What VGGT Knows About Overlap: Probing Geometric Foundation Models for Co-Visibility

Filippo Ziliotto[1,2], Luciano Serafini[2], Lamberto Ballan[1], and Tommaso Campari[2]

[1] University of Padova
[2] Fondazione Bruno Kessler (FBK)

Abstract. A fundamental challenge in 3D reconstruction and robotic localization is co-visibility: determining which image pairs share overlapping visible surfaces, particularly in scenarios with minimal overlap. We demonstrate that VGGT implicitly encodes co-visibility as an emergent behavior: without any supervision for this task, its internal representations exhibit a clear hierarchical structure mirroring that of large language models — early layers build a 3D-aware scene representation, while late layers act as dedicated co-visibility reasoners. In particular, we identify layer L17 as a negative anchor that consistently routes non-co-visible pairs for this backbone, regardless of the evaluation setting, providing task-grounded evidence of layer specialization in a geometry-grounded foundation model. Building on this, we introduce Co-VGGT, which freezes VGGT and trains only a lightweight layer-wise mixture-of-experts head (∼7.5M parameters) to classify co-visibility from RGB alone, treating each layer as a specialized expert whose geometric abstraction is adaptively weighted per input pair. On the Co-VisiON benchmark, Co-VGGT surpasses the human annotation baseline and improves over prior work by more than 25% pairwise and 10% multiview. Pairwise predictions are well-calibrated (ECE = 0.030), enabling direct use as edge weights in visibility graphs for downstream SfM and SLAM pipelines without post-hoc correction. Code and data available[1].

Keywords: Co-visibility · Multiview Geometry · Embodied perception

1 Introduction

Robotic perception and 3D reconstruction systems typically operate on sparse, imperfect image sets rather than isolated, perfect views. A fundamental challenge in these pipelines—whether for mapping, localization, or scene reconstruction—is determining co-visibility, the subset of surfaces jointly observed by multiple cameras. High co-visibility yields abundant geometric constraints and stable optimization. Conversely, when spatial overlap is limited or absent, matching becomes ambiguous and pose estimation drifts. In these scenarios, reconstruction pipelines often fail silently, generating plausible but geometrically inconsistent

[1] https://github.com/covisibility-probing

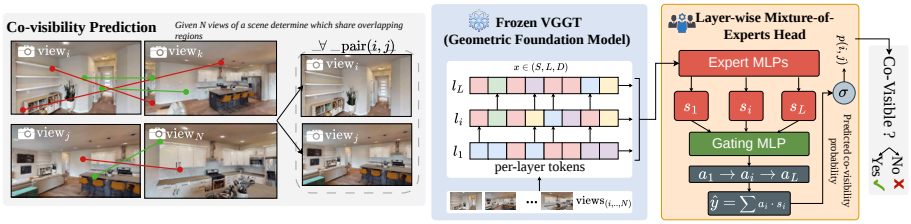

Fig. 1: Co-visibility Task. We study co-visibility prediction: given multiple RGB views of a scene, the goal is to determine which image pairs share overlapping visible regions. Our approach probes geometric consistency in a frozen VGGT foundation model, extracting layer-wise view embeddings and combining them through a lightweight mixture-of-experts head. Without modifying the backbone, the model aggregates information across transformer layers to produce a co-visibility probability for each pair, enabling efficient construction of scene-level visibility graphs.

structures. Operating in this sparse-view regime is not an edge case but rather a standard condition for embodied agents navigating complex, occluded environments.

Despite advances in multiview transformers and learned reconstruction models, performance drops significantly as spatial overlap decreases. Under these conditions, fusion modules misalign features, learned descriptors drift, and global reasoning degrades into spurious correlations. Recent benchmarks, such as Co-VisiON [8], explicitly isolate this co-visibility reasoning, exposing a critical performance gap between current models and human baselines in sparse, highly imbalanced scenarios. Bridging this gap is essential for robotics: robust co-visibility estimation dictates which view pairs to match, which constraints to trust, and when to flag reconstruction failures.

Concurrently, geometry-grounded foundation models present a promising alternative. The Visual Geometry Grounded Transformer (VGGT) [34] demonstrates strong emergent geometric reasoning; without explicit 3D supervision, it infers scene structure and reconstructs geometry even across widely separated viewpoints. Similar to emergent phenomena in large language models, VGGT's internal representations encode latent structures that transcend its immediate training objective. However, the exact nature of these spatial reasoning signals and the methodology to extract them for embodied decision-making remain poorly understood.

In this work, we show that frozen VGGT features contain a strong, directly usable signal for co-visibility estimation (see Fig. 1). Furthermore, we find that specific late VGGT layers act as consistent anchors for co-visibility decisions, while earlier layers encode broader scene-aware geometric representations. Leveraging this property, we propose Co-VGGT, an MoE head in which each layer acts as an "expert" whose geometric abstraction is adaptively weighted per image pair. Compared to single-layer probing, this routing improves accuracy and

calibration while exposing which layer cues drive each decision. Relying solely on RGB inputs, Co-VGGT achieves near human-level accuracy on the Co-VisiON benchmark. It surpasses the current state-of-the-art by over 25% in pairwise co-visibility prediction and nearly 10% in multiview inference. We further observe that co-visibility reasoning is dominated by late layers, whereas earlier layers contribute more general geometric context.

Our findings indicate that geometry-grounded foundation models encode spatial priors that are far richer than those obtained through conventional contrastive or matching-based pretraining. Furthermore, these signals can be efficiently distilled into modular predictors. Such predictors could be integrated into reconstruction pipelines: a dedicated co-visibility head can refine overlap prediction and serve as an auditing mechanism, identifying inconsistent constraints and detecting early-stage failures to support reliable embodied autonomy.

In summary, our contributions are as follows: (i) we introduce Co-VGGT, achieving near human-level co-visibility from RGB and improving the Co-VisiON SOTA by >25% (pairwise) and ∼10% (multiview); (ii) we provide evidence that VGGT implicitly encodes co-visibility structure across its layers, exposing emergent spatial priors that are absent in standard supervised baselines; and (iii) we show that Co-VGGT is robust to minimal overlap scenarios and its well-calibrated outputs (ECE = 0.030 on Gibson val) enable direct use as edge weights in visibility graphs, providing a practical auditing signal for SfM and SLAM pipelines without post-hoc correction.

