RAGR: REGION-ADAPTIVE GROUP ROUTING FOR MULTI-VIEW 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent visual geometry foundation models recover camera poses, depth, and dense 3D structure from multiple images, yet a single global prediction model struggles to accurately reconstruct every local region. We investigate whether applying multi-choice, region-level corrections on top of a global prediction can further enhance reconstruction quality. We propose RAGR, a region-adaptive group routing framework for post-prediction refinement. RAGR partitions an initial 3D reconstruction into spatial groups supported by multi-view co-visibility, and employs a lightweight router to decide for each group whether to retain the shared global prediction or apply a compact correction learned from diverse teacher models. In zero-shot evaluations, RAGR improves the shared global prediction on ETH3D, while its AbsRel gain does not transfer to TNT. Teacher-derived corrections remain useful across domains, and post-hoc stitching shows realizable gains from better regional action choices on both benchmarks. These results highlight reliable, deployment-aligned action selection as a key requirement for regional post-prediction refinement.

1 INTRODUCTION

Recent visual geometry foundation models have substantially advanced multi-view 3D reconstruction by jointly estimating camera geometry, depth, correspondences, and dense scene structure within unified feed-forward frameworks (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a). Their global reasoning provides a solid foundation for establishing cross-view consistency and recovering complete scenes from multi-view observations. However, a single global prediction often struggles to reconstruct all local regions with uniform quality. Because residual errors naturally vary with local geometry, appearance, visibility, and viewing conditions, indiscriminately scaling or retraining the global model is often inefficient. Instead, this work explores an alternative avenue: refining the established global reconstruction through multi-choice, localized residual corrections.

Realizing such a localized refinement strategy hinges on establishing an effective matching mechanism between regional errors and corrective actions. Across a 3D scene, geometric errors naturally vary across spatial regions, while distinct correction behaviors can exhibit complementary local advantages. Such complementarity implies that uniformly applying a single correction across the entire scene is rarely optimal, motivating adaptive action selection conditioned on regional geometric characteristics. In practical deployment where target-domain ground truth is unavailable, however, the system must not only construct transferable correction capabilities, but also reliably discriminate and select the optimal action for each region under unsupervised domain shift.

To this end, we introduce RAGR (Region-Adaptive Group Routing), a post-prediction refinement framework designed for efficient and localized geometric correction. RAGR executes a global visual geometry model once to establish a shared scene baseline, and subsequently partitions the resulting reconstruction into spatial groups grounded in multi-view co-visibility constraints. For each co-visible group, RAGR defines a four-way action space: retaining the shared global prediction, or applying one of three lightweight depth-residual corrections distilled from frozen geometry teachers. Operating directly on the spatial representations output by the global model, these correction modules eliminate the need for redundant, costly full-scene forward passes. Finally, a compact policy network adaptively routes each group to its appropriate action, seamlessly merging the regional outputs into a unified 3D reconstruction. Figure 1 illustrates these localized modifications.

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.

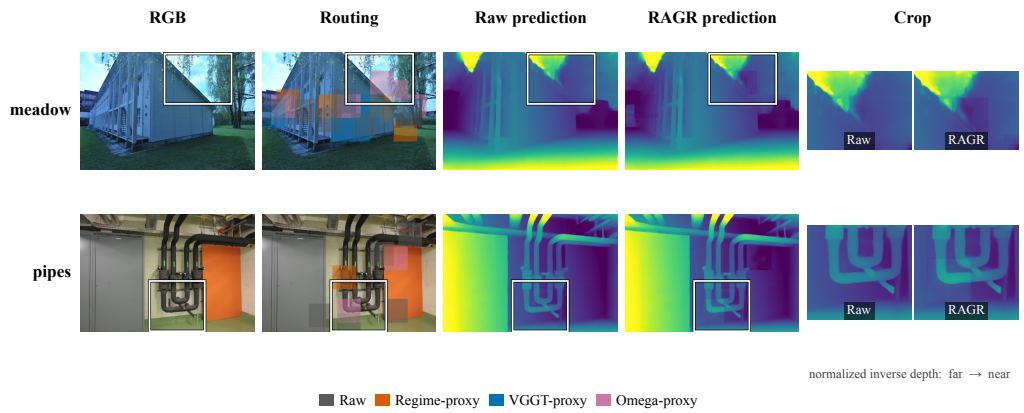

Figure 1: **Region-adaptive residual refinement on ETH3D.** Starting from one shared global prediction, RAGR assigns either Raw or one of three learned residual corrections to each multi-view-supported spatial group. The routing maps and matched crops illustrate localized modifications to the shared geometry. *Meadow* shows useful regional refinement, while *pipes* illustrates a case where the selected local corrections do not recover the strongest available alternative.

To dissect the mechanisms underlying localized post-prediction refinement, our evaluation decouples three critical dimensions: whether teacher supervision translates into generalizable residual corrections, whether distinct spatial regions genuinely exhibit complementary preferences for different corrective actions, and whether robust cross-domain action selection can be achieved without target-domain ground truth (Schops et al., 2017; Knapitsch et al., 2017).

Extensive cross-domain experiments provide clear empirical answers to these questions. First, regarding correction construction, residual proxies distilled from geometry teachers generalize substantially better than a same-capacity baseline trained directly on source ground-truth depth, markedly improving upon the shared base prediction and confirming the effective transfer of complementary priors. Second, post-hoc local diagnostics corroborate the existence of inter-region action complementarity, showing that diverse geometric structures indeed benefit from distinct corrective behaviors. However, translating this localized complementarity into tangible performance gains hinges critically on the discriminative stability of the routing policy under domain shift: on ETH3D, group-wise RAGR effectively enhances the baseline reconstruction; on Tanks and Temples, however, where distribution shifts are more severe, RAGR struggles to stably replicate these routing gains, achieving baseline improvements primarily through fixed residual executions. Together, these findings indicate that constructing transferable corrective actions and achieving reliable cross-domain spatial routing represent two distinct challenges, establishing key design principles for post-prediction adaptive refinement in multi-view 3D geometry models.

Contributions. Our primary contributions are summarized as follows:

- **Formulation of Post-Prediction Refinement:** We formulate post-prediction multi-view 3D reconstruction as an adaptive selection problem over localized residual actions, systematically decoupling spatial error heterogeneity, regional action complementarity, and cross-domain action selection.
- **The RAGR Framework:** We introduce RAGR, an efficient region-adaptive group routing framework. By reusing cached global representations, organizing reconstructions into multi-view covisible spatial groups, and deploying compact teacher-derived residual proxies, RAGR enables targeted refinement of intricate local geometries without repeatedly executing compute-intensive full-scene foundation models.
- **Decomposed Empirical Benchmarks and Insights:** We provide a comprehensive zero-shot evaluation protocol that disentangles residual transferability from cross-domain policy discriminability. Our findings reveal an intrinsic asymmetry between constructing transferable corrections and

executing reliable spatial selection under domain shift, establishing the key trade-offs and boundary conditions for localized post-prediction refinement.

2 Related Work

Visual geometry foundation models. Feed-forward geometry models have increasingly unified camera estimation, depth prediction, correspondence, and dense 3D reconstruction. DUSt3R (Wang et al., 2024) formulates geometric reconstruction through dense pointmap regression, MASt3R (Leroy et al., 2024) extends this framework with local feature prediction for image matching, and VGGT (Wang et al., 2025a) jointly estimates cameras, depth, pointmaps, and tracks from multiple views. Recent VGGT variants improve efficiency or reconstruction capability in different ways: FastVGGT (Shen et al., 2026) reduces inference computation through training-free token reduction, RegimeVGGT (You et al., 2026) exploits layer-dependent spatial redundancy, and VGGT-Ω (Wang et al., 2026) scales model capacity and training data with additional architectural changes. RAGR uses a global geometry model as a shared scene-level predictor and studies what can be gained by refining its remaining errors at a finer spatial granularity.

Global-to-local refinement and multi-view organization. Global and local processing have been combined in several forms. PatchFusion (Li et al., 2024a) fuses globally consistent monocular depth with high-resolution local predictions while enforcing consistency across overlapping tiles. PatchRefiner (Li et al., 2024b) and its subsequent extension PatchRefiner V2 (Li et al., 2026) further study patch-based monocular depth refinement and cross-domain generalization, while PRO (Kwon & Kim, 2025) studies seamless patchwise refinement for zero-shot high-resolution monocular depth estimation. For scalable multi-view geometry, Diversity-aware View Partitioning (Park et al., 2026) organizes input views into balanced subsets, while Co-VGGT (Ziliotto et al., 2026) probes frozen VGGT representations to estimate cross-view co-visibility. These works demonstrate the value of localized processing and structured multi-view organization. RAGR differs in operating after a shared multi-view reconstruction has been established: it uses co-visible 3D regions as decision units and selects among multiple output-space residual actions, including retaining the shared prediction unchanged.

