PRUNEBA: EFFICIENT VISUAL GEOMETRY INITIALIZATION WITH CHUNKED PROGRESSIVE BUNDLE ADJUSTMENT

ABSTRACT

Feed-forward visual geometry foundation models provide camera hypotheses for image collections that challenge feature-based Structure-from-Motion (SfM). Turning these hypotheses into supported sparse geometry requires efficient neural inference and effective geometric refinement. We introduce PruneBA, a training-free framework combining shared random patch-token pruning in the Visual Geometry Grounded Transformer (VGGT) with chunked progressive bundle adjustment (BA). Independently inferred overlapping windows undergo local triangulation and BA; a deterministic centrality criterion selects camera estimates for global re-triangulation and BA. Across seven ground–UAV, wide-baseline, and indoor evaluation cases, unpruned progressive BA retains all 157 case-level RGB cameras, compared with 58 for classical incremental SfM, and lowers internal mean reprojection error relative to direct global BA in every case. On Bus Ground–UAV 30, progressive refinement increases point yield from 96 to 298; adding 75% pruning yields 364 points at 1.364 pixels while retaining all 30 cameras. Synthetic ten-frame inference achieves a $2.63\times$ speedup with random weights and $2.20\times$ with pretrained weights, with lower GPU allocation in both studies. Exact attention-pair counts and finite-population sampling identities characterize the computational and spatial effects of pruning, while a centrality surrogate formalizes camera selection. These results establish sparse neural aggregation and staged geometric refinement as a practical combination for visual geometry initialization.

1 INTRODUCTION

Recovering camera poses and three-dimensional structure requires both a viable initialization and sufficient multi-view observations for geometric refinement. Classical Structure-from-Motion (SfM) combines local features, correspondence verification, triangulation, and bundle adjustment (BA) (Lowe, 2004; Schönberger & Frahm, 2016; Triggs et al., 2000). Wide baselines, repetitive textures, and large viewpoint changes can obstruct initial pair selection or fragment the reconstructed camera set.

Feed-forward visual geometry foundation models provide a complementary source of camera hypotheses. DUSt3R, MASt3R, and VGGT infer pointmaps, correspondences, or cameras from uncalibrated images (Wang et al., 2024b; Leroy et al., 2024; Wang et al., 2025a). A predicted camera set establishes an initialization even when incremental feature-based registration is incomplete. Its conversion into triangulated structure remains sensitive to image correspondences and the order of geometric refinement. At the same time, cross-view attention incurs a quadratic cost in the number of visual tokens, making the allocation of neural computation a central design choice.

PruneBA combines shared patch-token pruning with chunked progressive BA (Figure 1). After the complete DINOv2 image encoder, we retain a common subset of spatial patch tokens across images, together with every camera and register token. Independently inferred overlapping windows receive local triangulation and BA. A deterministic centrality criterion then selects one locally refined camera record per image, and global matching, re-triangulation, and BA establish the final sparse geometry. The pretrained network weights remain unchanged.

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.

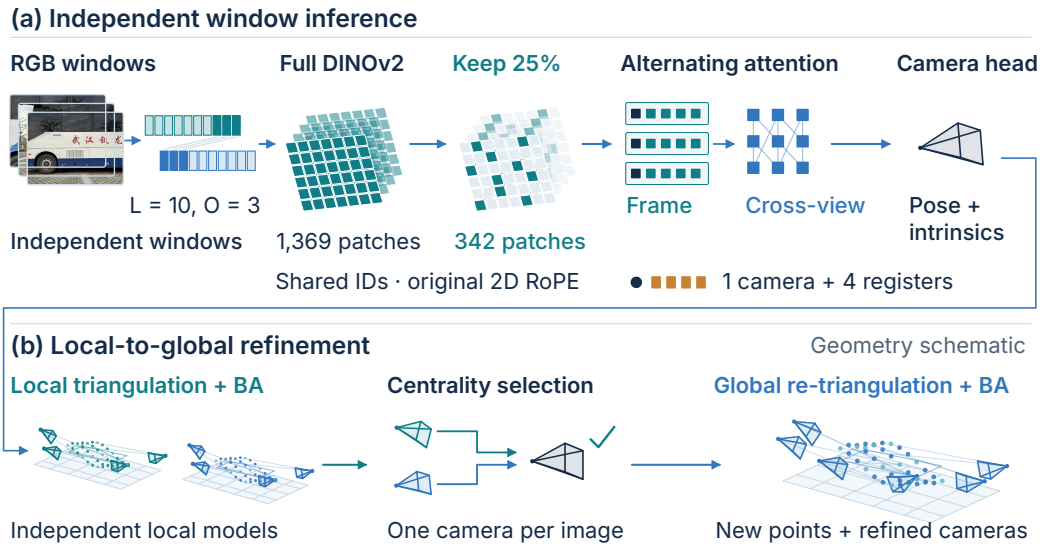

Figure 1: The PruneBA framework integrates shared patch-token pruning with independent overlapping-window inference, local triangulation, and bundle adjustment (BA), followed by centrality-based camera selection and global re-triangulation and BA. RGB inputs represent real evaluation examples; tokens, windows, and camera geometry are illustrated schematically. The unpruned progressive configuration operates with 0% pruning, while the accelerated PruneBA-75 configuration employs 75% patch-token pruning.

Our study separates camera initialization, geometric refinement, and neural efficiency. Across seven evaluation cases, unpruned progressive BA retains all 157 case-level RGB cameras and reduces internal mean reprojection error relative to direct global BA in every case. Its point yield increases in four cases. On Bus Ground–UAV 30, adding 75% pruning produces 364 points, compared with 96 for direct global BA, while decreasing internal error by 24.4%. Controlled synthetic inference measurements show $2.63\times$ and $2.20\times$ speedups with random and pretrained weights, respectively.

We contribute a training-free shared sampling strategy that preserves the original spatial coordinates and reduces the exact attention-pair count; a local-to-global BA schedule with deterministic camera selection and an explicit centrality surrogate; and a seven-case reconstruction study complemented by pruning and inference measurements. The analysis establishes finite-sample properties of token selection, and the experiments quantify camera coverage, point support, reprojection fit, and computational cost.

2 Related Work

Classical reconstruction and optimization. SIFT and COLMAP provide local feature matching and incremental camera registration (Lowe, 2004; Schönberger & Frahm, 2016). Hierarchical SfM partitions reconstruction into subproblems (Toldo et al., 2015), and BA jointly refines cameras and structure through reprojection least squares (Triggs et al., 2000). Dense multi-view stereo estimates depth and surfaces downstream of camera reconstruction (Schönberger et al., 2016). PruneBA uses sparse triangulation and BA with neural camera initialization and a local-to-global refinement schedule.

Learned visual geometry. DUSt3R predicts pairwise pointmaps (Wang et al., 2024b); MASt3R grounds matching in 3D (Leroy et al., 2024); and VGGSfM combines learned tracks with differentiable geometric estimation (Wang et al., 2024a). VGGT alternates frame and cross-view attention (Wang et al., 2025a), while Fast3R extends joint feed-forward reconstruction to large image sets (Yang et al., 2025). DROID-SLAM couples recurrent flow estimation with dense BA (Teed & Deng, 2021), and MASt3R-SLAM exploits learned reconstruction priors for tracking (Murai et al.,

2025). VGGT-SLAM, VGGT-SLAM 2.0, and VGGT-SLAM++ develop geometric optimization for VGGT-based mapping (Maggio et al., 2025; Maggio & Carlone, 2026; Mandal et al., 2026). We study sparse reconstruction from independently inferred windows and select camera estimates through a fixed centrality rule.

Token reduction in vision Transformers. DynamicViT learns token sparsification (Rao et al., 2021), ToMe merges similar tokens without additional training (Bolya et al., 2023), and PatchDropout randomly removes patches during training (Liu et al., 2023). PruneBA samples shared spatial positions at inference between the complete image encoder and VGGT's alternating attention. Camera tokens, register tokens, and original two-dimensional positions remain available to camera regression.

Efficient geometry Transformers. FastVGGT develops training-free token merging (Shen et al., 2025); HTTM introduces head-wise temporal merging (Wang et al., 2026); LiteVGGT combines geometry-aware tokens and cached merging (Shu et al., 2026); Co-Me uses confidence-guided merging (Chen et al., 2026); and Spark3R reduces tokens asymmetrically (Tang et al., 2026). PruneBA couples sparse neural aggregation with progressive geometric refinement and evaluates this combination through complete camera sets, sparse point and observation counts, internal reprojection fit, and measured inference cost. Additional context on correspondence learning, global reconstruction and attention implementations is provided in Appendix C.

