Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37027

Keyframe Sampling under Budget and Image Corruption for VGGT-Based 3D Reconstruction: A Controlled Comparison

Yunkai Du

Zhejiang University-University of Illinois Urbana-Champaign Institute, Zhejiang University, Haining, China
yunkai.23@intl.zju.edu.cn

Abstract. Feed-forward 3D reconstruction makes multi-view geometry accessible, but a video stream often contains more frames than a fixed memory or latency budget can admit. This paper presents a controlled comparison of keyframe sampling for the Visual Geometry Grounded Transformer (VGGT). Five ordinary strategies and one redundancy stress condition are tested on a 116-frame TartanAir trajectory at budgets of 4, 8, and 16 images. Reconstructions are registered to a fixed reference cloud through Sim(3) alignment and evaluated with accuracy, completeness, Chamfer-L1 distance, and F-score. No simple strategy dominates every budget on this trajectory. Sharpness-aware temporal sampling gives the highest observed ordinary F-score at budget 4, but its margin over the random mean lies within the random baseline's standard deviation and is not a deterministic advantage. Appearance diversity gives the highest observed ordinary F-score at budgets 8 and 16. Across three mixed-corruption seeds, its mean F-score falls by 84.7%, whereas sharpness-aware sampling selects no corrupted frames. Threshold and reference-order checks expose further evaluation sensitivity. These conclusions remain limited to one synthetic trajectory and do not establish a new online selector.

Keywords: keyframe sampling, VGGT, equal-budget evaluation, corruption sensitivity, reference-frame sensitivity

1. Introduction

Aerial and mobile mapping systems collect dense image streams with many redundant neighbors. Processing every frame increases memory and latency without guaranteeing new geometry. The VGGT predicts cameras, depth, point maps, and correspondences from image sets, but its output depends on the admitted views [1]. SwiftMap addresses related resource pressure through hierarchical content and token selection [2].

A keyframe is a retained image that represents useful information from a longer stream. Nearby frames provide overlap for correspondence and alignment. Wider spacing improves coverage and parallax but can reduce compatibility. Dark, blurred, or texture-poor frames consume the same budget as clear images. A rule based on one signal may therefore fail when the budget changes.

Comparisons can hide this interaction by changing image counts, reconstruction settings, or metrics. Camera-trajectory error does not measure surface coverage. Point-cloud accuracy can

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37027

reward a precise local fragment while the remaining flight is unseen. Reference-frame handling adds variation because feed-forward multi-view models require an internal coordinate frame.

This paper asks how transparent keyframe strategies behave when VGGT receives the same number of images. It compares uniform, boundary-anchored random, appearance-diversity, translation-path, and sharpness-aware sampling. Consecutive frames provide a redundancy stress condition. Corruption and reference-order checks extend the comparison without claiming an implemented adaptive selector.

The study contributes an equal-budget matrix with repeated random trials, a multi-seed image-corruption test, and a separate reference-frame control. Together, these components identify requirements for later online adaptation.

2. Related work

2.1. Keyframes in visual SLAM

Classical simultaneous localization and mapping (SLAM) systems retain sparse frames for tracking and optimization. Oriented FAST and Rotated BRIEF (ORB)-SLAM3 inserts keyframes in feature-based visual and visual-inertial maps [3]. Direct Sparse Odometry selects and marginalizes frames while optimizing photometric error [4]. These decisions serve a back end and are not equivalent to selecting one fixed VGGT input set.

Photogrammetric Key-frame Selection (PKS) replaces heuristic thresholds with photogrammetric geometry, adaptive thresholds, effective-point distribution constraints, and inertial cues in an ORB-SLAM3-based system [5]. Adaptive work also combines signals and refines thresholds during rapid motion [6]. The present study removes the iterative back end and fixes the admitted frame count so that each simple signal can be examined under the same reconstructor.