Conditional computation and routing for geometry. Conditional computation adapts model behavior by selecting among alternative computational paths. In geometry, SMoEStereo (Wang et al., 2025b) adapts frozen vision foundation models for stereo matching using selectively activated mixture-of-experts modules. RAGR instead routes over output-space residual actions after global multi-view geometry has been fixed, choosing for each spatial group between retaining the global prediction and applying one of several teacher-derived corrections. Beyond introducing this action space, our evaluation separates whether local action complementarity survives teacher-to-proxy transfer from whether a source-trained policy can exploit that complementarity reliably under domain shift.

3 Method

3.1 Overview and Action Space

Figure 2 summarizes the RAGR framework, spanning zero-shot target inference and source-only learning. At target inference, a single global pass of frozen FastVGGT produces shared geometry and cached features. The scene is organized into co-visible 3D groups, each routed to either Raw or one of three residual proxy actions, after which the regional predictions are stitched into the final reconstruction. Source-only learning proceeds in two sequential stages. First, selective teacher imitation constructs the residual action pool, utilizing source ground truth to identify teacher-improving groups. Second, after freezing the action pool, source ground truth provides best-action labels based on local depth error to train the group router. Neither teacher networks nor ground-truth depth are involved during target inference.

Formally, given V input views $\mathcal{I} = \{I_v\}_{v=1}^{V}$, RAGR performs regional refinement after a shared multi-view geometry prediction has been established. A frozen global model $\mathcal{M}_g$ is executed once

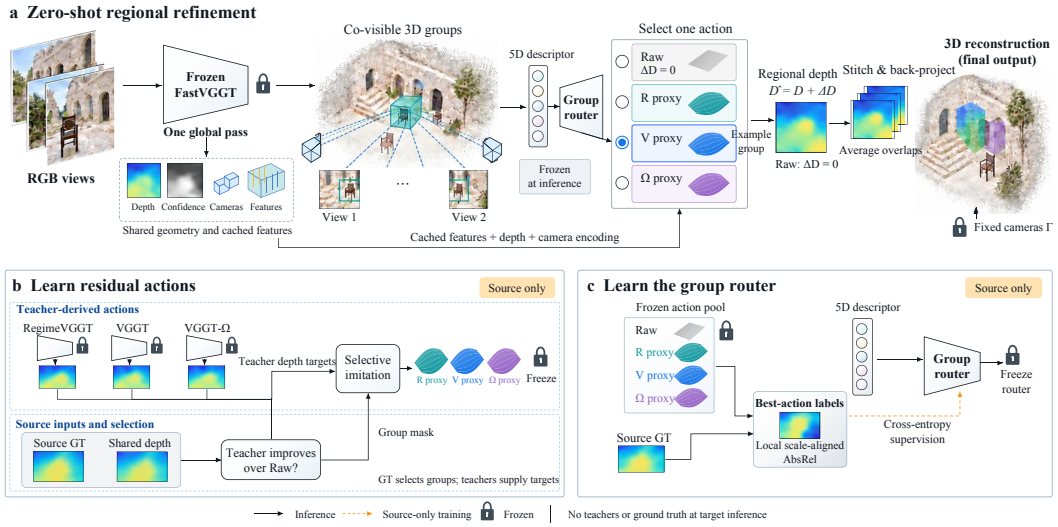

Figure 2: Overview of RAGR. (a) At zero-shot inference, a single frozen FastVGGT forward pass provides shared geometry and cached features.The scene is organized into co-visible 3D groups, and a frozen group router selects Raw or one of three lightweight residual actions for each group. The resulting regional depth predictions are stitched in the original views and back-projected with the fixed cameras to obtain the final reconstruction. (b) On source-domain data, frozen geometry teachers supervise the three residual proxies through selective imitation: source ground truth determines which groups receive supervision, while teacher depths provide the residual targets. (c) After freezing the action pool, source ground truth is used to construct best-action labels for training the four-way group router. Teachers and ground-truth depth are used only during source-domain training.

and produces

$$\mathcal{O}_p = \{D_p, C_p, \Gamma, T\}, \qquad \Gamma = \{K_v, R_v, t_v\}_{v=1}^{V}, \tag{1}$$

where D_p and C_p are the predicted depth and confidence maps, Γ contains the estimated cameras, and T denotes cached spatial representations.

The predicted scene is organized into co-visible spatial groups $\mathcal{G} = \{g_i\}_{i=1}^{M}$. For each group, RAGR selects one action from

$$\mathcal{A} = \{a_{\text{raw}}, a_R, a_V, a_\Omega\}, \tag{2}$$

corresponding to retaining the shared prediction or applying one of three teacher-derived residual proxies. Given the group descriptor $f(g_i)$, the deployed policy selects

$$\hat{a}_i = \arg\max_{a \in \mathcal{A}} \pi_\phi(a \mid f(g_i)). \tag{3}$$

If residual action a_k is selected, the regional depth is

$$\hat{D}(g_i) = \begin{cases} D_p(g_i), & \hat{a}_i = a_{\text{raw}}, \\ D_p(g_i) + \Delta D_k(g_i), & \hat{a}_i = a_k, \end{cases} \qquad k \in \{R, V, \Omega\}. \tag{4}$$

3.2 Global Prediction and Co-visible Groups

We instantiate M_g with FastVGGT (Shen et al., 2026). Its camera estimates remain fixed throughout regional refinement, providing a shared coordinate system for all subsequent group operations. Cached representations T are spatially sliced and reused by the residual proxies, so no regional action requires a new global forward pass.

Using the predicted depth and cameras, we construct localized 3D regions that are jointly supported by multiple views. Each group is represented as

$$g_i = (\mathcal{V}_i, \{\Omega_{i,v}\}_{v \in \mathcal{V}_i}), \tag{5}$$

where $\mathcal{V}_i$ is the supporting-view set and $\Omega_{i,v}$ is the corresponding image-space window. Exact anchor sampling, spatial clustering, merging, projection, and token-window parameters are given in Appendix A.2.

For routing, each group is summarized by the five-dimensional descriptor

$$f(g_i) = [r_i,\, d_i,\, \sigma_{D,i},\, E_{\mathrm{tex},i},\, \mu_{C,i}]^\top, \tag{6}$$

where r_i measures multi-view overlap, d_i is a robust median-depth statistic, $\sigma_{D,i}$ summarizes within-window depth variation, $E_{\mathrm{tex},i}$ measures image-gradient energy, and $\mu_{C,i}$ is the mean model confidence. All five quantities are computed from the input images and the shared global prediction; their exact definitions are provided in Appendix A.2.

3.3 Teacher-Derived Residual Proxies

The first source-training stage constructs the deployable residual action pool. We use frozen RegimeVGGT, VGGT, and VGGT-Ω models as offline teachers. Each teacher is paired with a lightweight depth-residual proxy h_{θ_k}, $k \in \{R, V, \Omega\}$. For group g_i, the proxy receives

$$X(g_i) = \big(T(g_i), D_p(g_i), \mathrm{PE}(\Gamma_i)\big) \tag{7}$$

and predicts

$$\Delta D_k(g_i) = h_{\theta_k}\big(X(g_i)\big). \tag{8}$$

where $T(g_i)$ and $D_p(g_i)$ denote the spatially sliced feature representations and initial depth from the shared global prediction, and $\mathrm{PE}(\Gamma_i)$ represents the positional encoding of the relative camera poses across supporting views $\mathcal{V}_i$. The three proxies share an identical architecture and differ solely in the teacher supervision they receive.

To prevent the proxies from absorbing potential failure modes or local artifacts from the teachers, we introduce a selective imitation mechanism. Let e_i^{raw} and $e_{i,k}^T$ denote the source-domain group AbsRel of the shared baseline and teacher k, respectively. Proxy k receives supervisory signal on group g_i only when the corresponding teacher demonstrates a verified geometric improvement:

$$m_{i,k} = \mathbb{I}\left[e_{i,k}^T < e_i^{\mathrm{raw}} - \epsilon_{\mathrm{mimic}}\right], \qquad \epsilon_{\mathrm{mimic}} = 0.01. \tag{9}$$

Here, source ground truth is accessed exclusively to construct this validity mask. On positive groups ($m_{i,k} = 1$), the proxy is optimized via relative depth imitation:

$$\mathcal{L}_{\mathrm{imit}}^{(k)} = \frac{1}{|\Omega_{i,k}^T|} \sum_{u \in \Omega_{i,k}^T} \frac{\left|\hat{D}_k(u) - D_T^{(k)}(u)\right|}{D_T^{(k)}(u)}, \tag{10}$$

where $\hat{D}_k = D_p + \Delta D_k$, and $\Omega_i = \bigcup_{v \in \mathcal{V}_i} \Omega_{i,v}$ comprises all valid projection pixels across supporting views. By strictly distilling relative geometric discrepancies from the teachers rather than fitting absolute source ground truth directly, the proxies retain generalizable geometric priors while remaining robust against domain-specific scale overfitting. Architectural details and training protocols are elaborated in Appendix A.3.

3.4 Group-Wise Action Policy

Once the pool of candidate residual actions is frozen, the second training stage learns a compact selection policy, parameterized by ϕ, to dynamically assign the most effective action to each co-visible group.

