# AVGGT: Rethinking Global Attention for Accelerating VGGT

## Supplementary Material

Table 1. Camera pose estimation on Sintel [2]. Best and second best are highlighted within each baseline block, excluding the baseline row.

| Baseline | Method | ATE ↓ | RPE trans ↓ | RPE rot ↓ | Time (s) ↓ |
|---|---|---|---|---|---|
| $\pi^3$ | $\pi^3$ | 0.073 | 0.038 | 0.288 | 0.960 |
| | Faster_$\pi^3$_25 | **0.076** | **0.044** | 0.318 | 1.020 |
| | Faster_$\pi^3$_75 | 0.096 | 0.071 | 0.508 | 0.809 |
| | A_$\pi^3$(2) | 0.091 | 0.046 | **0.304** | 0.725 |
| | A_$\pi^3$(4) | 0.127 | 0.061 | 0.345 | **0.687** |
| VGGT | VGGT | 0.169 | 0.061 | 0.476 | 1.348 |
| | FastVGGT | **0.169** | 0.072 | 0.525 | 1.584 |
| | FasterVGGT_25 | 0.171 | **0.063** | 0.505 | 1.348 |
| | FasterVGGT_75 | 0.199 | 0.095 | 0.847 | 1.107 |
| | AVGGT(2) | 0.183 | 0.068 | **0.497** | 1.069 |
| | AVGGT(4) | 0.199 | 0.087 | 0.539 | **1.006** |

Table 2. Camera pose estimation on ETH3D [12]. Best and second best are highlighted within each baseline block, excluding the baseline row.

| Baseline | Method | Racc@5 ↑ | Tacc@5 ↑ | AUC@5 ↑ | Racc@15 ↑ | Tacc@15 ↑ | AUC@15 ↑ | Racc@30 ↑ | Tacc@30 ↑ | AUC@30 ↑ | Time (s) ↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\pi^3$ | $\pi^3$ | 99.471 | 86.772 | 67.566 | 100.000 | 96.296 | 85.538 | 100.000 | 98.148 | 91.455 | 1.307 |
| | Faster_$\pi^3$_25 | **99.735** | **80.159** | **60.106** | **100.000** | 95.767 | **81.446** | **100.000** | 97.884 | **89.356** | 1.278 |
| | Faster_$\pi^3$_75 | 96.032 | 67.989 | 42.222 | **100.000** | 91.799 | 70.176 | **100.000** | **98.677** | 83.254 | 0.991 |
| | A_$\pi^3$(2) | 98.413 | 79.630 | 52.857 | **100.000** | **96.561** | 78.836 | **100.000** | 98.413 | 88.228 | 0.955 |
| | A_$\pi^3$(4) | 96.825 | 68.783 | 43.810 | **100.000** | 93.122 | 71.693 | **100.000** | 97.619 | 84.171 | **0.871** |
| VGGT | VGGT | 100.000 | 79.630 | 57.937 | 100.000 | 98.413 | 81.993 | 100.000 | 99.471 | 90.503 | 1.766 |
| | FastVGGT | 96.032 | 73.280 | 39.524 | **100.000** | 97.090 | 74.321 | **100.000** | 98.677 | 86.305 | 1.823 |
| | FasterVGGT_25 | **100.000** | **81.481** | **57.407** | **100.000** | **98.413** | **82.028** | **100.000** | 99.206 | 90.538 | 1.746 |
| | FasterVGGT_75 | **100.000** | 64.815 | 38.677 | **100.000** | 93.915 | 68.536 | **100.000** | 98.942 | 83.016 | 1.372 |
| | AVGGT(2) | **100.000** | 79.894 | 56.667 | **100.000** | 98.148 | 81.834 | **100.000** | **100.000** | **90.688** | 1.344 |
| | AVGGT(4) | **100.000** | 76.190 | 53.492 | **100.000** | 97.619 | 79.506 | **100.000** | 99.735 | 89.259 | **1.207** |

Table 3. Camera pose estimation on ScanNet [4]. Best and second best are highlighted within each baseline block, excluding the baseline row.

| Baseline | Method | ATE ↓ | RPE trans ↓ | RPE rot ↓ | Time (s) ↓ |
|---|---|---|---|---|---|
| $\pi^3$ | $\pi^3$ | 0.030 | 0.013 | 0.346 | 5.718 |
| | Faster_$\pi^3$_25 | **0.030** | **0.013** | **0.347** | 5.175 |
| | Faster_$\pi^3$_75 | 0.038 | 0.014 | 0.388 | 3.204 |
| | A_$\pi^3$(2) | 0.032 | **0.013** | 0.355 | 2.801 |
| | A_$\pi^3$(4) | 0.033 | **0.013** | 0.365 | **2.403** |
| VGGT | VGGT | 0.035 | 0.015 | 0.376 | 7.959 |
| | FastVGGT | 0.036 | 0.017 | 0.414 | 4.923 |
| | FasterVGGT_25 | **0.035** | **0.015** | **0.379** | 7.238 |
| | FasterVGGT_75 | 0.041 | 0.018 | 0.479 | 4.586 |
| | AVGGT(2) | 0.036 | 0.016 | 0.384 | 4.376 |
| | AVGGT(4) | 0.036 | 0.016 | 0.391 | **3.755** |

Table 4. Point map estimation on NRGBD [1]. Best and second best are highlighted within each baseline block, excluding the baseline row.

| Baseline | Method | Acc. ↓ | | Comp. ↓ | | NC. ↑ | | Time (s) ↓ |
|---|---|---|---|---|---|---|---|---|
| | | Mean | Med. | Mean | Med. | Mean | Med. | |
| $\pi^3$ | $\pi^3$ | 0.012 | 0.006 | 0.013 | 0.005 | 0.768 | 0.870 | 0.453 |
| | Faster_$\pi^3$_25 | **0.014** | **0.008** | **0.014** | **0.006** | **0.749** | **0.865** | 0.493 |
| | Faster_$\pi^3$_75 | 0.050 | 0.031 | 0.028 | 0.012 | 0.700 | 0.848 | 0.451 |
| | A_$\pi^3$(2) | 0.019 | 0.012 | 0.016 | 0.007 | 0.738 | **0.865** | 0.411 |
| | A_$\pi^3$(4) | 0.028 | 0.019 | 0.020 | 0.009 | 0.732 | 0.864 | **0.398** |
| VGGT | VGGT | 0.013 | 0.007 | 0.015 | 0.006 | 0.784 | 0.877 | 0.619 |
| | FastVGGT | 0.020 | 0.012 | 0.019 | 0.009 | 0.597 | 0.662 | 1.053 |
| | FasterVGGT_25 | **0.014** | **0.007** | **0.016** | **0.006** | 0.776 | 0.875 | 0.637 |
| | FasterVGGT_75 | 0.054 | 0.030 | 0.050 | 0.027 | 0.714 | 0.851 | 0.604 |
| | AVGGT(2) | **0.014** | **0.007** | **0.016** | **0.006** | **0.781** | **0.876** | 0.559 |
| | AVGGT(4) | 0.015 | 0.008 | 0.017 | **0.006** | 0.776 | 0.875 | **0.555** |

Table 5. Ablation results of not subsampling Query tokens for AVGGT on RealEstate10K [16].