3 METHOD

3.1 CAMERA INITIALIZATION AS THE INTERFACE

Given an ordered RGB collection $\mathcal{I} = \{I_i\}_{i=1}^{N}$, we estimate world-to-camera extrinsics (R_i, t_i), pinhole intrinsics K_i, and points $\{X_j\}_{j=1}^{M}$ supported by image observations. VGGT predicts rotations, translations, and fields of view; these predictions initialize camera models for feature-based triangulation and BA. Each local window supplies refined cameras. Global re-triangulation creates the final point set from the selected cameras and scene-wide correspondences.

3.2 SHARED PATCH-TOKEN PRUNING

The complete pretrained DINOv2 ViT-L/14 image encoder (Oquab et al., 2024) produces P patch embeddings per image. For a window of S inputs and pruning fraction $\rho \in [0, 1)$, sample uniformly without replacement an index set $J \subseteq \{1, \dots, P\}$ of size

$$P_\rho = \max\{1, \lfloor (1-\rho)P \rfloor\}. \tag{1}$$

The sorted set J is shared across all images in the forward call. Both patch embeddings and their original two-dimensional rotary position embedding (RoPE) coordinates (Su et al., 2024) are gathered using J. One camera token and four register tokens remain intact, giving $q = P_\rho + 5$ tokens per image. Shared sampling preserves a common grid subset across frames; it does not require a semantic score or confidence predictor.

One frame-attention block evaluates Sq^2 token pairs, and one global block evaluates S^2q^2. With the same number of blocks before and after pruning, their exact ratio is

$$\eta = \frac{(P_\rho + 5)^2}{(P + 5)^2}. \tag{2}$$

At 518×518 resolution and stride 14, $P = 1369$. With $\rho = 0.75$, $P_\rho = 342$ and $q = 347$, compared with 1374 unpruned tokens. Thus $\eta = 0.06378$, a $15.68\times$ reduction in attention pairs. The full image encoder, projections, and camera head also contribute to latency; Section 5.4 reports measured inference speedups.

For any fixed pre-aggregation spatial statistic a_p, uniform sampling gives $\mathbb{E}[\sum_{p\in J} a_p] = (P_\rho/P)\sum_{p=1}^{P} a_p$ and an unbiased sample mean. These analytic identities describe retained spatial information and are derived with finite-population variance in Appendix A. They are not used as a pose estimator or training loss.

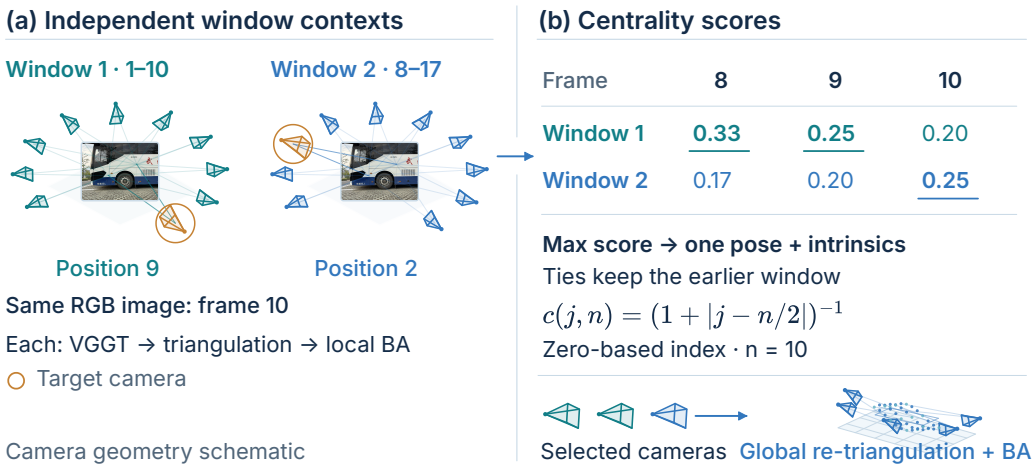

Figure 2: Centrality selects one locally refined pose and its associated intrinsics for each image. Two ten-input windows share positions 8–10; the score comparison identifies the selected camera at each position. This matches the nominal experimental window length and overlap. Selected cameras initialize global triangulation and BA without an estimated inter-window transformation.

3.3 Overlapping Local Reconstruction

Windows have nominal length L, overlap O, and stride $L - O$. Each window receives independent neural inference, followed by GPU SIFT extraction and exhaustive intra-window matching. The predicted fields of view determine pinhole intrinsics at native image resolution. Triangulation initializes local points, and local BA solves the reprojection least-squares objective

$$\min_{\{R_i,t_i,K_i\}_{i\in\mathcal{W}_k},\{X_j\}} \sum_{(i,j)\in\Omega_k} \|u_{ij} - \pi(R_i, t_i, K_i, X_j)\|_2^2, \tag{3}$$

where Ω_k is the observation set and π is pinhole projection (Triggs et al., 2000). Local and global BA use the installed COLMAP defaults with squared reprojection residuals. Local points support camera refinement; global structure is subsequently triangulated from scene-wide matches.

3.4 Centrality Selection and Global Refinement

An image in an overlap has camera estimates from multiple windows. Let $j_k(i) \in \{0, \ldots, n_k - 1\}$ be its zero-based position in a window of n_k enumerated inputs. The implemented criterion is

$$c_k(i) = \frac{1}{1 + |j_k(i) - n_k/2|}, \qquad k^*(i) = \arg\max_{k:\, i\in\mathcal{W}_k} c_k(i), \tag{4}$$

with earlier-window selection on a tie (Figure 2). Both extrinsics and intrinsics come from the selected record. The reference location is exactly $n_k/2$. It defines a deterministic preference for central positions without a learned score or rotation averaging.

The criterion exactly minimizes a shared affine surrogate of positional distance; Appendix A.3 states and proves this property.

The selected camera records initialize global matching, re-triangulation, and BA. Local point sets and tracks are replaced by globally triangulated tracks. This local-to-global schedule defines progressive refinement; Section 6 discusses coordinate-gauge considerations.

Table 1: Evaluated image collections and recorded progressive windows. Throughout all progressive reconstructions, window length $L = 10$ and overlap $O = 3$. RGB denotes the denominator for camera coverage; enumerated frames include historical depth inputs present in the three 7-Scenes cases.

Collection	Scene type	RGB	Enumerated	Windows
Bus Ground–UAV 30	Ground / UAV bus views	30	30	4
UAV Wide-Baseline 9	UAV wide baseline	9	9	1
Office14	Indoor / Office	14	20	3
Heads14	Indoor / Heads	14	20	3
Chess14	Indoor / Chess	14	20	3
Bus Ultra-Wide13	Ultra-wide bus views	13	13	2
Bus Ground–UAV 63	Ground / UAV bus views	63	63	9

4 Experimental Setup

4.1 Image collections and evaluation scope

We evaluate seven cases spanning ground–UAV bus views, wide-baseline aerial imagery, ultra-wide bus views, and indoor scenes (Table 1). Bus Ground–UAV 30 contains 15 ground and 15 UAV views; Bus Ground–UAV 63 contains all 38 ground and 25 UAV views from the same scene. UAV Wide-Baseline 9 samples every fourth image of an ordered aerial sequence. Bus Ultra-Wide13 contains 13 ultra-wide bus views. Office14, Heads14, and Chess14 use images from the corresponding 7-Scenes environments (Shotton et al., 2013).

The executed indoor input lists contain 20 files: 14 RGB images and six depth PNGs, all supplied to neural inference after RGB conversion. The six depth filenames are absent from the feature database and exported camera sets. The three neural windows contain 10, 10, and 6 inputs and retain 7, 7, and 4 RGB camera records. Camera coverage therefore uses 14 RGB images per indoor case. Across all cases the denominator sums to 157 case-level RGB images; the two bus cases share source imagery. Appendix B records the selection and window details.

4.2 Compared reconstruction pipelines

Incremental SfM evaluates the first COLMAP model (Model 0). Failure to initialize yields zero cameras and points with undefined reprojection error. **Direct triangulation** initializes COLMAP triangulation with full-sequence VGGT cameras, without a separate bundle-adjuster stage; triangulation retains its default internal refinement. **Direct global BA** independently obtains full-sequence VGGT initialization, triangulates, and applies global BA. **Progressive BA** uses unpruned independent windows, local triangulation and BA, centrality selection, and global re-triangulation and BA. **PruneBA-75** adds 75% shared patch-token pruning to progressive BA and is evaluated on Bus Ground–UAV 30.