To facilitate stable optimization, each group descriptor is standardized feature-wise using empirical statistics computed exclusively over the source-domain training groups:

$$\tilde{f}_i = \frac{f(g_i) - \mu_f}{\sigma_f + \epsilon_f}. \tag{11}$$

where $\epsilon_f = 10^{-6}$ ensures numerical stability. The normalization parameters (μ_f, σ_f) remain fixed across all held-out and target-domain evaluations. The policy network is instantiated as a compact

MLP with layer dimensions $5 \to 128 \to 128 \to 4$ and ReLU activations, containing only 17,796 trainable parameters, which outputs a categorical probability distribution $\pi_\phi(a|\tilde{f}_i)$ over the action space $\mathcal{A}$ via a final Softmax layer.

Supervisory labels for the policy are constructed offline on the source domain. For each source group g_i, we evaluate all four candidate actions against the ground-truth depth. To isolate intrinsic geometric fidelity from subtle regional scale fluctuations inherent to relative geometry models, the local action error $E_i(a)$ is evaluated using per-window median-ratio scale-aligned AbsRel, aggregated across supporting views weighted by valid pixel counts. The action achieving the lowest error designates the optimal supervision target:

$$a_i^* = \arg\min_{a \in \mathcal{A}} E_i(a), \tag{12}$$

The policy parameters ϕ are then optimized by minimizing the cross-entropy loss:

$$\mathcal{L}_{\text{router}} = -\frac{1}{M_{\text{tr}}} \sum_{i=1}^{M_{\text{tr}}} \log \pi_\phi \left(a_i^* \mid \tilde{f}_i \right). \tag{13}$$

where M_{tr} denotes the total number of source training groups. At test-time inference, the trained policy operates entirely in a zero-shot, feed-forward manner conditioned solely on the unlabelled descriptor $\tilde{f}_i$, requiring no ground-truth depth or auxiliary scale alignment. Detailed support filtering and weighting formulations are provided in Appendix A.3.

3.5 Regional Stitching and Reconstruction

Following action selection, the refined regional depth predictions are mapped back into their respective supporting-view image windows. For a pixel coordinate u in view v, let

$$\mathcal{N}_v(u) = \{i : v \in \mathcal{V}_i,\ u \in \Omega_{i,v}\}$$

denote the set of co-visible spatial groups covering that pixel. The composite stitched depth map is synthesized via:

$$\widehat{D}_v(u) = \begin{cases} \dfrac{1}{|\mathcal{N}_v(u)|} \displaystyle\sum_{i \in \mathcal{N}_v(u)} \widehat{D}_{i,v}(u), & |\mathcal{N}_v(u)| > 0, \\ D_{p,v}(u), & |\mathcal{N}_v(u)| = 0. \end{cases} \tag{14}$$

This formulation naturally preserves the coherent global structure in uncovered regions ($|\mathcal{N}_v(u)| = 0$) by retaining the base prediction $D_{p,v}(u)$. In densely covered regions, the aggregation across overlapping groups, facilitated by the boundary halo of each window, provides smooth transitions and effectively mitigates boundary seam artifacts across adjacent spatial patches.

The final dense 3D pointmaps are deterministically reconstructed by unprojecting the stitched depth maps using the frozen global camera geometry:

$$\widehat{P}_v = \mathcal{B}(\widehat{D}_v; K_v, R_v, t_v), \tag{15}$$

where $\mathcal{B}$ denotes standard pinhole back-projection parameterized by intrinsics K_v and camera pose (R_v, t_v). Any benchmark-level scale alignment is deferred until the entire scene has been fully stitched into a complete reconstruction (Sec. 4.1), strictly preserving internal geometric consistency throughout the refinement pipeline.

4 Experiments

We first evaluate final zero-shot reconstruction quality and then use controlled studies to analyze where the gains and limitations of regional action selection come from. All trainable components and deployment decisions are fixed using source-domain data before evaluation on the target benchmarks.

Table 1: **Main zero-shot reconstruction results on ETH3D and Tanks and Temples (TNT).** All methods are evaluated under the unified scene-level protocol in Sec. 4.1. Raw is also the fixed action selected using source-held-out data. Lower is better for both metrics.

	ETH3D		TNT	
Method	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓
Full-frame reference models				
RegimeVGGT	0.42136	2.66115	0.26770	1.44000
VGGT	0.43147	2.75081	0.27583	1.44730
VGGT-Ω	0.45274	2.85080	0.29751	1.46099
Shared prediction and fixed corrections				
Raw	0.43949	2.66911	0.23708	1.42487
Regime-derived	0.41405	2.42032	0.23400	**1.42324**
VGGT-derived	**0.40507**	**2.35661**	0.23408	1.42392
Omega-derived	0.45630	2.87332	0.24090	1.42589
Adaptive action selection				
Scene-level control	0.43668	2.67069	**0.23394**	1.42360
RAGR (group-wise)	0.42681	2.54528	0.23794	1.42440

4.1 Experimental Setup

Datasets and zero-shot protocol. All learnable components are trained on a frozen source split containing 21 training and 11 held-out sequences from 7-Scenes and ARKitScenes (Shotton et al., 2013; Baruch et al., 2021). The held-out source split is used for source-side validation and target-free deployment decisions. Feature-normalization statistics are estimated from the source training groups and then frozen. We evaluate zero-shot on 13 ETH3D scenes with valid depth (Schops et al., 2017) and six Tanks and Temples (TNT) scenes (Knapitsch et al., 2017). No target-domain ground truth is used before inference.

Compared methods. We compare three types of execution. First, RegimeVGGT, VGGT, and VGGT-Ω are evaluated directly as frozen full-frame references. Second, Raw and each teacher-derived residual correction are applied uniformly, which measures the quality of individual correction actions without routing. Third, we compare group-wise RAGR with a scene-level control built from the same trained router. RAGR may select a different action for each spatial group, whereas the scene-level control combines the group predictions and applies one common action to the whole scene. As a simple target-free fixed baseline, we also use the single action with the lowest error on the held-out source split for every target scene; this procedure selects Raw.

Evaluation protocol. All methods are evaluated on the same pixels within each scene: we retain only locations where the ground-truth depth and all compared predictions are valid. This common evaluation region still covers more than 98% of GT-valid pixels on average on both benchmarks. Regional predictions are fully stitched before evaluation, and each method receives one median-ratio scale estimate per scene. We report scene-level absolute relative error (AbsRel) and root mean squared error (RMSE), averaged uniformly across scenes. For uncertainty estimates, we resample complete scenes rather than individual pixels. Detailed confidence intervals and support statistics are provided in the appendix.

4.2 Main Zero-Shot Results

Transferred correction actions. Table 1 shows that useful teacher-derived correction behaviors can transfer into lightweight residual proxies. On ETH3D, the Regime- and VGGT-derived actions reduce AbsRel from 0.43949 for Raw to 0.41405 and 0.40507, respectively, with paired scene-bootstrap intervals excluding zero. On TNT, the same actions obtain slightly lower mean AbsRel than Raw (0.23400/0.23408 versus 0.23708), although the corresponding intervals include zero. Both actions also obtain lower AbsRel than their corresponding full-frame references on the two benchmarks, showing that useful teacher behavior can be retained in lightweight post-prediction corrections. The Omega-derived action degrades Raw on both domains.

Table 2: Local complementarity and routing gap under the post-hoc per-window scale-aligned AbsRel diagnostic. These values are computed before scene-level stitching.

Dataset	Best fixed ↓	Group oracle ↓	Router ↓	Oracle gain ↑	Routing gap ↓
ETH3D	0.32950	0.31519	0.33804	0.01431	0.02286
TNT	0.19764	0.19421	0.26314	0.00343	0.06893

Adaptive action selection. The adaptive executions exhibit domain-dependent behavior. On ETH3D, group-wise RAGR improves Raw from 0.43949 to 0.42681, with the paired scene-bootstrap interval excluding zero. On TNT, RAGR obtains slightly higher AbsRel than Raw, while the matched scene-level control attains the lowest mean AbsRel among the target-free adaptive executions. These results motivate the mechanism studies below, which separate the quality of transferred correction actions from the reliability of cross-domain action selection.

4.3 Mechanism Studies and Ablations

We examine four factors underlying localized post-prediction refinement: spatial error heterogeneity, regional action complementarity, residual capacity, and cross-domain action selection. Target-domain ground truth is used only post-inference for diagnosis.

Spatial Error Heterogeneity. Per-window scale-aligned AbsRel reveals substantial variation in Raw error across spatial groups. For ETH3D, the pooled 10th/50th/90th percentiles are 0.203/0.338/0.726 across 598 groups (mean 0.412 ± 0.258); for TNT they are 0.042/0.183/0.824 across 336 groups (mean 0.366 ± 0.518). Wide spreads also persist within individual scenes, showing that the pooled heterogeneity is not solely explained by scene-level difficulty (Appendix E).