| Method | AUC@5 ↑ | AUC@15 ↑ | AUC@30 ↑ |
|---|---|---|---|
| VGGT | 63.176 | 81.103 | 88.130 |
| AVGGT(4) | 59.583 | 79.188 | 87.045 |
| AVGGT(2+Q2Near) | 42.565 | 66.684 | 78.310 |
| AVGGT(2+Q2GM) | 26.973 | 53.198 | 68.382 |

Table 6. Ablation results of the subsampling strategy for AVGGT on RealEstate10K [16].

| Method | AUC@5 ↑ | AUC@15 ↑ | AUC@30 ↑ |
|---|---|---|---|
| AVGGT(2) | 61.959 | 80.443 | 87.758 |
| AVGGT(2_SIFT) | 55.438 | 77.415 | 86.034 |
| AVGGT(2_Random) | 57.849 | 78.233 | 86.441 |
| AVGGT(2_High) | 61.220 | 79.982 | 87.448 |
| AVGGT(2_Low) | 59.622 | 79.493 | 87.341 |
| AVGGT(2_Mean) | 56.990 | 77.491 | 85.886 |

# 1. More Analysis

## 1.1. Analyzing $\pi^3$ Global Attention

Following the analysis conducted for VGGT [14] in the main paper, we first briefly introduce the $\pi^3$ [15] architecture. Given multiple input images, $\pi^3$ uses a frozen DINOv2 [8] encoder. For each frame, five learnable register tokens [5] are appended to the patch tokens. All frame tokens are then passed through an aggregator composed of 36 transformer blocks that alternate between frame attention and global attention. After aggregation, all register tokens are removed, and the remaining tokens are fed into the camera and point heads to predict camera poses and point clouds. Unlike VGGT, $\pi^3$ discards camera tokens and adopts a fully permutation-equivariant architecture with respect to the input frames.

We visualize all global attention layers (indices 0-17) in Fig. 5 and Fig. 6, and analyze four representative layers (indices 1, 3, 11, and 17). Overall, the observations are highly consistent with those for VGGT. In the early global attention layers (indices 0-9), the maximum attention values are significantly smaller than those in the middle layers, indicating a more uniform distribution. The top activated entries reveal two characteristic patterns: in the very first layers (*e.g.*, layers 0-1), attention is dominated by po-


Figure 1. Rotation test on highly activated key-token subsets in VGGT. All input views are rotated by $180^\circ$ and fed through VGGT, and the outputs are rotated back for visualization. For each layer, we first collect the top 1000 attention entries in both the original and rotated runs, then select the top 50 entries that share the same query patch (following the ranking in the original run). Arrows start at the query patch and end at the corresponding key patch.

Table 7. Ablation results of the diagonal preservation for AVGGT on RealEstate10K [16] and 7-Scenes [13].

| Method | RealEstate10K (Sparse) | | | 7-Scenes (Dense) | | |
|---|---|---|---|---|---|---|
| | AUC@5 ↑ | AUC@15 ↑ | AUC@30 ↑ | AUC@5 ↑ | AUC@15 ↑ | AUC@30 ↑ |
| AVGGT(2) | 61.959 | 80.443 | 87.758 | 26.061 | 61.951 | 78.113 |
| AVGGT(2_WithDiagonal) | 62.039 | 80.647 | 87.926 | 24.532 | 60.551 | 77.247 |
| AVGGT(2_WithMean) | 59.427 | 79.030 | 86.930 | 26.227 | 62.171 | 78.170 |
| AVGGT(2_SubsampleOnly) | 61.908 | 80.539 | 87.838 | 24.539 | 60.564 | 77.262 |

Table 8. Ablation results on RealEstate10K [16] for keeping the first frame tokens in VGGT.

| Method | AUC@5 ↑ | AUC@15 ↑ | AUC@30 ↑ |
|---|---|---|---|
| AVGGT(2) | 61.959 | 80.443 | 87.758 |
| AVGGT(2_FullySubsample) | 60.931 | 79.962 | 87.499 |
| AVGGT(4) | 59.583 | 79.188 | 87.045 |
| AVGGT(4_FullySubsample) | 55.570 | 76.675 | 85.446 |

Table 9. Ablation results on RealEstate10K [16] for $\pi^3$.

| Method | AUC@5 ↑ | AUC@15 ↑ | AUC@30 ↑ |
|---|---|---|---|
| $\pi^3$ | 67.186 | 83.288 | 89.500 |
| $\pi^3$(G2F) | 66.790 | 83.006 | 89.313 |
| $\pi^3$(G2M) | 63.166 | 80.593 | 87.766 |
| A_$\pi^3$(2) | 64.832 | 82.306 | 89.008 |
| A_$\pi^3$(4) | 58.703 | 79.350 | 87.434 |
| A_$\pi^3$(6) | 51.780 | 75.037 | 84.718 |
| A_$\pi^3$(9) | 40.437 | 65.927 | 78.060 |

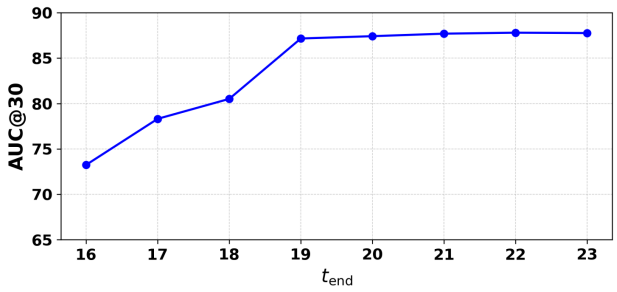


Figure 2. Ablation results on RealEstate10K for different VGGT $t_{end}$ choices.

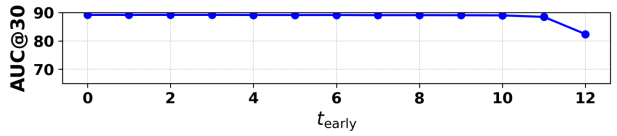


Figure 3. Ablation results on RealEstate10K for different $\pi^3$ $t_{early}$ choices.

sitional embeddings, whereas in layers 2-9, attention frequently focuses on a small and inconsistent subset of key tokens. These behaviors suggest that the early global attention layers contribute little to establishing cross-view correspondences.

In the middle global attention layers (indices 10-16), the attention becomes far more selective, with noticeably larger peak values. The highest responses predominantly correspond either to self-attention on the same patch or to cross-view patches at the same spatial location, indicating that these layers are responsible for building cross-view correlations. Finally, the last global attention layer (index 17) again tends toward a more uniform distribution. However,

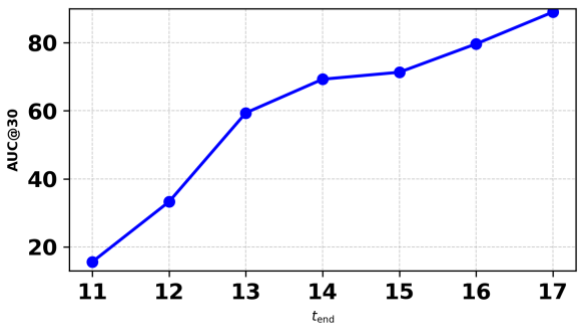


Figure 4. Ablation results on RealEstate10K for different $\pi^3$ $t_{end}$ choices.

unlike in VGGT, the top-activated token pairs in this layer still exhibit clear 3D-related structure. Therefore, in $\pi^3$, we consider that the last global attention layer still contributes to building cross-view correspondences.