2.2. Feed-forward geometry and view choice

DUSt3R predicts point maps from unconstrained image pairs and aligns them for multi-view geometry [7]. VGGT predicts several geometric attributes within one transformer [1]. These models simplify the pipeline, but unsuitable overlap, redundancy, distractors, and input limits still affect output.

Keyframe-Based Feed-Forward Visual Odometry learns a policy for a VGGT-based odometry pipeline and reports problems from redundancy or insufficient parallax [8]. Emergent Outlier View Rejection uses intermediate VGGT features to detect irrelevant views [9]. Outlier rejection and keyframe selection overlap, although a weak keyframe may remain relevant but redundant, blurred, or badly spaced. The baselines here test these cases without reproducing a learned policy.

2.3. Evaluation gap

Published systems usually optimize end-to-end performance, whereas a controlled comparison must separate selection quality from implementation sensitivity. Equal frame counts support budget claims. Accuracy and completeness jointly support mapping claims. Reference handling must also be controlled because presenting a different image first can change a fixed-set result.

DOI: 10.54254/2755-2721/2026.37027

3. Methodology

3.1. Dataset and experimental unit

The quantitative experiment uses the synthetic TartanAir dataset [10]. The ArchVizTinyHouseDay Data_easy P000 left-front-camera sequence contains 116 synchronized 640 x 640 Red, Green, Blue (RGB) images, dense depth, and poses. Its path is approximately 12.97 m and ends near its start. Every clean strategy selects from this same window.

EuRoC Micro Aerial Vehicle (MAV) supplies real drone imagery and motion ground truth [11]. An eight-frame MH_01_easy run confirmed that grayscale images pass through the VGGT implementation and yield a recognizable factory reconstruction. It remains qualitative because EuRoC lacks the matched dense-depth protocol used for the TartanAir tables.

3.2. Keyframe strategies

Uniform sampling rounds equally spaced temporal positions. Boundary-anchored random sampling keeps the first and last images and draws the remainder without replacement. Seeds 0 through 4 provide a mean and sample standard deviation for every budget.

Appearance diversity is an offline image-only baseline. Each RGB image is resized to 64 x 64 and flattened. Mean absolute appearance distance defines dissimilarity. Farthest-point sampling starts with the endpoints and repeatedly maximizes the distance to the retained set. It favors novelty without explicitly preserving geometric overlap.

Path-uniform sampling matches equal cumulative-translation targets to the nearest unique frames. It uses provided poses as an offline diagnostic, not as a deployable vision-only policy. An online version would require estimated motion.

Sharpness-uniform sampling anchors the endpoints. It divides the interior timeline into B-2 equal bins and selects the highest Laplacian-variance image in each bin. Laplacian variance is an edge-energy sharpness score. Temporal bins prevent all clear images from clustering in one interval.

The consecutive condition keeps frames 0 through B-1. It maximizes local overlap but minimizes full-window coverage, so it is a stress condition rather than an ordinary competitor. All six conditions use budgets of 4, 8, and 16.

3.3. Reconstruction and fixed-ground-truth evaluation

Each selected set is processed with the official VGGT-1B checkpoint at 518 x 518 resolution. Inference uses bfloat16 automatic mixed precision on an NVIDIA GeForce RTX 5090 Laptop Graphics Processing Unit. Every run records its selection, seed, software, timing, memory, metric settings, and output path.

A reference cloud is built from all 116 depth maps and poses. Valid depth is 0.1-20.0 m. Pixels are sampled with stride 4, transformed to the TartanAir world frame, and voxelized at 0.05 m to produce 74,532 points. Predicted points retain the top half of depth-confidence values before the same voxelization.