Action Complementarity vs. Routing Gap. Under the post-hoc local diagnostic, the group oracle improves over the best fixed action by 0.01431 on ETH3D and 0.00343 on TNT, while the deployed router remains 0.02286 and 0.06893 above the oracle, respectively (Table 2). Thus, measurable regional complementarity exists, but the source-trained router does not reliably recover it after domain shift. To test whether this opportunity survives the actual reconstruction pipeline, we replay the same local oracle labels through regional write-back, overlap averaging, and unified scene-level evaluation. This local-oracle-selected stitching reduces AbsRel from 0.42681 to 0.41828 on ETH3D and from 0.23794 to 0.23344 on TNT relative to RAGR, with paired scene-bootstrap intervals excluding zero on both domains. It is a realizable post-hoc diagnostic rather than a scene-optimal oracle; full protocol details are given in Appendix E.

Teacher-Derived Supervision vs. Model Capacity. A matched-capacity residual head trained directly on source ground-truth depth reduces source AbsRel from 0.5806 to 0.3439, but increases matched target error from 0.4156 to 0.5358 on ETH3D and from 0.2406 to 0.3698 on TNT. The teacher-derived formulation therefore transfers more reliably than this same-capacity GT-only formulation. Because the supervision target and selective-imitation mechanism change together, this comparison does not isolate their individual contributions.

Cross-Domain Policy Mismatch. On TNT, the VGGT-derived action is oracle-optimal for 50.3% of groups but selected by the router for only 5.1%, whereas the VGGT-Ω action is selected for 36.0% but oracle-optimal for only 3.6%. Router–oracle agreement is 14.6% on TNT and 26.6% on ETH3D. Moreover, local-scale and shared-scene-scale diagnostic labels agree on only 42.9% and 42.6% of groups, respectively, showing that the evaluation convention substantially affects action preference. Removing the overlap descriptor r_i yields no stable independent benefit, and broader router audits likewise provide no consistent advantage over matched fixed-action references (Appendix H). Together, these results identify both cross-domain selection reliability and selection-objective alignment as limitations of the current RAGR configuration.

4.4 Hyperparameters and Implementation Details

All headline experiments use a frozen FastVGGT global predictor, with camera parameters kept fixed during regional refinement. Spatial groups are constructed from stride-4 anchors after removing the lowest 10% confidence candidates, clustered on a 16^3 voxel grid, and retained only when supported by at least two views. Each group is represented by 14×14 token-aligned windows with a one-pixel halo.

The three deployed residual proxies share the same architecture and predict depth residuals only. Each proxy has 741,869 trainable parameters, hidden width 256, three 128-dimensional fusion blocks, and a maximum residual magnitude of 1.5. They are trained with selective teacher imitation using $\epsilon_{\text{mimic}} = 0.01$, Adam with learning rate 3×10^{-4}, 40 epochs, and gradient clipping at 1.0.

The deployed router is a $5\to128\to128\to4$ ReLU MLP with 17,796 trainable parameters. Its input features are normalized using statistics estimated from the source training split and then frozen. The router is trained with hard action labels using Adam with learning rate 3×10^{-3} for 200 full-batch epochs. Further optimization details, grouping thresholds, training-update counts, random seeds, and exact descriptor definitions are provided in the appendix.

5 Discussion and Limitations

Our results distinguish transferable correction construction from reliable regional action selection. Teacher-derived residual actions retain useful cross-domain behavior, and post-hoc diagnostics show both regional complementarity and realizable scene-level gains when local oracle choices are replayed through stitching. However, the deployed router does not consistently recover this opportunity, and action preferences change substantially with the local evaluation convention. Effective regional refinement therefore depends on both cross-domain selection reliability and alignment between local supervision and the final reconstruction objective.

Several limitations remain. The GT-only control does not separate teacher targets from selective imitation, and the current co-visible grouping is an operational rather than learned partition. The local routing objective also differs from the final stitched scene-level metric, motivating better-aligned supervision. Finally, current profiling shows that grouping, feature extraction, and routing dominate regional post-processing cost; reducing this overhead remains an important engineering direction.

6 Conclusion

We presented RAGR, a region-adaptive framework that composes lightweight teacher-derived residual corrections over co-visible spatial groups after one shared multi-view prediction. Across ETH3D and TNT, useful correction actions transfer across domains and different regions exhibit measurable action complementarity. Post-hoc stitching further shows that locally preferred actions can improve over the deployed router at the final scene level, while the source-trained policy does not consistently recover this opportunity. These results highlight cross-domain action selection and alignment between local supervision and scene-level reconstruction as key challenges for regional post-prediction refinement.

AI Use Statement

We used OpenAI's ChatGPT as an assistive tool during the revision and preparation of this manuscript. The tool was used to provide feedback on experimental design and control experiments, assist in the interpretation and organization of experimental results, and support the refinement of the paper's presentation and structure. It was also used for drafting, rewriting, language editing, and Chinese–English translation of portions of the manuscript, as well as for organizing discussions of related work. Generative AI was not used to generate or modify experimental measurements, benchmark results, or ground-truth data reported in this paper.

All experimental configurations, reported numerical results, methodological decisions, scientific claims, and citations were reviewed and verified by the authors against the underlying implementa-

tions, experimental artifacts, and original sources. The authors made all final research and writing decisions and take full responsibility for the content of this paper.

REFERENCES

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

Byeongjun Kwon and Munchurl Kim. One look is enough: Seamless patchwise refinement for zero-shot monocular depth estimation on high-resolution images. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8077–8087. IEEE, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10016–10025. IEEE, 2024a.

Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchrefiner: Leveraging synthetic data for real-domain high-resolution monocular metric depth estimation. In *European Conference on Computer Vision*, pp. 250–267. Springer, 2024b.

Zhenyu Li, Wenqing Cui, Shariq Bhat, and Peter Wonka. Patchrefiner v2: Fast and lightweight real-domain high-resolution metric depth estimation. In *International Conference on Learning Representations*, volume 2026, pp. 117738–117753, 2026.

Jinsoo Park, Donggyu Choi, Ahyun Seo, Minsu Cho, and Jeany Son. Diversity-aware view partitioning for scalable vggt. In *European Conference on Computer Vision*, pp. 547–564. Springer, 2026.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. Fastvggt: Fast visual geometry transformer. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2930–2937, 2013.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 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, 2026.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Yun Wang, Longguang Wang, Chenghao Zhang, Yongjian Zhang, Zhanjie Zhang, Ao Ma, Chenyou Fan, Tin Lun Lam, and Junjie Hu. Learning robust stereo matching in the wild with selective mixture-of-experts. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21276–21287. IEEE, 2025b.

Jinhao You, Shuo Lyu, Zhuohang Lyu, Tanxuan Li, Zibo Zhao, Jiaxiang Hu, Kai Tang, and Yichen Guo. Regimevggt: Layer-wise spatially preserving redundancy removal for visual geometry grounded transformer. *arXiv preprint arXiv:2606.18439*, 2026.

Filippo Ziliotto, Luciano Serafini, Lamberto Ballan, and Tommaso Campari. What vggt knows about overlap: Probing geometric foundation models for co-visibility. *arXiv preprint arXiv:2607.09503*, 2026.

A Implementation and Evaluation Details

A.1 Frozen Experimental Configuration

All headline experiments follow a frozen source-only protocol. The source split contains 21 training and 11 held-out sequences from 7-Scenes and ARKitScenes. The held-out split is used for source-side validation and target-free deployment decisions, including selection of the fixed deployment action. Neither ETH3D nor Tanks and Temples (TNT) is used for proxy training, router training, feature normalization, or any decision made before target-domain evaluation.

The shared global prediction is produced by FastVGGT. The action pool contains Raw and three depth-residual proxies derived from RegimeVGGT, VGGT, and VGGT-Ω. Each proxy contains 741,869 trainable parameters and predicts only a depth residual on top of the shared global prediction. Camera parameters remain fixed during regional refinement, and final pointmaps are obtained by back-projecting the stitched depth.

The deployed router is a $5 \to 128 \to 128 \to 4$ MLP with 17,796 trainable parameters. Feature-normalization statistics are estimated exclusively from source training groups and then frozen for source-held-out and target-domain evaluation.

A.2 Spatial Group Construction and Routing Descriptor

Group construction. Anchor locations are sampled from the shared prediction with stride 4. For each scene, among anchors with finite positive predicted depth, we compute the 10th percentile of global-prediction confidence,

$$\tau_C = \mathrm{Quantile}_{0.10}\{C(a)\}, \tag{16}$$

and retain anchors satisfying $C(a) \geq \tau_C$. Thus, approximately the lowest-confidence 10% of candidate anchors are removed on a scene-adaptive basis rather than by a fixed confidence threshold.

The retained anchors are back-projected using the shared camera estimates and clustered in a $16 \times 16 \times 16$ voxel grid. Clusters supported by fewer than two distinct views are discarded. Spatially adjacent clusters are merged when their implementation depth discrepancy is at most 0.15 and the Jaccard overlap between their supporting-view sets is at least 0.5. Each resulting group is projected into its supporting views and represented by token-aligned 14×14 windows. A one-pixel outer halo is included for regional write-back and stitching.