### 1.2. Rotation Test for Early Global Attention Layers

For the early global attention layers in VGGT/$\pi^3$, the attention matrices exhibit nearly uniform distributions, and the top activated entries frequently attend to a small subset of key tokens. To investigate whether this subset encodes any meaningful 3D or view-consistent information, we conduct the following test. Instead of directly feeding the original input images, we rotate all input views by $180^\circ$ and then pass them through VGGT. This operation is equivalent to rotating the camera while keeping the underlying 3D scene unchanged.

As illustrated in Fig. 1, we first collect the top 1000 attention entries (over all query–key pairs) in both the original and rotated runs, respectively. We then identify the entries that share the same query patch token in the two runs and, following the attention ranking in the original run, select the top 50 such entries for analysis. For visualization clarity, we rotate the outputs back to the original orientation. We observe that, for early global attention layers (*e.g.*, Layer 4), the highly activated key tokens change almost entirely after rotation. In contrast, for middle global attention layers (*e.g.*, Layer 11), the corresponding entries in the rotated case still largely point to the same spatial locations across both runs. This indicates that the highly activated tokens in the early global attention layers do not correspond to stable 3D structures or view-consistent relationships. Therefore, this rotation test further supports our conclusion that the early global attention layers do not contribute meaningfully to building cross-view correspondences.

## 2. More Experiments

### 2.1. Additional Dataset Results

Here, we provide additional results on more datasets. For camera pose estimation, we further evaluate on ScanNet [4] and Sintel [2] and ETH3D [12]. For point-map estimation, we additionally report results on NRGBD [1]. The results are summarized in Tables 1, 2, 3 and 4. We observe that on several datasets, FasterVGGT achieves better metrics than our method; however, this is mainly because its sparse ratio is significantly smaller than our minimum subsampling factor (a factor of 2 corresponds to keeping 50% of patch tokens), resulting in substantially higher computational cost. Under comparable runtime budgets, our method attains superior accuracy. Overall, our approach consistently achieves the best trade-off between accuracy and efficiency across all evaluated datasets.

### 2.2. More Ablation Studies on VGGT

**Effect of Not Subsampling Query Tokens.** We evaluated two training-free Query-subsampling variants on top of AVGGT(2) (with K/V already subsampled using $\sigma$=2): **AVGGT(2+Q2GM)** uses a global-mean Query token to approximate the missing attention weights induced by Query subsampling; **AVGGT(2+Q2Near)** similarly replaces each skipped Query with its nearest retained Query, without additional training. In terms of computational complexity, both AVGGT(2+Q2GM) and AVGGT(2+Q2Near) are comparable to AVGGT(4). We evaluate them on RealEstate10K. As shown in Table 5, both variants perform substantially worse than AVGGT(4). We believe this performance drop arises because, while global attention is functionally used to build cross-view correspondences, its effect is ultimately realized through attention computation that updates token values. Consequently, subsampling Query tokens inevitably reduces differences between tokens, which is clearly harmful for 3D tasks that require dense features. This interpretation is consistent with the results: AVGGT(2+Q2GM) performs worse than AVGGT(2+Q2Near), since using a global-mean Query makes different tokens even more similar.

**Effect of the Grid-Based Subsampling Strategy.** Our subsampling strategy is grid-based, but we also explored several alternative token-selection methods. Since our analysis shows that global attention builds cross-view correspondences by aligning tokens at the same spatial positions, we first draw inspiration from traditional SfM pipelines [3, 6, 9] such as COLMAP [10, 11], where SIFT [7] keypoints are used for feature matching. Although patch tokens and SIFT keypoints are intrinsically different, we test whether SIFT-based cues can help select more informative tokens. Specifically, we detect SIFT keypoints on all in-

put images, accumulate their scores over the patch grid, and select the top-scoring patch tokens. We denote this variant as AVGGT(2_SIFT), meaning a subsampling factor of 2 with SIFT-based token selection. As shown in Tab. 6, this variant performs worse than our grid-based strategy, suggesting that the tokens used by global attention for alignment do not align well with traditional handcrafted keypoint pipelines. Within the grid-based setting, we further evaluate whether selecting a fixed spatial index is optimal. We consider four variants: randomly selecting a token within each grid cell (AVGGT(2_Random)), selecting the token with the highest feature magnitude (AVGGT(2_High)), selecting the lowest-magnitude token (AVGGT(2_Low)), and using the mean value within each cell (AVGGT(2_Mean)). In AVGGT(2_Mean), we exclude both the diagonal and the mean components from our enhanced strategy. As shown in Tab. 6, the fixed grid-based selection remains the best-performing choice, whereas AVGGT(2_Mean) performs the worst. We attribute this to the fact that global attention constructs cross-view correspondences through spatially consistent tokens, while mean aggregation destroys these spatial anchors that are crucial for cross-view alignment.

**Effect of Diagonal Preservation.** We also study our enhanced subsampling strategy, which incorporates diagonal preservation and a mean component. We denote the variant without either enhancement (with subsampling factor 2) as AVGGT(2_SubsampleOnly), the variant with only the mean component as AVGGT(2_WithMean), and the variant with only diagonal preservation as AVGGT(2_WithDiagonal). As shown in Tab. 7, diagonal preservation yields a slight improvement in the sparse setting but slightly degrades performance under dense inputs, while the mean component exhibits the opposite trend. Despite these differences, combining both enhancements achieves the best overall performance, which is why we adopt this configuration as our default choice.

**Effect of the Last Global Attention Layers.** In the previous ablation studies, we observed that directly replacing the last global attention layers (indices 20–23) with frame attention results in only a slight performance drop. Here, we provide a more detailed analysis by introducing a parameter $t_{end}$, which specifies that all global attention layers with indices $> t_{end}$ are converted to frame attention. As shown in Fig. 2, the last global attention layers contribute very little to the final performance, and this effect becomes even more pronounced as $t_{end}$ increases. We interpret this behavior as evidence that, at deeper layers, the global feature maps are already well aligned across views, so additional global attention contributes less to cross-view consistency, leading to minimal performance difference.

**Effect of Not Subsampling the First Frame.** Our subsampling strategy differs slightly between VGGT and $\pi^3$. Since VGGT treats the first frame as the reference view, this frame generally plays a more important role than the others. Therefore, for VGGT we choose not to subsample the patch tokens of the first frame. We denote by AVGGT(2_FullySubsample) the variant that subsamples all frames with a factor of 2. As shown in Tab. 8, keeping the first frame uncompressed yields a small but consistent performance improvement.

### 2.3. More Ablation Studies on $\pi^3$

Similar to our ablation results on VGGT, as shown in Fig. 3, the early global attention layers in $\pi^3$ do not contribute to building cross-view correspondences. We observe that $t_{early} = 10$ provides the best trade-off between accuracy and speed. From Tab. 9, both $\pi^3$(G2F) and $\pi^3$(G2M) modify global attention layers with indices 0-9. The fact that $\pi^3$(G2F) achieves nearly the same performance as the original model indicates that no cross-view correspondences are formed in these layers, while the small performance gap of $\pi^3$(G2M) further suggests that these layers contain little meaningful selective attention. Regarding the subsampling factor, $\pi^3$ follows the same pattern as VGGT: it is more sensitive under sparse inputs and more robust when the input views are dense. A difference arises in the behavior of the last global attention layers. As shown in Fig. 4, in $\pi^3$ even Layer 17 still has a noticeable impact on the final performance. We attribute this to architectural differences: $\pi^3$ contains only 36 alternating transformer blocks, whereas VGGT contains 48. From this perspective, $\pi^3$ likely has less redundancy in its deeper layers, making its final global attention layer more influential than in VGGT.