2 Related Works

Co-visibility prediction sits at the intersection of 3D reconstruction, feature matching, and geometric reasoning. We review the most relevant lines of work, highlighting how each studies co-visibility as an explicit, predictable signal. In doing so, we ground our approach within the broader landscape and clarify what makes it fundamentally distinct from all prior work.

Co-visibility in Reconstruction and Matching Pipelines. Co-visibility is handled implicitly in classical SfM and SLAM through feature matching and geometric verification [6, 26], degrading under weak texture or large viewpoint changes. Learned SLAM methods [32, 41] and implicit neural mapping [29, 42] improve robustness but still require sufficient overlap, while systems with explicit graph structures — Kimera [24], Hydra [15], and situational graphs [4] — treat co-visibility as a derived post-hoc quantity. Learned matchers — SuperGlue [25], LightGlue [19], Efficient LoFTR [36], and OmniGlue [16] — estimate matchability conditioned on the existence of overlap and cannot flag non-overlapping pairs, while retrieval methods like NetVLAD [2] conflate semantic similarity with geometric overlap. We instead predict co-visibility directly from RGB as a dedicated, calibrated first-class signal before any reconstruction is attempted.

Geometric Foundation Models and Sparse-View Reasoning. DUSt3R [35], MUSt3R [5], MVSNeRF [7], and MV-DUSt3R+ [39] have advanced multiview geometry estimation; yet, all degrade in sparse-view regimes where co-visible

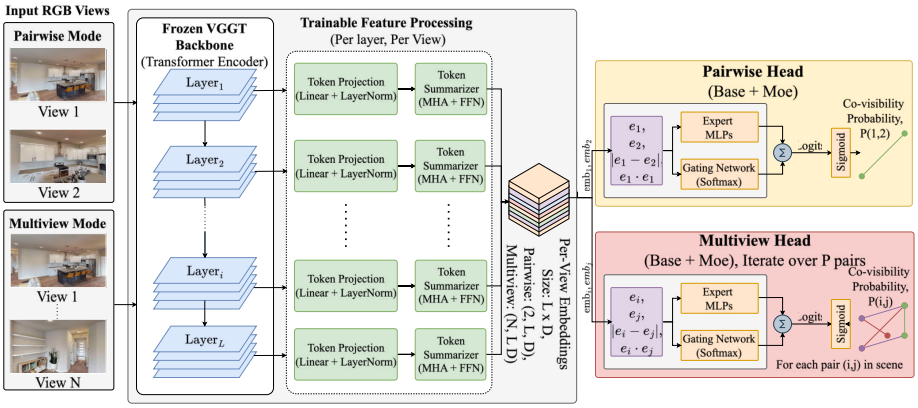

Fig. 2: Overview of the Co-VGGT method. Input RGB views are processed by the frozen VGGT backbone to extract layer-wise features. These features are projected, summarized, and formed into per-view embeddings. In both pairwise and multiview modes, these embeddings are used to construct pair features, which are then fed into a trainable Mixture-of-Experts (MoE) head to predict co-visibility probabilities, enabling the construction of scene-level visibility graphs. Tensor shapes are annotated for key intermediate representations.

surface support is limited. VGGT [34] achieves emergent 3D reasoning without explicit geometric supervision; we show its internal representations encode a hierarchically-organized co-visibility signal that can be distilled without modifying the backbone. The Co-VisiON benchmark [8] formalizes this as graph inference over sparse indoor views, exposing a critical performance gap that our approach substantially closes.

VLMs, Geometric Distillation, and Transformer Interpretability. VLMs such as GPT-4o [22], Gemini [31], and SpatialRGPT [9] show emergent spatial reasoning but remain unreliable for precise 3D consistency under significant viewpoint changes, as confirmed by our experiments. Recent distillation approaches inject geometric priors into frozen language backbones [1,17], but target general scene understanding rather than co-visibility. Meanwhile, probing work localizes capabilities to specific transformer layers [12,28,33], with reasoning concentrated in the late layers of LLMs [20] and global semantics in the late layers of vision transformers [40]. We provide the first task-grounded evidence of an analogous hierarchical structure in a geometry-grounded model, identifying the late VGGT layers — particularly L17 — as decisive co-visibility reasoners.

3 Method

We address co-visibility estimation simply using RGB inputs by training a lightweight binary classifier on top of a frozen geometric foundation model. Given a set of views from the same scene, the goal is to predict whether each view pair

shares any jointly visible surface region to reconstruct the scene-graph. Our model operates in two regimes: (i) pairwise, where each training example is a single pair of images; and (ii) multiview, where each example is an entire scene containing many views and labeled pairs. An overview of the method is shown in Fig. 2.

Importantly, pairwise inference is a special case of multiview with $N{=}2$ and a single pair. This unified formulation allows training in multiview mode and evaluation in both pairwise and multiview settings without architectural changes.

Problem Setup. Let $\mathcal{I} = \{I_s\}_{s=1}^{S}$ be a set of RGB views of the same scene, with $I_s \in \mathbb{R}^{3\times H\times W}$. For any pair (i, j), the target is a binary co-visibility label $y_{ij} \in \{0, 1\}$. In the pairwise setting we predict a single labeled pair per sample ($S{=}2$). In the multiview setting we process $S{>}2$ views jointly and predict a set of labeled pairs $\mathcal{P} = \{(i_p, j_p, y_p)\}_{p=1}^{P}$, which defines a co-visibility graph over the views.