These settings define the operational spatial partition used throughout the experiments; we do not claim that this grouping is optimal among alternative spatial organizations.

Routing descriptor. For group g_i with supporting-view set $\mathcal{V}_i$ and $K_i = |\mathcal{V}_i|$, the deployed descriptor is

$$f(g_i) = [r_i,\, d_i,\, \sigma_{D,i},\, E_{\mathrm{tex},i},\, \mu_{C,i}]^\top . \tag{17}$$

The overlap statistic r_i is computed from the complete shared-prediction depth maps rather than only the local group windows. For every ordered pair of distinct supporting views (p, q), each valid depth pixel in view p is back-projected to 3D and reprojected into view q. The projected location

is sampled at the nearest integer pixel. A reprojection is counted as geometrically matched if it lies inside the target image, both depths are valid, and

$$|z_{p\to q} - D_q(u')| \le 0.05 D_q(u'). \tag{18}$$

Let $O_{p\to q}$ be the number of matched pixels divided by the number of valid shared-prediction depth pixels in view p. We define

$$r_i = \frac{1}{K_i(K_i-1)} \sum_{\substack{p,q\in\mathcal{V}_i \\ p\neq q}} O_{p\to q}. \tag{19}$$

For the remaining statistics, let $W_{i,v}$ denote the token-aligned halo window of group i in supporting view v. The depth statistic is a median rather than a pixel mean:

$$d_i = \mathrm{median}_{v\in\mathcal{V}_i} \left[\mathrm{median}_{u\in W_{i,v},\, D_v(u)>0} D_v(u)\right]. \tag{20}$$

Depth variation is computed independently in each supporting view and then averaged without pixel-count weighting:

$$\sigma_{D,i} = \frac{1}{K_i} \sum_{v\in\mathcal{V}_i} \mathrm{std}_{u\in W_{i,v},\, D_v(u)>0} D_v(u). \tag{21}$$

For texture, the grayscale image is the channel-wise mean of RGB. For each window, we compute horizontal and vertical first-order absolute finite differences,

$$e_{i,v} = \mathrm{mean}\left(|\Delta_x I_v| \cup |\Delta_y I_v|\right), \tag{22}$$

where $\cup$ denotes concatenation of the two finite-difference sets. The group statistic is

$$E_{\mathrm{tex},i} = \frac{1}{K_i} \sum_{v\in\mathcal{V}_i} e_{i,v}. \tag{23}$$

This implementation does not use a Sobel operator or $\sqrt{G_x^2 + G_y^2}$ gradient magnitude.

Finally, confidence is averaged over all pixels in all group windows, weighted naturally by window pixel count,

$$\mu_{C,i} = \frac{\sum_{v\in\mathcal{V}_i} \sum_{u\in W_{i,v}} C_v(u)}{\sum_{v\in\mathcal{V}_i} |W_{i,v}|}. \tag{24}$$

No additional valid-depth mask is applied to this statistic. The implementation also computes confidence standard deviation, but it is not included in the deployed five-dimensional descriptor.

A.3 Proxy and Router Training

Residual proxies. The three residual proxies share the same architecture but are trained independently from their corresponding frozen teachers. Each proxy has hidden dimension 256, three 128-dimensional fusion blocks, stem width 16, and maximum residual magnitude 1.5. The source proxy-training pool contains 411 groups.

All three proxies use the selective teacher-imitation objective from Sec. 3. A group receives supervision from teacher k only when that teacher improves over Raw by at least $\epsilon_{\mathrm{mimic}} = 0.01$. The proxies are depth-only and use no ground-truth anchoring term.

Each proxy is optimized independently with Adam $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 10^{-8})$, learning rate 3×10^{-4}, zero weight decay, and no learning-rate scheduler for 40 epochs. One multi-view group forms one optimizer batch, with all supporting views of that group processed jointly. We use no gradient accumulation or mixed-precision training and clip the global ℓ_2 gradient norm to 1.0. All deployed proxies use seed 0.

Because the teacher-positive masks differ, the proxies use identical hyperparameters but receive different numbers of parameter updates. Regime-, VGGT-, and Omega-derived proxies contain 48, 42, and 134 positive groups per epoch, corresponding to 1,920, 1,680, and 5,360 Adam updates over 40 epochs, respectively.

Router supervision. The router is trained after the proxy pool has been frozen. For source-domain group g_i, candidate action $a \in \mathcal{A}$, and supporting view $v \in \mathcal{V}_i$, let $\Omega^a_{i,v}$ denote the pixels for which both the ground-truth depth and the prediction produced by action a are finite and positive. Each action is evaluated on its own valid support.

An independent median-ratio scale is estimated for each action within each supporting window,

$$s^a_{i,v} = \mathrm{median}_{u \in \Omega^a_{i,v}} \frac{D_{\mathrm{gt},v}(u)}{\max(D_{a,i,v}(u), 10^{-6})}. \tag{25}$$

The corresponding window-level scale-aligned AbsRel is

$$\mathrm{AbsRel}_{\mathrm{SA}}(a,i,v) = \frac{1}{|\Omega^a_{i,v}|} \sum_{u \in \Omega^a_{i,v}} \frac{\left| s^a_{i,v} D_{a,i,v}(u) - D_{\mathrm{gt},v}(u) \right|}{D_{\mathrm{gt},v}(u)}. \tag{26}$$

Supporting views are aggregated by valid-pixel count,

$$E_i(a) = \frac{\sum_{v \in \mathcal{V}_i} |\Omega^a_{i,v}| \, \mathrm{AbsRel}_{\mathrm{SA}}(a,i,v)}{\sum_{v \in \mathcal{V}_i} |\Omega^a_{i,v}|}. \tag{27}$$

The hard target is

$$a^*_i = \arg\min_{a \in \mathcal{A}} E_i(a). \tag{28}$$

Router optimization. The deployed router is a flat $5 \to 128 \to 128 \to 4$ MLP with ReLU nonlinearities. Each feature dimension is standardized using source-training statistics,

$$\tilde{f}_k = \frac{f_k - \mu_k}{\sigma_k + 10^{-6}}, \tag{29}$$

which are frozen for all held-out and target-domain evaluation.

The router is optimized with Adam $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 10^{-8})$, learning rate 3×10^{-3}, and no weight decay for 200 epochs using seed 0. Optimization is full-batch over the 410 groups in the frozen router-training set, giving one update per epoch. The 151 source-held-out groups are excluded from gradient optimization.

The deployed router uses hard labels. The recorded temperature $\tau = 0.01$ applies only to soft-label variants analyzed in Appendix H. The per-window scale normalization above is used only for source-domain label construction and post-hoc local diagnostics. No ground-truth depth or scale alignment is available to the router during target-domain inference.

A.4 Unified Scene-Level Evaluation

All methods in the main reconstruction comparison use one scene-specific comparison support,

$$\Omega_{\mathrm{cmp}} = \{u : D_{\mathrm{gt}}(u) \text{ and every evaluated prediction at } u \text{ are finite and positive}\}. \tag{30}$$

The same support is therefore used by Raw, all three fixed corrections, both adaptive executions, and the three full-frame reference models.

For regional methods, all local predictions are first assembled into complete view-wise depth maps. Overlapping predictions are uniformly averaged and uncovered pixels retain Raw. No scale normalization is performed before stitching. Each method m then receives exactly one scene-level median-ratio scale,

$$s^{(m)}_{\mathrm{scene}} = \mathrm{median}_{u \in \Omega_{\mathrm{cmp}}} \frac{D_{\mathrm{gt}}(u)}{\max(\hat{D}^{(m)}(u), 10^{-6})}. \tag{31}$$

AbsRel and RMSE are evaluated on the same Ω_{cmp}, and benchmark values are simple means over scenes.

The unified support retains 99.71% of GT-valid pixels on ETH3D on average, with a range of 99.04–99.99%. On TNT it retains 98.72% on average, with a range of 92.73–100.00%; the minimum occurs on Church. Thus, the common-support intersection removes only a small fraction of GT-valid pixels.

Table 3: Per-scene ETH3D AbsRel under unified comparison support.

Scene	Raw	Fix-R	Fix-V	Fix-Ω	Scene	RAGR	RegimeVGGT	VGGT	VGGT-Ω
courtyard	.27553	.27763	.27580	.28286	.27553	.27761	.30907	.30414	.29167
delivery area	.53400	.51423	.49349	.56643	.53400	.50409	.63932	.61356	.56774
electro	.45648	.42370	.41994	.47434	.47466	.44329	.46683	.46803	.48859
facade	.24400	.24417	.24340	.25712	.24400	.24513	.32685	.33828	.34300
kicker	.34892	.35553	.34929	.35486	.34892	.34959	.34584	.34536	.34474
meadow	.51997	.45463	.42641	.59409	.59443	.45070	.55315	.56492	.47169
office	.30878	.31200	.30931	.31220	.30878	.30959	.31184	.30077	.30485
pipes	.53266	.46621	.44330	.56232	.56280	.53542	.49569	.50105	.52476
playground	.66732	.61366	.60867	.66419	.61402	.63139	.55477	.67499	.71095
relief	.36232	.34068	.33909	.35795	.34073	.35220	.35764	.35587	.36747
relief 2	.40542	.37538	.37454	.43630	.37521	.41709	.32611	.35813	.41094
terrace	.38954	.36544	.36212	.39809	.36545	.38992	.38837	.38702	.41820
terrains	.66848	.63943	.62058	.67111	.63836	.64256	.40220	.39703	.64099
Mean	.43949	.41405	.40507	.45630	.43668	.42681	.42136	.43147	.45274

Table 4: Per-scene ETH3D RMSE under unified comparison support.