## 3. More Visualization Results of Global Attention for VGGT

In addition to Figs. 7–9, we provide further visualizations of the global attention layers in VGGT. Please refer to Figs. 10, 11, 12, and 13.

## References


[1] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *CVPR*, pages 6290–6301, 2022. 1, 3

[2] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. TransformerFusion: Monocular RGB scene reconstruction using transformers. *NeurIPS*, 34:1403–1414, 2021. 1, 3

[3] Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu. HSfM: Hybrid structure-from-motion. In *CVPR*, pages 1212–1221, 2017. 3

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet:

Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 1, 3

[5] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 1

[6] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge university press, 2003. 3

[7] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3

[8] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[9] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *ECCV*, pages 58–77. Springer, 2024. 3

[10] Johannes L Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 3

[11] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, pages 501–518. Springer, 2016. 3

[12] Thomas Schöps, Johannes L Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 3260–3269, 2017. 1, 3

[13] 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 *CVPR*, pages 2930–2937, 2013. 2

[14] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 1

[15] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. $\pi^3$: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 1

[16] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 1, 2

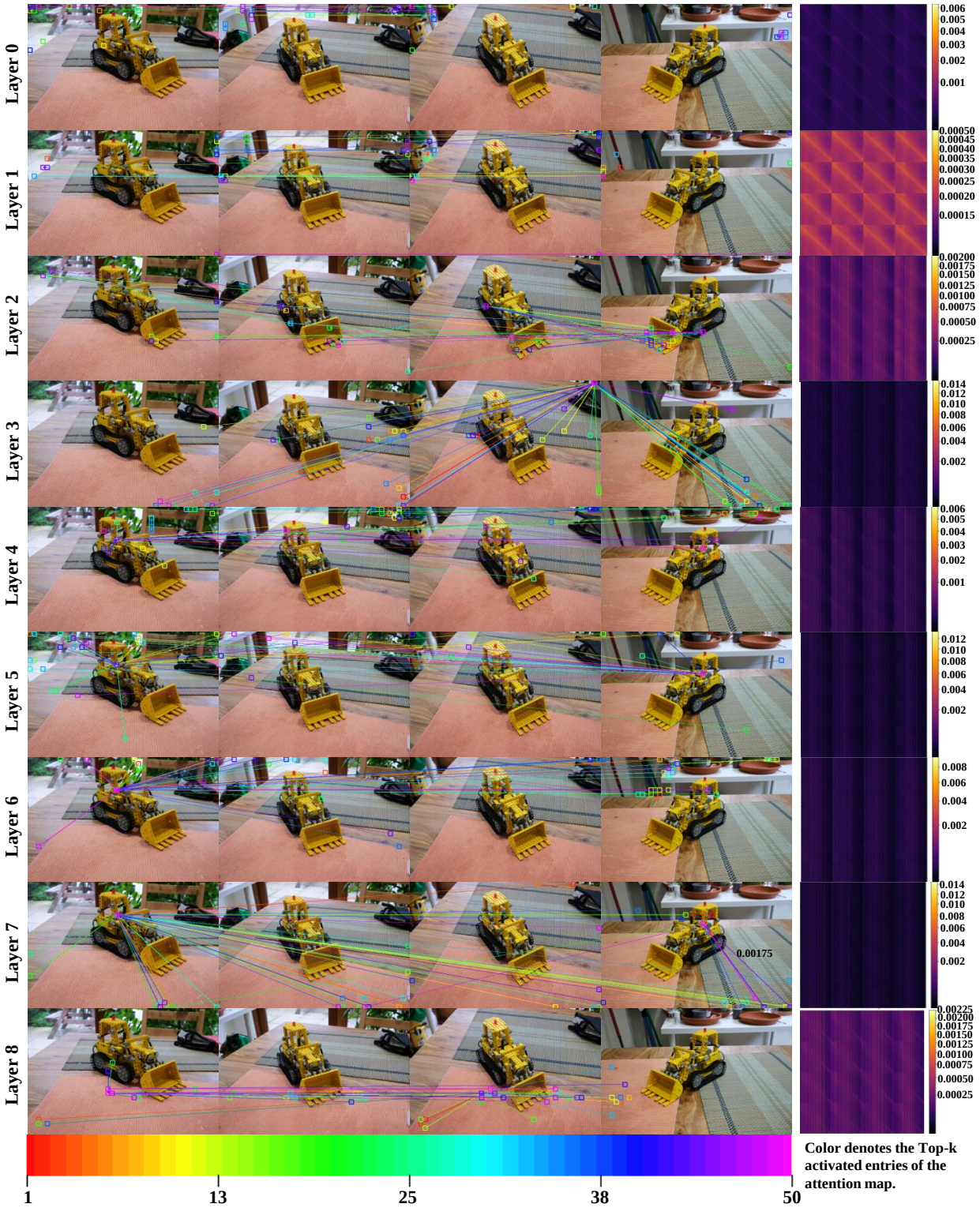


Figure 5. Visualization of global attention for layers 0-8 in $\pi^3$.

