CALIBADAPTER: ADAPTING 3D FOUNDATION MODELS TO CALIBRATED CAMERAS

Anonymous authors
Paper under double-blind review

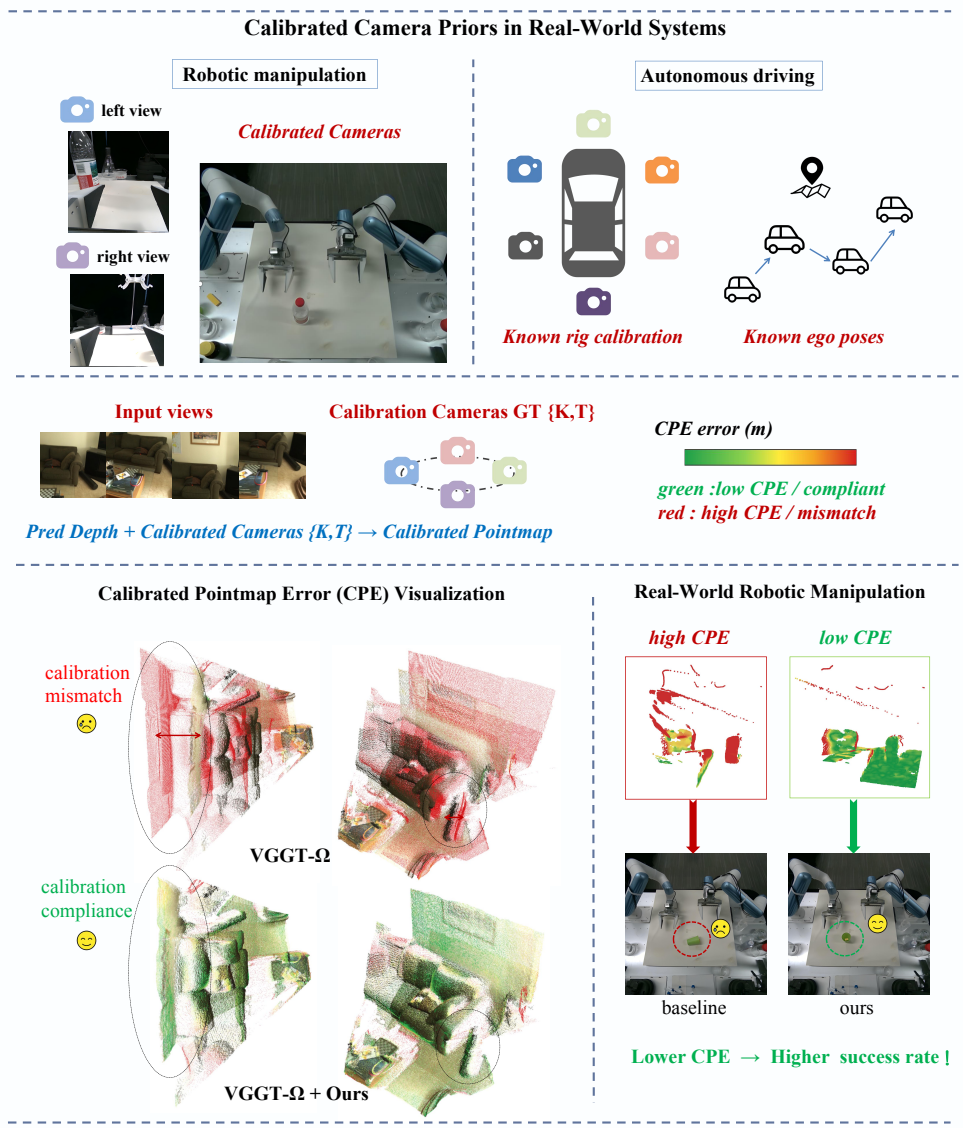

Figure 1: **Calibrated geometry in practical 3D perception.** CalibAdapter reduces CPE with provided camera calibration, improving metric geometry for downstream tasks.

ABSTRACT

Recent feed-forward 3D foundation models have demonstrated strong capabilities in reconstructing scene geometry, supporting a wide range of downstream tasks. In practical settings such as robotic manipulation and autonomous driving, calibrated camera intrinsics and extrinsics are usually readily available through standard sensor-calibration pipelines. Existing works like MapAnything, DepthAnything3, and PromptDA have already shown that utilizing calibrated camera parameters can improve reconstruction accuracy. However, these methods do not guar-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

antee that the predicted depth will back-project to geometry consistent with the provided calibration. We quantify this deployment discrepancy using **Calibrated Pointmap Error (CPE)**, which measures the resulting pixel-corresponded metric 3D error. To address this discrepancy, we propose **CalibAdapter**, a lightweight adaptation mechanism for incorporating known camera calibration into 3D foundation models. CalibAdapter attaches a projective sparse residual branch to a selected cross-view interaction block while keeping all pretrained parameters frozen. CalibAdapter consistently reduces CPE across **VGGT-Ω**, **VGGT**, **DA3**, and π^3, achieving an average relative CPE reduction of **28.3%** together with overall improvements in conventional reconstruction metrics. It further yields gains on several held-out benchmarks. Beyond conventional reconstruction benchmarks, CalibAdapter improves geometric accuracy in **autonomous driving with known rig calibration and ego poses**, and increases **real-world robotic manipulation** success rate by 18.4 percentage points over the parameter-matched baseline.

1 INTRODUCTION

Feed-forward 3D foundation models have demonstrated strong capabilities in scene reconstruction and 3D perception (Wang et al., 2024; 2025a; Lin et al., 2026; Wang et al., 2026b), and are increasingly used as general-purpose backbones for robotic manipulation and autonomous driving (Guo et al., 2025a; Peng et al., 2026b; Jia et al., 2025). In many such deployment settings, calibrated camera intrinsics and extrinsics are readily available through standard sensor-calibration pipelines. Recent methods therefore exploit camera parameters either as conditioning signals or through post-hoc geometric alignment (Jang et al., 2025; Keetha et al., 2026; Peng et al., 2026a; Lin et al., 2026), often improving conventional depth and point-cloud metrics. Yet these gains do not necessarily ensure accurate metric geometry under calibrated back-projection after a single trajectory-level scale recovery and without further alignment. As illustrated in Figure 1, such calibrated camera priors are naturally available in practical systems such as autonomous driving (Li et al., 2022; Wei et al., 2023; Huang et al., 2023b; Zhang et al., 2023b) and robotic manipulation (Shridhar et al., 2023; Goyal et al., 2023; Gervet et al., 2023; Ke et al., 2025; Shridhar et al., 2022; Lu et al., 2024; Ze et al., 2023; Chen et al., 2023; Huang et al., 2023a; Yuan et al., 2023; Mo et al., 2021), making this discrepancy particularly relevant in real-world deployment.

We use *camera compliance* to denote whether scale-restored predicted depth yields accurate metric geometry when directly back-projected with the provided calibration, and introduce *Calibrated Pointmap Error (CPE)* to quantify the resulting 3D error. CPE measures the pixel-corresponded 3D discrepancy between scale-restored prediction and ground truth after back-projection with the same provided camera intrinsics and extrinsics, allowing only a single trajectory-level scale shared across all views. Unlike common evaluation protocols that use per-image scale–shift fitting or post-hoc similarity alignment, CPE keeps the provided camera intrinsics and extrinsics unchanged and directly evaluates the resulting back-projected geometry. This raises a central question: *in settings with known camera calibration, how can pretrained 3D foundation models be efficiently adapted to improve metric geometry under direct calibrated back-projection?*

To answer this question, we propose **CalibAdapter**, a lightweight framework for parameter-efficient adaptation of 3D foundation models to known camera calibration. Rather than modifying the pretrained reconstruction pathway, we freeze all pretrained parameters and introduce a projective sparse residual branch at a selected cross-view interaction block, where calibrated camera geometry is injected into cross-view feature interactions. Our layer-wise analysis further shows that early-to-middle interaction layers provide favorable insertion points compared with late layers. Across VGGT-Ω, VGGT, DA3, and π^3, CalibAdapter consistently reduces CPE while maintaining competitive reconstruction quality. It further yields gains on several held-out benchmarks and improves geometric accuracy in autonomous driving with known rig calibration and ego poses. In a controlled real-world robotic manipulation study, CalibAdapter increases task success by 18.4 percentage points over the parameter-matched baseline.

Our main contributions are summarized as follows:

- Calibrated Pointmap Error (CPE), an operational metric for calibrated deployment that measures the pixel-corresponded metric 3D error obtained after a single trajectory-level scale recovery and direct back-projection with the provided calibration.

- CalibAdapter, a lightweight framework for parameter-efficient adaptation of 3D foundation models to known camera calibration, using a zero-initialized projective sparse residual branch for calibrated cross-view interaction.
- Consistent CPE reductions across diverse 3D foundation models, together with cross-dataset transfer while maintaining competitive reconstruction quality.
- Experiments on autonomous driving and real-world robotic manipulation further demonstrate the practical utility of CalibAdapter, improving geometric reconstruction while increasing manipulation task success by 18.4 percentage points over the baseline.

2 Related Work

2.1 Feed-Forward 3D Reconstruction

Classical SfM jointly recovers camera poses and scene structure through feature matching and geometric optimization (Schönberger & Frahm, 2016), whereas learned MVS estimates dense geometry from calibrated multi-view images (Yao et al., 2018; 2019; Gu et al., 2020; Wang et al., 2021; Ding et al., 2022; Cao et al., 2024; Ji et al., 2017; Chen et al., 2019; Cheng et al., 2020; Yang et al., 2020). Feed-forward approaches instead infer 3D structure directly from images (Charatan et al., 2024; Chen et al., 2024; Szymanowicz et al., 2024; Ye et al., 2025; Xu et al., 2025a; Zhang et al., 2025b; Xu et al., 2024). DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) establish point-map prediction and matching, with subsequent work extending this paradigm toward scalable, stateful, streaming, and dynamic reconstruction (Duisterhof et al., 2025; Yang et al., 2025; Tang et al., 2025; Cabon et al., 2025; Wang & Agapito, 2025; Wang et al., 2025b; Liu et al., 2025; Murai et al., 2025; Wu et al., 2025; Zhang et al., 2025a; Lu et al., 2025). Recent feed-forward 3D foundation models further unify geometric prediction, metric-scale reconstruction, and large-scale reconstruction (Wang et al., 2025a; 2026c;b; Wang & Agapito, 2026; Elflein et al., 2026). These models are commonly evaluated after resolving gauge ambiguity through scale, scale–shift, or similarity alignment, which does not directly report metric 3D error under the provided calibration without further alignment.