Scene	Raw	Fix-R	Fix-V	Fix-Ω	Scene	RAGR	RegimeVGGT	VGGT	VGGT-Ω
courtyard	3.21953	3.28702	3.22114	3.40869	3.21953	3.27547	3.55616	3.50955	3.53278
delivery area	3.72583	3.55612	3.39406	4.02871	3.72583	3.46352	4.30846	4.23770	4.08963
electro	3.72109	3.29876	3.25217	4.01478	4.02112	3.55594	3.82594	3.80234	4.64646
facade	2.99450	2.99922	2.98594	3.32115	2.99450	3.03580	3.78179	3.82696	3.99528
kicker	1.33095	1.35995	1.33850	1.37327	1.33095	1.33929	1.29274	1.29614	1.30652
meadow	5.37410	4.30416	4.00774	6.46803	6.47548	4.33443	5.85690	6.11945	4.12332
office	.62442	.63343	.62658	.63254	.62442	.62680	.62288	.60744	.61435
pipes	1.44507	1.23733	1.15793	1.54159	1.54456	1.45480	1.32077	1.33655	1.42716
playground	2.80578	2.48687	2.47569	2.73817	2.48777	2.59005	2.16120	2.75801	2.99154
relief	1.68715	1.47876	1.47416	1.67928	1.47913	1.62699	1.58778	1.62231	1.76770
relief 2	1.97691	1.71591	1.70746	2.29406	1.71438	2.13789	1.27130	1.66334	2.10283
terrace	3.52896	3.04243	3.01802	3.65432	3.04239	3.54822	3.57449	3.56308	4.28029
terrains	2.26413	2.06417	1.97655	2.19856	2.05888	2.09948	1.43451	1.41767	2.18254
Mean	2.66911	2.42032	2.35661	2.87332	2.67069	2.54528	2.66115	2.75081	2.85080

All paired uncertainty intervals treat scenes, rather than pixels or local groups, as the independent bootstrap units. Under this protocol, the ETH3D differences Fixed-R−Raw, Fixed-V−Raw, and Fixed-O−Raw are -0.02544 ($[-0.03902, -0.01288]$), -0.03442 ($[-0.05168, -0.01891]$), and $+0.01680$ ($[+0.00697, +0.02912]$), respectively. On TNT the corresponding differences are -0.00308 ($[-0.00857, +0.00058]$), -0.00300 ($[-0.00890, +0.00134]$), and $+0.00382$ ($[+0.00199, +0.00715]$).

Relative to the source-selected Raw deployment baseline, RAGR−Raw is -0.01268 ($[-0.02541, -0.00227]$) on ETH3D and $+0.00086$ ($[+0.00025, +0.00138]$) on TNT. The matched scene-level control differs from Raw by -0.00281 ($[-0.01833, +0.01495]$) and -0.00314 ($[-0.00839, +0.00080]$), respectively. Finally, RAGR−Scene is -0.00987 ($[-0.03704, +0.00985]$) on ETH3D and $+0.00400$ ($[+0.00016, +0.00961]$) on TNT.

B Additional Reconstruction Results

B.1 Per-Scene Unified-Support Results

Tables 3–6 report per-scene results under exactly the same unified comparison support as the main table. For compactness, Fix-R, Fix-V, and Fix-Ω denote fixed Regime-, VGGT-, and Omega-derived corrections, and Scene denotes the matched scene-level control.

B.2 Auxiliary Point-Cloud Diagnostics

The point-cloud results are retained as auxiliary geometric diagnostics under their original matched reconstruction support. They should not be interpreted as part of the unified pixel-level comparison in Table 1. No adaptive execution strategy is uniformly dominant across these geometric metrics.

Table 5: Per-scene TNT AbsRel under unified comparison support.

Scene	Raw	Fix-R	Fix-V	Fix-Ω	Scene	RAGR	RegimeVGGT	VGGT	VGGT-Ω
Barn	.17064	.17092	.17119	.17325	.17328	.17144	.17096	.17148	.17352
Caterpillar	.19475	.19209	.19192	.19661	.19207	.19424	.17619	.17823	.17496
Church	.23765	.23891	.24099	.23984	.23765	.23822	.44986	.47495	.51464
Courthouse	.29716	.29854	.29822	.29959	.29716	.29882	.36252	.36048	.36435
Ignatius	.27470	.25902	.25767	.28672	.25897	.27620	.20944	.22435	.30286
Truck	.24759	.24452	.24449	.24938	.24452	.24875	.23724	.24550	.25474
Mean	.23708	.23400	.23408	.24090	.23394	.23794	.26770	.27583	.29751

Table 6: Per-scene TNT RMSE under unified comparison support.

Scene	Raw	Fix-R	Fix-V	Fix-Ω	Scene	RAGR	RegimeVGGT	VGGT	VGGT-Ω
Barn	.59891	.59720	.59705	.59919	.59920	.59840	.58583	.58647	.58765
Caterpillar	3.01106	3.00912	3.00910	3.01050	3.00912	3.00973	2.99991	3.00004	3.00222
Church	1.12187	1.12190	1.12716	1.12630	1.12187	1.12298	1.27976	1.31707	1.37382
Courthouse	.43252	.43229	.43208	.43344	.43252	.43252	.42766	.42823	.43063
Ignatius	1.71871	1.71396	1.71336	1.71946	1.71393	1.71631	1.69342	1.69679	1.71845
Truck	1.66615	1.66496	1.66479	1.66643	1.66495	1.66643	1.65340	1.65518	1.65318
Mean	1.42487	1.42324	1.42392	1.42589	1.42360	1.42440	1.44000	1.44730	1.46099

Table 7: Aggregate point-cloud diagnostics. Accuracy and completeness are distances in meters and are lower when better; F-scores are higher when better.

Domain	Method	Acc.↓	Comp.↓	F@5cm↑	F@10cm↑
ETH3D	Raw	.3146	.5379	.0667	.2100
	Fixed-R	.2817	.5181	.0670	.2138
	Fixed-V	.2733	.5232	.0672	.2144
	Fixed-O	.3465	.5016	.0647	.2093
	RAGR	.2937	.5229	.0676	.2126
TNT	Raw	.07525	.05595	.8501	.9140
	Fixed-R	.07508	.05606	.8479	.9126
	Fixed-V	.07592	.05606	.8476	.9127
	Fixed-O	.07670	.05593	.8475	.9135
	RAGR	.07577	.05592	.8482	.9132

Table 8: Same-capacity GT-only residual control. Comparisons use their own matched within-run support.

Domain	Raw	GT-only	Difference	95% CI
ETH3D	.4156	.5358	+.1202	[+.0617,+.1899]
TNT	.2406	.3698	+.1292	[+.0970,+.1712]

C SAME-CAPACITY RESIDUAL CONTROL

The GT-only control uses the same depth-residual architecture as one deployed proxy: hidden dimension 256, three 128-dimensional fusion blocks, stem width 16, and maximum residual magnitude 1.5. It is trained for 40 epochs with learning rate 3×10^{-4} and seed 0 using the same 21/11 source split. Teacher-derived supervision is disabled, and the head is trained directly from source-domain ground-truth depth.

On source-held-out data, this control reduces AbsRel from 0.5806 for Raw to 0.3439. The source-side gain does not transfer to either target domain. Because this control was evaluated within its own matched run, its Raw values should be compared only with the corresponding GT-only values below and should not be numerically mixed with the unified-support headline table.

This control supports the narrower conclusion that the teacher-derived proxy formulation transfers more reliably than the matched same-capacity GT-only formulation. It does not isolate whether the difference arises from teacher targets, selective imitation, or their combination.

D Source-Selected Deployment and Selection Granularity

Source-selected fixed action. Using the frozen source-held-out count-weighted scores, the four actions obtain 0.07617 for Raw, 0.08068 for Regime-proxy, 0.07831 for VGGT-proxy, and 0.09521 for Omega-proxy. Raw is therefore the target-free source-selected fixed deployment action.

This ordering does not match the target-domain means: Regime- and VGGT-derived corrections become stronger fixed alternatives on both target benchmarks. Thus, domain shift affects not only local routing but also the ranking of candidate actions at the global level.