VGGT has an arbitrary similarity gauge. Predicted camera centers are aligned to their ground-truth centers with a Sim(3) transform that estimates scale, rotation, and translation. The transform is then applied to the predicted cloud. Accuracy is the mean prediction-to-reference nearest-neighbor distance. Completeness reverses the direction. Chamfer-L1 is their sum. Precision and recall use 0.10 m as the primary nearest-neighbor threshold, with 0.05 and 0.15 m used for a sensitivity check,

DOI: 10.54254/2755-2721/2026.37027

and their harmonic mean is the F-score. The threshold check reuses the same saved Sim(3)-aligned point clouds and does not rerun VGGT. Lower distances and higher F-scores are better.

3.4. Sensitivity factors

The clean matrix contains 30 VGGT runs: three runs for each deterministic condition and 15 random runs. A mixed-corruption extension adds 12 runs at B=8. Three fixed 25% corruption patterns use seeds 20260822, 20260823, and 20260824. Each pattern degrades 29 interior candidates, alternating a 15-pixel horizontal blur with a 0.30 intensity multiplier. Uniform, appearance-diversity, sharpness-uniform, and path-uniform sampling use each stream, and every manifest records which corrupted frames were retained.

Reference-order sensitivity uses three fixed eight-frame sets: P000 discovery, held-out P000, and cross-trajectory P001. Each image becomes the first input once. Five controls keep the default first image and shuffle only the tail. Best-to-worst reference ratios are compared with tail-shuffle ranges. Figure 1 summarizes the design.

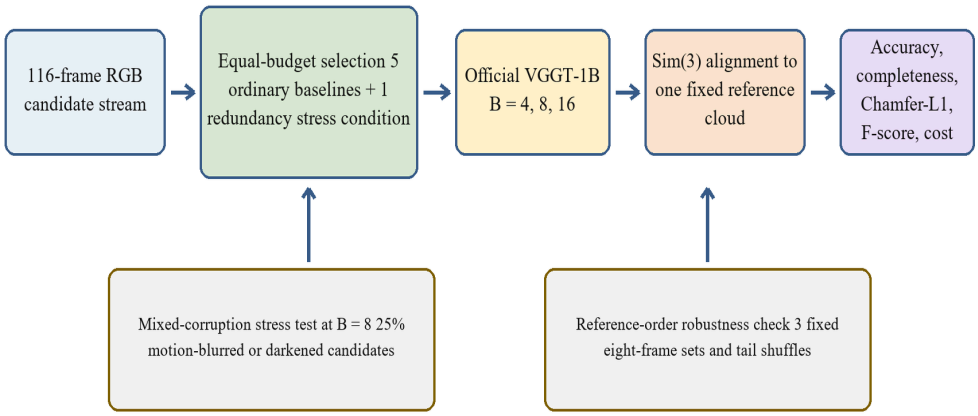

Figure 1. Multi-factor evaluation workflow (photo credit: original)

4. Results and analysis

4.1. Clean equal-budget comparison

Table 1. Clean-condition reconstruction results

Strategy	B	Accuracy	Completeness	Chamfer-L1	F-score@0.10
Uniform	4	0.4988	5.8851	6.3839	0.0139
Random mean	4	0.2820	4.8215	5.1035	0.0396 +/- 0.0263
Appearance diversity	4	0.2893	6.2341	6.5233	0.0075
Path uniform	4	0.7023	5.1955	5.8978	0.0221
Sharpness uniform	4	0.2897	4.1937	4.4833	0.0475
Consecutive stress	4	0.1609	4.5439	4.7048	0.0671
Uniform	8	0.2563	4.8312	5.0875	0.0364

DOI: 10.54254/2755-2721/2026.37027

Table 1. (continued)

Random mean	8	0.2964	5.1060	5.4023	0.0358 +/- 0.0128
Appearance diversity	8	0.2903	4.6519	4.9422	0.0714
Path uniform	8	0.3190	5.7885	6.1075	0.0121
Sharpness uniform	8	0.3400	5.8255	6.1655	0.0111
Consecutive stress	8	0.0628	4.3995	4.4623	0.1768
Uniform	16	0.2936	5.3495	5.6431	0.0500
Random mean	16	0.2921	5.1686	5.4607	0.0451 +/- 0.0307
Appearance diversity	16	0.2726	4.4647	4.7373	0.1079
Path uniform	16	0.2802	5.8982	6.1784	0.0277
Sharpness uniform	16	0.2605	5.8921	6.1527	0.0325
Consecutive stress	16	0.0687	4.3703	4.4390	0.1782