2.2 Camera-Conditioned Reconstruction and Geometric Encoding

Camera and geometric priors have increasingly been incorporated into feed-forward reconstruction. Existing approaches condition reconstruction on camera parameters, depth, or partial geometry through prompting, feature modulation, or geometry-aware adaptation (Jang et al., 2025; Keetha et al., 2026; Lin et al., 2025; Peng et al., 2026a; Khafizov et al., 2025). Pi3X extends π^3 (Wang et al., 2026c) with optional camera, intrinsic, and depth conditioning. Other methods explicitly exploit known poses or structured camera rigs to constrain multi-view geometry (Lin et al., 2026; Li et al., 2025b; Philion & Fidler, 2020; Li et al., 2023a;b). Related monocular metric-depth methods address camera-dependent scale variation through geometric normalization or camera-aware representations, enabling metric prediction across heterogeneous camera models (Yin et al., 2023; Guizilini et al., 2023; Piccinelli et al., 2024; Guo et al., 2025b).

At a finer representation level, camera and ray geometry can be incorporated into feature interactions or positional representations, enabling multi-view features to respect projective relationships (Wang et al., 2022; Liu et al., 2022; 2023; Wang et al., 2023; Miyato et al., 2024; Li et al., 2025a; Zhang et al., 2024; Wu et al., 2026; Zhang et al., 2026; Li et al., 2026; Yin et al., 2026).

2.3 Efficient Interaction and Model Adaptation

Sparse cross-view interaction has been explored to reduce the computational cost of multi-view geometry transformers (Wang et al., 2026a). Beyond interaction efficiency, parameter-efficient fine-tuning adapts pretrained models through lightweight modules, low-rank updates, prompts, or feature modulation (Houlsby et al., 2019; Hu et al., 2022; Zhang et al., 2023a; Jia et al., 2022; Chen et al., 2022; Jie et al., 2024; Lian et al., 2022; Jie & Deng, 2023; Zhang et al., 2025c; Luo et al., 2023; He et al., 2023; Dong et al., 2023). In feed-forward 3D reconstruction, Fin3R (Ren et al., 2025) combines monocular geometric distillation with parameter-efficient tuning to improve reconstruction while preserving multi-view matching. Test-time optimization provides another adaptation regime, with TCO (Zhou et al., 2026) specializing reconstruction models separately for each test scene under geometric priors. CalibAdapter instead adapts frozen 3D foundation models to known camera calibration with fixed, single-pass inference.

3 METHOD

Figure 2 summarizes CalibAdapter and its zero-initialized projective sparse residual branch. We first formalize camera compliance and CPE, then describe calibrated residual adaptation, projective camera conditioning, and geometry-conditioned sparse routing.

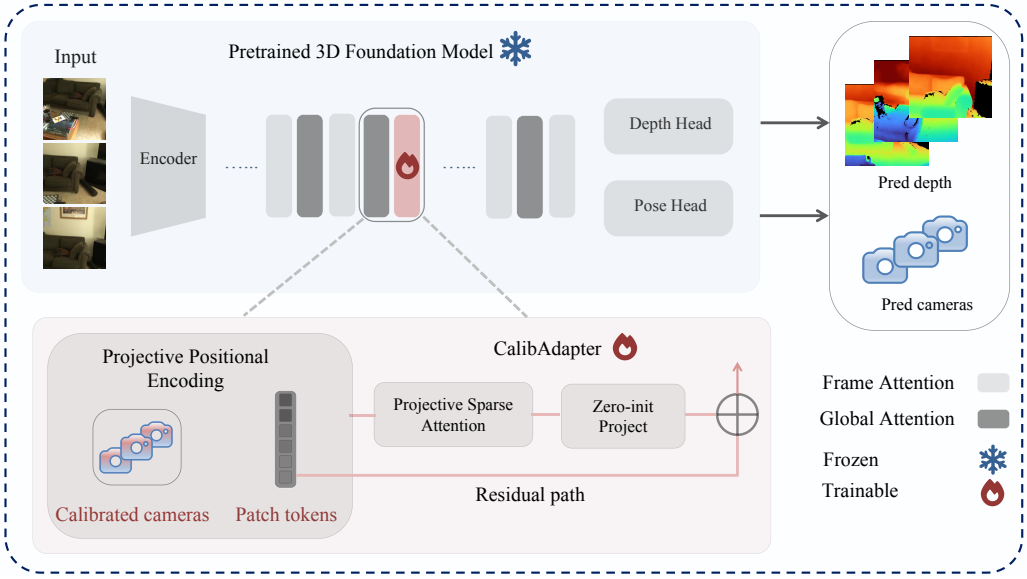

Figure 2: **Overview of CalibAdapter.** The frozen reconstruction pathway is augmented by a camera-conditioned projective sparse residual branch that injects the provided camera geometry as a zero-initialized residual correction.

3.1 CAMERA COMPLIANCE AND CPE

Given a calibrated multi-view sequence $\mathcal{D} = \{\mathbf{I}_s, \mathbf{K}_s, \mathbf{E}_s\}_{s=0}^{S-1}$ with $S \geq 2$ views, the homogeneous world-to-camera extrinsic is $\mathbf{E}_s = \begin{bmatrix} \mathbf{R}_s & \mathbf{t}_s \\ \mathbf{0}^\top & 1 \end{bmatrix} \in \mathrm{SE}(3)$, with camera center $\mathbf{c}_s = -\mathbf{R}_s^\top \mathbf{t}_s$. We define *camera compliance* operationally as whether scale-restored predicted depth can be directly back-projected with the provided camera intrinsics and extrinsics to recover accurate metric geometry. Unlike protocols that remove the full similarity ambiguity, we allow only a single trajectory-level scale shared across all views, without further pose alignment.

When the prediction contains unresolved global scale ambiguity, we estimate a similarity transform between predicted and provided camera centers using Umeyama (Umeyama, 1991), where $\hat{\mathbf{c}}_s$ is predicted by the backbone and $\mathbf{c}_s$ is defined by the provided calibration:

$$(a^\star, \mathbf{R}^\star, \mathbf{t}^\star) = \underset{a>0, \mathbf{R}\in SO(3), \mathbf{t}}{\operatorname{argmin}} \sum_{s=0}^{S-1} \|a\mathbf{R}\hat{\mathbf{c}}_s + \mathbf{t} - \mathbf{c}_s\|_2^2 \tag{1}$$

Following the posed-reconstruction protocol of DA3 (Lin et al., 2026), we recover scale from the predicted and provided camera trajectories. Only the estimated scale $a^\star$ is retained and applied uniformly to all predicted depths, whereas $R^\star$ and $t^\star$ are discarded; consequently, no rotation or translation alignment is introduced into the final geometry. Let $\bar{d}_s(u) = a^\star \hat{d}_s(u)$ denote the scale-restored predicted depth. Figure 3 summarizes the complete evaluation protocol. The scale-restored depth is back-projected using the provided calibration:

$$\hat{\mathbf{p}}_s(\mathbf{u}) = \left[\mathbf{E}_s^{-1} \begin{pmatrix} \bar{d}_s(\mathbf{u}) \mathbf{K}_s^{-1} \tilde{\mathbf{u}} \\ 1 \end{pmatrix} \right]_{1:3} \tag{2}$$

where $\tilde{\mathbf{u}} = [u_x, u_y, 1]^\top$ is the homogeneous pixel coordinate. Ground-truth points $\mathbf{p}_s^\star(\mathbf{u})$ are obtained by back-projecting ground-truth depth with the same provided camera calibration. For the set of valid pixels Ω, we define the Calibrated Pointmap Error (CPE) as:

$$\mathrm{CPE} = \frac{1}{|\Omega|} \sum_{(s,\mathbf{u})\in\Omega} \|\hat{\mathbf{p}}_s(\mathbf{u}) - \mathbf{p}_s^\star(\mathbf{u})\|_2 \tag{3}$$

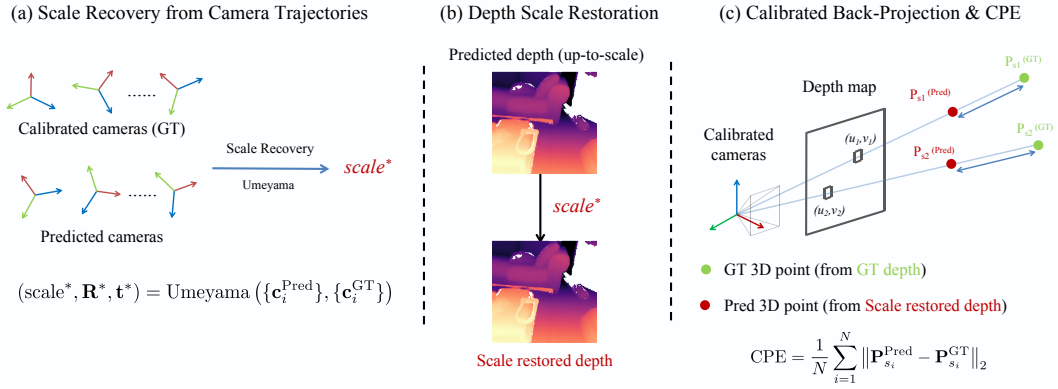

Figure 3: **CPE evaluation protocol.** (a) Recover one trajectory-level scale from the predicted and provided camera trajectories, (b) restore the predicted depth, and (c) back-project prediction and ground truth with the provided calibration to compute pixel-corresponded 3D error.