4.3 Implementation and measurement criteria

The experiments use pretrained VGGT, COLMAP 3.11.1, GPU SIFT, and an NVIDIA RTX 4090 D with 24 GB VRAM. Images are square-padded and resized to 518×518 for inference; predicted intrinsics are mapped to native resolution. Incremental SfM uses shared pinhole intrinsics, while neural configurations use per-image pinhole models. Progressive windows use $L = 10$, $O = 3$.

Camera coverage is $N_{\mathrm{cam}}/N_{\mathrm{RGB}}$. Point count P and observation count O_{obs} quantify sparse support. COLMAP's model analyzer supplies internal mean reprojection error e in pixels on each model's own points and observations. Table 2 reports all saved-model measurements and original end-to-end wall times t in seconds. Synthetic neural measurements separately report pose-path latency and GPU allocation, with their distinct weight and measurement settings retained in Section 5.4 and Appendix B.3.

Table 2: Complete saved-model results for the seven-case comparison. Incremental SfM is COLMAP Model 0; direct triangulation uses VGGT initialization; direct global BA adds a separate global BA stage; progressive BA uses our unpruned local-to-global schedule. N_{cam} denotes retained cameras / total RGB images, P is reconstructed 3D points, O_{obs} is visual observations, e is internal mean reprojection error (pixels), and t is wall time (seconds). An em dash denotes failure to initialize a reconstruction. Bold highlights progressive BA; each error is evaluated on the corresponding model's own geometric support.

Collection	Method	$N_{\text{cam}} \uparrow$	$P \uparrow$	$O_{\text{obs}} \uparrow$	$e \downarrow$	t
Bus Ground–UAV 30	Incremental SfM	7/30	430	1,207	0.591	4.4
	Direct triangulation	30/30	99	200	1.746	36.0
	Direct global BA	30/30	96	194	1.806	18.9
	Progressive BA	**30/30**	**298**	**601**	**1.662**	**32.8**
UAV Wide-Baseline 9	Incremental SfM	2/9	1,187	2,374	0.876	6.2
	Direct triangulation	9/9	10	20	2.160	11.6
	Direct global BA	9/9	10	20	2.160	11.5
	Progressive BA	**9/9**	**32**	**64**	**1.108**	**16.5**
Office14	Incremental SfM	14/14	914	8,444	0.533	17.5
	Direct triangulation	14/14	1,067	9,911	0.852	3.0
	Direct global BA	14/14	1,186	10,783	0.854	5.0
	Progressive BA	**14/14**	**164**	**583**	**0.553**	**6.5**
Heads14	Incremental SfM	2/14	4	8	0.000	3.5
	Direct triangulation	14/14	1,399	9,915	0.789	3.0
	Direct global BA	14/14	1,379	9,924	0.799	4.2
	Progressive BA	**14/14**	**1,314**	**5,240**	**0.475**	**7.8**
Chess14	Incremental SfM	0/14	0	0	—	8.3
	Direct triangulation	14/14	22	57	2.300	4.1
	Direct global BA	14/14	15	42	1.604	4.1
	Progressive BA	**14/14**	**223**	**467**	**1.480**	**7.2**
Bus Ultra-Wide13	Incremental SfM	0/13	0	0	—	1.6
	Direct triangulation	13/13	28	56	2.014	9.0
	Direct global BA	13/13	28	56	2.014	9.3
	Progressive BA	**13/13**	**13**	**26**	**1.919**	**16.2**
Bus Ground–UAV 63	Incremental SfM	33/63	2,436	6,994	1.078	24.7
	Direct triangulation	63/63	296	596	1.828	50.0
	Direct global BA	63/63	299	602	1.848	54.9
	Progressive BA	**63/63**	**506**	**1,016**	**1.324**	**92.8**

5 RESULTS

5.1 COMPLETE CAMERA COVERAGE

Progressive BA retains all 157 case-level RGB cameras across seven cases, as do the direct neural baselines (Table 2). Incremental SfM retains 58/157 cameras (36.9%), giving a 63.1 percentage-point coverage gain for the neural configurations. Its first model contains 7/30 cameras on Bus Ground–UAV 30, 2/9 on UAV Wide-Baseline 9, 2/14 on Heads14, and 33/63 on Bus Ground–UAV 63; initialization fails on Chess14 and Bus Ultra-Wide13. Office14 is its only fully covered case. Neural camera initialization therefore supplies complete camera sets in collections where incremental registration produces a fragment or no model.

5.2 PROGRESSIVE REFINEMENT IMPROVES INTERNAL FIT

Progressive BA lowers internal mean reprojection error relative to direct global BA in all seven cases, with reductions from 4.7% to 48.7% (Figure 3). On Bus Ground–UAV 30, points increase

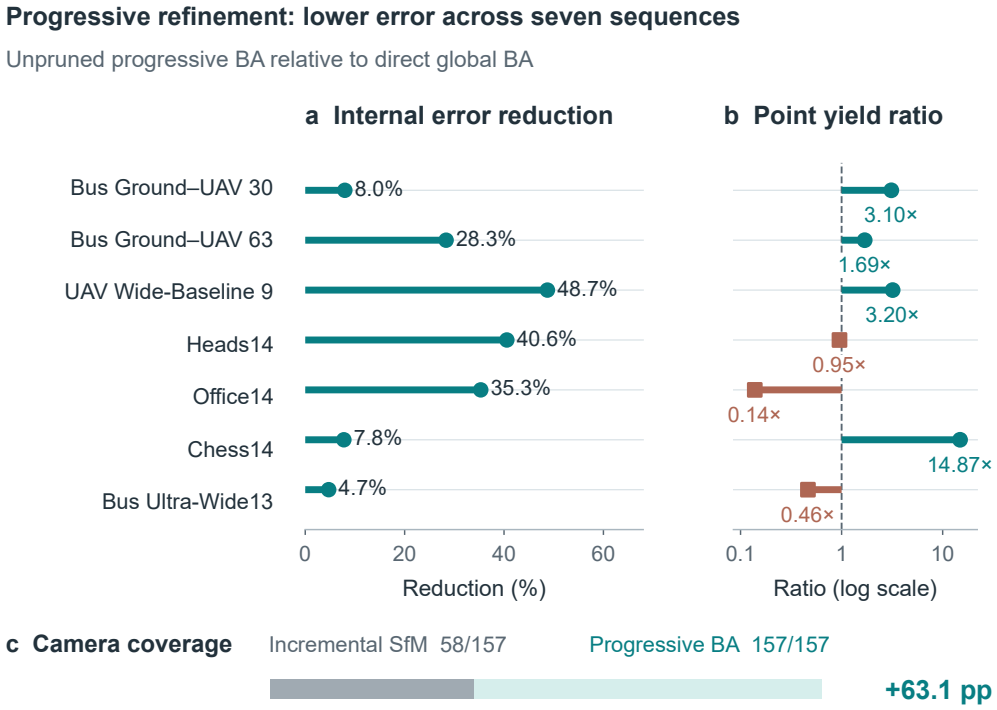

Figure 3: Seven-case reconstruction comparison. Camera coverage, point yield, and internal error are shown together. Progressive BA retains full RGB coverage and lowers internal error relative to direct global BA in all cases, with more points in four. The point-yield ratio uses a logarithmic axis; each error is computed on that method's own saved support. The aggregate coverage difference is 63.1 percentage points, with the denominator summed across evaluation cases.

from 96 to 298 ($3.10\times$), observations from 194 to 601, and error decreases from 1.806 to 1.662 pixels. On Bus Ground–UAV 63, points increase from 299 to 506 and observations from 602 to 1016, while error decreases from 1.848 to 1.324 pixels.

On UAV Wide-Baseline 9, progressive BA produces 32 points at 1.108 pixels, compared with 10 points at 2.160 pixels for both direct baselines. This case uses one window and evaluates staged refinement without an overlap-selection effect. On Chess14, points increase from 15 to 223 ($14.9\times$) and observations from 42 to 467 relative to direct global BA, while error decreases from 1.604 to 1.480 pixels.