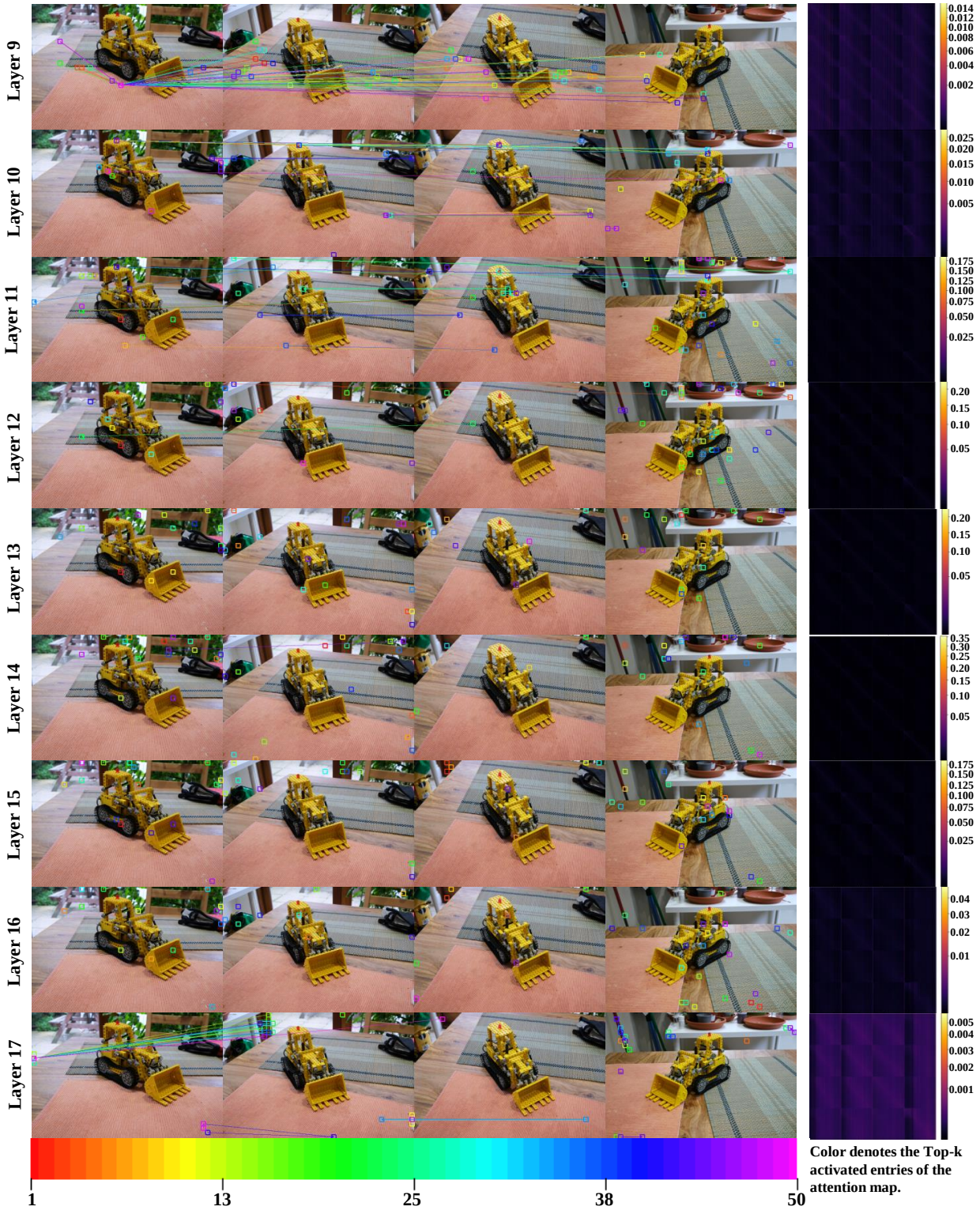


Figure 6. Visualization of global attention for layers 9-17 in $\pi^3$.

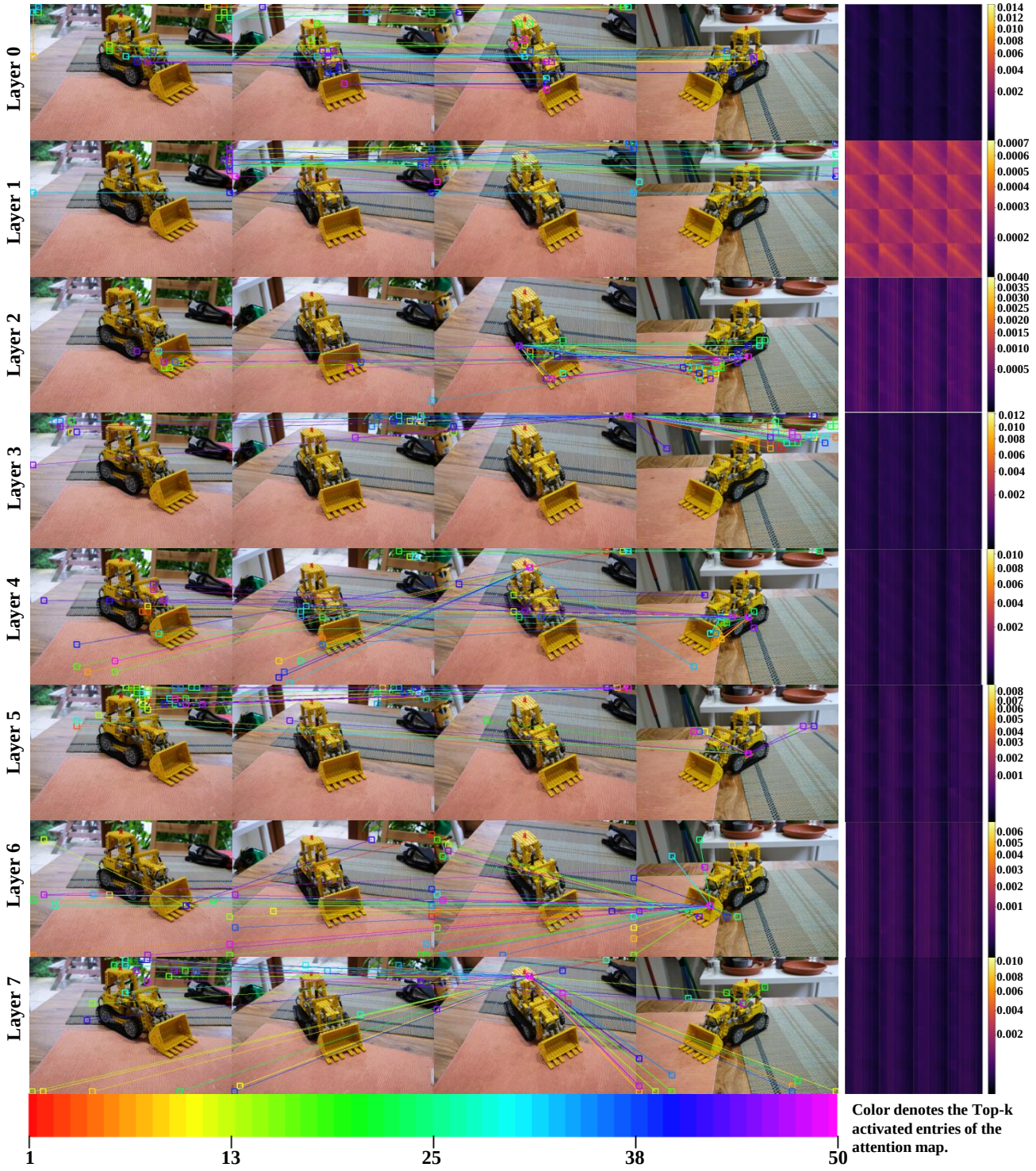


Figure 7. Visualization of global attention for layers 0-7 in VGGT.