CPE evaluates the pixel-corresponded metric 3D error obtained under the provided calibration after a single trajectory-level scale recovery. No per-image scale–shift fitting or further pose, $\mathrm{Sim}(3)$, or ICP alignment is applied; its relation to scale-restored depth error is derived in Appendix A.1.

3.2 Calibrated Residual Adaptation

Rather than modifying the pretrained cross-view interaction itself, CalibAdapter introduces an auxiliary projective residual pathway alongside a selected interaction block. Let $\mathbf{X}_\ell \in \mathbb{R}^{B\times S\times N\times C}$ denote the patch representation at layer ℓ. Given a frozen backbone block $\mathcal{F}_\ell$ and a trainable residual branch $\mathcal{A}_\ell$, the adapted representation is computed as:

$$\mathbf{X}_{\ell+1} = \mathcal{F}_\ell(\mathbf{X}_\ell) + \Gamma_\ell \mathcal{A}_\ell\left(\mathbf{X}_\ell, \{\mathbf{K}_s, \mathbf{E}_s\}_{s=0}^{S-1}\right) \tag{4}$$

where Γ_ℓ is the identity for backbones without native residual scaling and the frozen native LayerScale otherwise, introducing no trainable gate.

The branch initializes its internal projections from the corresponding native interaction layer and its output projection to zero, thereby recovering the pretrained function at initialization while learning only a calibration-dependent residual. We study insertion depth and branch capacity in Sec. 4.4.

3.3 Projective Camera Transform

The residual branch first normalizes the provided camera parameters to reduce dependence on image resolution and global scene scale. After image preprocessing, the intrinsics are written as:

$$\bar{\mathbf{K}}_s = \begin{bmatrix} f_{x,s}/W & 0 & c_{x,s}/W - \frac{1}{2} \\ 0 & f_{y,s}/H & c_{y,s}/H - \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

Taking view 0 as reference, we normalize relative translations by the median camera baseline:

$$\mathbf{T}_{s0} = \mathbf{E}_s\mathbf{E}_0^{-1} = \begin{bmatrix} \mathbf{R}_{s0} & \mathbf{t}_{s0} \\ \mathbf{0}^\top & 1 \end{bmatrix} \quad \beta = \mathrm{median}_{i<j}\,\|\mathbf{c}_i - \mathbf{c}_j\|_2 \quad \bar{\mathbf{T}}_{s0} = \begin{bmatrix} \mathbf{R}_{s0} & \mathbf{t}_{s0}/(\beta+\epsilon) \\ \mathbf{0}^\top & 1 \end{bmatrix} \tag{6}$$

Following GTA and PRoPE (Miyato et al., 2024; Li et al., 2025a), we apply the normalized camera transform to 4D homogeneous subspaces of residual $Q/K/V$:

$$\mathbf{P}_s = \begin{bmatrix} \bar{\mathbf{K}}_s & 0 \\ 0^\top & 1 \end{bmatrix} \bar{\mathbf{T}}_{s0} \qquad \mathbf{\Phi}_s = \mathbf{I}_{d_h/4} \otimes \mathbf{P}_s \tag{7}$$

$$\mathbf{Q}_s^g = \mathbf{\Phi}_s^\top \mathbf{Q}_s \quad \mathbf{K}_s^g = \mathbf{\Phi}_s^{-1}\mathbf{K}_s \quad \mathbf{V}_s^g = \mathbf{\Phi}_s^{-1}\mathbf{V}_s \quad (\mathbf{q}_s^g)^\top \mathbf{k}_t^g = \mathbf{q}_s^\top \mathbf{\Phi}_s \mathbf{\Phi}_t^{-1} \mathbf{k}_t \tag{8}$$

The resulting query–key relation incorporates relative camera geometry into the residual interaction. After sparse attention, the output is mapped back to the native feature coordinates as $\mathbf{O}_s = \mathbf{\Phi}_s \mathbf{O}_s^g$ before the zero-initialized output projection. The projective transform remains confined to the auxiliary residual pathway, leaving the pretrained attention and positional representation unchanged; Fig. 4 illustrates the complete projective sparse interaction.

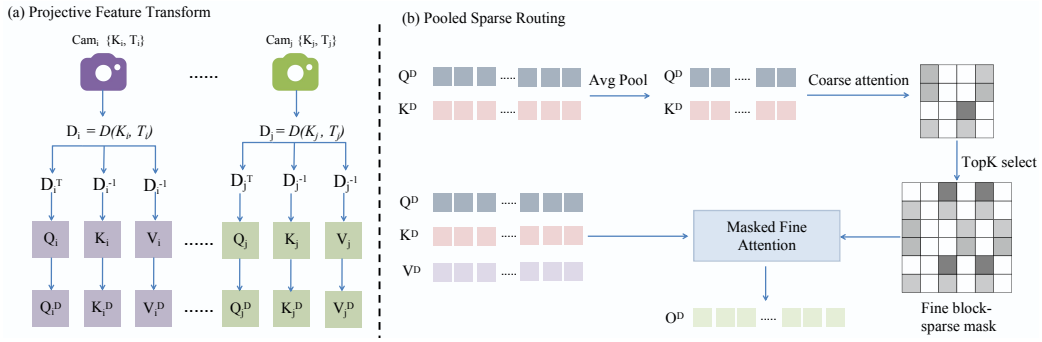

Figure 4: **Projective sparse interaction.** (a) Camera-conditioned transformation of $Q/K/V$. (b) Pooled routing constructs a block-sparse mask for residual attention.

3.4 Projective Sparse Routing

To limit the computational overhead of the auxiliary projective interaction, we adopt pooled block routing inspired by FasterVGGT (Wang et al., 2026a). Patch tokens are flattened in view-major order and average-pooled into query and key blocks. Since routing is performed on the projectively transformed features, the resulting affinities are conditioned on the provided camera geometry. For each head h, the coarse affinity and retained key blocks are:

$$r_{ab}^{h} = \frac{(\bar{\mathbf{q}}_a^{g,h})^\top \bar{\mathbf{k}}_b^{g,h}}{\sqrt{d_h}} \qquad \mathcal{R}_a^h = \mathrm{TopK}\left(\{r_{ab}^h\}_{b=1}^{M_k}, k_r\right) \qquad k_r = \max(1, \lfloor \rho M_k \rfloor) \tag{9}$$

Here, $\bar{\mathbf{q}}_a^{g,h}$ and $\bar{\mathbf{k}}_b^{g,h}$ denote pooled transformed features, M_k is the number of key blocks, and ρ is the routing budget. Fine-grained attention is then performed only within the selected blocks:

$$\mathbf{o}_a^{g,h} = \mathrm{softmax}\left(\frac{\mathbf{Q}_a^{g,h}(\mathbf{K}_{\mathcal{R}_a^h}^{g,h})^\top}{\sqrt{d_h}}\right)\mathbf{V}_{\mathcal{R}_a^h}^{g,h} \tag{10}$$

We use $\rho = 0.25$ by default. Sparse routing and fine attention operate on projectively transformed $Q/K/V$ within the auxiliary residual branch. Table 4 reports the accuracy–efficiency trade-off.

4 Experiments

4.1 Experimental Setup

Implementation details. All pretrained backbones and prediction heads are frozen, and only the added adaptation modules are optimized under the fixed-backbone setting. For cross-backbone evaluation, each adapter is trained separately on ScanNetV2; for cross-dataset evaluation, a single VGGT-Ω adapter is trained on mixed-domain data and evaluated unchanged across all held-out datasets. Unless otherwise specified, we use balanced 2–8-view sampling. Full optimization and implementation details are provided in Appendix A.

Metrics. We use CPE as the primary metric for calibrated metric geometry and follow DA3 (Lin et al., 2026) for depth and point-cloud evaluation. All metrics are computed using the provided camera intrinsics and extrinsics after a single trajectory-level scale recovery shared across all views, without per-image scale–shift fitting or post-hoc $\mathrm{Sim}(3)$/ICP alignment.

4.2 Camera Compliance Across 3D Foundation Models

We compare representative camera-conditioning paradigms on the same frozen VGGT-Ω backbone under a shared 5.248M trainable-parameter budget and evaluation protocol. MapAnything-style fuses ray and pose features with input image tokens, DA3-style introduces camera tokens into the input sequence, and PromptDA-style conditions decoder features. In contrast, CalibAdapter incorporates relative camera geometry directly into an auxiliary cross-view attention branch. Since calibration describes the projective relationships between views, this design allows it to guide cross-view feature comparison and aggregation while retaining the pretrained interaction pathway. Table 1(a) shows that CalibAdapter achieves the lowest CPE and depth errors among the evaluated variants, together with the best point-cloud accuracy and F-score. Decoder-stage conditioning provides only

Table 1: **Camera-conditioning paradigms and cross-backbone evaluation.** (a) Matched VGGT-Ω paradigms summarized by camera representation, injection stage, and direct target; (b) cross-backbone evaluation. ΔParams is reported in millions.