Backbone and Summarization. We use a pretrained VGGT backbone Φ as a frozen feature extractor. Given a batch $\mathbf{X} \in \mathbb{R}^{B\times S\times 3\times H\times W}$, we extract patch-token features from selected layers $\ell \in \{1, \dots, L\}$: $\mathbf{T}^{(\ell)} = \Phi^{(\ell)}(\mathbf{X}) \in \mathbb{R}^{B\times S\times P_{\text{tok}}\times C_{\text{raw}}}$. All parameters of Φ remain frozen. To obtain compact per-view embeddings, we (i) project token channels and (ii) summarize tokens with learned queries. We first apply a shared linear projection with LayerNorm: $\widehat{\mathbf{T}}^{(\ell)} = \mathrm{LN}\big(\mathbf{T}^{(\ell)}W\big)$, where $W \in \mathbb{R}^{C_{\text{raw}}\times C_{\text{proj}}}$. Then, for each view we summarize P_{tok} tokens into T summary tokens via cross-attention with learned queries $\mathbf{Q} \in \mathbb{R}^{T\times C_{\text{proj}}}$:

$$\mathbf{S}_s^{(\ell)} = \mathrm{Attn}\Big(\mathbf{Q}, \widehat{\mathbf{T}}_s^{(\ell)}, \widehat{\mathbf{T}}_s^{(\ell)}\Big) \in \mathbb{R}^{T\times C_{\text{proj}}}, \quad \mathbf{e}_s^{(\ell)} = \mathrm{vec}\Big(\mathbf{S}_s^{(\ell)}\Big) \in \mathbb{R}^{D}, \tag{1}$$

with $D = T\,C_{\text{proj}}$. This yields per-view, per-layer embeddings $\{\mathbf{e}_s^{(\ell)}\}_{\ell=1}^{L}$.

Pair Representation & MoE Head. Following [10, 13, 21], to enhance the embedding representation for each labeled pair $(i, j) \in \mathcal{P}$ and layer ℓ, we build a symmetric pair feature:

$$\mathbf{f}_{ij}^{(\ell)} = \Big[\mathbf{e}_i^{(\ell)},\ \mathbf{e}_j^{(\ell)},\ \Big|\mathbf{e}_i^{(\ell)} - \mathbf{e}_j^{(\ell)}\Big|,\ \mathbf{e}_i^{(\ell)} \odot \mathbf{e}_j^{(\ell)}\Big] \in \mathbb{R}^{4D}. \tag{2}$$

Based on the assumption that VGGT mirrors the hierarchical reasoning structure of LLMs, in our method, each layer acts as an "expert" that predicts a logit $z_{ij}^{(\ell)}$ from $\mathbf{f}_{ij}^{(\ell)}$, while a gating network assigns mixture weights across layers:

$$z_{ij}^{(\ell)} = \mathrm{MLP}_{\text{exp}}\Big(\mathrm{LN}(\mathbf{f}_{ij}^{(\ell)})\Big), \quad \alpha_{ij}^{(\ell)} = \mathrm{softmax}_{\ell}\Big(\mathrm{MLP}_{\text{gate}}(\mathrm{LN}(\mathbf{f}_{ij}^{(\ell)}))\Big). \tag{3}$$

The final logit and probability are

$$z_{ij} = \sum_{\ell=1}^{L} \alpha_{ij}^{(\ell)}\, z_{ij}^{(\ell)}, \qquad p_{ij} = \sigma(z_{ij}), \tag{4}$$

as illustrated in Fig. 3.

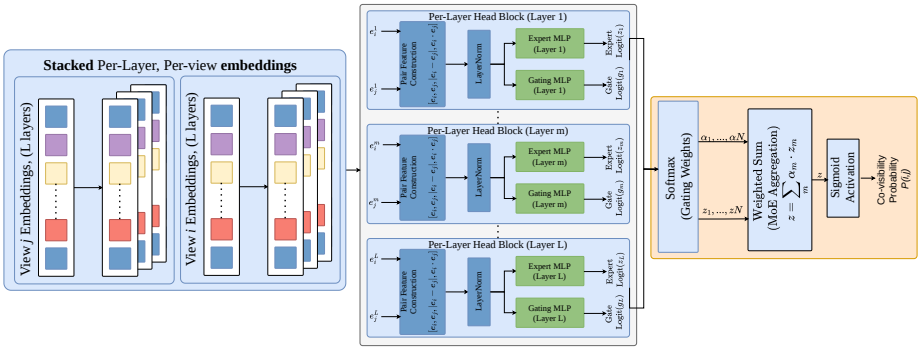

Fig. 3: Co-visibility Estimation Head. Given a set of sparse RGB views (left), our method leverages the frozen Visual Geometry Grounded Transformer (VGGT) to extract layer-wise view embeddings. A lightweight, trainable Mixture-of-Experts (MoE) head (center) adaptively aggregates these multi-scale features to predict pairwise co-visibility probabilities. The resulting predictions form a dense scene-level visibility graph (right), identifying overlapping regions between disjoint viewpoints to guide robust 3D reconstruction and robotic perception.

Training Objective. We optimize the projector, summarizer, expert, and gate parameters with binary cross-entropy on logits:

$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(i,j,y_{ij})\in\mathcal{P}} \ell_{\text{BCELogits}}(z_{ij}, y_{ij}). \tag{5}$$

In pairwise mode $|\mathcal{P}|$=1, while in multiview mode, we average over all labeled pairs in the scene. Although co-visibility resembles image similarity, naive embedding matching is brittle under viewpoint changes, occlusion, and texture ambiguity. This structure empirically reduces false positives in low-overlap regimes while improving calibration.

3.1 Zero-Shot Evaluation

We introduce a zero-shot baseline that predicts co-visibility without any trainable parameters. The classification head and mixture-of-experts are removed, and scores are computed directly from frozen VGGT features using cosine similarity. Given a pair of RGB images, we extract layer activations $\{\mathbf{T}^{(\ell)}\}_{\ell=1}^{L}$ from the frozen backbone and apply a pooling operator $\mathcal{P}(\cdot)$ to obtain per-view embeddings $\mathbf{e}^{(\ell)} \in \mathbb{R}^D$. Embeddings are then computed for each of the considered layers (see Tab. 6 for details).

For a pair (i, j), we compute the cosine similarity per layer and average across layers:

$$c_{ij} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{\langle \mathbf{e}_i^{(\ell)}, \mathbf{e}_j^{(\ell)} \rangle}{\|\mathbf{e}_i^{(\ell)}\|_2 \|\mathbf{e}_j^{(\ell)}\|_2} \in [-1, 1]. \tag{6}$$

Finally, we rescale to $[0, 1]$ to obtain a pseudo-probability $p_{ij}^{\mathrm{ZS}} = (c_{ij} + 1)/2$, which is evaluated using the same metrics as the trained models.

4 Experiments

We evaluate our method on the Co-VisiON benchmark across both HM3D and Gibson environments. We report results for the *pairwise* and *multiview* settings and include the zero-shot method described in Sec. 3. All experiments use AdamW with a learning rate 10^{-4}, a batch size of 32, and a weight decay 10^{-4}, trained for 50 epochs using the results corresponding to the best AUC value. In the multiview setting, peak performance is reached around 30 epochs, whereas in the pairwise setting, it converges within 10 epochs.