5.3 Point support and scene dependence

Point yield increases in four of the seven progressive-versus-direct-global comparisons. Heads14 retains 1314 points versus 1379 for direct global BA and lowers error from 0.799 to 0.475 pixels, a 40.6% reduction using the saved precision. Office14 yields 164 points at 0.553 pixels versus 1186 at 0.854 pixels. Bus Ultra-Wide13 retains 13 cameras with 13 points and 26 observations. Section 6 examines the relationship between support and internal fit in these scenes.

Progressive BA takes 32.8 seconds on Bus Ground–UAV 30 and 92.8 seconds on Bus Ground–UAV 63, compared with 18.9 and 54.9 seconds for direct global BA. These times include local matching and BA followed by global matching and re-triangulation.

5.4 Pruning efficiency and reconstruction quality

Synthetic inputs of shape $1 \times 10 \times 3 \times 518 \times 518$ measure the neural pose path. With random weights and FP16 inference, 75% pruning reduces mean latency from 649.8 to 246.9 ms ($2.63\times$),

Table 3: Bus Ground–UAV 30 component comparison. All configurations retain 30/30 cameras. PruneBA-75 adds 75% token pruning to progressive BA. Reconstruction times describe their original executions; they do not measure the isolated effect of pruning.

Method	Pruning	$P \uparrow$	$O_{\text{obs}} \uparrow$	$e \downarrow$ (px)	t (s)
Direct triangulation	0%	99	200	1.746	36.0
Direct global BA	0%	96	194	1.806	18.9
Progressive BA	0%	298	601	1.662	32.8
PruneBA-75	**75%**	**364**	**729**	**1.364**	**46.0**

and incremental peak allocation from 2596.1 to 1973.6 MiB (24.0%). The 50% and 90% settings take 328.3 and 279.5 ms. Thus 75% is the fastest measured setting in this study. The pretrained study records 663.7 to 301.8 ms (2.20×), with absolute peak allocation decreasing from 7863.9 to 7106.4 MiB (9.6%). Figure 5 and Appendix B.3 preserve all ratios, allocations, and measurement definitions.

On real Bus Ground–UAV 30 images, PruneBA-75 retains all 30 cameras and yields 364 points, 729 observations, and 1.364 pixels internal error (Table 3). Relative to unpruned progressive BA, points increase by 22.1% and error decreases by 17.9%. Relative to direct global BA, the point multiplier is 3.79× and the error reduction is 24.4%. The saved reconstruction demonstrates that substantial token reduction is compatible with complete camera coverage and improved sparse support in this case. Its original 46.0-second total and the unpruned 32.8-second total include different checkpoint-loading conditions; the synthetic studies quantify the isolated neural acceleration.

5.5 Qualitative reconstruction

Figures 4 and 6 link original RGB examples to saved 3D points, image observations, and camera distributions for Bus Ground–UAV 30, Heads14, and Office14. PruneBA-75 retains 30 bus cameras and 364 points, compared with the seven-camera incremental fragment in Table 2. On Heads14, progressive BA retains all 14 cameras and 1314 points, compared with two cameras and four points for incremental SfM. Complete views and magnified details expose global support and local point distributions in each model's own coordinate frame. The separate indoor reference-motion panels describe the source image sequence.

6 Discussion

PruneBA separates neural token allocation, local camera refinement, and global structure estimation. The attention-pair analysis quantifies the reduction in aggregator computation, while the saved reconstructions show lower internal error after progressive refinement in all seven cases. Random sampling retains spatial statistics in expectation; the nonlinear camera predictor and geometric solver determine the resulting reconstruction empirically.

Independent windows introduce distinct coordinate gauges. PruneBA transfers the centrality-selected camera parameters directly to global triangulation and BA. Explicit similarity alignment would complement positional selection by accounting for inter-window scale, rotation and translation, extending the framework to broader acquisition geometries.

On Office14, the lower residual accompanies a smaller reconstructed support (164 versus 1186 points). This emphasizes joint reporting of coverage, point yield and internal fit; filtering of ambiguous tracks provides a geometric interpretation to examine in further analysis. Bus Ultra-Wide13 retains all 13 cameras with 13 points, motivating further assessment of image overlap and camera-model expressiveness. Calibrated equidistant or polynomial models and adaptive non-parametric calibration, as in GenSfM, provide complementary directions for these acquisition conditions (Wang et al., 2025b).

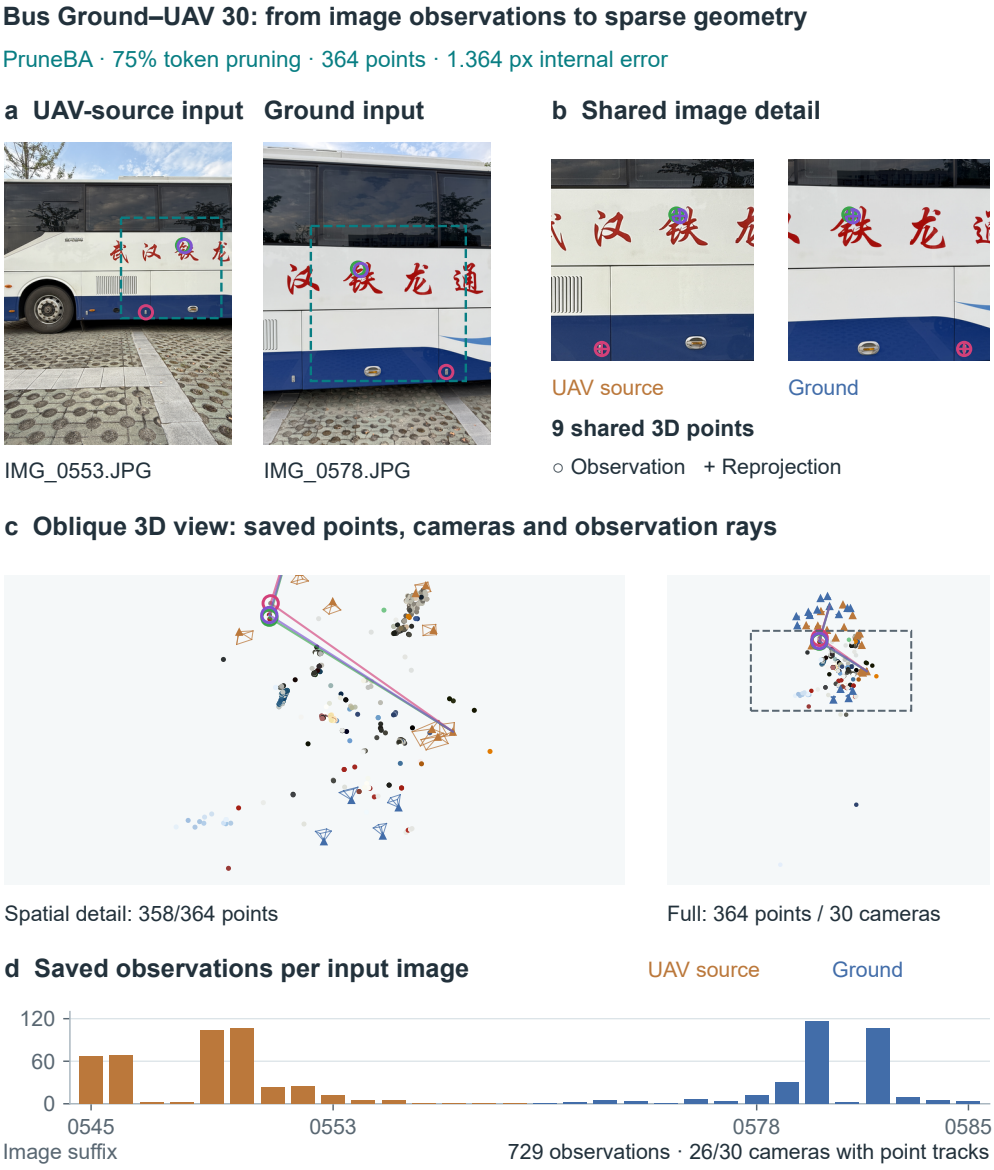

Figure 4: Real RGB observations and saved PruneBA-75 geometry on Bus Ground–UAV 30. Circles mark observed features and plus signs their reprojections; camera frusta and rays come from the saved model. The full view contains 364 points and 30 cameras, with 729 observations attached to 26 cameras. Details are enlarged in the native reconstruction frame.