Table 1 shows a budget-dependent ranking. At B=4, sharpness-uniform sampling has the highest observed ordinary F-score, 0.0475, and the lowest ordinary Chamfer-L1, 4.4833 m. The random mean is 0.0396 with a standard deviation of 0.0263. The gap lies within random variation and is not a deterministic advantage. Appearance diversity is lowest at 0.0075, suggesting that four widely separated views can lose necessary overlap.

Random entries in Table 1 are five-seed means, and their F-score cells report sample standard deviations. Consecutive sampling remains a stress condition rather than a full-window competitor.

At B=8, appearance diversity gives the highest ordinary F-score, 0.0714, versus 0.0364 for uniform and 0.0358 for random. It also gives the best ordinary completeness and Chamfer-L1 at this budget. Path-uniform and sharpness-uniform reach only 0.0121 and 0.0111, showing that motion spacing or individual sharpness alone does not guarantee a compatible set.

At B=16, appearance diversity remains highest at 0.1079. Uniform, random, sharpness-uniform, and path-uniform reach 0.0500, 0.0451, 0.0325, and 0.0277. Table 2 lists the leading ordinary strategy at each budget under the three evaluation thresholds. Sharpness-uniform leads at B=4 with F-scores of 0.019816, 0.047541, and 0.078638 at 0.05, 0.10, and 0.15 m. Appearance diversity leads at B=8 with 0.033737, 0.071393, and 0.100770, and at B=16 with 0.062808, 0.107877, and 0.158830. Figure 2 shows the budget interaction and random variability.

Table 2. F-score sensitivity of the leading ordinary strategy at each budget

Threshold (m)	Sharpness uniform, B=4	Appearance diversity, B=8	Appearance diversity, B=16
0.05	0.019816	0.033737	0.062808
0.10	0.047541	0.071393	0.107877
0.15	0.078638	0.100770	0.158830

These values are recomputed from the same saved Sim(3)-aligned point clouds; the sensitivity check changes only the nearest-neighbor threshold and does not rerun VGGT.

DOI: 10.54254/2755-2721/2026.37027

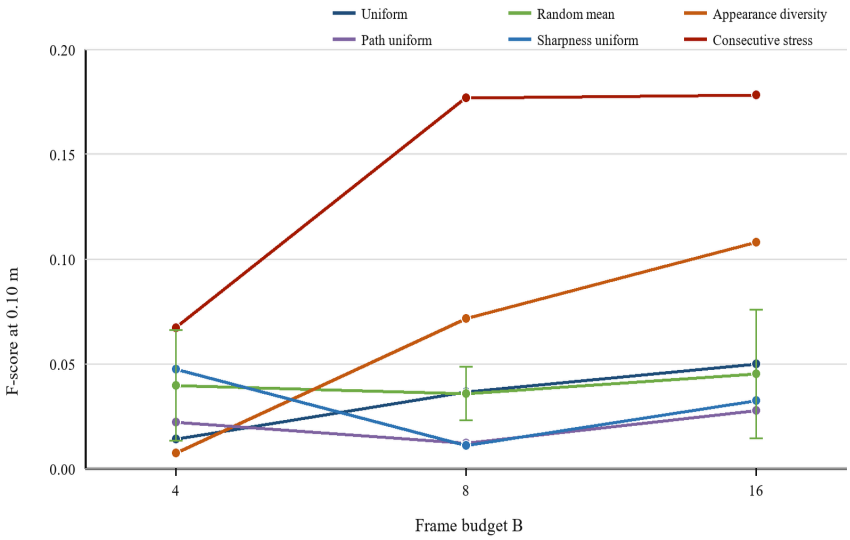

Figure 2. Equal-budget F-score trends (photo credit: original)