4.1 Dataset & Metrics

Co-VisiON [8] evaluates co-visibility reasoning in sparse indoor view sets. Each sample consists of a small collection of RGB views from the same scene, and the task is to predict a binary co-visibility graph, where an edge (i, j) is positive iff the two views share non-zero co-visible surface area. The benchmark spans Gibson [38] (80/20 split) and HM3D [23] (90/10 split), comprising 85/755 scenes and 33,849/210,008 labeled pairs, respectively. Moreover, all validation scenes are disjoint from training scenes, so performance reflects generalization to unseen environments, akin to a zero-shot setting. For detailed specifics, refer to [8].

The dataset also reports a human baseline for the Gibson multiview setting. However, this should not be interpreted as a definitive measure of human-level performance, but rather as an indicative reference point. Given the presence of artifacts in Gibson, the true performance achievable by humans under cleaner conditions is likely higher.

We use the same protocol for *pairwise* and *multiview* settings since both produce pairwise co-visibility scores that define an adjacency matrix. For a scenario with N views, the model outputs scores $d_{ij} \in [0, 1]$, forming $\mathbf{D} \in [0, 1]^{N \times N}$, which are compared to the ground-truth adjacency $\mathbf{A} \in \{0, 1\}^{N \times N}$. The only difference is how scores are computed: independently per pair in the pairwise setting, or jointly from multiple views in the multiview setting.

Given a threshold τ, we binarize $\mathbf{D}$ as $\hat{\mathbf{A}}(\tau) = \mathbb{I}[\mathbf{D} \geq \tau]$ (symmetric, zero diagonal) and compute the *Graph IoU*: $\mathrm{IoU}(\tau) = \frac{|\mathcal{E} \cap \hat{\mathcal{E}}(\tau)|}{|\mathcal{E} \cup \hat{\mathcal{E}}(\tau)|}$. We report the best-threshold IoU (denoted as IoU^* in the tables), $\mathrm{IoU}^* = \max_{\tau \in [0,1]} \mathrm{IoU}(\tau)$, and the area under the IoU-threshold curve over $\tau \in [0, 1]$, defined as $\mathrm{AUC} = \int_0^1 \mathrm{IoU}(\tau)\, d\tau$. Metrics are computed per scenario and averaged over the split. Additional metrics results (e.g., F1 score, Accuracy) are provided in Sec. A.6 of the Supplementary material.

4.2 Results

Tabs. 1–2 report co-visibility prediction performance on Co-VisiON in the pairwise and multiview settings, measured with Graph IoU* and AUC. Our Co-

VGGT consistently achieves the best results across datasets and evaluation protocols. In the pairwise setting (Tab. 1), Co-VGGT attains 0.85/0.78 IoU*/AUC on Gibson and 0.84/0.78 on HM3D, substantially outperforming strong learned baselines such as Covis [8] (0.56/0.54 on Gibson; 0.53/0.51 on HM3D) and reconstruction based DUSt3R [35] (0.54/0.54 on Gibson; 0.40/0.40 on HM3D). Prompting large VLMs yields competitive results relative to earlier learned models (e.g., GPT-4o [22] reaches 0.58/0.58 on Gibson and 0.54/0.54 on HM3D), but remains far behind Co-VGGT, indicating that explicit geometric specialization is critical for reliable overlap reasoning under sparse viewpoints.

In the multiview setting (Tab. 2), Co-VGGT again leads with 0.74/0.72 on Gibson and 0.76/0.74 on HM3D, surpassing Covis/Covis-freeze and multiview reconstruction (MV-DUSt3R+ [30]). Interestingly, performance in multiview is lower than in pairwise, which is counterintuitive for reconstruction-style models but can be explained by our current multiview embedding extraction, which aggregates features across many views and yields noisier per-view vectors for the downstream pair classifier.

On Gibson, Co-VGGT also exceeds the reported Human Annotation baseline (0.72/0.72) [8], suggesting that the learned head on top of frozen VGGT features captures geometric cues that are difficult to assess reliably from sparse RGB images presented to humans.

We note that for sparse view sets (small N), running the pairwise model exhaustively over all $\binom{N}{2}$ pairs and assembling the graph post-hoc yields superior calibration and accuracy (0.85 vs. 0.74 IoU*) at negligible additional cost, and is therefore the recommended inference strategy in practice. The multiview mode offers a scalable alternative for larger view sets where the quadratic cost becomes prohibitive.

Tab. 3 reports validation performance under two complementary difficulty protocols that stress different sparsity regimes of the co-visibility graph. Edge-level difficulty bins individual pairs by their overlap ratio: *easy* if overlap $\geq 50\%$, *medium* if $10\% \leq$ overlap $< 50\%$, and *hard* if overlap $< 10\%$. While VLM-based baselines are near-saturated in the easy/medium regimes, they degrade sharply when overlap becomes minimal; in particular, on the hard split Co-VGGT achieves substantially higher Graph-IoU than the strongest VLM baseline (e.g., 0.84 vs. 0.34 for GPT-4o [22]), indicating markedly better robustness to low-overlap pairs. We further report graph-level difficulty by binning entire scenes according to scene sparsity, defined as the average pairwise overlap within the scenario: *easy* if $\geq 10\%$, *medium* if $4\% \leq \cdot < 10\%$, and *hard* if $< 4\%$. Even in globally sparse scenes, Co-VGGT maintains strong performance (hard: 0.84 Graph-IoU), outperforming both Covis-based baselines and GPT-4o (hard: 0.57), showing that the method remains reliable when the overall graph connectivity is low, rather than just a few pairs being difficult. Overall, these results suggest that Co-VGGT encodes more stable geometric cues than VLM prompting or learned baselines, and that its advantage concentrates exactly where co-visibility is most failure-prone: extremely low pairwise overlap and scene-wide sparse connectivity. AUC result tables are reported in the Supplementary material.

Table 1: Pairwise co-visibility prediction. Results on Gibson and HM3D datasets. Baselines results are reported from [8], no human baseline available in this case.