(a) Camera-conditioning paradigms on VGGT-Ω

	Conditioning design			Budget	Calib. geom.	Depth		Point Cloud		
Method	Camera repr.	Injection stage	Direct target	ΔParams	CPE ↓	AbsRel↓	RMSE↓	Acc.↓	Comp.↓	F-score↑
VGGT-Ω	–	–	–	0	0.2240	0.1054	0.2877	0.1790	0.1194	0.4563
+ MapAnything-style	Ray + pose feat.	Input	Image tokens	5.248	0.1875	0.0898	0.2230	0.1221	0.1199	0.4514
+ DA3-style	Camera token	Input	Token sequence	5.248	0.2049	0.0965	0.2481	0.1419	0.1233	0.4434
+ PromptDA-style	Proj. K/T feat.	Decoder	Decoder feat.	5.248	0.2216	0.1052	0.2803	0.1718	**0.1185**	0.4474
+ Ours	Proj. K/T	Cross-view	Attn. relation	5.248	**0.1668**	**0.0812**	**0.2013**	**0.1014**	0.1201	**0.4730**

(b) Cross-backbone evaluation

	Calibrated Geometry		Depth		Point Cloud		
Method	CPE_{mean} ↓	CPE_{P90} ↓	AbsRel↓	RMSE↓	Acc.↓	Comp.↓	F-score↑
VGGT	0.2237	0.4220	0.1068	0.2604	0.1596	0.1348	0.4675
VGGT+Ours	**0.1188**	**0.2685**	**0.0588**	**0.1526**	**0.0767**	**0.0910**	**0.5509**
π^3	0.1383	0.2774	0.0644	0.1664	0.0883	0.0872	0.5865
π^3+Ours	**0.1172**	**0.2714**	**0.0559**	**0.1463**	**0.0742**	**0.0770**	**0.6090**
DA3-Small	0.2116	0.5150	0.1009	0.2608	0.1376	0.1649	0.4141
DA3-Small+Ours	**0.1574**	**0.3550**	**0.0767**	**0.2054**	**0.1008**	**0.1281**	**0.4603**

a marginal CPE gain over the frozen backbone, whereas the interaction-stage design yields substantially lower CPE and depth errors in the matched fixed-backbone setting.

Table 1(b) extends the evaluation to three additional independently pretrained backbones: VGGT, π^3, and DA3-Small. Across both panels, CalibAdapter consistently reduces CPE on all four backbones while maintaining or improving depth and point-cloud quality. Notably, the improvement on DA3-Small, which already exploits camera priors, indicates that the proposed adaptation complements native camera-conditioned reconstruction. Additional robustness and scale-source analyses are provided in Appendix Tables 8–10.

4.3 Evaluation on Additional Datasets

We further evaluate cross-dataset generalization using a single VGGT-Ω CalibAdapter trained on a mixture of ScanNetV2 (Dai et al., 2017), ARKitScenes (Baruch et al., 2021), Hypersim (Roberts et al., 2021), TartanAir (Wang et al., 2020), and BlendedMVS (Yao et al., 2020). The same checkpoint is evaluated unchanged on all target benchmarks without target-domain adaptation. We additionally compare against alternative camera-conditioning designs trained under the same mixed-domain setting. Table 2 summarizes the results.

Overall, our cross-view conditioning shows more consistent cross-dataset generalization than the alternative camera-conditioning designs, while generally preserving complementary depth and point-cloud quality. ETH3D exhibits a different trend, suggesting that this benchmark presents a more challenging cross-domain transfer setting, where all adapted variants degrade relative to the frozen backbone. Nevertheless, our method shows the smallest degradation among the adapted variants. These results suggest that injecting calibrated geometry at cross-view interaction provides a more robust way to transfer camera conditioning across diverse domains. Additional object-centric and scene-level CPE visualizations are provided in Appendix Figs. 8 and 9.

Table 2: **Cross-dataset evaluation with VGGT-Ω.** † denotes the in-domain ScanNetV2 reference.

(a) Calibrated geometry: CPE↓

Method	**7Scenes**	**DTU**	**ScanNet++**	**ScanNetV2†**	**ETH3D**
VGGT-Ω	0.1322	0.0140	0.1044	0.2240	**0.1511**
+ MapAnything-style	0.1243	0.0117	0.0953	0.1960	0.2776
+ DA3-style	0.1182	0.0102	0.1562	0.2092	0.5094
+ PromptDA-style	0.1119	0.0110	0.0923	0.2009	0.1734
+ Ours	**0.0964**	**0.0091**	**0.0856**	**0.1807**	0.1608

(b) Complementary depth and point-cloud quality

	7Scenes			**DTU**			**ScanNet++**			**ScanNetV2†**			**ETH3D**		
Method	**AbsRel↓**	**CD↓**	**F↑**	**AbsRel↓**	**CD↓**	**F↑**	**AbsRel↓**	**CD↓**	**F↑**	**AbsRel↓**	**CD↓**	**F↑**	**AbsRel↓**	**CD↓**	**F↑**
VGGT-Ω	0.2174	0.1466	0.5488	0.0201	0.0374	**0.9328**	0.0953	0.1623	0.6772	0.1059	0.2984	0.4563	**0.0226**	**0.1840**	**0.4871**
+ MapAnything-style	0.2208	0.1443	0.5467	0.0171	**0.0320**	0.9296	0.0886	0.1214	0.6778	0.0917	0.2575	0.4687	0.0382	0.3195	0.4285
+ DA3-style	0.2321	0.1368	0.5834	0.0137	0.0327	0.9268	0.1117	0.1597	0.4848	0.1013	0.2751	0.4299	0.0662	0.5057	0.1738
+ PromptDA-style	0.2149	0.1399	0.6122	0.0158	0.0339	0.9272	0.0867	0.1194	0.6861	0.0940	0.2645	0.4544	0.0252	0.2160	0.4336
+ Ours	**0.1987**	**0.1298**	**0.6389**	**0.0131**	0.0329	0.9282	**0.0783**	**0.1062**	**0.6998**	**0.0851**	**0.2394**	**0.4824**	0.0241	0.2001	0.4678

4.4 Design Analysis

We analyze projective conditioning, branch placement and capacity, and sparse routing. All studies use VGGT-Ω on ScanNetV2.

Table 3: **Controlled design analysis of CalibAdapter.** (a) Projective conditioning and routing; (b) single-branch placement; (c) multi-branch scaling. ΔParams is reported in millions.

(a) Projective conditioning and routing

	Design		Budget	Calib. geometry	Depth	Point Cloud	
Variant	Proj. K/T	Pooled Routing	ΔParams	CPE↓	AbsRel↓	CD↓	F-score↑
Frozen	–	–	0	0.2240	0.1054	0.2984	0.4563
Plain	No	No	5.248	0.2139	0.1008	0.2923	0.4630
Plain + Sparse	No	Yes	5.248	0.2042	0.0998	0.2856	0.4576
Projective	Yes	No	5.248	0.1710	0.0823	**0.2110**	0.4712
CalibAdapter	Yes	Yes	5.248	**0.1668**	**0.0812**	0.2215	**0.4730**

(b) Single-branch placement

Position	Block	CPE↓	AbsRel↓	F-score↑
Early	B3	0.1753	**0.0808**	0.4647
Early-Mid	B7	**0.1668**	0.0812	**0.4730**
Mid-Late	B15	0.1739	0.0825	0.4638
Late	B21	0.2042	0.0965	0.4452

(c) Multi-branch scaling

Variant	Branches	ΔParams	CPE↓	CPE gain↑	AbsRel↓	F-score↑
Frozen	0	0	0.2240	–	0.1054	0.4563
1-branch (Default)	1	5.248	0.1668	25.55%	0.0812	0.4730
2-branch	2	10.497	0.1506	32.78%	0.0741	0.4981
4-branch	4	20.993	**0.1439**	**35.75%**	**0.0702**	**0.5223**

Effect of projective camera conditioning. Table 3(a) isolates the contributions of generic adaptation capacity, projective K/T conditioning, and sparse routing under matched trainable-parameter budgets. A parameter-matched Plain Adapter yields only modest improvements over the frozen backbone, while adding sparse routing provides a further small gain. In contrast, projective camera conditioning substantially reduces CPE and improves reconstruction quality. Combining projective conditioning with sparse routing further improves CPE and F-score while reducing the additional computation. These results demonstrate that explicit camera conditioning provides substantial benefits beyond those obtained from generic adaptation capacity alone.

Placement and adaptation capacity. Table 3(b,c) shows that early-to-middle interaction blocks outperform late-stage placement, while additional branches improve accuracy at the cost of more parameters. We therefore adopt a single early-to-middle branch as the lightweight default.

Table 4: **Accuracy–efficiency analysis of CalibAdapter.** (a) Reconstruction quality under different routing budgets; (b) additional adapter FLOPs; (c) end-to-end inference latency.

(a) Routing budget

ρ	CPE↓	AbsRel↓	F-score↑
1.00	0.1710	0.0823	0.4712
0.50	0.1691	0.0821	0.4612
0.25	**0.1668**	**0.0812**	**0.4730**
0.125	0.1810	0.0869	0.4476

(b) Adapter FLOPs (GFLOPs)

Views	Dense Projective	CalibAdapter	Reduction
2	26.013	**18.623**	28.41%
4	71.731	**42.174**	41.20%
8	222.287	**104.058**	53.19%
16	759.869	**286.953**	62.24%
32	2780.924	**889.260**	68.02%

(c) Inference latency (ms)

Views	Frozen	Dense Projective	CalibAdapter
2	62.04	68.25	**64.86**
4	98.14	101.31	**101.02**
8	193.79	200.59	**199.34**
16	415.55	439.00	**426.78**
32	960.07	1013.86	**975.96**

Notes. FLOPs denote the additional computation over the frozen backbone. Latency is measured end-to-end with batch size 1 in BF16. Dense Projective denotes dense projective interaction.

Sparse routing and computational scaling. Table 4 shows that moderate sparsification preserves accuracy while substantially reducing the additional FLOPs of dense projective interaction, with end-to-end latency remaining close to the frozen backbone. We use $\rho = 0.25$ by default. Together, these controls suggest that calibrated camera geometry is most effectively introduced as a projective residual during early-to-middle cross-view interaction, rather than through generic adaptation or late-stage correction.