7 Conclusion

PruneBA combines shared patch-token pruning with chunked progressive BA. Independent overlapping windows, local camera refinement, and centrality selection initialize global re-triangulation and BA. Across seven cases, unpruned progressive BA retains 157/157 case-level RGB cameras versus 58 for incremental SfM, and lowers internal mean reprojection error relative to direct global BA in every case. On Bus Ground–UAV 30, 75% pruning yields 3.79× as many points and 24.4% lower error. Synthetic inference accelerates by 2.63× with random weights and 2.20× with pretrained weights. Attention counts and sampling identities characterize the computation. Future directions include explicit gauge alignment, nonlinear camera models, and robotic visual SLAM.

AI USE STATEMENT

AI tools were used solely to aid and polish the writing of the manuscript, including language refinement and translation.

REFERENCES

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token Merging: Your ViT But Faster. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/pdf?id=JroZRaRw7Eu.

Yutian Chen, Yuheng Qiu, Ruogu Li, Jay Patrikar, and Sebastian Scherer. Co-Me: Confidence Guided Token Merging for Visual Geometric Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14590–14599, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_Co-Me_Confidence_Guided_Token_Merging_for_Visual_Geometric_Transformers_CVPR_2026_paper.html.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022. doi: 10.52202/068431-1189. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018_workshops/w9/html/DeTone_SuperPoint_Self-Supervised_Interest_CVPR_2018_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive Token Sampling For Efficient Vision Transformers. In *European Conference on Computer Vision*, pp. 396–414, 2022. doi: 10.1007/978-3-031-20083-0_24. URL https://doi.org/10.1007/978-3-031-20083-0_24.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding Image Matching in 3D with MASt3R. In *European Conference on Computer Vision*, pp. 71–91, 2024. doi: 10.1007/978-3-031-73220-1_5. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/9080_ECCV_2024_paper.php.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BjyvwnXXVn_.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17627–17638, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Lindenberger_LightGlue_Local_Feature_Matching_at_Light_Speed_ICCV_2023_paper.html.

Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, and Kevin Smith. PatchDropout: Economizing Vision Transformers Using Patch Dropout. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3953–3962, 2023. URL https://openaccess.thecvf.com/content/WACV2023/html/Liu_PatchDropout_Economizing_Vision_Transformers_Using_Patch_Dropout_WACV_2023_paper.html.

David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94. URL https://link.springer.com/article/10.1023/B:VISI.0000029664.99615.94.

Dominic Maggio and Luca Carlone. VGGT-SLAM 2.0: Real-time Dense Feed-forward Scene Reconstruction, 2026. URL https://arxiv.org/abs/2601.19887v2.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM Optimized on the SL(4) Manifold. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4324. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bc65ab11abfbad890171109686233f4e-Abstract-Conference.html.

Avilasha Mandal, Rajesh Kumar, Sudarshan Sunil Harithas, and Chetan Arora. VGGT-SLAM++. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1008–1018, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/VOCVALC/html/Mandal_VGGT-SLAM_CVPRW_2026_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2025/html/Murai_MASt3R-SLAM_Real-Time_Dense_SLAM_with_3D_Reconstruction_Priors_CVPR_2025_paper.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L. Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision*, pp. 58–77, 2024. doi: 10.1007/978-3-031-73661-2_4. URL https://link.springer.com/chapter/10.1007/978-3-031-73661-2_4.

Linfei Pan, Johannes Schönberger, and Marc Pollefeys. Global Structure-from-Motion Meets Feedforward Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21880–21890, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Pan_Global_Structure-from-Motion_Meets_Feedforward_Reconstruction_CVPR_2026_paper.html.

Luis Pineda, Taosha Fan, Maurizio Monge, Shobha Venkataraman, Paloma Sodhi, Ricky T. Q. Chen, Joseph Ortiz, Daniel DeTone, Austin Wang, Stuart Anderson, Jing Dong, Brandon Amos, and Mustafa Mukadam. Theseus: A Library for Differentiable Nonlinear Optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3801–3818, 2022. doi: 10.52202/068431-0275. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/185969291540b3cd86e70c51e8af5d08-Abstract-Conference.html.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/747d3443e319a22747fbb873e8b2f9f2-Abstract.html.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching With Graph Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Sarlin_SuperGlue_Learning_Feature_Matching_With_Graph_Neural_Networks_CVPR_2020_paper.html.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.html.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision*, pp. 501–518, 2016. doi: 10.1007/978-3-319-46487-9_31. URL https://link.springer.com/chapter/10.1007/978-3-319-46487-9_31.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. FastVGGT: Training-Free Acceleration of Visual Geometry Transformer, 2025. URL https://arxiv.org/abs/2509.02560v2.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2930–2937, 2013. URL https://openaccess.thecvf.com/content_cvpr_2013/html/Shotton_Scene_Coordinate_Regression_2013_CVPR_paper.html.

Zhijian Shu, Cheng Lin, Tao Xie, Wei Yin, Ben Li, Zhiyuan Pu, Weize Li, Yao Yao, Xun Cao, Xiaoyang Guo, and Xiao-Xiao Long. LiteVGGT: Boosting Vanilla VGGT via Geometry-aware Cached Token Merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36422–36432, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Shu_LiteVGGT_Boosting_Vanilla_VGGT_via_Geometry-aware_Cached_Token_Merging_CVPR_2026_paper.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-Free Local Feature Matching With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Sun_LoFTR_Detector-Free_Local_Feature_Matching_With_Transformers_CVPR_2021_paper.html.

Chengzhou Tang and Ping Tan. BA-Net: Dense Bundle Adjustment Networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1gabhRcYX.

Zecheng Tang, Jiaye Fu, Qiankun Gao, Haijie Li, Yanmin Wu, Jiaqi Zhang, Siwei Ma, and Jian Zhang. Spark3R: Asymmetric Token Reduction Makes Fast Feed-Forward 3D Reconstruction, 2026. URL https://arxiv.org/abs/2605.06270v2.

Zachary Teed and Jia Deng. DeepV2D: Video to Depth with Differentiable Structure from Motion. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJeO7RNKPr.

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, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/89fcd07f20b6785b92134bd6c1d0fa42-Abstract.html.

Roberto Toldo, Riccardo Gherardi, Michela Farenzena, and Andrea Fusiello. Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision and Image Understanding*, 140:127–143, 2015. doi: 10.1016/j.cviu.2015.05.011. URL https://arxiv.org/abs/1506.00395.

Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle Adjustment — A Modern Synthesis. In *Vision Algorithms: Theory and Practice*, pp. 298–372, 2000. doi: 10.1007/3-540-44480-7_21. URL https://link.springer.com/chapter/10.1007/3-540-44480-7_21.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_VGGSfM_Visual_Geometry_Grounded_Deep_Structure_From_Motion_CVPR_2024_paper.html.

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*, pp. 5294–5306, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

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*, pp. 20697–20709, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DUSt3R_Geometric_3D_Vision_Made_Easy_CVPR_2024_paper.html.

Weitian Wang, Lukas Meiner, Rai Shubham, Cecilia De La Parra, and Akash Kumar. HTTM: Head-wise Temporal Token Merging for Faster VGGT. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26379–26388, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_HTTM_Head-wise_Temporal_Token_Merging_for_Faster_VGGT_CVPR_2026_paper.html.

Yihan Wang, Linfei Pan, Marc Pollefeys, and Viktor Larsson. Structure-from-Motion with a Non-Parametric Camera Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1040–1049, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Structure-from-Motion_with_a_Non-Parametric_Camera_Model_CVPR_2025_paper.html.

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*, pp. 21924–21935, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Fast3R_Towards_3D_Reconstruction_of_1000_Images_in_One_Forward_CVPR_2025_paper.html.

Runze Zhang, Siyu Zhu, Tian Fang, and Long Quan. Distributed Very Large Scale Bundle Adjustment by Global Camera Consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 29–38, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Zhang_Distributed_Very_Large_ICCV_2017_paper.html.

A Analytical Properties of Shared Sampling and Centrality

This appendix derives three properties of the specified computation: exact attention-pair counts, finite-population statistics of shared spatial sampling, and minimization of a centrality surrogate. The token vectors and spatial statistics are conditioned on the full image-encoder output, before the sampled aggregator is evaluated.