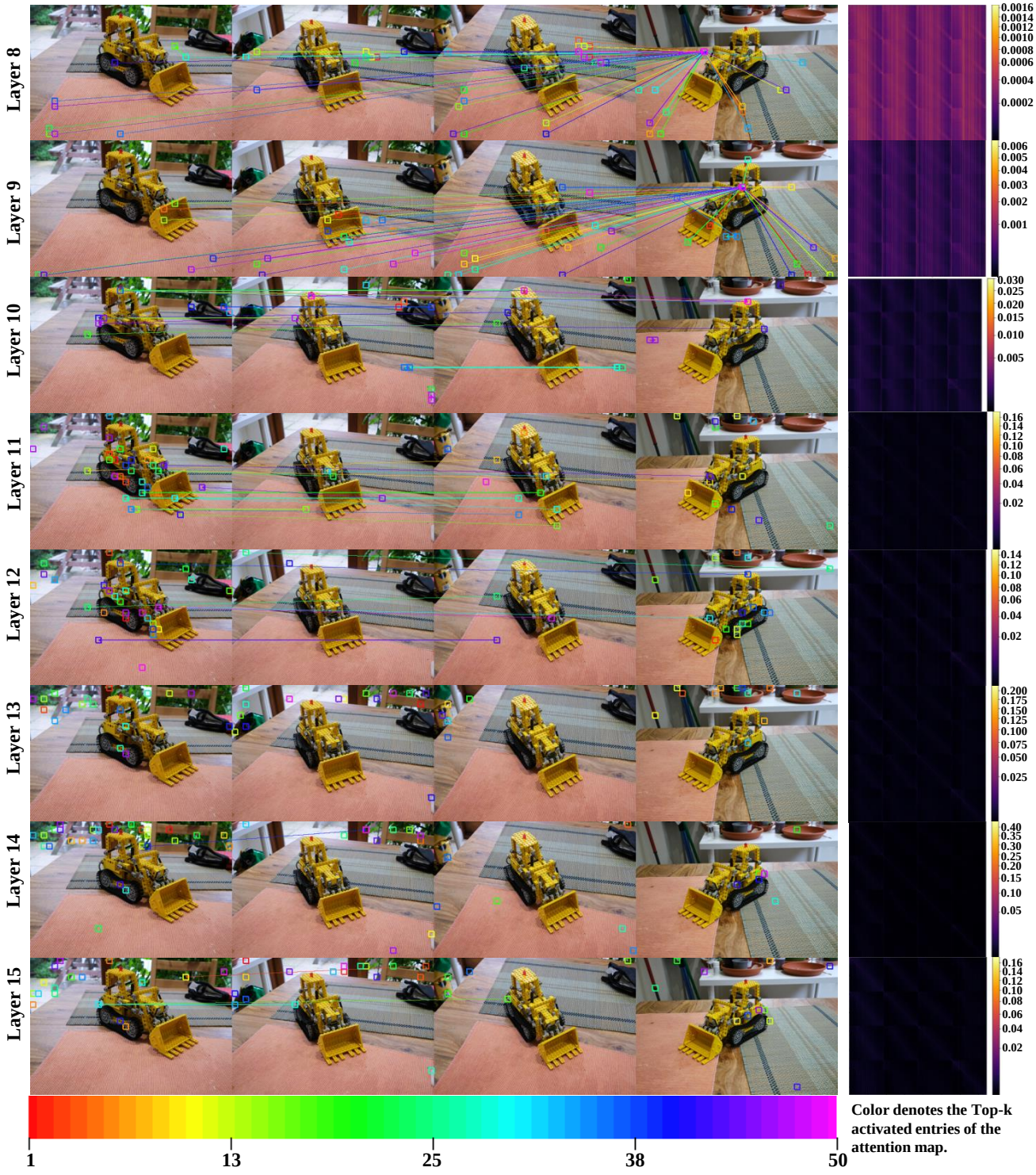


Figure 8. Visualization of global attention for layers 8-15 in VGGT.

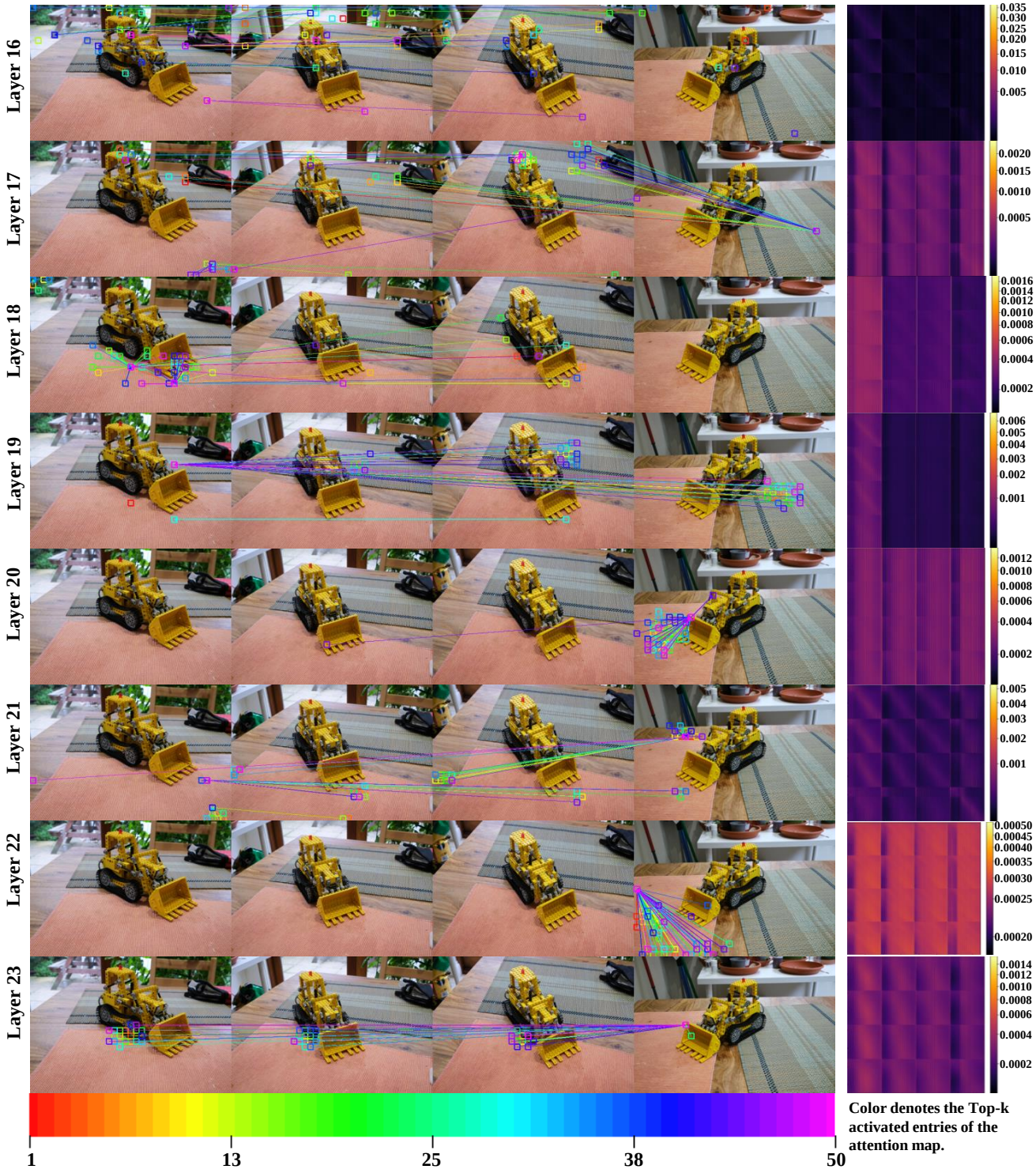


Figure 9. Visualization of global attention for layers 16-23 in VGGT.