4.5 Autonomous Driving Evaluation

We evaluate CalibAdapter on nuScenes following DriveVGGT's data-processing and evaluation protocol (Jia et al., 2025). The known six-camera rig calibration is combined with timestamped ego poses to express all extrinsics in a common reference frame. Under matched trainable-parameter and training budgets, Table 5 shows that CalibAdapter consistently improves calibrated geometry and reconstruction quality over the evaluated camera-conditioning variants, with a modest trade-off in completeness. Figure 5 provides a qualitative comparison.

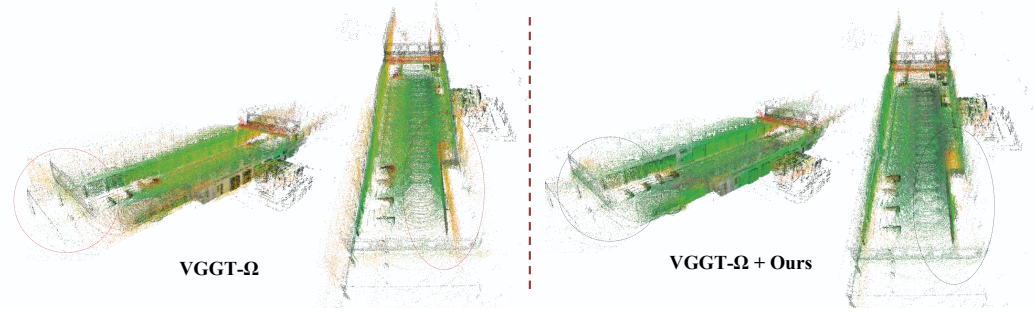

Figure 5: **CalibAdapter improves multi-camera reconstruction on nuScenes.**

4.6 Real-World Robotic Evaluation

We evaluate real-world manipulation on a dual-arm Dobot X-Trainer with wrist-mounted RealSense D405 cameras and calibrated poses. CalibAdapter and the parameter-matched Plain Adapter share the same backbone, downstream pipeline, and camera calibration, isolating the effect of calibration-aware adaptation. Table 6 reports 20 trials per task, with qualitative examples in Fig. 6. CalibAdapter increases overall success from 68.3% to 86.7% (**+18.4 percentage points**), while RGB-D depth serves as an oracle, showing that improved calibrated geometry translates into better real-world manipulation performance.

Table 5: Multi-camera driving on nuScenes. We follow the DriveVGGT protocol (Jia et al., 2025). All adapted variants use a matched trainable-parameter budget (5.248M) and the same training budget.

Method	ΔParams	CPE↓	AbsRel↓	Acc.↓	Comp.↓	F-score @10 cm↑
VGGT-Ω	0	7.2260	0.2893	0.9582	**1.0789**	0.1166
+ MapAnything-style	5.248	6.3692	0.2476	0.7908	1.2519	0.1184
+ DA3-style	5.248	6.5890	0.2563	0.8239	1.3761	0.1170
+ PromptDA-style	5.248	6.8510	0.2307	0.8632	1.2684	0.1099
+ Ours	5.248	**5.9087**	**0.2255**	**0.6283**	1.1618	**0.1208**

Table 6: Robot manipulation success rates (%). Twenty trials per task; RGB-D depth is the oracle. Plain Adapter and CalibAdapter are parameter-matched (5.248M).

Method	Bottle Grasp↑	Bottle→ Beaker↑	Block→ Cup↑	Overall↑
RGB-D Depth (Oracle)	100.0	100.0	100.0	100.0
VGGT-Ω + Plain Adapter	80.0	75.0	50.0	68.3
VGGT-Ω + Ours	**95.0**	**90.0**	**75.0**	**86.7**

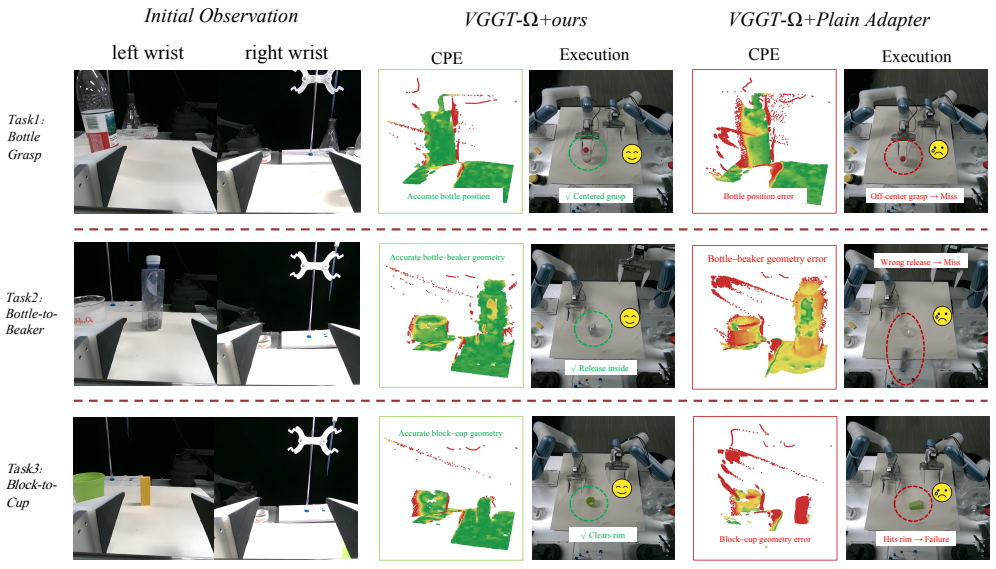

Figure 6: **Real-world robot manipulation.** CPE visualization and execution for three tasks.

5 Conclusion

In this work, we introduce CPE for calibrated metric 3D evaluation and present CalibAdapter, a lightweight camera-conditioned adaptation framework for feed-forward 3D reconstruction. Experiments across multiple 3D foundation models, autonomous driving, and real-world robotic manipulation demonstrate consistent CPE reductions and practical effectiveness.

AI USE STATEMENT

Generative AI tools were used to assist with code implementation and debugging, English-language editing and translation, literature search, experimental design, result interpretation, and LaTeX formatting. All AI-assisted outputs were reviewed and validated by the authors, who take full responsibility for the content and conclusions of this work.

REPRODUCIBILITY STATEMENT

The method and evaluation protocol are specified in Sections 3 and 4. Appendix A provides the dataset construction, optimization settings, loss functions, hardware, checkpoint policy, metric computation, robustness analyses, and diagnostic evaluations used for the reported results. Additional qualitative video demos are provided in the supplementary material.

REFERENCES

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jérôme Revaud, and Vincent Leroy. 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.

Chenjie Cao, Xinlin Ren, and Yanwei Fu. MVSFormer++: Revealing the devil in transformer's details for multi-view stereo. In *International Conference on Learning Representations (ICLR)*, 2024.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D gaussian splats from image pairs for scalable generalizable 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19457–19467, 2024.

Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-Based Multi-View Stereo Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1538–1547, 2019.

Shizhe Chen, Ricardo Garcia Pinel, Cordelia Schmid, and Ivan Laptev. PolarNet: 3D point clouds for language-guided robotic manipulation. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1761–1781, 2023.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 16664–16678, 2022.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision (ECCV)*, pp. 370–386, 2024.

Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep Stereo Using Adaptive Thin Volume Representation With Uncertainty Awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2524–2534, 2020.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017.

Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. TransMVSNet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8585–8594, 2022.

Wei Dong, Dawei Yan, Zhijun Lin, and Peng Wang. Efficient adaptation of large vision transformer via adapter re-composing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 52548–52567, 2023.

Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. MASt3R-SfM: A fully-integrated solution for unconstrained structure-from-motion. In *International Conference on 3D Vision (3DV)*, pp. 1–10, 2025. doi: 10.1109/3DV66043.2025.00008.

Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. VGG-T^3: Offline feed-forward 3D reconstruction at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36464–36474, 2026.

Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3D: 3D feature field transformers for multi-task robotic manipulation. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3949–3965, 2023.

Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. RVT: Robotic view transformer for 3D object manipulation. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 694–710, 2023.

Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2495–2504, 2020.

Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareş Ambruş, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9233–9243, 2023.

Minghao Guo, Meng Cao, Jiachen Tao, Rongtao Xu, Yan Yan, Xiaodan Liang, Ivan Laptev, and Xiaojun Chang. GLaD: Geometric latent distillation for vision-language-action models. *arXiv preprint arXiv:2512.09619*, 2025a.

Yuliang Guo, Sparsh Garg, S. Mahdi H. Miangoleh, Xinyu Huang, and Liu Ren. Depth any camera: Zero-shot metric depth estimation from any camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26996–27006, 2025b.

Haoyu He, Jianfei Cai, Jing Zhang, Dacheng Tao, and Bohan Zhuang. Sensitivity-aware visual parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11825–11835, 2023.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3D value maps for robotic manipulation with language models. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 540–562, 2023a.

Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9223–9232, 2023b.

Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jérôme Revaud. Pow3R: Empowering unconstrained 3D reconstruction with camera and scene priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1071–1081, 2025.

Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. SurfaceNet: An end-to-end 3D neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2307–2315, 2017.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pp. 709–727, 2022.

Xiaosong Jia, Yanhao Liu, Yu Hong, Renqiu Xia, Junqi You, Bin Sun, Zhihui Hao, and Junchi Yan. DriveVGGT: Calibration-constrained visual geometry transformers for multi-camera autonomous driving. *arXiv preprint arXiv:2511.22264*, 2025.

Shibo Jie and Zhi-Hong Deng. FacT: Factor-tuning for lightweight adaptation on vision transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):1060–1068, 2023. doi: 10.1609/aaai.v37i1.25187.

Shibo Jie, Zhi-Hong Deng, Shixuan Chen, and Zhijuan Jin. Convolutional bypasses are better vision transformer adapters. In *European Conference on Artificial Intelligence (ECAI)*, pp. 202–209, 2024.

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3D diffuser actor: Policy diffusion with 3D scene representations. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1949–1974, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision (3DV)*. IEEE, 2026.

