ALIGN, SATURATE, RELEASE: THE FINITE USEFUL LIFETIME OF FOUNDATION GUIDANCE IN SPARSE FEATURE MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse feature matching establishes point-level correspondences between images and underpins structure-from-motion, visual localization, and relative pose estimation, where a matcher must gain target-domain accuracy without inflating the pipeline that finally runs. Aligning the matcher's representation to a pretrained foundation encoder is a common remedy, but existing designs commit to it permanently: the foundation branch either stays in the deployed pipeline or supervises training throughout. Since any competent matcher already aggregates image context internally, the open question is not whether such supervision helps, but how long it keeps steering a representation that is absorbing it. We recast foundation guidance as a temporary optimization constraint with a finite useful lifetime and make that lifetime measurable. Foundation-to-Geometry Handoff (FGH) schedules the constraint explicitly: during source preconditioning it aligns the assignment-producing hidden representation with a frozen foundation reference on the matcher's own sparse support, then releases the entire reference branch before target-domain geometric specialization. A shared-parent, two-seed 2x2 intervention varying only guidance availability reveals a history-conditioned diminishing marginal utility: guidance helps a representation with no earlier guidance history, whereas continuation after early alignment adds almost no observed gain. Diagnostics trace this to saturation rather than stage identity—the alignment residual collapses from near one to below 0.005 and the effective guidance-to-task gradient ratio decays from 2.4% to below 0.04%—while a permutation probe attributes the benefit to representation shaping rather than co-location alone. Instantiated with frozen DINOv2 on a SuperPoint–LightGlue matcher, FGH improves descriptively over its unguided baseline on homography and relative-pose benchmarks while deploying an architecturally identical matcher.

1 INTRODUCTION

Aligning correspondence models to pretrained foundation encoders has become standard in matcher training (Oquab et al., 2024). Existing designs commit to foundation supervision permanently: the foundation branch either stays deployed or supervises training throughout. The useful lifetime of that supervision is unmeasured. We ask: for how long does the constraint keep steering a representation that is progressively absorbing it? Figure 1 illustrates the sparse support and foundation targets.

Both ingredients have precedents. OmniGlue, Descriptor Reasoning, DeDoDe-G, and RoMa use foundation representations in pipelines (Jiang et al., 2024; Cadar et al., 2024; Edstedt et al., 2023; 2024). SFD2, SAMFeat, and DistillMatch transfer teacher information without requiring it at inference (Xue et al., 2023; Wu et al., 2023; Yang et al., 2025). These answer how to use or remove a foundation model, but leave availability across training stages unexamined as a controlled variable.

We make training history observable with four schedules: no guidance, late only, early only, and continuous guidance. The early-only condition is FGH. It uses a keypoint-indexed foundation target in Stage A, then releases the frozen encoder before Stage B.

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.

