VGGT-Det: Mining VGGT Internal Priors for Sensor-Geometry-Free Multi-View Indoor 3D Object Detection

Supplementary Material

A. More Ablation Studies

Performance under different λ_{dist}. In this subsection, we study the impact of varying λ_{dist} in Eqn. (8) on the performance of our proposed Attention-Guided Query Generation (AG). As shown in Tab. 1, the best performance in our evaluation is obtained around $\lambda_{\text{dist}} = 0.8$. Intuitively, λ_{dist} modulates the balance between attention guidance and spatial dispersion. When λ_{dist} is too small, the model relies heavily on attention guidance, leading to excessive focus on specific high-attention regions while neglecting the global spatial structure. When λ_{dist} is too large, the spatial dispersion term dominates, reducing the influence of attention guidance. In summary, λ_{dist} controls an intuitive balance between attention guidance and spatial dispersion. The observed optimal performance around $\lambda_{\text{dist}} = 0.8$ in our ablations aligns well with our goal of initializing queries that focus on semantic regions while preserving the global spatial structure.

Methods	mAP@0.25
w/ AG (λ_{dist}=0.9)	44.9
w/ AG (λ_{dist}=0.8)	46.9
w/ AG (λ_{dist}=0.5)	40.7

Table 1. Performance under different λ_{dist} in Eqn. (8).

The impact of varying levels of noise. In this subsection, we evaluate the robustness of our method to noise in VGGT-predicted point clouds. Because the quality of the pretrained VGGT [4] outputs is not directly controllable, we introduce controlled noise by adding Gaussian noise to the VGGT point clouds. Specifically, the noise is sampled from a normal distribution $n \sim \mathcal{N}(0, \sigma^2)$, where σ represents the standard deviation of the noise. To ensure precise and scalable noise intensities, σ is defined as: $\sigma = \text{range} \cdot \text{noise\_level}$, where $\text{range} = \max(P) - \min(P)$ represents the value range of the point clouds, and noise_level $\in [0, 1]$ is a user-defined parameter controlling the noise intensity. The noisy point cloud is generated as: $P' = P + N$, where N is the noise matrix. This method enables fine-grained noise control, facilitating precise evaluation of noise effects. We evaluate VGGT-Det on ScanNet using noisy point clouds. To provide a fair comparison, we also evaluate the representative point cloud–based method FCAF3D [3], which likewise takes noisy point clouds as input. The results are summarized in Tab. 2. As expected, the performance of both methods degrades as the noise level increases, confirming the adverse impact of noise on detection. However, FCAF3D starts to degrade significantly at a noise level of 0.001 and drops to 0.0 mAP@0.25 at 0.1, whereas VGGT-Det maintains robustness and only starts degrading at 0.1, achieving 34.1 mAP@0.25 even at 0.3. At the same noise level of 0.01, VGGT-Det significantly outperforms FCAF3D by **28.3 points** (47.0 vs. 18.7), demonstrating its superior robustness.

Noise Level	FCAF3D	VGGT-Det (Ours)
0.30	0.0	34.1
0.20	0.0	39.4
0.10	0.0	44.5
0.01	18.7	47.0
0.005	33.4	47.0
0.001	39.7	46.9
0.000	40.6	46.9

Table 2. Robustness under different noise levels: comparison between FCAF3D [3] and our VGGT-Det.

This robustness stems from our Attention-Guided Query Generation, which efficiently leverages VGGT's internal semantic priors to generate object queries that focus on semantically relevant regions, as illustrated in Fig. 1. By reducing reliance on the precise geometric details of the predicted point clouds from VGGT, this mechanism effectively mitigates the impact of noise. In contrast, FCAF3D relies heavily on the geometric integrity of the input point clouds to extract features, making it significantly more susceptible to noise, as evidenced by its rapid performance degradation in Tab. 2.

Performance across different numbers of input frames. For a fair comparison, we follow the standard practice in indoor multi-view 3D object detection (e.g., the official MVSDet implementation) by using 80 input frames. Nevertheless, our approach is not restricted to this configuration. Across different numbers of input frames, it consistently surpasses the strongest competing method, MVSDet, as shown in Tab. 3.