A.1 Exact attention-pair count

Proposition 1 (Alternating-attention pair count)**.** *Consider a window of $S \geq 1$ images with q tokens each, including special tokens. A stage consisting of one full frame-attention block and one full cross-view block evaluates $(S + S^2)q^2$ query–key pairs per attention head. For B windows in a batch, D stages, and h heads of the same token dimensions, the total is $BDh(S + S^2)q^2$.*

Proof. A frame block has S attention matrices of size $q \times q$, hence Sq^2 pairs. A cross-view block has one attention matrix of size $Sq \times Sq$, hence S^2q^2 pairs. Addition gives the stage count; multiplication by independent batch elements, stages, and heads gives the total. Substituting $q = P_\rho + 5$ and $q_0 = P + 5$ and cancelling the common factors proves Equation (2). □

The count includes ordered pairs and diagonal entries. It describes attention matrices rather than total floating-point operations or latency. Image encoding is performed before pruning, and projection, normalization, and camera-head costs remain. With $P = 1369$ and $P_{0.75} = 342$, the reduction is $(1374/347)^2 = 15.6789$. Both fixed-S counts remain quadratic in the retained token count; pruning changes their magnitude.

For windows $\mathcal{W}_k$ with $n_k \leq L$, the corresponding pair count per stage and head is $q^2 \sum_k (n_k + n_k^2) \leq q^2 W(L + L^2)$ for W windows. This expression includes overlap because each window receives independent inference. It accounts for the neural attention schedule only; scene-wide feature matching and global BA retain their own dependence on the collection size.

A.2 Expected retained spatial statistic

Let $P \geq 1$, $1 \leq K \leq P$, and choose J uniformly among the $\binom{P}{K}$ subsets of size K. Define indicators $Z_p = \mathbf{1}[p \in J]$. For any fixed collection of scalars or vectors $a_1, \ldots, a_P$, set

$$A = \sum_{p=1}^{P} a_p, \qquad A_J = \sum_{p \in J} a_p, \qquad \bar{a} = \frac{A}{P}, \qquad \widehat{a} = \frac{A_J}{K}. \tag{5}$$

Proposition 2 (Unbiased spatial sample mean)**.** *Under uniform sampling without replacement, $\mathbb{E}[A_J] = (K/P)A$ and $\mathbb{E}[\widehat{a}] = \bar{a}$. If $P > 1$ and a_p are scalar, then*

$$\mathrm{Var}(\widehat{a}) = \frac{1 - K/P}{K}\, s_a^2, \qquad s_a^2 = \frac{1}{P-1} \sum_{p=1}^{P} (a_p - \bar{a})^2. \tag{6}$$

For vectors, the same identity holds with variance replaced by covariance and s_a^2 by the finite-population covariance matrix. For $P = K = 1$, the sample mean is deterministic.

Proof. Counting subsets containing an index gives

$$\mathbb{E}[Z_p] = \frac{\binom{P-1}{K-1}}{\binom{P}{K}} = \frac{K}{P}. \tag{7}$$

Thus $\mathbb{E}[A_J] = \sum_p \mathbb{E}[Z_p] a_p = (K/P)A$, and division by K gives unbiasedness. For $p \neq r$ and $P > 1$, counting subsets containing both indices gives

$$\mathbb{E}[Z_p Z_r] = \frac{K(K-1)}{P(P-1)}, \quad \mathrm{Var}(Z_p) = \frac{K(P-K)}{P^2}, \quad \mathrm{Cov}(Z_p, Z_r) = -\frac{K(P-K)}{P^2(P-1)}. \tag{8}$$

Set $d_p = a_p - \bar{a}$, so $\sum_p d_p = 0$ and $\sum_{p \neq r} d_p d_r = -\sum_p d_p^2$. Since $\widehat{a} - \bar{a} = K^{-1} \sum_p Z_p d_p$, substitution yields

$$\mathrm{Var}(\widehat{a}) = \frac{1}{K^2} \left[\frac{K(P-K)}{P^2} + \frac{K(P-K)}{P^2(P-1)} \right] \sum_p d_p^2 \tag{9}$$

$$= \frac{P-K}{KP(P-1)} \sum_p d_p^2 = \frac{1 - K/P}{K} s_a^2. \tag{10}$$

Applying the calculation to centered outer products gives the vector result. □

Choosing $a_p = \mathbf{1}[p \in \mathcal{R}]$ for a fixed spatial region $\mathcal{R}$ proves $\mathbb{E}[|J \cap \mathcal{R}|] = K|\mathcal{R}|/P$: expected retained region occupancy is proportional to region size. Choosing a fixed feature component or fixed spatial-coordinate statistic gives the corresponding mean identity. The value $K \geq 1$ in Equation (1) ensures the mean is defined at every permitted pruning fraction.

Corollary 1 (Shared sampling across images)**.** *For fixed spatial statistics $a_{i,p}$ and a shared index set J, each image's mean is unbiased. The sampled difference between any two images is also unbiased:*

$$\mathbb{E}\left[\frac{1}{K} \sum_{p \in J} (a_{i,p} - a_{r,p}) \right] = \frac{1}{P} \sum_{p=1}^{P} (a_{i,p} - a_{r,p}). \tag{11}$$

Proof. Apply Proposition 2 to the fixed population $a_{i,p} - a_{r,p}$. The shared index set also retains the identical grid positions deterministically in every image. □

Identical grid positions need not depict corresponding scene points across views. These results apply to the fixed pre-aggregation statistic. In particular, nonlinear attention, camera regression, and BA do not inherit unbiasedness merely by composition. The sample mean above is an analytical quantity; PruneBA does not replace its camera head with that statistic.

A.3 Exact minimization of the implemented centrality surrogate

For image i, let $\mathcal{K}_i$ be its nonempty ordered set of candidate windows. Define $d_k(i) = |j_k(i) - n_k/2|$ and $c_k(i) = (1 + d_k(i))^{-1}$.

Proposition 3 (Centrality and a monotone positional surrogate)**.** *Let $a_i \geq 0$ and $b_i > 0$ be constants shared across the candidates for image i. Then maximizing $c_k(i)$ is equivalent to minimizing $b_k(i) = a_i + b_i d_k(i)$. Resolving either optimization by the earliest candidate gives the same unique selected index.*

Proof. The function $d \mapsto (1+d)^{-1}$ is strictly decreasing for $d \geq 0$, whereas $d \mapsto a_i + b_i d$ is strictly increasing. Therefore both optimizations select exactly the candidates with minimum $d_k(i)$. The candidate order resolves any equality identically. □

The surrogate is constructed from the actual implemented position rule, including the reference $n_k/2$. If a task-specific analysis supplies an error quantity $\epsilon_k(i)$ satisfying $\epsilon_k(i) \leq a_i + b_i d_k(i)$ for all candidates with common coefficients, Proposition 3 minimizes that upper bound. Such a positional error bound is an additional assumption, not an empirical conclusion from the saved reconstructions. Minimizing the bound also does not generally minimize the realized error $\epsilon_k(i)$.

For two windows spanning one-based input positions 1–10 and 8–17, the three shared inputs are 8, 9, and 10. Their first-window zero-based positions are 7, 8, and 9, with scores $1/3$, $1/4$, and $1/5$; their second-window positions are 0, 1, and 2, with scores $1/6$, $1/5$, and $1/4$. Selection therefore uses the first window for inputs 8 and 9 and the second for input 10. Window lengths include every enumerated neural input, including the historical indoor depth PNGs.

B Experimental Details and Complete Inference Measurements

B.1 Input selection and image formation

Bus Ground–UAV 30 uses the first 15 sorted ground images and first 15 sorted UAV images, while Bus Ground–UAV 63 uses all 38 ground and 25 UAV images from the same scene. The 30-image case has fifteen 5712×4284 and fifteen 4032×3024 images. The 63-image case has thirty-seven 5712×4284, one 2912×3932, and twenty-five 4032×3024 images. UAV Wide-Baseline 9 uses every fourth sorted image of an aerial sequence, with nine 8000×6000 inputs. Bus Ultra-Wide13 contains all 13 images of its ultra-wide bus-view collection.

The three indoor cases use sequence 01 of Office, Heads, and Chess. Their 14 RGB indices are 0, 1, 4, 5, 8, 9, 12, 13, 16, 17, 20, 21, 24, and 25 at 640×480 resolution. The executed 20-file lists additionally include depth PNGs at indices 2, 6, 10, 14, 18, and 22. Those files participated in neural input enumeration after RGB conversion, but feature-based camera export retains the 14 RGB images. The available source pose and depth references are not used to compute the reported sparse-model metrics.