Ramil Khafizov, Artem Komarichev, Ruslan Rakhimov, Peter Wonka, and Evgeny Burnaev. G-CUT3R: Guided 3D reconstruction with camera and depth prior integration. *arXiv preprint arXiv:2508.11379*, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, pp. 71–91, 2024. doi: 10.1007/978-3-031-73220-1_5.

Chunyang Li, Yuanbo Yang, Jiahao Shao, Hongyu Zhou, Katja Schwarz, and Yiyi Liao. ReRoPE: Repurposing RoPE for relative camera control. In *Proceedings of the ACM SIGGRAPH Conference Papers*, pp. 64:1–64:11, 2026. doi: 10.1145/3799902.3811226.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 15984–16009, 2025a. doi: 10.52202/085713-0540.

Samuel Li, Pujith Kachana, Prajwal Chidananda, Saurabh Nair, Yasutaka Furukawa, and Matthew A. Brown. Rig3R: Rig-aware conditioning and discovery for 3D reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 27349–27373, 2025b. doi: 10.52202/085713-0815.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3D object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1477–1485, 2023a.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision (ECCV)*, pp. 1–18, 2022.

Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M. Alvarez. FB-BEV: BEV representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6919–6928, 2023b.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 109–123, 2022.

Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4K resolution accurate metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17070–17080, 2025.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 141261–141285, 2026.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *European Conference on Computer Vision (ECCV)*, pp. 531–548, 2022.

Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. PETRv2: A unified framework for 3D perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3262–3272, 2023.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time dense scene reconstruction from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16651–16662, 2025.

Guanxing Lu, Shiyi Zhang, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. ManiGaussian: Dynamic gaussian splatting for multi-task robotic manipulation. In *European Conference on Computer Vision (ECCV)*, pp. 349–366, 2024.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3R: Aligned monocular depth estimation for dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22820–22830, 2025.

Gen Luo, Minglang Huang, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106*, 2023.

Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. GTA: A geometry-aware attention mechanism for multi-view transformers. In *International Conference on Learning Representations*, 2024.

Kaichun Mo, Leonidas J. Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2Act: From pixels to actions for articulated 3D objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6813–6823, 2021.

Riku Murai, Eric Dexheimer, and Andrew J. Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16695–16705, 2025.

Haosong Peng, Hao Li, Yalun Dai, Yushi Lan, Yihang Luo, Tianyu Qi, Zhengshen Zhang, Yufeng Zhan, Junfei Zhang, Wenchao Xu, and Ziwei Liu. OmniVGGT: Omni-modality driven visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36485–36497, 2026a.

Yue Peng, Yongzhe Zhao, Artur Habuda, Khuyen Pham, Yanheng Zhu, Tran Nguyen Le, Fares Abu-Dakka, and Li Guo. G^3VLA: Geometric inductive bias for vision-language-action models. *arXiv preprint arXiv:2606.24472*, 2026b.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *European Conference on Computer Vision (ECCV)*, pp. 178–193, 2020.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10106–10116, 2024.

Weining Ren, Hongjun Wang, Xiao Tan, and Kai Han. Fin3r: Fine-tuning feed-forward 3d reconstruction models via monocular knowledge distillation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10912–10922, 2021.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. CLIPort: What and where pathways for robotic manipulation. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 894–906, 2022.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 785–799, 2023.

Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10208–10217, 2024.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5283–5293, 2025.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

Chung-Shien Brian Wang, Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Block-sparse global attention for efficient multi-view geometry transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14546–14555, 2026a.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. PatchmatchNet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14194–14203, 2021.

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *International Conference on 3D Vision (3DV)*, pp. 78–89, 2025.

Hengyi Wang and Lourdes Agapito. AMB3R: Accurate feed-forward metric-scale 3D reconstruction with backend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14612–14625, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21486–21499, 2026b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025b.

Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3621–3631, 2023.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916, 2020.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 10481–10497, 2026c.

Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3D object detection from multi-view images via 3D-to-2D queries. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 180–191, 2022.

Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21729–21740, 2023.

Yu Wu, Minsik Jeon, Jen-Hao Rick Chang, Oncel Tuzel, and Shubham Tulsiani. RayRoPE: Projective ray positional encoding for multi-view attention. In *European Conference on Computer Vision (ECCV)*, 2026.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3R: Streaming 3D reconstruction with explicit spatial pointer memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 77686–77710, 2025.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. DepthSplat: Connecting gaussian splatting and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16453–16463, 2025a.

Qi Xu, Dongxu Wei, Lingzhe Zhao, Wenpu Li, Zhangchi Huang, Shunping Ji, and Peidong Liu. SIU3R: Simultaneous scene understanding and 3D reconstruction beyond feature alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 122851–122881, 2025b.

Yinghao Xu, Zifan Shi, Yifan Wang, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. GRM: Large gaussian reconstruction model for efficient 3D reconstruction and generation. In *European Conference on Computer Vision (ECCV)*, pp. 1–20, 2024.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935, 2025.

Jiayu Yang, Wei Mao, Jose M. Alvarez, and Miaomiao Liu. Cost Volume Pyramid Based Depth Inference for Multi-View Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4877–4886, 2020.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 785–801, 2018.

Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent MVSNet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5520–5529, 2019. doi: 10.1109/CVPR.2019.00567.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1790–1799, 2020.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3D gaussian splats from sparse unposed images. In *International Conference on Learning Representations (ICLR)*, 2025.

Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Ying Shan, and Kai Han. SCoPE: Sightline-coordinate positional encoding for video diffusion transformers. *arXiv preprint arXiv:2606.27345*, 2026.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3D prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9043–9053, 2023.

Wentao Yuan, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. M2T2: Multi-task masked transformer for object-centric pick and place. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3619–3630, 2023.

Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, and Xiaolong Wang. GNFactor: Multi-task real robot learning with generalizable neural feature fields. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 284–301, 2023.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Q. Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38027–38037, 2026.

Jason Y. Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. MonST3R: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations (ICLR)*, 2025a.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023a.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. FLARE: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21936–21947, 2025b.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5268–5280, 2025c.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9433–9443, 2023b.

Lei Zhou, Haoyu Wu, Akshat Dave, and Dimitris Samaras. Learning 3D reconstruction with priors in test time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36550–36560, 2026.

A Additional Details and Analyses

Cross-backbone training. Each adapter in the cross-backbone comparison is trained separately on the ScanNetV2 (Dai et al., 2017) training split for 10K optimizer steps. We use balanced 2–8-view sampling and construct multi-view tuples by frame overlap, following SIU3R (Xu et al., 2025b). All pretrained parameters are frozen, and only CalibAdapter is optimized with the backbone-specific objective and learning-rate schedule below. Unless stated otherwise, we report fixed final checkpoints without validation-based selection. The strict312 evaluation uses 312 four-view tuples at 378×504 resolution or the backbone's native equivalent. Table 7 summarizes the reported configurations.

Table 7: **Training configurations for the reported cross-backbone checkpoints.** All models are trained with AdamW and weight decay 0.05. LR denotes the effective learning-rate schedule.

Backbone	Insertion	Params (M)	Effective LR	Objective	Steps
VGGT-Ω	Inter-frame B7	5.248	$10^{-8} \to 5\times10^{-5} \to 10^{-8}$	Eq. (11)	10K
VGGT	Global B7	5.248	$10^{-8} \to 5\times10^{-5} \to 10^{-8}$	Eq. (11)	10K
π^3	B13 pre-block	5.248	$5\times10^{-5} \to 5\times10^{-10}$	Official objective	10K
DA3-Small	Global B7	0.739	$10^{-8} \to 5\times10^{-5} \to 10^{-8}$	Eq. (12)	10K

Common optimization details. Unless otherwise specified, all adapters are trained with AdamW (weight decay 0.05) in BF16 on four NVIDIA A800 80GB GPUs, using random seed 42 and gradient-norm clipping at 1.0. VGGT-Ω, VGGT, and DA3-Small use a per-GPU budget of 30 images, yielding a global batch of $4\lfloor 30/V \rfloor$ tuples for $V \in \{2, \dots, 8\}$, while π^3 uses one tuple per GPU and follows its native OneCycle schedule. Learning-rate schedules and objectives are summarized in Table 7. We report the final 10K checkpoints, with validation loss used only as a training sanity check.

Cross-backbone objectives. VGGT-Ω and VGGT use

$$\begin{aligned} \mathcal{L} &= 5\mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{conf-depth}} + \mathcal{L}_{\text{reg-depth}} + \mathcal{L}_{\text{grad-depth}}, \\ \mathcal{L}_{\text{cam}} &= \mathcal{L}_T + \mathcal{L}_R + 0.5\mathcal{L}_{\text{FoV}}. \end{aligned} \tag{11}$$

The camera loss operates on the native translation (*absT*), quaternion rotation (*quaR*), and field-of-view (*FoV*) parameterization. We use confidence coefficient $\alpha = 0.2$ and temporal factor $\gamma = 0.6$ for iterative camera predictions.

π^3 retains its official objective $\mathcal{L}_{\text{point}} + 0.1\mathcal{L}_{\text{cam}}$, where $\mathcal{L}_{\text{point}}$ contains the scale-aligned local-point and surface-normal terms, and $\mathcal{L}_{\text{cam}} = 100\,\text{Huber}(\hat{t}, t) + \mathcal{L}_{\text{geo-rot}}$.

DA3-Small uses

$$\mathcal{L} = \mathcal{L}_{\text{conf-depth}} + \mathcal{L}_{\text{grad-depth}} + \mathcal{L}_{\text{ray}} + \mathcal{L}_{\text{pointmap}} + \mathcal{L}_{\text{cam}}, \tag{12}$$

with all terms weighted equally and confidence coefficient $\alpha = 0.2$.