4.2. Why is the consecutive result not a winning selector

Consecutive frames produce the highest numerical F-scores: 0.0671, 0.1768, and 0.1782. Dense overlap supports stable local alignment, but these sets span only 2.6%, 6.1%, and 13.0% of candidate-frame intervals. They omit most later viewpoints, so a long non-revisiting route would leave later regions unmapped after the budget was exhausted.

This outcome exposes a metric limitation rather than a deployable policy. Nearest-neighbor measures can reward a precise local cloud despite weak temporal coverage. The full reference penalizes missing geometry through completeness, yet repeated structure and loop closure can still favor a prefix.

4.3. Mixed-corruption sensitivity

Table 3 aggregates three corruption seeds, each degrading 29 of 116 candidates. Percentage changes are calculated from unrounded source values, and F-scores are displayed to six decimal places for reproducibility. Appearance diversity selects one or two corrupted images, and its mean F-score falls from 0.071393 to 0.010948, an 84.7% loss. Uniform selects zero to two corrupted images and averages 0.027497, below its clean score of 0.036408. A dark or blurred frame can appear novel, so the appearance distance may lead to a harmful observation.

Table 3. Three-seed mixed-corruption sensitivity at B=8

Strategy	Selected corrupt, mean (range)	Clean F-score	Corrupt F-score, mean +/- SD	Change from clean	Corrupt Chamfer, mean (m)
Uniform	1.33 (0-2)	0.036408	0.027497+/-0.009165	-24.5%	5.4099
Appearance diversity	1.33 (1-2)	0.071393	0.010948+/-0.002789	-84.7%	6.3284

DOI: 10.54254/2755-2721/2026.37027

Table 3. (continued)

Sharpness uniform	0.00 (0-0)	0.011073	0.018107+/- 0.006351	+63.5%	6.1688
Path uniform	1.00 (0-2)	0.012120	0.012252+/- 0.002077	+1.1%	6.1102

Sharpness-uniform selects no corrupted frame in any seed. Its mean F-score is 0.018107 +/- 0.006351, above its clean value of 0.011073 by 63.5%. This increase does not show that corruption helps; quality changes redirect selection, and some replacements happen to be more compatible. Path-uniform averages 0.012252 +/- 0.002077 and sometimes includes corrupted frames. Three seeds reduce dependence on one pattern but do not establish cross-scene robustness.

4.4. Reference-order sensitivity

Table 4 reports reference-order sensitivity with the selected images fixed. Best-to-worst F-score ratios are 1.83 for P000 discovery, 3.92 for held-out P000, and 8.34 for P001. Tail-only shuffling changes F-score by just 0.0005, 0.0008, and 0.0001 across the three sets. Reference choice, therefore, dominates ordinary tail-order variation in these controls. These trials are not keyframe baselines because the frame sets do not change. They isolate an implementation factor that could otherwise be mistaken for selector quality. A production pipeline should define the reference deterministically or apply a validated reference-selection rule.

Table 4. Reference-frame and tail-order sensitivity at B=8

Fixed eight-frame set	Best/worst reference F-score ratio	Tail-shuffle F-score range
P000 discovery	1.83	0.0363-0.0368
P000 held-out frames	3.92	0.0239-0.0247
P001 cross-trajectory	8.34	0.0238-0.0239

4.5. Cost, interpretation, and limitations

Peak allocated memory depends on B rather than selector identity: 7.762 GiB at B=4, 9.106 GiB at B=8, and 9.518 GiB at B=16. Clean inference ranges from about 0.9-1.3 s at B=4 to 3.2-3.9 s at B=16. Once B is fixed, selection changes quality without changing the VGGT tensor size.