# of Frames	20	40	60	80	100
MVSDet [6]	35.9	39.7	40.9	42.5	42.5
Ours	**42.6**	**45.3**	**46.2**	**46.9**	**47.3**

Table 3. Performance across different numbers of input frames

Performance gap to methods that utilize sensor geometry. In Tab. 5, M1, M2, and M3 correspond to ImVoxelNet [2], NeRF-Det [5], and MVSDet [6], respectively, while "+SG" indicates methods that leverage sensor geometry. Our method achieves the highest performance in both settings—SG-Free and SG-Based. A noticeable gap remains between the two, which is expected since sensor-provided ge-

ometric priors offer strong cues for detection. Nevertheless, we regard the SG-Free setting as an important and practical regime, given that sensor geometry is often unavailable or impractical in real-world scenarios.

Method	**M1**	**M2**	**M3**	**Ours**
mAP	35.2	41.2	42.5	**46.9**
Method	**M1+SG**	**M2+SG**	**M3+SG**	**Ours+SG**
mAP	48.1	49.5	56.2	**58.8**

Table 4. Performance gap to methods that leverage sensor geometry.

More methods in efficiency analysis. To ensure fair comparison under the same SG-Free setting, we train and evaluate SG-Free variants of all models, rather than their original sensor-geometry-dependent versions. Specifically, multi-view RGB images are first processed by VGGT to estimate camera poses, which are then provided to each detector. Consequently, the SG-Free pipeline inherently includes VGGT's runtime and memory costs, ensuring fairness in comparison. In Tab. 5, M1, M2, and M3 represent ImVoxelNet [2], NeRF-Det [5], and MVSDet [6], respectively. As shown, our designs (+Ours) achieve inference time comparable to the strongest competitor (+MVSDet), while attaining the lowest GPU memory usage among all methods.

Method	**VGGT**	**+M1**	**+M2**	**+M3**	**+Ours**
T (s)	0.68	0.09	**0.07**	0.21	0.23
M (GB)	11.71	12.95	6.76	13.81	**3.57**

Table 5. Efficiency comparison with different mothods in SG-Free setting with 40 input frames.

Impact of view number on efficiency. As shown in Tab. 6, the time and GPU memory cost increase with more views, which is a common challenge in multi-view 3D detection.

# of Frames	**20**	**40**	**60**	**80**	**100**
T (s)	0.49	0.91	1.48	2.20	3.06
M (GB)	14.46	15.28	17.50	19.48	21.66

Table 6. Efficiency comparison under different input numbers.

Visualization analysis. To further study the effectiveness of our Attention-Guided Query Generation, we conduct a visualization analysis of both the attention maps and the positions of the generated object queries. As shown in Fig. 1, the 2nd column illustrates the attention maps, which clearly highlight the object regions. This observation supports our motivation: attention can provide semantic guidance for query generation. In the 3rd column, we compare the generated object query positions. The red points represent the object queries generated without attention guidance, while the green points correspond to the queries generated by our Attention-Guided Query Generation. Notably, within the object regions (highlighted by green boxes), our method generates significantly more object queries (green points) compared to the baseline (red points). This demonstrates that our Attention-Guided Query Generation effectively focuses on object regions, leading to a more accurate query distribution. The effectiveness of our approach is further reflected in an improvement of **2.8 points** in mAP@0.25, as reported in Tab. 2a of the main paper. This result highlights the direct contribution of Attention-Guided Query Generation to the overall performance.

B. Comparison with Alternatives

Qualitative Comparison. We provide a qualitative comparison with the best-performing competitive method, MVSDet [6]. To adapt to the Sensor-Geometry-Free (SG-Free) setting and ensure a fair comparison, MVSDet is trained and tested using multi-view poses predicted by VGGT [4]. As shown in Fig. 2, our VGGT-Det detects more objects with higher accuracy, which is consistent with the significant performance improvement of **4.4 points** reported in Tab. 1 of the main paper. The higher performance benefits from the effective utilization of internal VGGT priors by the proposed Attention-Guided Query Generation (AG) and Query-Driven Feature Aggregation (QD) modules.

C. Training and Testing Time

All the ablation experiments are conducted on eight H800 GPUs. Training our model on the ScanNet dataset [1] takes approximately 2 days to complete. For testing, the entire ScanNet testing set can be processed in about 1 minute.

D. Limitation and Future work

While VGGT-Det achieves significant improvements over strong alternative methods, several limitations remain for further exploration. Across current SG-free pipelines, VGGT incurs noticeable runtime and memory overhead. Besides, because VGGT produces normalized predictions, the scales from datasets are utilized to denormalize predictions of VGGT in all the current SG-free pipelines. Looking ahead, introducing a lighter VGGT-like model with metric-scale predictions could further advance this direction.