Metric computation. All evaluated backbones predict depth up to scale. For each tuple, we recover a single scale from the predicted camera trajectory and apply it uniformly to all views. Predicted and ground-truth depth are then back-projected using the same ground-truth intrinsics and extrinsics. We do not use per-image scale–shift fitting or subsequent $\text{Sim}(3)$/ICP alignment. CPE mean and P90 are computed after pooling all valid pixels across the evaluation tuples; the mean is therefore sensitive to rare trajectory-scale failures. Depth metrics use the same restored scale. Point-cloud accuracy, completeness, Chamfer distance, and F-score@5 cm are measured directly in the calibrated frame, unless a different threshold is stated explicitly. All geometric distances are reported in meters.

A.1 Additional Analysis

CPE interpretation. For OpenCV z-depth, back-projection with the provided world-to-camera extrinsic $\mathbf{E}_s$, whose rotation and translation are $(\mathbf{R}_s, \mathbf{t}_s)$, gives

$$\hat{\mathbf{p}}_s(\mathbf{u}) = \mathbf{R}_s^\top \left(\bar{d}_s(\mathbf{u}) \mathbf{K}_s^{-1} \tilde{\mathbf{u}} - \mathbf{t}_s \right)$$

and analogously

$$\mathbf{p}_s^\star(\mathbf{u}) = \mathbf{R}_s^\top \left(d_s^\star(\mathbf{u}) \mathbf{K}_s^{-1} \tilde{\mathbf{u}} - \mathbf{t}_s \right)$$

Therefore,

$$\|\hat{\mathbf{p}}_s(\mathbf{u}) - \mathbf{p}_s^\star(\mathbf{u})\|_2 = \left|\bar{d}_s(\mathbf{u}) - d_s^\star(\mathbf{u})\right| \left\|\mathbf{K}_s^{-1}\tilde{\mathbf{u}}\right\|_2$$

Thus, under shared calibrated cameras, the shared translation cancels, while the shared rotation preserves the pointwise Euclidean distance, while the camera ray determines the metric 3D displacement induced by a depth residual. CPE is therefore used as an operational calibrated-deployment metric rather than as an independent camera-pose or cross-view-consistency metric under the protocol above.

Statistical robustness. We evaluate robustness to both training randomness and evaluation-scene sampling on ScanNetV2 strict312. Repeating the default B7, $\rho = 0.25$ configuration with three random seeds yields consistent CPE improvements over the frozen VGGT-Ω backbone (Table 8).

Table 8: **Multi-seed robustness of VGGT-Ω+CalibAdapter.** ScanNetV2 strict312; $\pm$ denotes the sample standard deviation ($n = 3$), and CPE reduction is relative to the frozen backbone.

Seed	CPE mean↓	CPE P90↓	AbsRel↓	RMSE↓	F-score↑	CPE reduction↑
42	0.1668	0.3778	0.0812	0.2013	0.4730	25.55%
123	0.1689	0.3857	0.0823	0.2036	0.4600	24.59%
2026	0.1736	0.3942	0.0843	0.2081	0.4566	22.49%
Mean $\pm$ std	0.1698 ± 0.0035	0.3859 ± 0.0082	0.0826 ± 0.0016	0.2043 ± 0.0035	0.4632 ± 0.0086	24.21%

For evaluation-scene uncertainty, we perform a paired bootstrap over the 312 strict312 scenes using the default seed-42 checkpoint (100,000 resamples). CPE and AbsRel are re-aggregated by their valid sample counts, while F-score is averaged across scenes. The 95% intervals for CPE and AbsRel reductions exclude zero, whereas the F-score interval crosses zero (Table 9).

Table 9: **Paired scene-bootstrap analysis** on ScanNetV2 strict312 using the default seed-42 checkpoint with 100,000 resamples.

Metric	Estimate	Paired 95% CI
CPE reduction	0.0572 m	$[0.0183, 0.1084]$ m
AbsRel reduction	0.0247	$[0.0088, 0.0448]$
F-score increase	0.0166	$[-0.0155, 0.0490]$

Together, these analyses indicate that the primary CPE improvement is robust to both training randomness and evaluation-scene sampling.

Scale-source diagnostic. To test whether the CPE improvement mainly comes from trajectory-scale estimation, we compare the default predicted-trajectory scale with an oracle global scale on ScanNetV2 strict312. For OpenCV z-depth, the oracle scale is defined as

$$a_{\text{oracle}} = \frac{\sum_{(s,\mathbf{u})\in\Omega} w_s(\mathbf{u})^2 \hat{d}_s(\mathbf{u}) d_s^\star(\mathbf{u})}{\sum_{(s,\mathbf{u})\in\Omega} w_s(\mathbf{u})^2 \hat{d}_s(\mathbf{u})^2}, \qquad w_s(\mathbf{u}) = \left\|\mathbf{K}_s^{-1}\tilde{\mathbf{u}}\right\|_2. \tag{13}$$

This is the least-squares global scale minimizing the ray-weighted squared 3D error, with one shared scalar per tuple and no per-view scale, scale–shift, pose, $\mathrm{Sim}(3)$, ICP, or point-cloud alignment.

Table 10: **Scale-source diagnostic on ScanNetV2 strict312.** Oracle uses one least-squares global scale per tuple.

Method	Scale source	CPE mean↓	CPE P90↓
Frozen	Predicted trajectory	0.2240	0.4628
CalibAdapter	Predicted trajectory	**0.1668**	**0.3778**
Frozen	Oracle global	0.0850	0.1554
CalibAdapter	Oracle global	**0.0544**	**0.1009**

CalibAdapter remains better under oracle global scaling, with CPE decreasing from 0.0850 to 0.0544 m. This indicates that the improvement is not solely due to trajectory-scale estimation. The main evaluation protocol continues to use the predicted-camera trajectory scale defined in Sec. 3.1.

Design-analysis controls. All variants use the same frozen VGGT-Ω backbone, 5.248M trainable-parameter budget, ScanNetV2 data, 10K-step training budget, and strict312 evaluation protocol.

Sensitivity to camera calibration. We evaluate CalibAdapter under perturbed camera inputs to assess its dependence on the provided calibration. For the noisy-camera condition, focal lengths are perturbed with Gaussian noise of log-standard deviation 0.05, principal points by 2% of the image dimensions, rotations by a standard deviation of 5°, and camera centers to an RMS perturbation of 10% of the tuple median baseline. Perturbations affect only the adapter input; trajectory-scale recovery and all evaluation metrics use the original ground-truth (K, T). Table 11 shows that mild noise causes limited degradation, whereas identity or shuffled cameras substantially degrade performance.

Table 11: **Sensitivity to camera calibration.** Perturbations affect only the adapter input; evaluation uses the original ground-truth (K, T).

	Compliance		Depth		Point Cloud		
Camera input	$\text{CPE}_{\text{mean}}\downarrow$	$\text{CPE}_{\text{P90}}\downarrow$	AbsRel↓	RMSE↓	Acc.↓	Comp.↓	F-score↑
Correct (K, T)	**0.1668**	**0.3778**	**0.0812**	**0.2013**	**0.1014**	**0.1201**	**0.4730**
Noisy (K, T)	0.1730	0.4017	0.0851	0.2057	0.1053	0.1239	0.4526
Identity	0.5107	1.0445	0.2422	0.5226	0.3722	0.2470	0.2328
Shuffled (K, T)	0.7087	1.6290	0.3370	0.7152	0.5417	0.3199	0.1992

Placement, capacity, and efficiency. Placement evaluates B3, B7, B15, and B21 under the same single-branch budget; branch scaling uses the same training and evaluation setting. Efficiency uses query/key block sizes of 128/64 and reports adapter FLOPs relative to the frozen backbone and end-to-end latency with batch size 1 in BF16. The results show that sparse routing reduces the added projective-interaction cost while keeping end-to-end overhead small.

A.2 Additional Analysis and Qualitative Results

Feature effects of different injection paradigms. The key distinction among camera-conditioning paradigms is not merely *when* calibration is introduced, but *what computation* it directly conditions. MapAnything-style and DA3-style encode camera information into input image/camera representations that subsequently participate in cross-view reasoning, whereas PromptDA-style conditions decoder features downstream of the main inter-view interaction. In contrast, CalibAdapter injects projective cues derived from calibrated intrinsics and extrinsics directly into the cross-view interaction, thereby conditioning the inter-view relations through which geometric correspondence is established.

To examine how these different conditioning paradigms are reflected in the learned representation, we probe all variants at an identical downstream feature level. Specifically, we extract the 256-dimensional spatial features from the final depth decoder immediately before the prediction projection. For a query pixel in the source view, we first construct its reference correspondence geometrically: the query is unprojected using its ground-truth depth and the calibrated source camera, transformed into the target-camera frame, and reprojected using the calibrated target camera. Target-view ground-truth depth is used only to verify visibility and reject occluded correspondences.

Given the resulting source query, we L2-normalize its feature and all spatial features in the target view and compute cosine similarity over the *entire* target feature map:

$$S(i, j) = \frac{f_s^\top f_t(i, j)}{\|f_s\|_2 \, \|f_t(i, j)\|_2}$$

No epipolar mask, ground-truth correspondence constraint, or predicted depth is used during feature matching. The peak response,

$$(\hat{u}, \hat{v}) = \arg\max_{i,j} S(i, j)$$

defines the predicted correspondence, and its Euclidean distance to the geometry-derived reference correspondence is reported as the 2D correspondence localization error.

As shown in Fig. 7, the different conditioning strategies lead to distinct downstream correspondence-response structures. In the displayed examples, CalibAdapter yields responses that are more localized around the reference correspondence and substantially smaller localization errors. These observations are consistent with our design hypothesis that directly conditioning cross-view relations can propagate calibrated geometric constraints into more discriminative downstream correspondence features.