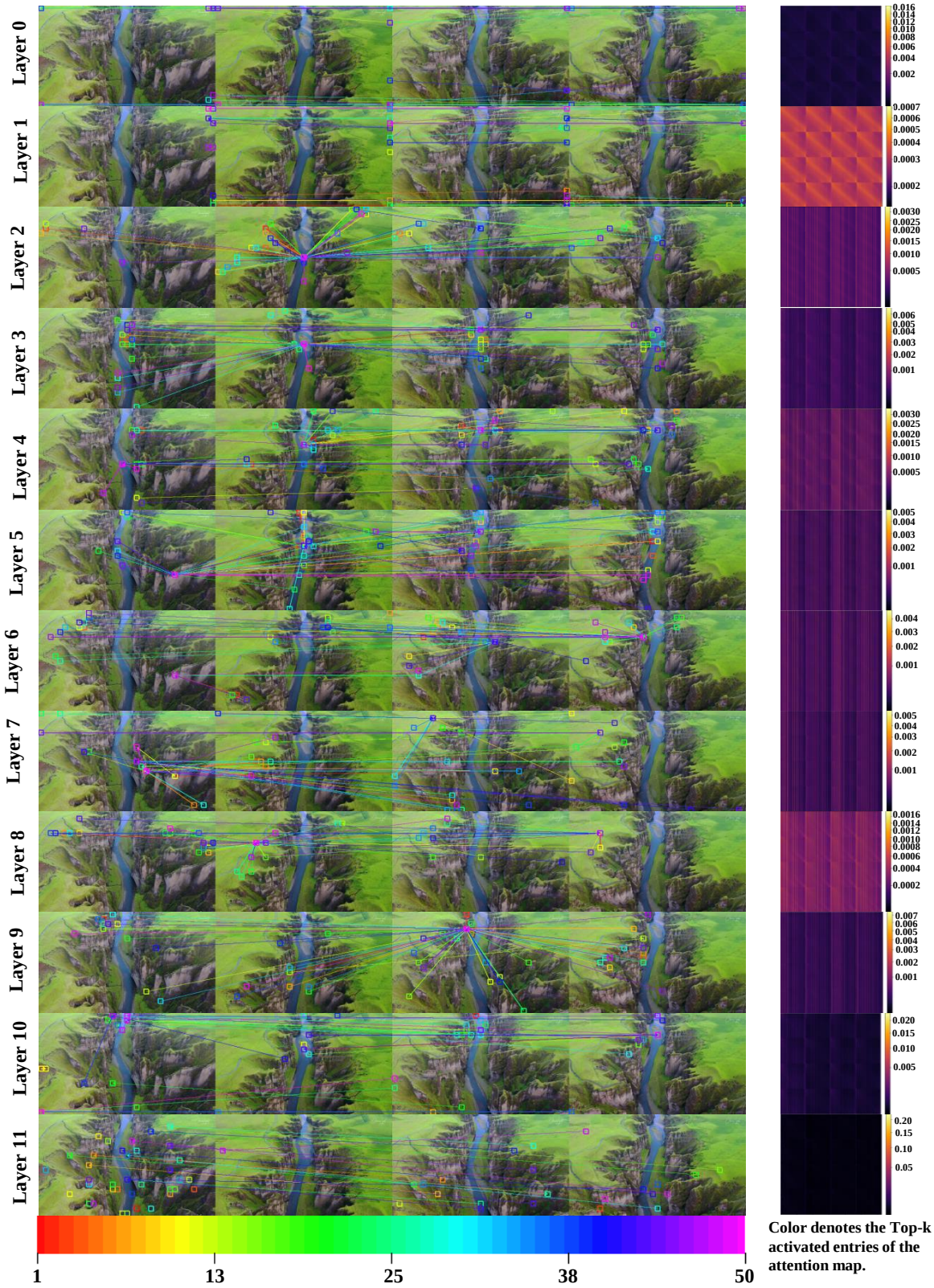


Figure 10. Visualization of global attention for layers 0-11 in VGGT.

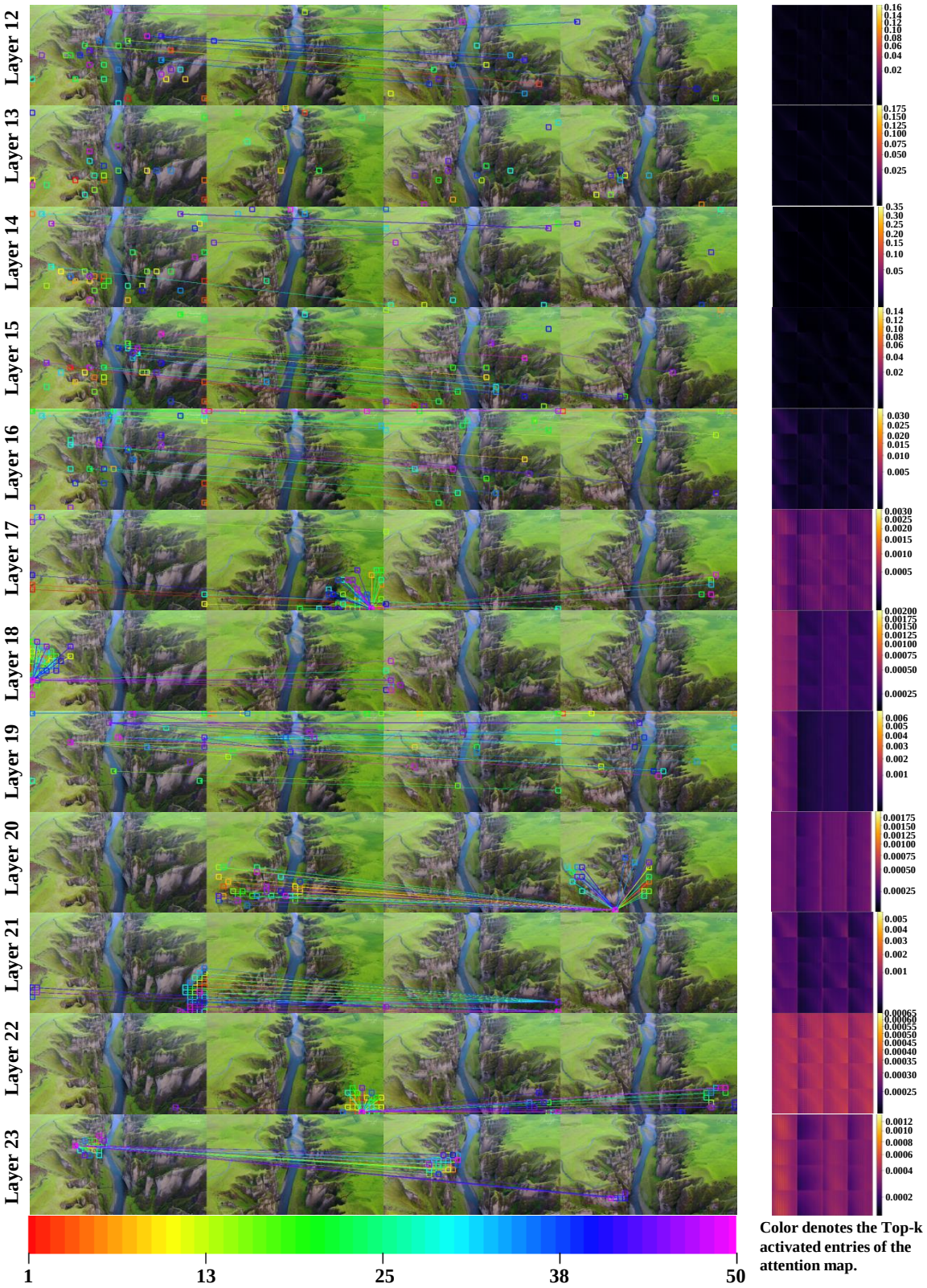


Figure 11. Visualization of global attention for layers 12-23 in VGGT.