Method	Gibson		HM3D	
	IoU* (↑)	**AUC** (↑)	**IoU*** (↑)	**AUC** (↑)
SuperGlue [25]	0.47	0.11	0.38	0.10
SIFT + RANSAC [18]	0.35	0.05	0.34	0.05
NetVLAD [2]	0.35	0.33	0.35	0.29
DUSt3R [35]	0.54	0.54	0.40	0.40
ViT [11]	0.47	0.43	0.48	0.46
VGG [27]	0.35	0.30	0.34	0.21
ResNet18 [14]	0.38	0.36	0.48	0.46
CroCo v2 [37]	0.50	0.45	0.43	0.38
Covis [8]	0.56	0.54	0.53	0.51
Qwen2.5-VL 72B [3]	0.41	0.41	0.39	0.39
Gemini-2.0-Flash [31]	0.42	0.42	0.39	0.39
SpatialRGPT [9]	0.49	0.49	0.37	0.37
GPT-4o [22]	0.58	0.58	0.54	0.54
Co-VGGT (Zero-shot)	0.50	0.31	0.46	0.31
Co-VGGT (Ours)	**0.85**	**0.78**	**0.84**	**0.78**

Even without any training, the zero-shot variant achieves 0.50 IoU* on the Gibson pairwise setting—already competitive with several supervised baselines (Tab. 6)—confirming that VGGT's frozen representations carry a meaningful co-visibility signal that supervised fine-tuning fully unlocks.

4.3 Pair Aggregation

Tab. 4 studies how the choice of symmetric pair features impacts performance. Using only raw embeddings (simple) is weakest, while adding multiplicative interactions ($e_i \odot e_j$) improves results (product). Including absolute differences ($|e_i - e_j|$) provides the largest single gain, suggesting that relative feature offsets capture key overlap cues. Our full aggregation ($[e_i, e_j], |e_i - e_j|, e_i \odot e_j$) combined with the layer-wise MoE yields the best IoU*/AUC, outperforming the same feature set under the standard (non-MoE) aggregation baseline.

Notably, the `absolute` variant matches Co-VGGT in IoU* (0.74) but lags in AUC (0.71 vs. 0.73), suggesting that the element-wise product $\mathbf{e}_i \odot \mathbf{e}_j$ contributes primarily to probability calibration rather than threshold-optimal classification: a meaningful distinction when predicted scores are used as continuous edge weights in downstream visibility graphs.

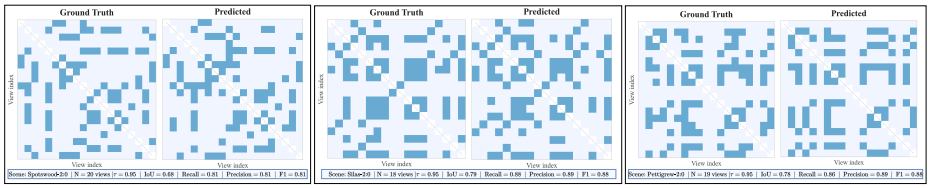

Fig. 4: Co-visibility Scene-Graph Examples (Multiview). Given N input views, we visualize the ground-truth adjacency matrix (left), the predicted binary graph obtained by thresholding scores at the IoU-optimal τ^*. Each matrix is symmetric and entry (i, j) denotes the relation between image i and image j.

Table 2: Multiview co-visibility prediction. Results on Gibson and HM3D datasets. Baselines and Human results are reported from [8].

Method	Gibson		HM3D	
	IoU* (↑)	**AUC** (↑)	**IoU*** (↑)	**AUC** (↑)
Human Annotation [8]	0.72	0.72	–	–
NetVLAD [2]	0.38	0.32	0.42	0.39
MV-DUSt3R+ [30]	0.56	0.56	0.45	0.45
CroCo v2 [37]	0.48	0.41	0.41	0.37
Covis [8]	0.59	0.57	0.57	0.56
Covis-freeze [8]	0.61	0.57	0.58	0.56
GPT-4o [22]	0.63	0.63	0.59	0.59
Co-VGGT (Zero-shot)	0.37	0.30	0.36	0.30
Co-VGGT (Ours)	**0.74**	**0.73**	**0.76**	**0.74**

4.4 Hierarchical Information

As shown in Fig. 6, co-visibility reasoning is concentrated in VGGT's late layers. Gate mass is essentially zero for layers 0–12, confirming that early representations encode geometric primitives rather than overlap cues. Among late layers, we observe a strong class asymmetry. Negative pairs consistently route to L17 (peak ≈0.37–0.41) with low gating entropy across both pairwise and multiview settings, making L17 a negative anchor that confidently rejects geometrically disjoint pairs.

This is supported by attention maps (see supplementary material): for non-co-visible pairs, the first view yields focused, structured attention, while the second produces diffuse, unanchored responses — a signature of failed cross-image correspondence at a geometrically mature layer. Positive pairs, by contrast, show higher gating entropy and a setting-dependent preference: multiview routing sharpens to L15 (peak ≈0.41), while pairwise spreads across L18–L23 (peak ≈0.19–0.23), reflecting greater uncertainty about which layer carries the overlap signal when no auxiliary views are available.

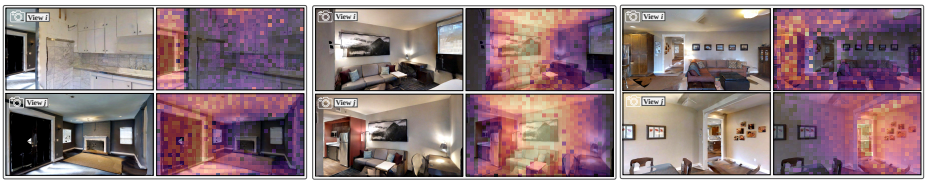

Fig. 5: Cross-view Similarity Matrix Examples. Examples of Co-VGGT attention token similarity over pair of co-visibile images. (left) input RGB image, (right) the attention output mask. Features are extracted from signal belonging to layer 17 of the Pairwise evaluation.

Table 3: Performance across difficulty levels. Left: edge-level difficulty defined by pairwise image overlap (Easy $\geq$ 50%, Med. 10–50%, Hard $<$ 10%). Right: graph-level difficulty defined by scene sparsity via average overlap (Easy $\geq$ 10%, Med. 4–10%, Hard $<$ 4%). Metrics report *Graph IoU* at the best threshold.

(a) Image overlap (edge-level).

