric cascade, the proposed approach preserves their parallel intermediate predictions and introduces explicit conditioning only at the final point-refinement stage.

There are several remaining limitations. Firstly, the rigid-consistency residual does not serve as a motion segmentation signal: a large residual may also be produced due to occlusions, incorrect correspondences or wrong provisional geometry. Secondly, the camera gains still have a data-dependent nature, which implies that the current residual construction and attention modulation do not solve the problem of dynamic scene camera estimation. Finally, the point-refinement network adds only a bounded local correction to the initial point map, but does not enforce global multi-view consistency nor geometric optimization. In addition, the current evaluation focuses primarily on short sampled sequences. Extending the approach to longer videos and stronger temporal reasoning may provide more stable geometric evidence and improve reconstruction in highly dynamic scenes.

7 Conclusion

We introduced R^2, a geometry-guided extension of VGGT that incorporates explicit cross-view consistency into camera attention and uses predicted camera and depth information to refine the final 3D point map. The cross-view rigid-consistency residual acts as a patch-level camera-support prior and is selectively incorporated into camera-query attention, while a simple camera- and depth-conditioned point refinement module directly improves the point prediction. Evaluating our system on different datasets, we show that geometric conditioning may improve selected camera and depth metrics and that point refinement substantially reduces point reconstruction error. These results support a broader view that outputs of feed-forward reconstruction models can serve not only as separate predictions, but also as mutually informative signals.

References

Berta Bescos, José M. Fácil, Javier Civera, and José Neira. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters*, 3(4):4076–4083, 2018.

doi: 10.1109/LRA.2018.2860039.

Chenjie Cao, Xinlin Ren, and Yanwei Fu. MVSFormer++: Revealing the devil in transformer's details for multi-view stereo. In *International Conference on Learning Representations (ICLR)*, 2024.

Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. LEAP-VO: Long-term effective any point tracking for visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19844–19853, 2024.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3R: Estimating disentangled motion from DUSt3R without training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9158–9168, 2025.

Jan Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J. Davison. DeepFactors: Real-time probabilistic dense monocular slam. *IEEE Robotics and Automation Letters*, 5(2):721–728, 2020. doi: 10.1109/LRA.2020.2965415.

Jisang Han, Sunghwan Hong, Jaewoo Jung, Wooseok Jang, Honggyu An, Qianqian Wang, Seungryong Kim, and Chen Feng. Emergent outlier view rejection in visual geometry grounded transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–437, 2026.

Xiankang He, Peile Lin, Ying Cui, Dongyan Guo, Chunhua Shen, and Xiaoqin Zhang. GeoMotion: Rethinking motion segmentation via latent 4D geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yu Hu, Chong Cheng, Sicheng Yu, Xiaoyang Guo, and Hao Wang. VGGT4D: Mining motion cues in visual geometry transformers for 4d scene reconstruction, 2025.

Hanyue Jia, Wei Zhou, Wenbo Zhou, Yanan Li, Hao Lu, and Tingting Wu. The turning point of 3D plant phenotyping: 3D foundation models enable minute-to-second cross-crop reconstruction and beyond. *arXiv preprint arXiv:2607.01753*, 2026.

Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4D: Leveraging video generators for geometric 4D scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20658–20671, 2025.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with MASt3R. In *European Conference on Computer Vision (ECCV)*, pp. 71–91, 2024. doi: 10.1007/978-3-031-73220-1_5.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10486–10496, 2025.

Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual SLAM. In *European Conference on Computer Vision (ECCV)*, pp. 424–440, 2024. doi: 10.1007/978-3-031-72627-9_24.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3R: Aligned monocular depth estimation for dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22820–22830, 2025a.

Yawen Lu, Jui-Cheng Chiu, Nabin Khanal, Sizheng Cailean Chen, Qianyu Guo, Dongfang Liu, Songlin Fei, and Yingjie Victor Chen. VidFin: A video-based monocular UAV pipeline for automatic forest inventory in diverse scenes. *Smart Agricultural Technology*, 12:101504, 2025b. doi: 10.1016/j.atech.2025.101504.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16695–16705, 2025.

Linfei Pan, Johannes Schönberger, and Marc Pollefeys. Global structure-from-motion meets feed-forward reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016.

Chengzhou Tang and Ping Tan. BA-Net: Dense bundle adjustment networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Zachary Teed and Jia Deng. DeepV2D: Video to depth with differentiable structure from motion. In *International Conference on Learning Representations (ICLR)*, 2020.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and rgb-d cameras. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16558–16569, 2021.

Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39033–39051, 2023.

Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pp. 298–372. Springer, 2000. doi: 10.1007/3-540-44480-7_21.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. PatchmatchNet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14194–14203, 2021.

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

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21686–21697, 2024a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024b.

Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jerome Revaud. 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 (ICCV)*, pp. 17969–17980, 2023.

Quanyun Wu, Kyle Gao, Wentao Sun, Zhengsen Xu, Hudson Sun, Linlin Xu, Yuhao Chen, David A. Clausi, and Jonathan Li. Rapid forest fuel load estimation via virtual remote sensing and metric-scale feed-forward 3D reconstruction. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2026.

Tian-Xing Xu, Xiangjun Gao, Wenbo Hu, Xiaoyu Li, Song-Hai Zhang, and Ying Shan. GeometryCrafter: Consistent geometry estimation for open-world videos with diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6632–6644, 2025.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935, 2025.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, pp. 767–783, 2018.

Albert J. Zhai, Xinlei Wang, Kaiyuan Li, Zhao Jiang, Junxiong Zhou, Sheng Wang, Zhenong Jin, Kaiyu Guan, and Shenlong Wang. CropCraft: Complete structural characterization of crop plants from images. In *International Conference on 3D Vision*, 2026.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations (ICLR)*, 2025a.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. FLARE: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21936–21947, 2025b.

Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. ParticleSfM: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European Conference on Computer Vision (ECCV)*, 2022.

Kaichen Zhou, Yuhan Wang, Grace Chen, Xinhai Chang, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. PAGE-4D: Disentangled pose and geometry estimation for 4D perception. In *International Conference on Learning Representations (ICLR)*, 2026.

AI Use Statement

Generative AI tools were used to assist with writing refinement and code development. All AI-assisted content was reviewed and verified by the authors.

A Appendix

A.1 Training Details

We initialize R^2 from the pretrained VGGT checkpoint and train the final model for 18,600 optimization steps using an effective batch size of 64 on 8 H100 GPUs and a weight decay of 0.01. Training uses a mixture of PointOdyssey, DynamicReplica, Kubric MOVi-E, Spring, CO3D, and Waymo. Unless otherwise specified, each sample contains ten input views resized to a width of 518 pixels, and all architectural comparisons use the same data sampling, preprocessing, and loss configuration.

Most pretrained VGGT parameters remain frozen. We fine-tune global-attention blocks 8–12 together with the camera-query adapters, layer-specific residual-guidance parameters, and the point-refinement network. The DINO backbone, all frame-attention blocks, global-attention blocks outside 8–12, and the original VGGT prediction heads remain fixed. The pretrained global-attention blocks use a smaller learning rate than the newly introduced modules.

Table 4: Complete camera evaluation using ten endpoint-sampled frames per sequence. *Seq* denotes the number of successfully evaluated sequences. Lower is better for error metrics and higher is better for AUC. Bold indicates the best result on each dataset.