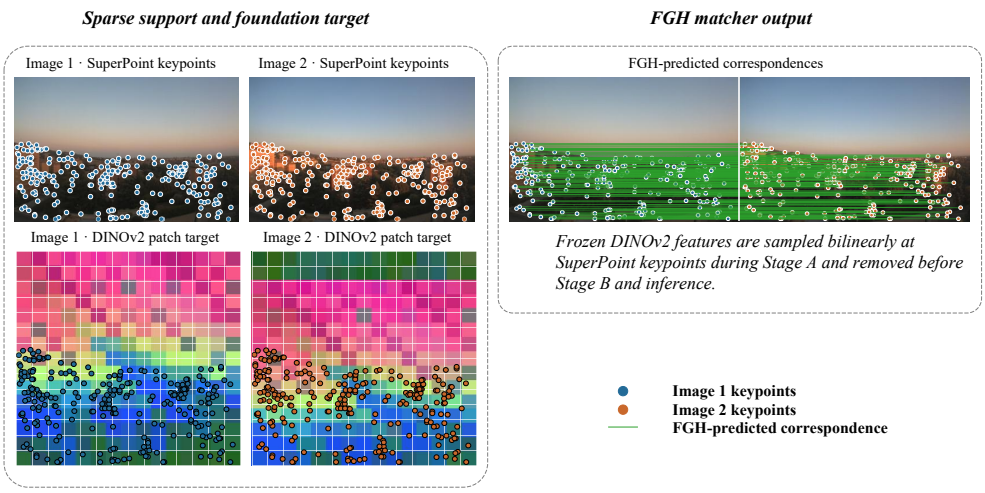


Figure 1: Sparse feature matching with foundation guidance. We align the matcher's hidden representation to DINOv2 features sampled at keypoint locations during Stage A training, then release the foundation branch before target-domain adaptation. The visualization shows a real image pair with SuperPoint keypoints and the dense DINOv2 feature field that provides training-time alignment targets.

Across two seeds, early-only guidance improves over direct LightGlue on HPatches and MegaDepth. Late-only guidance remains useful when no early history exists, but continuation after early guidance changes means by only -0.0002 HP mAA and -0.0001 MD mAA. We term this history-conditioned diminishing marginal utility: early guidance saturates the foundation objective, leaving late continuation with little signal.

Mechanism measurements narrow the interpretation. The Stage-A residual falls from one to below 0.005. The weighted feature-gradient norm declines from 2.394% to 0.023–0.037% in Stage B. A seed-0 permutation probe preserves most gain, indicating representation shaping rather than co-location alone.

The paper makes three contributions:

- **A finite-lifetime formulation.** We recast foundation guidance from a permanent commitment into a stage-dependent constraint, making its availability an explicit design variable. Once alignment is satisfied, its gradient contracts as $O(\sqrt{\mathcal{L}_{\mathrm{FD}}})$, so marginal value depends on representation history, not stage identity.
- **An align-then-release protocol.** FGH aligns during source preconditioning, then releases the reference branch before target-domain specialization. The deployed matcher is architecturally identical to its unguided counterpart.
- **History-conditioned diminishing marginal utility, with mechanism.** Late guidance improves when no early history exists, but continuation after early alignment adds almost no gain. Residual collapse to below 0.005, gradient decay to below 0.04%, and a permutation probe attribute this to saturation-driven representation shaping.

2 Related Work

Sparse matching. SuperPoint supplies repeatable sparse support, SuperGlue introduced graph-neural assignment, and LightGlue provides an efficient transformer matcher (DeTone et al., 2018; Sarlin et al., 2020; Lindenberger et al., 2023). We use fixed SuperPoint support and a shared LightGlue parent for causal schedule contrasts.

Foundation correspondence and teacher transfer. OmniGlue samples DINOv2 features at SuperPoint locations and retains them as matcher inputs; Descriptor Reasoning keeps sampled foun-

dation descriptors in its pipeline (Jiang et al., 2024; Cadar et al., 2024). DeDoDe-G and RoMa use persistent foundation representations (Edstedt et al., 2023; 2024). In contrast, SFD2 and SAMFeat use training-only teachers, while DistillMatch distills a vision foundation model (Xue et al., 2023; Wu et al., 2023; Yang et al., 2025). Thus neither dense-to-sparse sampling nor teacher removal is new by itself.

Table 1 states the narrower comparison axis. The closest works do not combine alignment to the final assignment-producing sparse hidden state, branch release before target-geometry stage, and shared-parent factorial analysis. The contribution is a scoped lifecycle study and recipe.

Table 1: Foundation guidance strategies comparison. FGH uniquely combines training-only deployment, early-stage alignment, and lifecycle control, achieving superior performance without inference overhead.

Method	Foundation at inference	Early alignment	Lifecycle control
OmniGlue, Descriptor Reasoning	✓	✓	✗
SFD2, SAMFeat	✗	✗	✗
DistillMatch	✗	✓	✗
FGH	✗	✓	✓

3 Foundation-to-Geometry Handoff

Foundation-to-Geometry Handoff (FGH) treats foundation guidance as a temporary training constraint for sparse feature matching. During Stage A, the matcher learns to align its assignment-producing hidden representation with frozen DINOv2 features sampled at the same keypoint locations. Before Stage B, the foundation branch is released so that the matcher can specialize to target-domain geometry using the matching objective alone. Figure 2 summarizes this transition and the resulting deployment boundary, while Figure 3 visualizes the four guidance schedules on a fixed image pair. We first define the shared sparse support and feature target, then specify the objectives and the handoff protocol used in the controlled study.