References

[1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 2

[2] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3D object detection in point clouds with image votes. In *CVPR*, 2020. 1, 2

[3] Danila Rukhovich, Anna Vorontsova, and Anton Konushin.

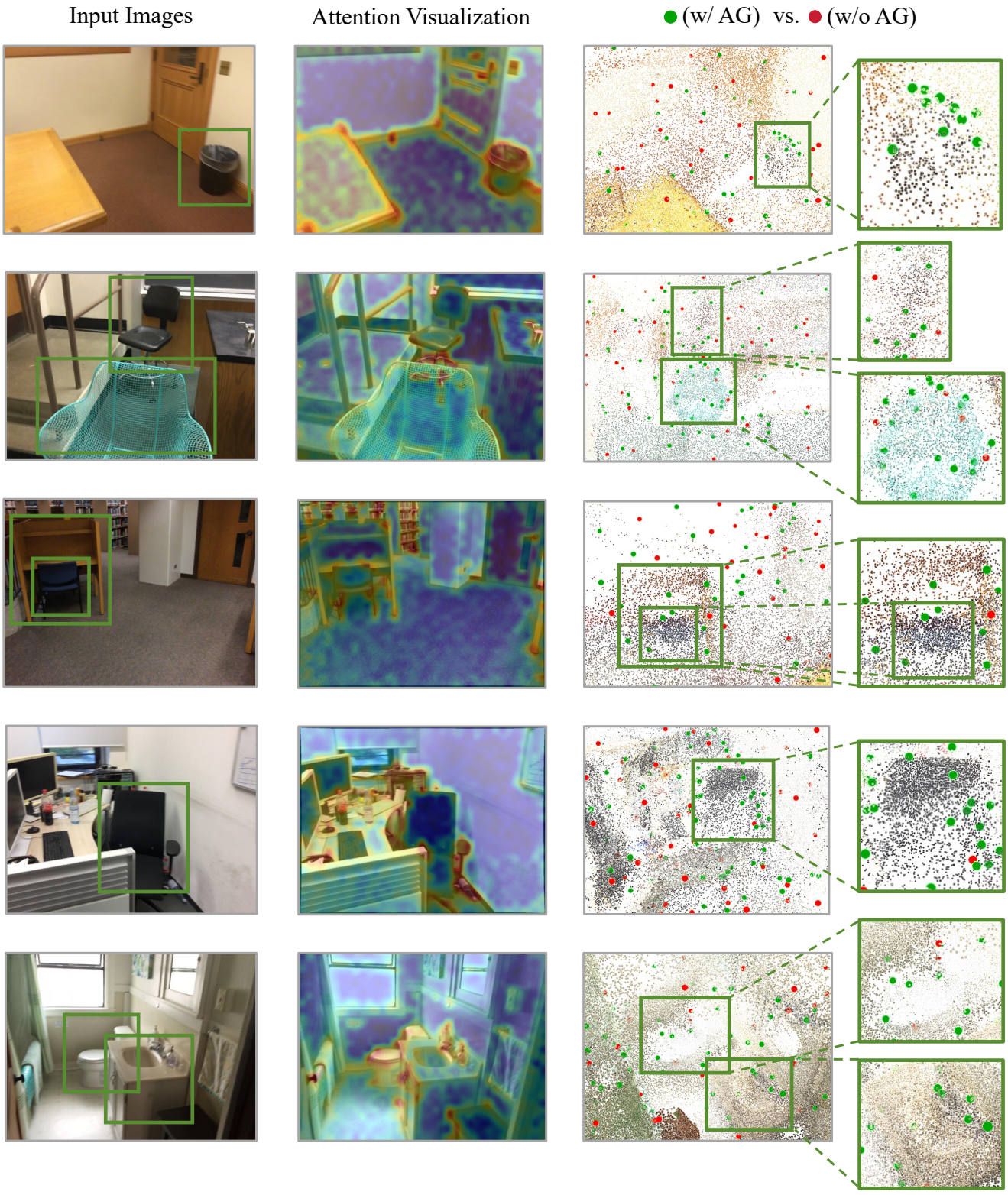

Figure 1. Visualization of attention and generated object query positions. Compared to farthest point sampling without guidance (red points), the points sampled by AG (green points) are more concentrated in object regions (labeled by green boxes), resulting in more green points than red points in those areas. For clarity, we recommend viewing the figure in color and zooming in.