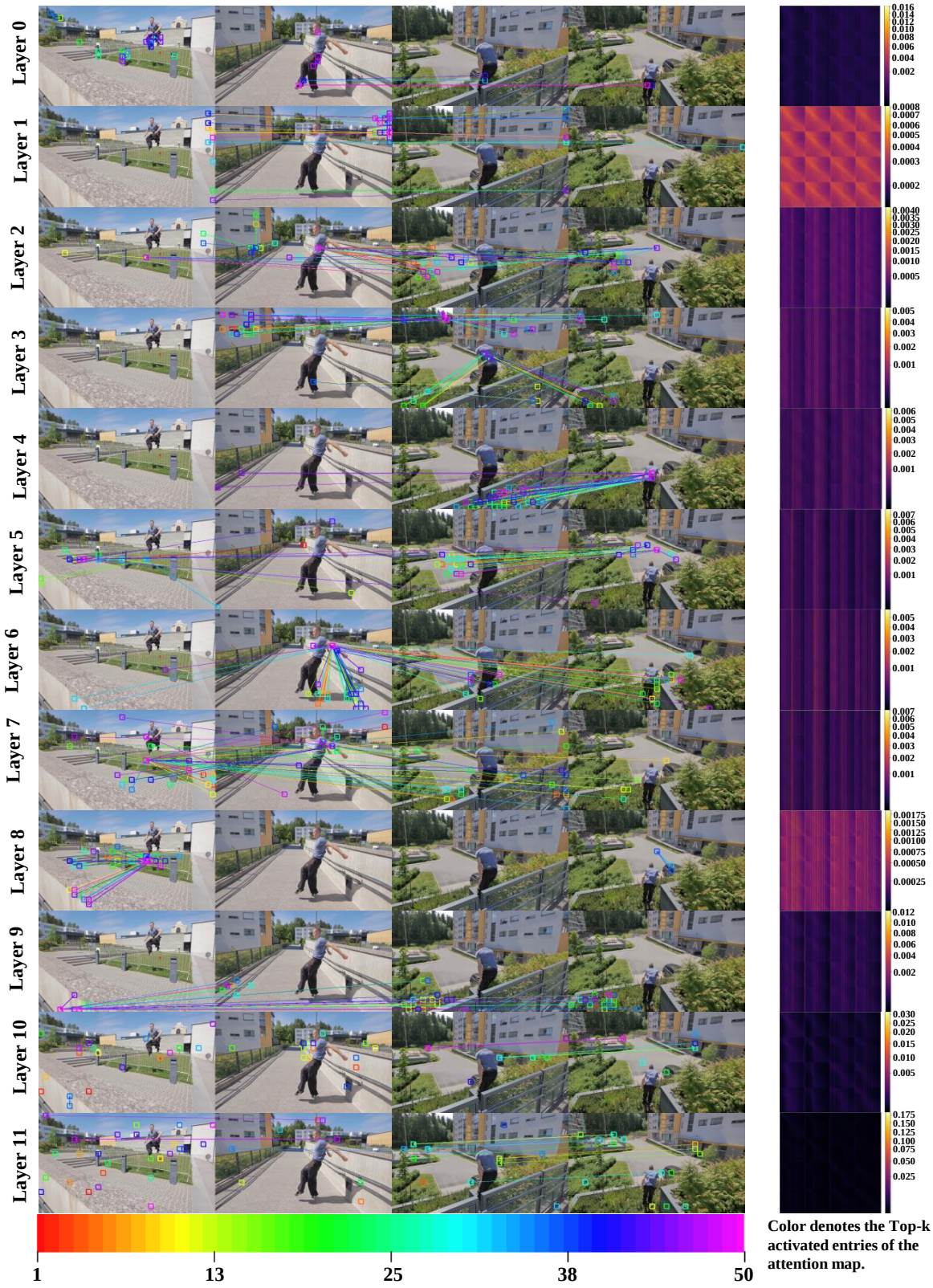


Figure 12. Visualization of global attention for layers 0-11 in VGGT.

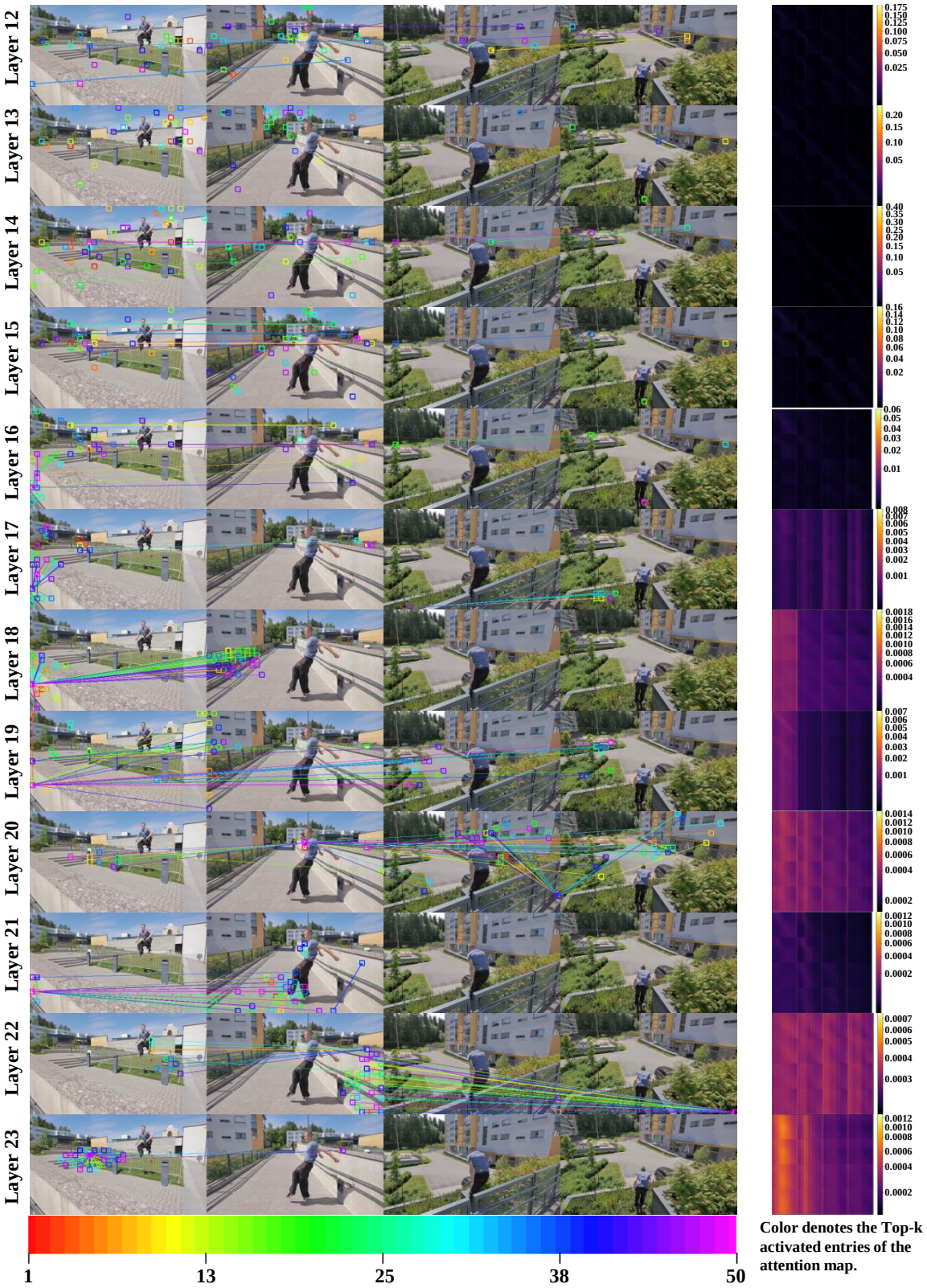


Figure 13. Visualization of global attention for layers 12-23 in VGGT.