Matched scene-level control. To isolate decision granularity without changing the action pool or learned router, we aggregate the frozen group-level policy to one action per scene. For scene s,

$$\hat{a}_s = \arg\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{G}_s} w_i \log \pi_\phi(a \mid f(g_i)), \tag{32}$$

where w_i is the normalized support weight of group i. The resulting action is applied to all eligible groups. Proxy checkpoints, group construction, router parameters, features, and stitching remain unchanged. No target-domain ground truth is used to construct this decision.

On ETH3D, RAGR$-$Scene is -0.00987 with a 95% scene-bootstrap interval of $[-0.03704, +0.00985]$. The mean favors group-wise selection but the interval includes zero. On TNT, RAGR$-$Scene is $+0.00400$ with interval $[+0.00016, +0.00961]$, establishing an advantage for the matched scene-level control. These results do not establish a stable independent benefit from finer group-wise decision granularity across domains.

E Detailed Local Reconstruction and Selection Diagnostics

E.1 Diagnostic Protocol

The analyses in this section use the same per-window median-ratio scale-aligned AbsRel employed to construct source-domain routing supervision. For each candidate action and supporting-view window, the prediction receives an independent median-ratio scale estimate on its valid pixels before AbsRel is computed. Group-level quantities are then aggregated using the same valid-support weighting as the local routing diagnostic.

These measurements are used only for post-hoc diagnosis. They do not affect proxy training, router supervision, checkpoint selection, or any target-domain deployment decision. Target-domain ground truth is accessed only after the frozen router has produced its predictions.

The ETH3D diagnostic contains 598 valid groups and the TNT diagnostic contains 336. These counts differ from the scene-level reconstruction comparison because the local diagnostic uses its own support filtering and aggregation. Consequently, local action rankings and errors need not coincide with the stitched scene-level metrics in Table 1.

E.2 Raw-Error Heterogeneity

We first quantify whether the shared Raw prediction exhibits heterogeneous error across spatial groups. For each valid group, we retain only the Raw error from the frozen local diagnostic matrix and analyze its distribution; no action identity or router decision is involved.

Table 9 shows a broad distribution of local Raw errors on both target domains. On ETH3D, the 90th-percentile group error is 0.726 compared with 0.203 at the 10th percentile. On TNT, the corresponding range is 0.042 to 0.824. The large TNT ratio is partly amplified by its small 10th-percentile error, so we report the underlying quantiles together with the ratio rather than using the ratio alone as evidence.

Table 9: Post-hoc distribution of Raw local reconstruction error. Errors are per-window median-ratio scale-aligned AbsRel and are separate from the unified stitched evaluation in Table 1.

Dataset	Groups	Mean ± std	p10	p50	p90	p90/p10
ETH3D	598	0.412 ± 0.258	0.203	0.338	0.726	$3.58\times$
TNT	336	0.366 ± 0.518	0.042	0.183	0.824	$19.64\times$

Table 10: Scene-bootstrap 95% confidence intervals for local selection diagnostics.

Quantity	ETH3D	TNT
Routing regret	[0.01624, 0.03623]	[0.02396, 0.17107]
Best-fixed-to-oracle opportunity	[0.00595, 0.01778]	[0.00246, 0.00479]
Router minus best fixed	[0.00770, 0.02101]	[0.02057, 0.16955]

The effect is also visible within individual scenes rather than arising only from differences in overall scene difficulty. For example, the within-scene p10/p50/p90 values are 0.323/0.614/1.382 for ETH3D `playground`, 0.232/0.493/1.028 for `relief`, and 0.273/0.575/1.033 for `relief_2`, corresponding to p90/p10 ratios of $4.28\times$, $4.43\times$, and $3.78\times$. On TNT, `Caterpillar` gives 0.073/0.229/0.872 ($11.97\times$), `Courthouse` gives 0.030/0.114/0.528 ($17.75\times$), and `Ignatius` gives 0.023/0.256/1.450 ($62.27\times$).

These statistics establish substantial variation in the local error of the shared prediction. They do not by themselves establish that different correction actions can exploit that variation. The oracle analysis below separately tests whether the action pool exhibits regional complementarity.

E.3 Action Complementarity, Margin, and Wrong-Selection Damage

For each group, we evaluate all four actions under the local diagnostic and identify the lowest-error action. The difference between the best and second-best action measures how strongly the locally preferred action is separated from its nearest alternative. Nearly half of the groups have a best-versus-second-best error margin below 0.005 on both target domains, indicating that many local decisions involve only small differences between the leading actions.

Despite these small margins, the group oracle improves over the strongest fixed action on both datasets, as reported in Table 2. The resulting oracle gains are 0.01431 on ETH3D and 0.00343 on TNT. This establishes measurable regional action complementarity beyond the heterogeneity of Raw error alone.

We additionally measure wrong-selection damage as the increase in local error caused by selecting an action different from the post-hoc oracle. TNT exhibits a substantially heavier wrong-selection damage distribution, consistent with its larger router–oracle gap. Thus, TNT combines relatively limited oracle headroom with greater cost when the selector chooses an unsuitable action.

E.4 Scene-Bootstrap Uncertainty

The pooled point estimates for local opportunity and routing gap are reported in Table 2. To assess their stability across scenes, we bootstrap complete scenes rather than individual groups.

The best-fixed-to-oracle interval excludes zero on both domains, supporting the presence of local action complementarity across scenes. At the same time, the router-minus-best-fixed intervals are positive on both domains under this local diagnostic. The source-trained router therefore does not recover enough of the available local opportunity to outperform the strongest fixed action in this post-hoc metric.

Table 11: Post-hoc oracle-selected stitching under the unified scene-level evaluation protocol. Local-oracle-selected stitching replays the original per-window diagnostic winners, while shared-scene-scale-label stitching uses post-hoc action labels constructed under the scene-level scale convention. Both variants use target ground truth only for diagnostic label construction and are not deployable policies.

	ETH3D		TNT	
Method	AbsRel ↓	RMSE ↓	AbsRel ↓	RMSE ↓
Raw	0.43949	2.66911	0.23708	1.42487
RAGR	0.42681	2.54528	0.23794	1.42440
Local-oracle-selected stitching	0.41828	2.46344	0.23344	1.42296
Shared-scene-scale-label stitching	0.40382	2.35003	0.23237	1.42276

E.5 Local-Oracle-Selected Stitching

The local oracle in Table 2 is defined under the per-window median-ratio scale-aligned AbsRel diagnostic used for router supervision. Because its action rankings are obtained before regional write-back and scene-level evaluation, the corresponding oracle gain need not translate directly to the final stitched reconstruction. We therefore perform an additional post-hoc diagnostic that replays the local oracle choices through the actual deployment pipeline without retraining any model.

Using the frozen seed-0 action pool and cached predictions, we assign each runtime group its existing locally preferred action, write the resulting regional predictions back to the supporting views, uniformly average overlapping predictions, and evaluate the fully stitched reconstruction under the same comparison support and scene-level scale convention as Table 1. Groups without a valid local diagnostic label fall back to Raw; this occurs for seven runtime groups in the reported evaluation (one on ETH3D and six on TNT). The resulting reconstructions retain at least 99.98% of the unified comparison support in every scene. Target-domain ground truth is used only to construct the post-hoc action labels and does not alter the frozen predictions, cameras, or stitching procedure.

Local-oracle-selected stitching improves over the deployed RAGR after the regional predictions pass through the actual stitching pipeline. The paired scene-level AbsRel difference is -0.00853 on ETH3D, with a 95% bootstrap confidence interval of $[-0.01755, -0.00188]$, and -0.00450 on TNT, with interval $[-0.01079, -0.00045]$. Thus, part of the local action-selection opportunity remains realizable at the final scene level. On TNT, however, the corresponding difference relative to Raw has a 95% interval of $[-0.00983, +0.00029]$ and therefore does not establish a stable improvement over Raw.

We additionally examine the effect of the scale convention used to construct post-hoc action labels. The original local-scale labels and the shared-scene-scale labels select the same action for only 255/598 ETH3D groups (42.6%) and 144/336 TNT groups (42.9%). The latter labels produce stitched AbsRel values of 0.40382 on ETH3D and 0.23237 on TNT. This limited agreement shows that the scale and evaluation conventions used to define the selection objective can substantially alter action rankings, providing additional evidence that local supervision and final scene-level evaluation are not fully aligned.

Neither diagnostic should be interpreted as a scene-optimal oracle or a strict upper bound on achievable reconstruction quality. Overlapping groups interact through write-back and averaging, AbsRel is nonlinear, and the final scene-level scale depends on the complete action assignment. For the same reason, the opportunity–regret decomposition in Table 2 applies only within the common local scoring matrix used by that diagnostic and should not be numerically mixed with the scene-level metrics in Table 1.

E.6 Action-Level Selection Diagnosis

Table 12 compares the router's selected action distribution with the post-hoc oracle action distribution. Precision and recall treat oracle action identity as a diagnostic target; wrong-action damage is averaged only over incorrect selections of the corresponding action.

Table 12: Action-level router/oracle diagnosis. Precision and recall use post-hoc oracle action identity as the diagnostic target. Wrong-action damage is averaged only over incorrect selections of that action.