Dataset	Method	Seq	ATE ↓	RPE-T ↓	RPE-R ↓	Rot. ↓	Trans. ↓	AUC@3 ↑	AUC@5 ↑	AUC@10 ↑	AUC@15 ↑	AUC@30 ↑
Sintel	VGGT	10	0.200536	0.135324	1.322223	3.235	17.943	0.1548	0.2298	0.3602	0.4539	0.6011
Sintel	PAGE4D	10	0.168138	0.164221	1.607194	3.741	22.605	0.0726	0.1289	0.2580	0.3507	0.5144
Sintel	Easi3R	10	0.235034	0.257437	4.974138	9.131	29.068	0.0481	0.1111	0.2129	0.2827	0.4193
Sintel	LEAP-VO	10	0.447669	0.251729	6.027791	12.860	24.102	0.1067	0.1516	0.2309	0.2824	0.3997
Sintel	MonST3R	10	0.192099	0.205477	4.257992	12.740	31.410	0.0726	0.1342	0.2144	0.2655	0.3750
Sintel	VGGT4D	10	0.202911	0.133599	1.359555	3.300	17.893	0.1370	0.2116	0.3573	0.4538	0.6019
Sintel	VGGT-Ω	10	**0.080760**	**0.079508**	**0.780726**	**2.058**	**10.067**	**0.3341**	**0.4240**	**0.5569**	**0.6329**	**0.7514**
Sintel	Ours	10	0.188077	0.137474	1.392261	3.244	18.589	0.1052	0.1507	0.2882	0.3840	0.5640
TUM	VGGT	8	0.010892	0.015797	0.461763	0.541	7.438	0.2685	0.3883	0.5497	0.6409	0.7883
TUM	PAGE4D	8	0.016190	0.022170	0.635481	0.775	14.064	0.0991	0.1989	0.3572	0.4415	0.6001
TUM	Easi3R	8	0.034936	0.055385	1.438698	2.026	14.706	0.0731	0.1600	0.3122	0.4185	0.6059
TUM	MonST3R	8	0.082542	0.156828	9.116100	14.469	24.281	0.0435	0.0878	0.1564	0.2096	0.3440
TUM	VGGT4D	8	0.011159	0.016271	0.453426	0.531	7.552	0.2602	0.3789	0.5444	0.6356	0.7841
TUM	VGGT-Ω	8	**0.009690**	**0.014111**	0.436785	0.542	**6.831**	**0.3148**	**0.4367**	**0.5917**	**0.6831**	**0.8150**
TUM	Ours	8	0.010705	0.016129	**0.425038**	**0.470**	7.838	0.2861	0.3867	0.5392	0.6350	0.7806
Bonn	VGGT	10	0.020677	0.585083	34.981841	41.214	46.404	–	–	0.0013	0.0047	0.0362
Bonn	PAGE4D	10	0.038977	0.584361	34.591028	40.720	46.636	–	–	0.0004	0.0024	0.0317
Bonn	MonST3R	10	0.148696	**0.537977**	**29.658704**	**35.758**	47.900	–	–	0.0002	0.0021	0.0287
Bonn	VGGT4D	10	0.020929	0.585265	34.955329	41.186	46.324	–	–	**0.0020**	0.0050	0.0361
Bonn	VGGT-Ω	10	**0.012773**	0.585643	34.969557	41.186	**45.937**	–	–	0.0016	**0.0053**	**0.0367**
Bonn	Ours	10	0.033702	0.581412	35.157318	41.403	46.256	–	–	0.0002	0.0039	0.0364
DyCheck	VGGT	10	0.012745	0.019087	1.068782	1.213	6.560	0.2281	0.3676	0.5796	0.6841	0.8184
DyCheck	PAGE4D	10	0.016324	0.024192	1.558721	1.890	9.190	0.0978	0.1982	0.4069	0.5400	0.7321
DyCheck	Easi3R	10	0.025609	0.040219	3.001516	3.612	11.805	0.0881	0.1893	0.3740	0.4867	0.6610
DyCheck	MonST3R	10	0.079202	0.111582	13.163112	15.123	28.003	0.0785	0.1382	0.2347	0.2932	0.3999
DyCheck	VGGT4D	10	0.010878	0.016303	0.659894	0.742	5.770	0.3163	0.4556	0.6460	0.7338	0.8454
DyCheck	VGGT-Ω	10	**0.004917**	**0.007722**	**0.439855**	**0.538**	**2.462**	**0.5941**	**0.7080**	**0.8280**	**0.8766**	**0.9340**
DyCheck	Ours	10	0.014923	0.019932	1.546136	1.898	7.645	0.1304	0.2658	0.4911	0.6135	0.7756

A.2 Complete Camera Evaluation

Table 4 reports the complete camera evaluation. Each evaluated sequence uses ten endpoint-sampled frames, and ground truth is used only for scoring. All trajectories produced by the evaluated methods are re-scored using the same evaluator. The *Seq* column reports the number of sequences successfully evaluated for each method.

VGGT-Ω uses its 416-pixel reproduction checkpoint and official preprocessing. Results may differ from those reported in the original papers because of differences in sequence selection, temporal sampling, preprocessing, camera intrinsics, or evaluation protocol.

A.3 Complete Depth Evaluation

Table 5 reports the complete depth evaluation. All predicted depth maps are scored using the same PAGE4D depth evaluator. A single median-based scale factor is estimated for each evaluated sequence before computing the metrics. When an official implementation center-crops its input, the corresponding ground-truth depth is cropped identically before scoring.

As with the camera evaluation, results can differ from those reported in the original papers because of differences in sequence selection, frame sampling, preprocessing, camera intrinsics, or depth-alignment protocol.

A.4 Layer-Placement Ablation

We study where the geometric camera-support prior should be introduced within VGGT's global-attention stack. We compare guidance in early blocks (0–4), middle blocks (8–12), and late blocks (19–23) under matched training settings. In this experiment, all original VGGT parameters remain frozen and only five layer-specific camera-guidance scalars are optimized. We freeze the backbone in this placement ablation to isolate the effect of where the guidance itself is introduced. Results are reported at training step 1,000 using the online student weights.