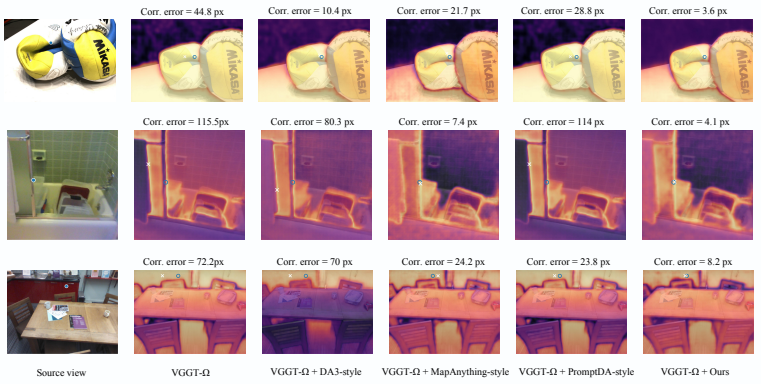

Figure 7: **Feature-level comparison of camera-conditioning paradigms.** Correspondence responses are computed from the same 256-D spatial features extracted from the final depth decoder immediately before the prediction projection for all variants. For each source-view query, the blue circle denotes the reference correspondence obtained by lifting the query with ground-truth depth and calibrated cameras and reprojecting it into the target view. The heatmap shows cosine similarity between the source query feature and all target-view spatial features; no ground-truth-guided search or epipolar mask is applied. The white cross denotes the global similarity maximum, and the reported value is its 2D pixel distance to the reference correspondence. CalibAdapter produces more localized correspondence responses in the shown examples.

Object-centric qualitative comparison. Figure 8 presents close-range, object-centric CPE comparisons under a shared visualization protocol and identical color thresholds. Points on the common GT-support geometry are colored by per-pixel CPE, ranging from green for lower error to red for higher error, while gray points indicate the ground-truth reference geometry. CalibAdapter exhibits larger and more spatially coherent low-error regions and fewer high-error regions across these examples, providing qualitative evidence consistent with the quantitative comparisons.

Scene-level qualitative comparison. Figure 9 extends the same visualization protocol to larger and more complex indoor scenes. CalibAdapter continues to show more spatially continuous low-CPE regions in these examples, indicating that the observed trend is not restricted to close-range objects. These qualitative results complement the cross-dataset evaluation in Table 2, where CalibAdapter achieves lower CPE than the alternative camera-conditioning paradigms across the evaluated datasets.

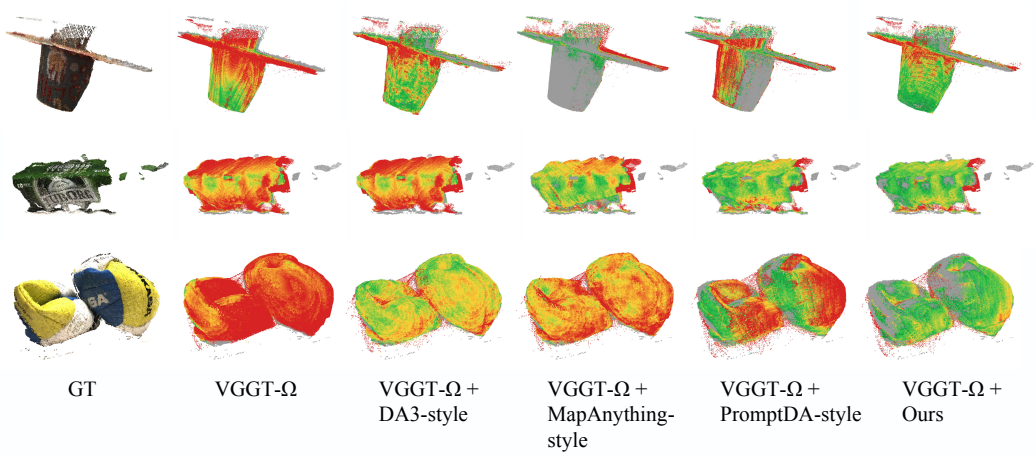

Figure 8: **Object-centric CPE comparison.** All methods use the same CPE visualization protocol and color thresholds; green and red denote lower and higher CPE, respectively, and gray points show the ground-truth reference geometry.

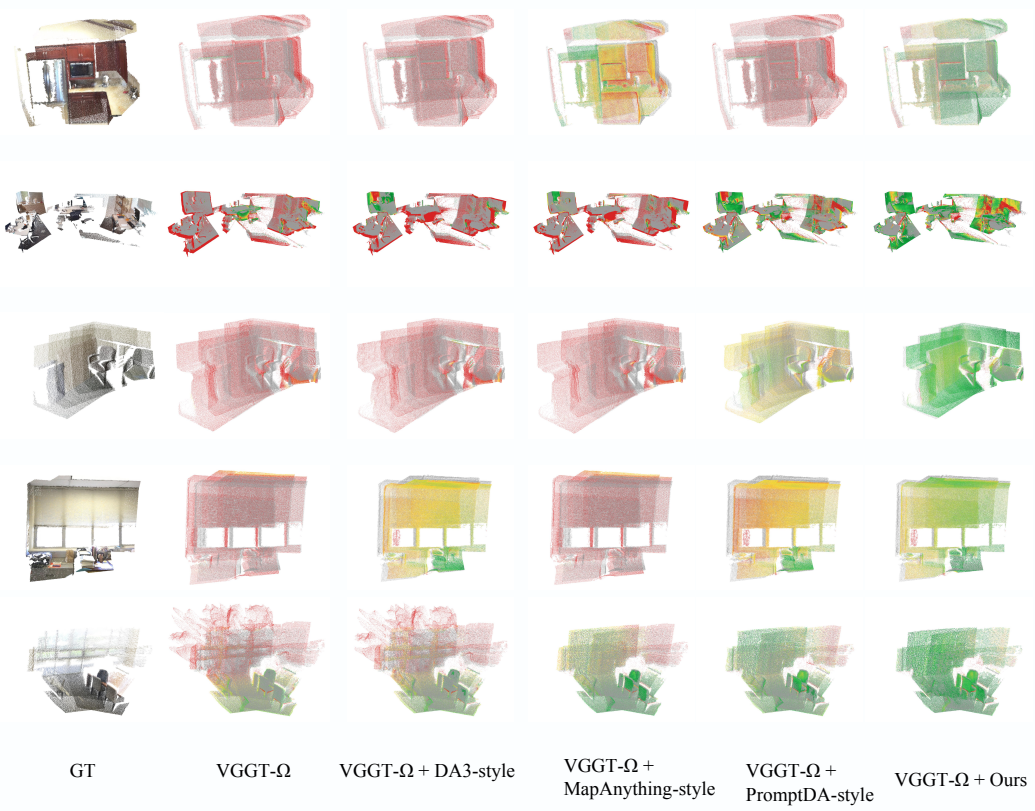

Figure 9: **Scene-level indoor CPE comparison.** Larger indoor scenes are visualized with the same protocol and color thresholds as Fig. 8; gray points show the ground-truth reference geometry.

A.3 Additional Experimental Details

Five-domain training. For cross-dataset evaluation, a single VGGT-Ω+CalibAdapter checkpoint is trained on ARKitScenes, ScanNetV2, Hypersim, TartanAir, and the connected subset of BlendedMVS. We use dataset sampling weights of 56K, 38.4K, 11.2K, 56K, and 32K, corresponding to sampling probabilities of 28.93%, 19.83%, 5.79%, 28.93%, and 16.53%, respectively. We follow the common optimization settings above and use balanced 2–8-view sampling with the view count synchronized across GPUs and a per-GPU image budget of 30. The learning rate is warmed linearly to 2×10^{-5} over 420 updates and cosine-decayed to 4×10^{-7} over 20 epochs of 1,000 updates. The native VGGT-Ω objective is

$$\mathcal{L} = 5\mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{depth}} + 0.5\mathcal{L}_{\text{point}}. \tag{14}$$

Here, $\mathcal{L}_{\text{cam}}$ is an L1 loss on the native 9D camera encoding, $\mathcal{L}_{\text{point}}$ supervises world points constructed from predicted depth and predicted cameras, and confidence weighting uses $\alpha = 0.2$. We evaluate the fixed 20K checkpoint unchanged across all target datasets, without target-domain adaptation.

Autonomous driving. We use the nuScenes v1.0-trainval split and follow the data-processing and evaluation protocol of DriveVGGT (Jia et al., 2025). The six surround-view cameras are calibrated by composing the static camera-to-ego extrinsic with the timestamped ego pose. Training samples continuous sequences of 3–10 keyframes; evaluation uses 15 keyframes (90 images) resized to 288×512, with the first-keyframe `CAM_FRONT` as the reference gauge. All adapted variants use 5.248M trainable parameters and are trained for 10K steps on the same data and objective with the optimizer and hardware above. Six views per tuple and a 30-image-per-GPU budget yield a global batch of 20 tuples. The learning rate follows $10^{-8} \to 5 \times 10^{-5} \to 10^{-8}$ with 500 warmup updates and cosine decay. The objective is $5\mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{conf-depth}} + \mathcal{L}_{\text{reg-depth}}$ with $\alpha = 0.2$; only valid static LiDAR-supported pixels supervise depth.

Robot manipulation. For the robotic experiments, the Plain Adapter and CalibAdapter are trained on the same robot-specific dataset collected with the Dobot X-Trainer setup, using the same training split and optimization budget. The pretrained backbone is kept frozen, and only the corresponding adapter parameters are optimized. The training data are disjoint from the evaluation trials, and no further adaptation is performed at test time. Camera intrinsics are obtained from offline calibration, while per-frame extrinsics combine hand–eye and dual-base calibration with time-synchronized robot forward kinematics. Each method is evaluated in 20 trials per task with matched initial layouts. The RGB-D oracle follows the same downstream pipeline and differs only in the depth source.

A.4 Limitations

CPE measures metric 3D error under calibrated back-projection rather than camera-pose error itself. CalibAdapter assumes reliable test-time calibration and access to a cross-view interaction block. Performance may degrade under camera-motion distributions substantially outside the adaptation data, as observed on ETH3D.