Each input is converted to RGB, centered in a black square, resized bicubically to 1024×1024, and then bilinearly to 518×518 with unaligned interpolation corners. Image normalization uses mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. Reconstruction inference uses evaluation mode without gradients and BF16 on the RTX 4090 D. The checkpoint is loaded once per process. The decoder produces world-to-camera rotations and translations and horizontal and vertical fields of view. With original width W, height H, and $M = \max(W, H)$, the exported intrinsics are

$$f_x = \frac{M}{2\tan(\mathrm{FoV}_w/2)}, \quad f_y = \frac{M}{2\tan(\mathrm{FoV}_h/2)}, \quad c_x = W/2, \quad c_y = H/2. \tag{12}$$

Quaternions are represented in scalar-first order for COLMAP. Neural depth, dense point, tracking, and confidence outputs are not used as geometric observations.

B.2 Window and camera selection details

Using half-open, zero-based input ranges, Bus Ground–UAV 30 has windows $[0, 10)$, $[7, 17)$, $[14, 24)$, and $[21, 30)$. UAV Wide-Baseline 9 has $[0, 9)$. Each indoor case has $[0, 10)$, $[7, 17)$, and $[14, 20)$, retaining 7, 7, and 4 RGB camera records locally. Bus Ultra-Wide13 has $[0, 10)$ and $[7, 13)$. Bus Ground–UAV 63 has starts at 0, 7, 14, 21, 28, 35, 42, 49, and 56; its first eight windows have length 10 and its last length 7.

The complete DINOv2 encoder precedes pruning. Retained patch positions keep their original grid coordinates, shifted by one for the aggregator; special-token coordinates are zero. No compact grid renumbering is applied. Frame attention processes tensors of shape $[BS, P_\rho + 5, 1024]$, and global attention processes $[B, S(P_\rho + 5), 1024]$. VGGT uses 24 alternating stages. Concatenated frame and global features have shape $[B, S, P_\rho + 5, 2048]$. The camera head reads the final stage's camera token and performs four iterative refinements; its last nine-component prediction supplies the camera initialization. Separate learned embeddings distinguish the first image from subsequent images in each independently inferred window.

Each local reconstruction performs SIFT extraction, exhaustive matching, pinhole camera initialization, triangulation, and BA. A strict score improvement replaces a selected camera record, preserving the earlier window on equal centrality. Selected rotations, translations, and intrinsics initialize cameras indexed in the global image set. Local points and tracks are not concatenated; global triangulation starts with the selected cameras and an empty point set, then BA refines the global result. The same COLMAP BA settings are used locally and globally, with no supplied robust loss or learned observation weights.

B.3 Synthetic timing studies

Table 4 preserves both completed synthetic measurements. Each uses a random tensor of shape $1 \times 10 \times 3 \times 518 \times 518$ on the RTX 4090 D. The random-weight study disables depth, point, and track heads, uses three warmups and ten timed iterations, and synchronizes CUDA after every measured

Table 4: Complete synthetic ten-frame pose-path measurements. Random-weight memory is incremental peak allocation above the resident tensors; pretrained memory is absolute peak allocation including the resident model. Both inputs are random tensors. MiB uses a binary divisor.

Weights	Pruning	Patches	Time (ms)	Peak (MiB)	Speedup
Random	0%	1369	649.8	2596.1	1.00×
Random	50%	684	328.3	2181.1	1.98×
Random	75%	342	246.9	1973.6	2.63×
Random	90%	136	279.5	1858.3	2.32×
Pretrained	0%	1369	663.7	7863.9	1.00×
Pretrained	50%	684	393.5	7233.8	1.69×
Pretrained	75%	342	301.8	7106.4	2.20×
Pretrained	90%	136	280.4	7045.7	2.37×

iteration. FP16 autocast covers inference except the prediction-head portion, where the model's forward method disables autocast. Incremental peak allocation subtracts the pre-existing allocated tensors after warmup and peak reset. The recorded PyTorch build did not include FlashAttention.

The pretrained study allocates the complete model but evaluates only the image encoder, alternating aggregator, and camera head. It uses two warmups and five timed iterations, synchronization per iteration, and FP16 autocast over that path. Its memory statistic is absolute peak allocated memory, including the resident model. The two studies therefore retain distinct memory definitions and precision scopes. Binary division by 1024^2 gives MiB.

A separate pretrained component measurement records the complete DINOv2 encoder at 113.5 ms and the encoder-plus-aggregator computation at 631.8 ms unpruned and 253.3 ms with 75% pruning ($2.49\times$); the camera head is excluded. This measurement is separate from the two pose-path studies in Table 4.

B.4 Saved pruning reconstruction

The PruneBA-75 reconstruction on Bus Ground–UAV 30 records 30 cameras, 364 points, 729 observations, 1.364332 pixels internal mean error, and 46.01 seconds total. Its four locally refined windows retain 10, 10, 10, and 9 cameras and contain 21, 4, 3, and 25 points. Recorded stage times are 27.39 seconds for neural-initialization stages, 6.64 seconds for local feature extraction and matching, 0.47 seconds for local BA, 5.27 seconds for global features and matching, and 0.24 seconds for global BA. These components omit triangulation and other processing and do not sum to the total.

The unpruned progressive result records 298 points and 1.661889 pixels at 32.8 seconds. Its checkpoint was already resident after the preceding direct baselines. The separate pruned reconstruction includes checkpoint loading in its first window, whose initialization stage takes 16.9 seconds. Table 3 retains the original complete reconstruction times; Table 4 characterizes the neural pruning effect under its controlled conditions.

C Additional Context and Reconstruction Comparisons

Correspondence and geometric optimization. Learned local features and matchers complement camera priors: SuperPoint learns interest points and descriptors (DeTone et al., 2018), SuperGlue uses context and assignment optimization (Sarlin et al., 2020), LoFTR estimates detector-free correspondences (Sun et al., 2021), and LightGlue adapts matching computation to image-pair difficulty (Lindenberger et al., 2023). GLOMAP organizes camera and point estimation globally (Pan et al., 2024); distributed BA couples local problems through camera consensus (Zhang et al., 2017). BA-Net, DeepV2D, and Theseus integrate learned representations with differentiable geometric optimization (Tang & Tan, 2019; Teed & Deng, 2020; Pineda et al., 2022). GLUEMAP combines local feed-forward inference, global motion averaging, and augmented BA (Pan et al., 2026). PruneBA retains SIFT observations and couples reduced neural token computation to independently refined camera windows.

Attention representation and implementation. The Transformer and ViT establish attention over token and image-patch sequences (Vaswani et al., 2017; Dosovitskiy et al., 2021); DINOv2 supplies transferable pretrained patch features (Oquab et al., 2024). RoPE incorporates relative positional structure into attention (Su et al., 2024). EViT reorganizes tokens using attention scores (Liang et al., 2022), while Adaptive Token Sampling chooses an input-dependent subset (Fayyaz et al., 2022). FlashAttention reduces memory traffic for exact attention (Dao et al., 2022). PruneBA's shared uniform subset reduces the token dimension of alternating attention after full image encoding. Its measured acceleration therefore depends on both token counts and the executed attention implementation.

Direct and progressive geometric refinement. The contrast is visible even when a small classical fragment has many points: incremental SfM reconstructs 430 points from seven bus views and 1187 points from two aerial views. Coverage and structural support jointly characterize these outcomes.

Adding a separate global BA stage to direct triangulation yields identical reported point counts and errors on UAV Wide-Baseline 9 and Bus Ultra-Wide13. On Bus Ground–UAV 30, the independently initialized direct global BA configuration yields 96 points, compared with 99 for direct triangulation. The progressive schedule changes the camera estimates supplied to final triangulation through local geometric refinement.

Evaluation scope. Coverage counts retained RGB camera records, and internal error measures consistency with each model's own observations. Pose and depth references exist for the underlying 7-Scenes sequences, but the reported evaluation measures sparse reconstruction support and internal fit. Joint assessment with external pose accuracy, explicit gauge alignment, and matched-support residuals would extend this characterization. The real-image pruning comparison covers one case; the two synthetic timing studies characterize their respective neural configurations.

Bus reconstruction figure details. Figure 4 shows original RGB images and the actual saved PruneBA-75 sparse reconstruction for Bus Ground–UAV 30. Image details link observed features (circles) to reprojections (plus signs) and shared 3D points. Point colors come from the model, and camera frusta locate the predicted viewpoints. The full view retains all 364 points and 30 cameras; dashed regions locate magnified details with displayed point counts. The lower panel reports 729 observations, with point tracks attached to 26 of the 30 retained cameras. The model is visualized in its own coordinate frame without a ground-truth alignment. Figure 6 provides corresponding indoor views.