As shown in Table 6, conditioning blocks 8–12 provides the strongest overall camera performance. On Sintel, the middle placement achieves the lowest ATE, RPE-T, RPE-R, pairwise rotation error, and translation-direction error, while also obtaining the highest AUC@10, AUC@15, and AUC@30 among the three placements. On TUM, the late placement obtains slightly lower ATE, RPE-T, and RPE-R. Blocks 8–12 provide the lowest pairwise rotation and translation-direction errors and the

Table 5: Complete depth evaluation. *Seq* denotes the number of successfully evaluated sequences. Lower is better for error metrics and higher is better for threshold accuracy. Bold indicates the best result on each dataset.

Dataset	Method	Seq	Abs Rel ↓	Sq Rel ↓	RMSE ↓	Log RMSE ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Sintel	VGGT	10	0.303145	2.326201	6.832123	0.509535	0.574288	0.764495	0.873161
Sintel	PAGE4D	10	**0.266954**	1.892005	**4.287417**	**0.402985**	**0.681670**	**0.832818**	**0.916877**
Sintel	Easi3R	10	0.305538	1.983252	6.187035	0.486316	0.533623	0.725080	0.845901
Sintel	MonST3R	10	0.289412	1.946139	6.213693	0.471855	0.553663	0.746627	0.859050
Sintel	VGGT4D	10	0.303204	2.326961	6.833545	0.509896	0.574359	0.764783	0.873230
Sintel	VGGT-Ω	10	0.287431	**1.609912**	5.257597	0.501092	0.603519	0.740592	0.824496
Sintel	Ours	10	0.274007	1.799023	5.922952	0.471565	0.591045	0.795124	0.894185
TUM	VGGT	8	0.090255	0.184387	0.643425	0.162512	0.936373	0.979395	0.987797
TUM	PAGE4D	8	**0.046739**	**0.080979**	**0.424289**	**0.115862**	**0.971684**	**0.986892**	**0.991414**
TUM	Easi3R	8	0.128375	0.298794	0.998452	0.255246	0.856746	0.942524	0.964090
TUM	MonST3R	8	0.134311	0.319476	1.043914	0.243885	0.854161	0.944084	0.966458
TUM	VGGT4D	8	0.090441	0.184014	0.642877	0.162574	0.936361	0.979417	0.987793
TUM	VGGT-Ω	8	0.159915	0.819042	1.005817	0.233616	0.925290	0.960272	0.970367
TUM	Ours	8	0.076969	0.105151	0.680385	0.153974	0.930448	0.979220	0.988265
Bonn	VGGT	10	0.037611	0.016001	0.172370	0.078775	0.983350	0.992827	0.996020
Bonn	PAGE4D	10	**0.029027**	**0.010905**	**0.125713**	**0.057971**	**0.991630**	**0.997172**	**0.998398**
Bonn	MonST3R	10	0.055057	0.024602	0.199424	0.105931	0.952566	0.985661	0.994642
Bonn	VGGT4D	10	0.037561	0.015969	0.172343	0.078791	0.983344	0.992820	0.995995
Bonn	VGGT-Ω	10	0.053029	0.049103	0.205739	0.103439	0.971284	0.984253	0.988941
Bonn	Ours	10	0.036982	0.014675	0.146042	0.069742	0.989262	0.994790	0.997130
DyCheck	VGGT	10	0.095184	0.051864	0.332549	0.192088	0.880865	0.953541	0.974685
DyCheck	PAGE4D	10	0.077773	0.036338	0.288543	0.151049	0.924029	0.975947	0.985831
DyCheck	Easi3R	10	0.107124	0.053263	0.328682	0.184936	0.839368	0.955519	0.979565
DyCheck	MonST3R	10	0.108648	0.052707	0.325680	0.183350	0.837068	0.951749	0.980884
DyCheck	VGGT4D	10	0.085502	0.038868	0.276136	0.170185	0.896308	0.963225	0.981788
DyCheck	VGGT-Ω	10	**0.071551**	**0.032625**	**0.254568**	**0.136173**	**0.931606**	**0.977231**	**0.988429**
DyCheck	Ours	10	0.100759	0.060159	0.366818	0.213898	0.882383	0.945702	0.967390

Table 6: Placement of residual-guided camera attention at training step 1,000. All original VGGT parameters are frozen and only the layer-specific guidance scalars are optimized. Results use the online student weights.