Method	Easy	Med.	Hard	Avg.
GPT-4o [22]	0.97	0.92	0.34	0.63
Gemini-2.0-Flash [31]	**1.00**	**0.99**	0.14	0.42
Qwen2.5-VL 72B [3]	**1.00**	**0.99**	0.14	0.41
SpatialRGPT [9]	**1.00**	**0.99**	0.13	0.35
Covis [8]	0.99	0.88	0.24	0.59
Covis (freeze) [8]	**1.00**	0.89	0.30	0.61
Co-VGGT (Multi)	0.80	0.88	0.70	0.74
Co-VGGT (Pair)	**1.00**	0.93	**0.84**	**0.85**

(b) Scene sparsity (graph-level).

Method	Easy	Med.	Hard	Avg.
GPT-4o [22]	0.83	0.64	0.57	0.63
Gemini-2.0-Flash [31]	0.75	0.39	0.28	0.42
Qwen2.5-VL 72B [3]	0.77	0.40	0.28	0.41
SpatialRGPT [9]	0.72	0.38	0.26	0.35
Covis [8]	0.80	0.63	0.52	0.59
Covis (freeze) [8]	0.81	0.65	0.54	0.61
Co-VGGT (Multi)	0.99	0.92	0.73	0.74
Co-VGGT (Pair)	**1.00**	**0.97**	**0.84**	**0.85**

This structure directly explains the calibration gap reported below: multi-view exhibits sharp, consistent routing but noisier embeddings from compressing multiview context into fixed-size vectors; pairwise produces cleaner embeddings and better calibration but has flatter routing due to the absence of additional scene context.

4.5 Calibration Measures

In Fig. 7, we assess probability calibration on Gibson val using Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Brier score. In the pairwise setting, Co-VGGT is well calibrated (ECE = 0.030, Brier = 0.043): predicted scores behave as empirical frequencies (e.g., $p{\approx}0.7$ implies $\sim$70% true positives), making them directly usable as edge weights for downstream filtering keeping $p{>}0.7$ for matching and discarding $p{<}0.2$ (to suppress spurious constraints) without any post-hoc correction. In multiview, calibration degrades (ECE = 0.074, Brier = 0.085), consistent with the noisier embeddings produced by compressing multiview context into fixed-size vectors, as discussed in Sec. 4.6. Both settings exhibit high MCE (0.223 pairwise, 0.347 multiview); however, this

Table 4: Aggregator ablation. All aggregators use MoE except the standard aggregation baseline. "Standard" denotes concatenation of $([e_i, e_j], |e_i - e_j|, e_i \odot e_j)$.

	Pair feature set			MoE	**Gibson**		**HM3D**	
Pair Aggregator	$\lvert e_i - e_j \rvert$	$e_i \odot e_j$	$[e_i, e_j]$	w/ or w/o	IoU*	AUC	IoU*	AUC
simple	✗	✗	✓	✓	0.67	0.70	0.69	0.70
product	✗	✓	✓	✓	0.71	0.70	0.72	0.71
absolute	✓	✗	✓	✓	0.74	0.71	0.73	0.72
standard	✓	✓	✓	✗	0.69	0.68	0.72	0.70
Co-VGGT (Ours)	✓	✓	✓	✓	**0.74**	**0.73**	**0.76**	**0.74**

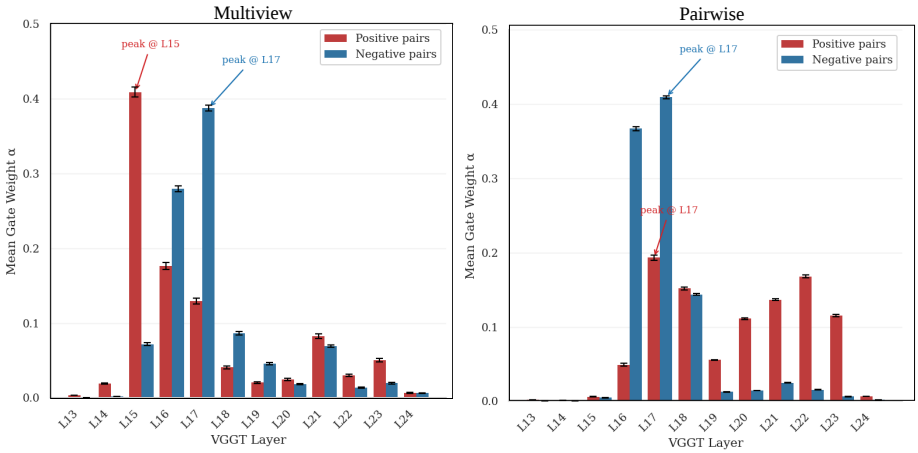

Fig. 6: MoE Gating weights. Average α parameter per-layer vs. layer id. on the multiview (left) and pairwise (right) task. We observe that specific layers are decisive for the final co-visibility analysis.

is driven by sparsely populated mid-confidence bins, since over 90% of predictions fall in [0, 0.1) or [0.9, 1]. Furthermore, we added a proof-of-concept COLMAP experiment against different methods to show this downstream capability (see Supplementary material).

4.6 Layer Ablation

Tab. 5 shows that restricting the MoE expert pool to the last 12 layers preserves nearly full performance (0.74/0.72 IoU*/AUC on Gibson), while using fewer layers causes consistent degradation. This is directly corroborated by the gating analysis (Fig. 6), where layers 0–12 receive near-zero average weight across both classes, confirming that the model has learned to ignore early layers. Since VGGT inference dominates runtime (50ms/150ms per 2/10 images on an H100) and the MoE head is lightweight, retaining all layers adds negligible overhead — we therefore do so in our final model. Together, these findings provide task-grounded empirical evidence for the hierarchical geometric reasoning hypothe-

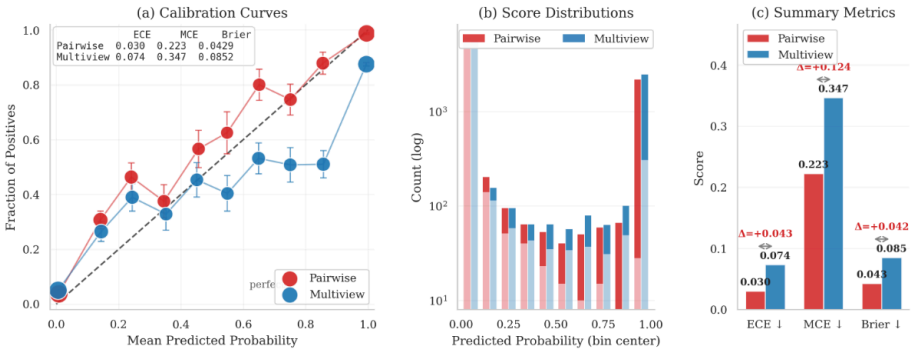

Fig. 7: Calibration Measures. Calibration plot (left), score distribution (center) and summary metrics (right) on Gibson validation for pairwise and multiview tasks.