Domain	Action	Router share	Oracle share	Precision	Recall	Wrong-action damage
ETH3D	Raw	25.6%	17.4%	20.9%	30.8%	0.03142
	Regime	29.8%	21.1%	18.0%	25.4%	0.05612
	VGGT	21.9%	45.8%	52.7%	25.2%	0.02705
	Omega	22.7%	15.7%	19.1%	27.7%	0.05881
TNT	Raw	31.0%	14.0%	9.6%	21.3%	0.20892
	Regime	28.0%	32.1%	33.0%	28.7%	0.01690
	VGGT	5.1%	50.3%	29.4%	3.0%	0.00292
	Omega	36.0%	3.6%	2.5%	25.0%	0.14758

Table 13: Teacher-to-proxy action fidelity on held-out source data.

Split	Winner agreement ↑	Proxy headroom ↑
7-Scenes	46.2%	.00589
ARKitScenes	82.8%	.00043
Combined	53.3%	.00498

The largest action-distribution mismatch occurs on TNT. The oracle chooses the VGGT-derived correction for 50.3% of groups, whereas the router chooses it for only 5.1%. Conversely, the router chooses the Omega-derived action for 36.0% of groups although it is oracle-optimal for only 3.6%. The overall router–oracle agreement is 26.6% on ETH3D and 14.6% on TNT.

The action-level damage statistics further show that selection frequency alone does not determine the impact of an error. For example, incorrect Raw and Omega selections on TNT incur substantially larger average damage than incorrect VGGT selections. This asymmetry helps explain why modest action-classification errors can translate into a much larger routing gap on TNT.

Taken together, the local diagnostics separate three properties of regional refinement. The Raw-error distribution establishes that the shared reconstruction has spatially varying residual error. The group oracle establishes that the available correction actions possess measurable regional complementarity. The deployed-router results show that these preferences are not recovered reliably enough after domain shift.

F Teacher-to-Proxy Transfer

For each source-held-out group, we separately evaluate the action preferred by the frozen teacher pool and by the deployed proxy pool. Winner agreement measures whether the locally minimum-error action identity is preserved after transfer. Proxy headroom measures the reduction from the strongest fixed proxy to the post-hoc proxy group oracle.

High action-identity agreement does not necessarily imply substantial multi-action headroom. In particular, ARKitScenes reaches 82.8% agreement while offering only 0.00043 additional local opportunity. The archived fidelity matrices contain scalar group errors rather than per-pixel residual vectors, so we use them only to analyze action identity and headroom and do not infer residual-vector correlation.

G Routing-Feature Ablation

The overlap-feature experiment retrains only the router. The proxy pool, source split, router capacity, and action-supervision protocol remain fixed.

Across three seeds, the full descriptor obtains 0.07723 ± 0.00063 AbsRel on source-held-out data and 0.33968 ± 0.00206 on the ETH3D local diagnostic. Removing r_i gives 0.07716 ± 0.00029 and

Table 14: Per-seed ETH3D ablation of r_i. Difference is (without r_i) − (full 5D).

Seed	Full 5D	Without r_i	Difference	95% CI
0	.33807	.34172	+.00365	[-.00205,+.01110]
1	.33896	.33824	-.00073	[-.00436,+.00447]
2	.34199	.34211	+.00011	[-.00597,+.01329]

0.34069 ± 0.00213, respectively. The target-domain direction is inconsistent across seeds and every paired scene-bootstrap interval contains zero. We therefore find no evidence of a stable independent benefit from r_i.

H Router-Variant Audit

The main experiments use a flat four-action router with the compact five-dimensional descriptor. We additionally audited source-frozen alternatives spanning decision structure, supervision targets, feature representations, model capacity, training-free selection, and sparse mixtures. These experiments were performed in several matched proxy pools and diagnostic subsets. Absolute errors from different configurations are therefore not directly comparable; each result should be interpreted only relative to the fixed reference from its own matched setting. Target-domain ground truth was used only after inference for diagnosis and was never used to choose the reported router.

Within the main flat-router family, the representative ETH3D diagnostic obtains 0.33674 compared with 0.32577 for its strongest matched fixed action. Across 12 matched runs, the router-minus-fixed gap is $+0.01097 \pm 0.00240$. An alternate matched proxy/router pool yields 0.34940 for the router and 0.34616 for the corresponding fixed action; six of twelve runs show small router improvements, but the direction is not stable. Increasing hidden width from 32 to 128 or 512 likewise does not establish a consistent target-domain advantage.

Changing the supervision or decision structure does not resolve the gap. A hierarchical Raw/refine decision followed by three-way proxy selection fits the source routing task but does not establish improved cross-domain stability; a representative TNT diagnostic reaches 0.45594. Hard-label margins $\delta \in \{0, 0.02, 0.05\}$ do not produce a stable target improvement. Soft-label KL supervision with $\tau \in \{0.002, 0.005, 0.01\}$ yields representative TNT errors between 0.4044 and 0.4268 across seeds without a stable advantage over the matched fixed reference. Direct regression of all four action errors gives representative ETH3D/TNT errors of 0.3159/0.4306 and likewise does not establish a matched fixed-action advantage.

Richer representations also fail to provide a consistent solution. A 9D statistic descriptor produces representative ETH3D/TNT errors of approximately 0.33/0.55. Token-based routers using pooled shared-representation features improve source-held-out routing, reaching 0.07465 in the best tested configuration, but obtain ETH3D errors of 0.34235 and 0.34119 compared with 0.33486 for the matched fixed action. Source-selected feature gating similarly gives 0.33530 and 0.33546 versus 0.33486 for the same matched fixed reference.

Finally, a training-free selector based on reprojection consistency obtains 0.07947 on the source diagnostic compared with 0.07617 for Raw and reaches 0.33668 on ETH3D versus 0.32934 for the matched fixed VGGT-derived action. Sparse Top-2 mixtures change local error by only 0.00075 and 0.00130 in the two evaluated settings while adding 0.754 and 0.647 proxy calls per group.

These variants do not exhaust the space of possible selectors. They instead show that the observed cross-domain selection gap persists across several straightforward changes in capacity, supervision, features, and decision structure and is unlikely to be explained by the default MLP alone.

I Deployment Profiling

We profile global prediction and regional post-processing separately. The global stage contains the FastVGGT forward pass. Starting from cached global outputs, the RAGR post-stage includes group

Table 15: Representative deployment latency. Values are p50/p95 seconds after one warm-up. Fixed-V denotes the representative fixed VGGT-derived execution.

Scene	Stage	p50 / p95 (s)
ETH3D courtyard	Global prediction	1.618 / 1.753
	Fixed-V post-stage	6.883 / 7.020
	RAGR post-stage	27.461 / 85.204
TNT Barn	Global prediction	5.303 / 5.792
	Fixed-V post-stage	19.850 / 20.384
	RAGR post-stage	41.705 / 180.944

Table 16: Stage-level p50 profiling in seconds. Component medians are measured independently and need not sum exactly to the overall post-stage median.

Scene	Views	Global	Group	Feature	Route	Proxy	Stitch
courtyard	24	1.618	4.764	2.986	19.497	.0379	.0331
Barn	100	5.303	12.483	9.601	19.166	.0409	.0767

construction, routing-feature extraction, action selection, execution of the selected residual proxies, and regional stitching.

Each configuration is run once for warm-up followed by five timed repetitions. We report median (p50) and observed 95th-percentile (p95) latency. These measurements are intended as stage-level diagnostics rather than a hardware-normalized end-to-end efficiency benchmark.

The selected residual proxies themselves account for only a small fraction of the adaptive post-stage latency. Their p50 execution time is 0.0379 s in courtyard and 0.0409 s in Barn. The frozen group-wise policy invokes a residual proxy for 29 of 40 groups in courtyard and 31 of 58 groups in Barn.

Group construction, feature extraction, and the routing path dominate the measured post-prediction cost; the routing path alone takes 19.497 s in courtyard and 19.166 s in Barn. Sparse proxy invocation therefore reduces residual evaluations but does not currently translate into low post-processing latency. We do not claim negligible deployment overhead for RAGR.

J Qualitative Evaluation Protocol

The examples in Fig. 1 are selected from the complete ETH3D evaluation using a frozen scene-level rule. *Meadow* is selected as a scene where regional refinement improves the Raw reconstruction, while *pipes* provides a matched case where the routed result does not recover the strongest available correction.

For both scenes, view 0 and the same center-crop rule are fixed before inspection of local crop errors. No local target-domain ground-truth crop metric is used to choose the displayed region. The routing overlay visualizes the frozen decisions and is not treated as additional quantitative evidence.

K Reproducibility Details

All headline comparisons use one frozen experimental stack. The shared predictor is a fixed FastVGGT implementation, each deployed residual proxy contains 741,869 trainable parameters, and the deployed router contains 17,796 trainable parameters.

The accompanying reproducibility artifacts provide the exact source-sequence split, checkpoint identifiers, cryptographic hashes for the frozen experimental components, software versions, teacher/backend identifiers, and evaluation manifests used to produce the reported results.