Dataset	Guided blocks	ATE ↓	RPE-T ↓	RPE-R ↓	Trans. ↓	AUC@5 ↑	AUC@30 ↑
Sintel	0–4	0.206883	0.145901	1.651763	17.566	**0.2500**	0.6031
Sintel	8–12	**0.206065**	**0.145607**	**1.648416**	**17.256**	0.2486	**0.6082**
Sintel	19–23	0.206926	0.146139	1.652558	17.513	0.2379	0.6002
TUM	0–4	0.012184	0.018773	0.489924	7.686	0.3821	0.7874
TUM	8–12	0.012170	0.018837	0.484941	**7.438**	**0.3902**	**0.7939**
TUM	19–23	**0.011993**	**0.018456**	**0.484424**	7.695	0.3750	0.7871

strongest AUC metrics. In contrast, the depth metrics remain nearly unchanged across placements, suggesting that the effect of the geometric prior is concentrated primarily in the camera pathway.

Based on this comparison, we apply residual-guided camera attention to global blocks 8–12 in the final model.

A.5 Qualitative Analysis of Residual-Guided Camera Attention

We qualitatively examine both the geometric prior and its effect on camera-query attention using dynamic TUM RGB-D sequences.

Figure 3 visualizes the signed camera-support prior derived from the cross-view rigid-consistency residual. Positive values indicate patches that are more consistent with the provisional rigid reconstruction, while negative values indicate stronger disagreement between the observed cross-view correspondence and the rigid reprojection. The residual exhibits clear spatial structure and produces strong negative responses around several visibly dynamic regions, including moving people, while large portions of the static scene receive positive support. This suggests that the residual provides a meaningful geometric signal rather than an unstructured perturbation.

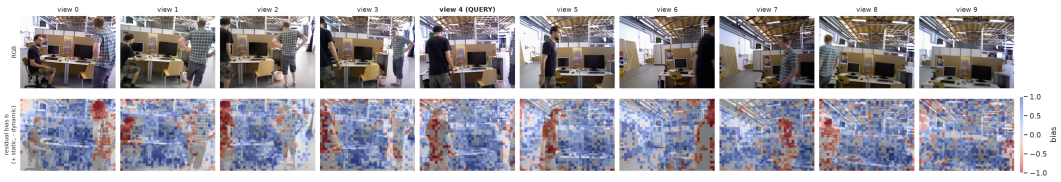

Figure 3: **Cross-view rigid-consistency camera-support prior.** The top row shows the input RGB views and the bottom row shows the signed patch-level camera-support prior derived from the rigid-consistency residual. Negative responses indicate stronger disagreement with the provisional rigid reconstruction, while positive responses indicate greater consistency. Several visibly dynamic regions receive strong negative support.

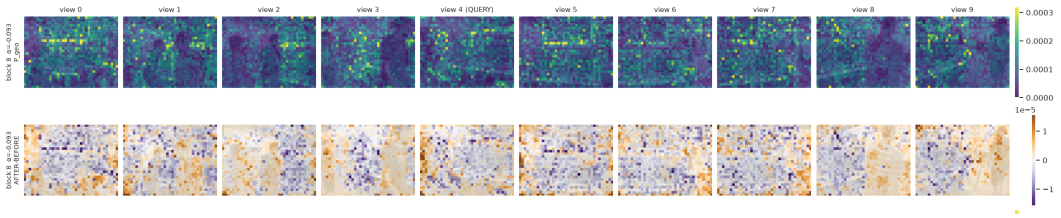

Figure 4: **Effect of geometric conditioning on camera-query attention.** The top row shows camera-query attention after residual-guided conditioning for a guided global-attention block, and the bottom row shows the corresponding change relative to the unconditioned attention. Conditioning produces spatially localized positive and negative corrections while largely preserving the pretrained attention structure.

Figure 4 shows how this geometric signal modifies the camera-query attention within a guided global-attention block. The top row shows the camera attention after conditioning, while the bottom row shows the difference relative to the unconditioned attention. The overall attention distribution remains largely preserved, but the difference maps reveal spatially localized positive and negative corrections. This indicates that the geometric prior acts as a targeted modification to the pretrained camera attention rather than replacing its existing attention pattern.

We emphasize that the residual is not a ground-truth motion mask. Large responses may also arise from occlusions, correspondence errors, or inaccurate provisional geometry. Nevertheless, the spatial correspondence between strong residual responses and visibly dynamic regions, together with the localized attention changes, provides qualitative evidence that the proposed rigid-consistency signal is incorporated into the camera pathway.