sized in [34]: early layers carry no discriminative signal for co-visibility, while late layers encode the high-level judgment of accepting or rejecting shared surface support, with positive and negative pairs peaking at distinct layers (L15 and L17, respectively). For completeness, we evaluate our approach using only the first 12 layers of the VGGT backbone (last row of Tab. 5). The resulting performance drops substantially below the state of the art, confirming that early layers encode more generic and diverse representations that are less informative for co-visibility reasoning.

4.7 Cross-domain transfer

To probe whether the learned co-visibility signal is tied to a single training distribution, we evaluate Co-VGGT trained on one Co-VisiON environment and tested zero-shot on the other (Tab. 5). Transferring across Gibson-HM3D costs at most 0.04 IoU* and 0.03 AUC in both pairwise and multiview settings, while still exceeding the in-domain state of the art on the target dataset. Combined with Co-VisiON's scene-disjoint train/val protocol, this indicates the extracted signal does not overfit to a single environment, though both datasets remain indoor Habitat-rendered scenes and broader cross-domain generalization is left to future work.

4.8 Zero-Shot Ablation

As shown in Tab. 1, the zero-shot performance of Co-VGGT is unexpectedly strong, indicating that the backbone already encodes a substantial signal for identifying co-visible pairs.

To further investigate this behavior, Tab. 6 presents a zero-shot ablation where we vary the subset of VGGT backbone layers used for similarity computation and subsequent IoU and AUC evaluation (analogous to Tab. 5). Performance remains within a relatively narrow range across different subsets. However, across

Table 5: Ablation and cross-domain transfer results. Left: ablation on the number of VGGT expert-pool layers used in the MoE head, evaluated on Gibson. Right: cross-domain transfer between Gibson and HM3D.

Layer-count ablation				
	Multiview		Pairwise	
MoE Range	IoU* ↑	AUC↑	IoU* ↑	AUC↑
[1, 24] (Ours)	**0.74**	**0.73**	**0.84**	**0.78**
[14, 24]	**0.74**	0.72	**0.84**	0.77
[19, 24]	0.71	0.70	0.83	0.75
[21, 24]	0.70	0.67	0.83	0.75
[24]	0.69	0.66	0.82	0.74
[1, 12]	0.53	0.50	0.30	0.11

Cross-domain transfer					
Train	Test	Pairwise		Multiview	
		IoU* ↑	AUC↑	IoU* ↑	AUC↑
Gibson	Gibson	**0.85**	**0.78**	**0.74**	**0.73**
HM3D	HM3D	**0.84**	**0.78**	**0.76**	**0.74**
Gibson	HM3D	0.81	0.75	0.72	0.71
HM3D	Gibson	0.83	0.76	0.72	0.72

Table 6: Zero-shot layer ablation. Ablation on the number of VGGT output feature layers used for cosine similarity scoring in zero-shot evaluation, reported on both Gibson and HM3D datasets for Multiview and Pairwise tasks.

Layer Range	Gibson				HM3D			
	Multiview		Pairwise		Multiview		Pairwise	
	IoU*(↑)	AUC(↑)	IoU*(↑)	AUC(↑)	IoU*(↑)	AUC(↑)	IoU*(↑)	AUC(↑)
[1, 24] (Ours)	0.37	0.30	0.50	0.31	0.36	**0.30**	0.46	0.31
[9, 24]	0.37	0.30	0.50	0.32	0.36	**0.30**	0.46	0.31
[19, 24]	**0.38**	**0.31**	**0.52**	**0.36**	**0.37**	**0.30**	**0.47**	**0.33**
[21, 24]	0.36	0.30	0.49	0.34	0.36	**0.30**	0.45	0.30

all tasks and datasets, the strongest results are consistently achieved when using embeddings from the last five layers. Additional analyzes and extended experiments are provided in Sec. A.1 of the Supplementary material.

5 Limitations

While Co-VGGT achieves strong results on the Co-VisiON benchmark, some limitations remain. We frame co-visibility prediction as a binary signal since the focus of this work is primarily on interpretability and understanding the overlap signal already present in VGGT, given that a simple co-visibility estimator already achieves strong performance. Nevertheless, we also conducted an auxiliary experiment that predicts, via a regression head, a graded co-visibility strength (see Supplementary material). We do not further develop this direction here; instead, we leave it as future work, where richer overlap estimates could provide more informative geometric constraints for downstream SfM and SLAM pipelines.

Moreover, the multiview setting exposes a structural limitation of our architecture: compressing multiview context into fixed-dimensional per-view embeddings introduces noise, and the model scores each pair independently without enforcing global graph consistency — leading to a performance gap relative to

the pairwise setting. Further failure examples and analysis are provided in the Supplementary material.

Finally, the mechanistic interpretation of layer-wise gating — while empirically grounded — remains correlational: we identify *which* layers are decisive, but not entirely *why* specific geometric abstractions emerge at those depths in VGGT's training. However, we argue that the emergent behavior of VGGT in 3D vision may mirror that of LLMs in NLP; consequently, several insights and solutions could potentially be transferred from that domain.

6 Conclusion

We presented Co-VGGT, a lightweight co-visibility predictor on frozen VGGT features, with well-calibrated pairwise probabilities (ECE=0.030) that are usable directly as visibility-graph edge weights. Our layer-wise mixture-of-experts head reveals that co-visibility reasoning emerges exclusively in VGGT's late transformer blocks: L17 acts as a negative anchor that confidently rejects non-overlapping pairs, while positive pairs rely on a broader set of late layers depending on the available context. This provides task-grounded evidence for hierarchical geometric reasoning in a geometry-grounded foundation model, mirroring the layer-specialization behavior observed in large language models.