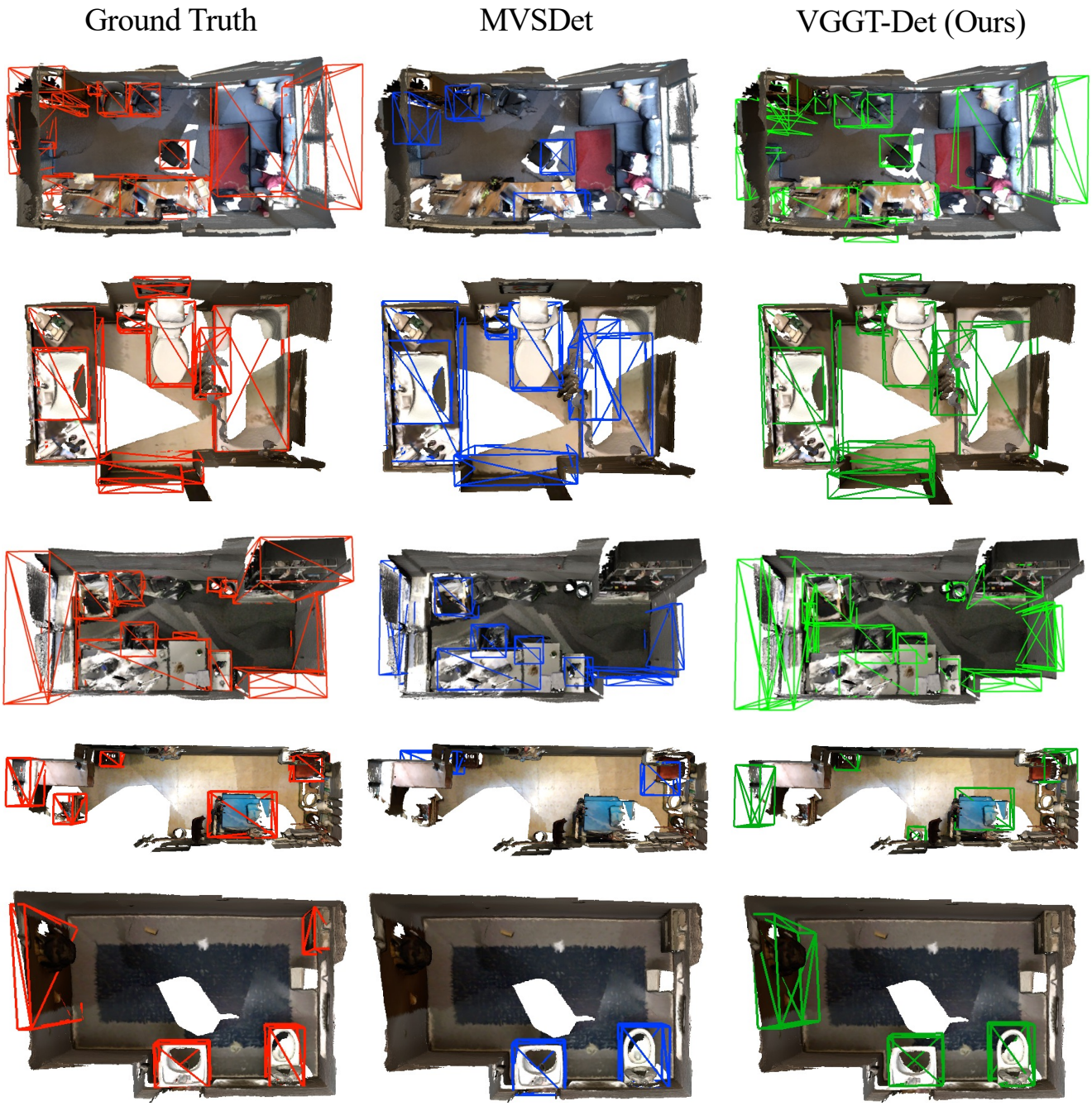

Figure 2. Qualitative comparison with MVSDet [6]. To achieve the Sensor-Geometry-Free (SG-Free) setting and ensure a fair comparison, MVSDet is trained with multi-view poses predicted by VGGT [4]. The mesh here is not utilized in the methods and is only for visualization.

Fcaf3d: Fully convolutional anchor-free 3d object detection. In *ECCV*, 2022. 1

[4] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 2, 4

[5] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, et al. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In *ICCV*, 2023. 1, 2

[6] Yating Xu, Chen Li, and Gim Hee Lee. Mvsdet: Multi-view indoor 3d object detection via efficient plane sweeps. In *NeurIPS*, 2024. 1, 2, 4