Centrality surrogate in the main formulation. Constructively, assign each candidate the surrogate $b_k(i) = a_i + b_i|j_k(i) - n_k/2|$, with $a_i \geq 0$ and $b_i > 0$ shared across that image's candidate windows. Equation (4) exactly minimizes $b_k(i)$. Appendix A.3 proves this equivalence and states the additional assumption needed to interpret the surrogate as an error bound.

D Additional Visual Results

The additional figures complement the main-text camera-selection, seven-case and bus-reconstruction views with complete pruning measurements and indoor reconstruction details. Synthetic timing and real-image reconstruction results remain distinct. The reconstruction views preserve all model points and camera records in the full views, with details linking image observations to local point subsets. Reference camera motion for the indoor sources is displayed separately from reconstructed geometry.

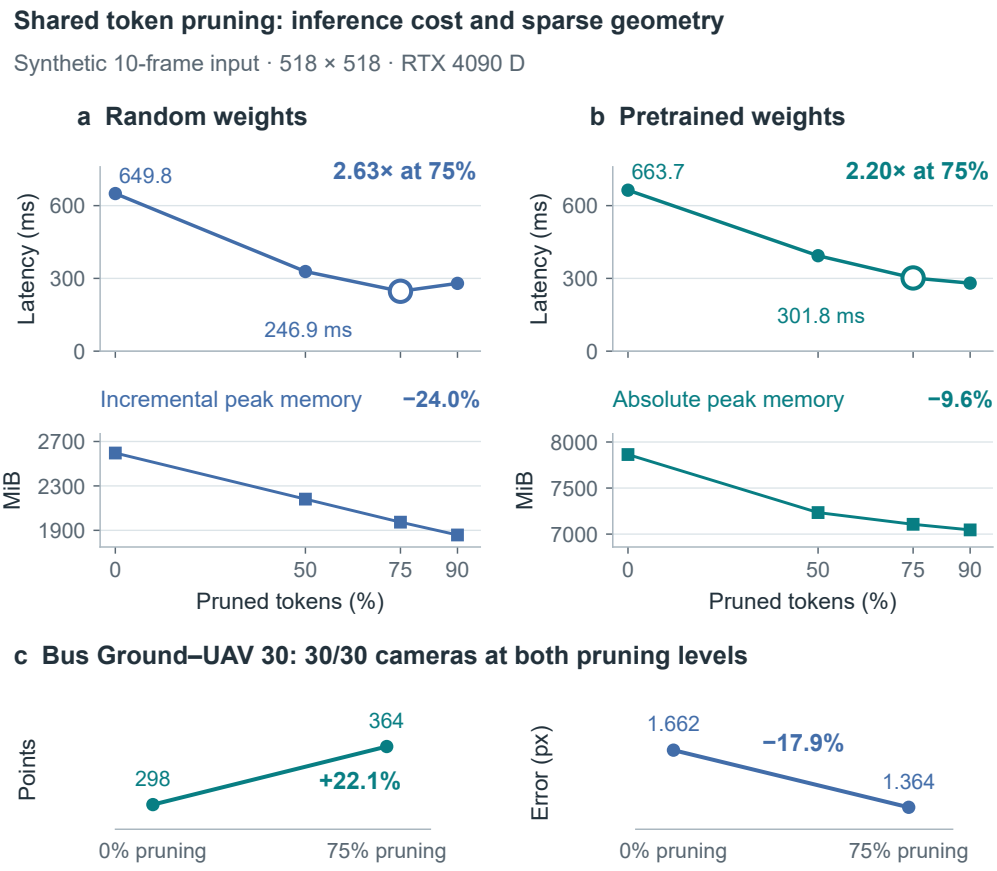

Figure 5: Token pruning reduces neural computation and remains compatible with sparse reconstruction. Synthetic ten-frame timing and allocation measurements are distinct from the saved real-image Bus Ground–UAV 30 results. Random-weight measurements use incremental allocation; checkpoint-loaded measurements use absolute allocation. At 75% pruning, the saved reconstruction retains 30/30 cameras with 364 points and 1.364 pixels internal error.

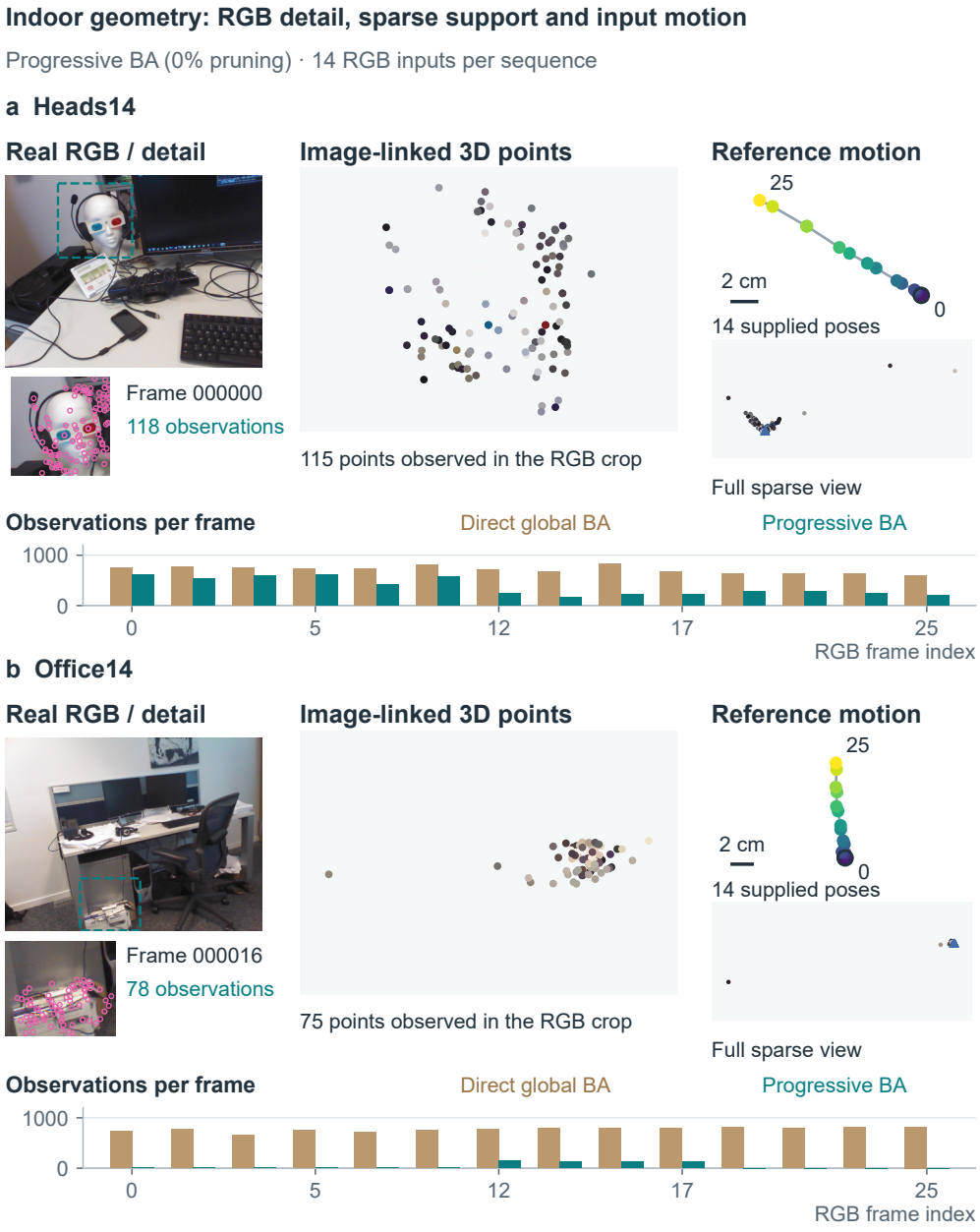

Figure 6: Original RGB images, image-linked sparse points, and observation support for Heads14 and Office14 under unpruned progressive BA. The full sparse views retain all saved points, while RGB crops identify local point subsets with displayed counts. Observation plots compare direct global BA with progressive BA for each RGB frame. Reference-motion panels show the 14 supplied source poses separately from reconstructed geometry; they are not an alignment or a pose-error evaluation. Each sparse model retains its own coordinate frame.