These findings suggest that geometry-grounded foundation models encode spatial priors that are richer and more structured than previously understood — and that these priors can be efficiently distilled for downstream tasks without modifying the backbone. Future work will pursue four directions: (i) moving beyond binary prediction toward continuous overlap ratio estimation to supply richer geometric constraints for SfM and SLAM pipelines; (ii) replacing the current per-pair multiview loop with a unified aggregation mechanism — such as set transformers or token-level routing — to enforce global graph consistency and close the pairwise–multiview calibration gap; and (iii) integrating Co-VGGT within embodied mapping systems to enable uncertainty-aware loop closure and next-best-view planning grounded in geometric connectivity rather than appearance alone. (iv) Investigate whether interpretability findings from NLP transfer to geometric foundation models, enabling principled interpretability methods for this domain.

Acknowledgements

We thank Fondazione Bruno Kessler (FBK) and the University of Padua, Department of Mathematics "Tullio Levi-Civita", for providing the computational resources used in this work. TC and LS were supported by the PNRR project Future Artificial Intelligence Research (FAIR, PE00000013) under the NRRP MUR program, funded by NextGenerationEU.

References

1. An, V.D., Vu, M.N., Reid, I.D.: Improving robotic manipulation with efficient geometry-aware vision encoder. arXiv preprint arXiv:2509.15880 (2025). https://doi.org/10.48550/arXiv.2509.15880
2. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: CVPR (2016)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
4. Bavle, H., Sanchez-Lopez, J.L., Civera, J., Voos, H.: Situational graphs for robot navigation in structured indoor environments. arXiv preprint arXiv:2202.12197 (2022)
5. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1050–1060 (2025). https://doi.org/10.1109/CVPR52734.2025.00106
6. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. In: T-RO. vol. 37, pp. 1874–1890 (2021)
7. Chen, A., Xu, Z., Zhao, J., Yu, J., Su, H., Zhang, J., Yu, J.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: ICCV (2021)
8. Chen, C., Dang, N., Zhang, J., Sun, W., Zheng, P., He, X., Ye, Y., Zhang, J., Srinivas, T., Feng, C.: Co-vision: Co-visibility reasoning on sparse image sets of indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4802–4812 (2025)
9. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatial-rgpt: Grounded spatial reasoning in vision-language models. In: NeurIPS (2024)
10. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 conference on empirical methods in natural language processing. pp. 670–680 (2017)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Geva, M., Katz, U., Ben-Arie, A., Berant, J.: What's in your head? emergent behaviour in multi-task transformer models. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 8201–8215 (2021)
13. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 855–864 (2016)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Hughes, N., Chang, Y., Carlone, L.: Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. arXiv preprint arXiv:2201.13360 (2022)

16. Jiang, H., Karpur, A., Cao, B., Huang, Q.: Omniglue: Generalizable feature matching with foundation model guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19865–19875 (2024). https://doi.org/10.1109/CVPR52733.2024.01878
17. Lee, S., Choi, J., Kang, I., Kim, J., Park, J., Shim, H.: 3d-aware vision-language models fine-tuning with geometric distillation. arXiv preprint arXiv:2506.09883 (2025). https://doi.org/10.48550/arXiv.2506.09883
18. Lindeberg, T.: Scale invariant feature transform. Scholarpedia **7**, 10491 (05 2012). https://doi.org/10.4249/scholarpedia.10491
19. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17581–17592 (2023). https://doi.org/10.1109/ICCV51070.2023.01616
20. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt. Advances in neural information processing systems **35**, 17359–17372 (2022)
21. Mou, L., Men, R., Li, G., Xu, Y., Zhang, L., Yan, R., Jin, Z.: Natural language inference by tree-based convolution and heuristic matching. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 130–136 (2016)
22. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
23. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. arXiv preprint arXiv:2109.08238 (2021)
24. Rosinol, A., Abate, M., Chang, Y., Carlone, L.: Kimera: an open-source library for real-time metric-semantic localization and mapping. IEEE International Conference on Robotics and Automation (ICRA) (2020)
25. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR (2020)
26. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Stary, M., Gaubil, J., Tewari, A., Sitzmann, V.: Understanding multi-view transformers. arXiv preprint arXiv:2510.24907 (2025)
29. Sucar, E., Liu, S., Ortiz, J., Davison, A.J.: imap: Implicit mapping and positioning in real-time. In: ICCV (2021)
30. Tang, Z., Fan, Y., Wang, D., Xu, H., Ranjan, R., Schwing, A., Yan, Z.: Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. arXiv preprint arXiv:2412.06974 (2024)
31. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., et al.: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
32. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In: NeurIPS (2021)
33. Tenney, I., Das, D., Pavlick, E.: Bert rediscovers the classical nlp pipeline. In: Proceedings of the 57th annual meeting of the association for computational linguistics. pp. 4593–4601 (2019)
34. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotný, D.: Vggt: Visual geometry grounded transformer. In: IEEE/CVF Conference on Computer

Vision and Pattern Recognition (CVPR) 2025, Nashville, TN, USA, June 11-15, 2025. pp. 5294–5306. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.00499
35. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20697–20709 (2024). https://doi.org/10.1109/CVPR52733.2024.01956
36. Wang, Y., He, X., Peng, S., Tan, D., Zhou, X.: Efficient loftr: Semi-dense local feature matching with sparse-like speed. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21666–21675 (2024). https://doi.org/10.1109/CVPR52733.2024.02047
37. Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J.: Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17969–17980 (2023)
38. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9068–9079 (2018)
39. Xu, Z., Fan, Y., Bao, J., Zhang, D., Li, H., Zhang, D.: Mvfusion: A multi-view diffusion model for 3d reconstruction. In: CVPR (2023)
40. Xue, M., Huang, Q., Zhang, H., Cheng, L., Song, J., Wu, M., Song, M.: Protopformer: Concentrating on prototypical parts in vision transformers for interpretable image recognition. arXiv preprint arXiv:2208.10431 (2022)
41. Zhang, T., Usenko, V., Engel, J., Cremers, D.: Bad-slam: Bundle adjusted direct rgb-d slam. In: CVPR (2021)
42. Zhu, Z., Peng, S., Larsson, V., Lin, C.H., Bao, H., Dai, A., Nießner, M., Zhou, X.: Nice-slam: Neural implicit scalable encoding for slam. In: CVPR (2022)