All quantitative selection results use one synthetic trajectory, and corruption is applied to that sequence. Random seeds quantify sampling and corruption variation, not scene variation. Path-uniform uses ground-truth translation, while appearance diversity sees the full window. Sim(3) registration may favor locally consistent subsets. Confidence filtering, voxel size, and depth range remain fixed. EuRoC supplies only a qualitative smoke test.

The paper is therefore a controlled comparative baseline, not evidence of universal superiority or a finished AdaMap method. Its protocol combines equal budgets, dense geometry metrics, repeated trials, corruption, threshold checks, and reference controls. The next step is an online selector that balances coverage, overlap, quality, and motion on held-out synthetic and real drone sequences.

DOI: 10.54254/2755-2721/2026.37027

5. Conclusion

This study compared five transparent keyframe strategies and one redundancy stress condition for VGGT under equal frame budgets. The 30-run clean matrix shows that the observed ranking depends on capacity. Sharpness-aware temporal sampling has the highest ordinary F-score at B=4. However, its margin over the random mean is smaller than the random baseline's standard deviation, so the result is not a deterministic advantage. Appearance diversity has the highest observed ordinary F-score at B=8 and B=16. Recalculation at 0.05, 0.10, and 0.15 m preserves these leading strategies. Consecutive frames score strongly because dense overlap supports a coherent local cloud, but their narrow temporal span prevents a global mapping interpretation. Across three mixed-corruption seeds, appearance diversity loses 84.7% of its clean F-score on average. Sharpness-uniform sampling selects no corrupted frames, although its reconstruction score still varies across patterns. The fixed-set control also shows that changing the reference image affects F-score far more than shuffling the remaining inputs. These findings support evaluation and design requirements rather than a new selector: equal budgets, surface completeness, image-quality checks, repeated corruption trials, threshold sensitivity, and deterministic reference handling should be considered together. The evidence is limited to one synthetic trajectory plus a qualitative EuRoC smoke test. Held-out environments, real drone sequences, and an implemented online policy are required before broader performance or publication claims.

References

[1] Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., & Novotny, D. (2025). VGGT: Visual geometry grounded transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5294–5306.

[2] Xu, J., Chen, X., Bala, M., Eiszler, T., Chanana, A., Harkes, J., Pillai, P., & Satyanarayanan, M. (2026). Towards fast and fully automatic drone mapping. *Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services*, 418–431.

[3] Campos, C., Elvira, R., Gómez Rodríguez, J. J., Montiel, J. M. M., & Tardós, J. D. (2021). ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM. *IEEE Transactions on Robotics, 37*(6), 1874–1890.

[4] Engel, J., Koltun, V., & Cremers, D. (2018). Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(3), 611–625.

[5] Azimi, A., Hosseininaveh Ahmadabadian, A., & Remondino, F. (2022). PKS: A photogrammetric key-frame selection method for visual-inertial systems built on ORB-SLAM3. *ISPRS Journal of Photogrammetry and Remote Sensing, 191*, 18–32.

[6] Chen, H., Wang, B., Gu, D., & Ye, W. (2025). A novel adaptive keyframe selection method with multi-source joint constraints for visual SLAM. *Intelligent Service Robotics, 18*(3), 513–527.

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

[8] Dai, W., Su, W., Kong, D., Ming, Y., & Kong, W. (2026). Keyframe-based feed-forward visual odometry. *arXiv preprint arXiv:2601.16020*.

[9] Han, J., Hong, S., Jung, J., Jang, W., An, H., Wang, Q., Kim, S., & Feng, C. (2026). Emergent outlier view rejection in visual geometry grounded transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 427–437.

[10] Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., & Scherer, S. (2020). TartanAir: A dataset to push the limits of visual SLAM. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 4909–4916.

[11] Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M. W., & Siegwart, R. (2016). The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research, 35*(10), 1157–1163.