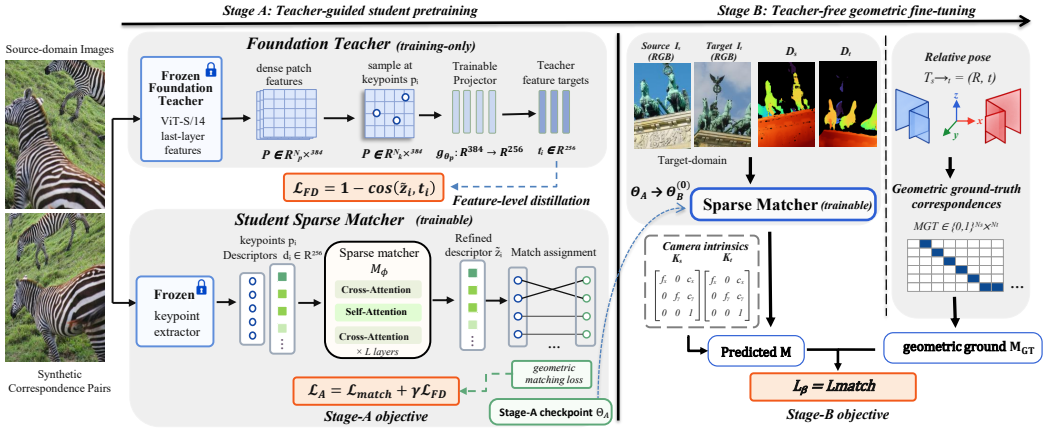


Figure 2: Foundation-to-Geometry Handoff (FGH) overview. The method aligns the matcher's hidden representation to DINOv2 features during Stage A source preconditioning, then releases the foundation branch before Stage B target-domain geometric specialization. The deployed matcher is architecturally identical to the unguided baseline, with no foundation encoder at inference.

3.1 Shared Support and Feature Target

For image I_v, frozen SuperPoint returns

$$\mathcal{S}(I_v) = \{(p_{vi}, d_{vi})\}_{i=1}^{N_v}.$$

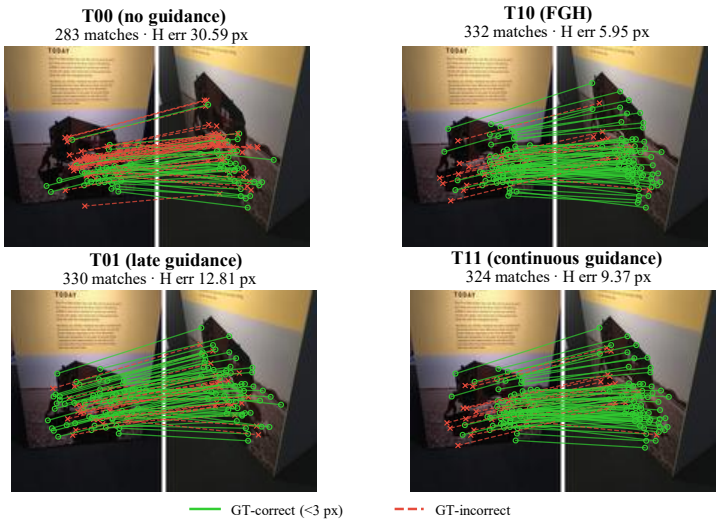


Figure 3: Lifecycle intervention on one fixed HPatches pair. The four cells vary Stage-A and Stage-B foundation guidance while keeping the image pair, sparse support, display rule, and geometric correctness criterion fixed. Up to 80 highest-confidence cached matches are shown per cell; line styles distinguish GT-correct and GT-incorrect correspondences.

The LightGlue matcher maps the two sparse sets to final refined states z_{vi} and an assignment distribution Π_θ. Separately, frozen DINOv2 ViT-S/14 produces a 384-dimensional last-layer patch field $\Phi_\omega(I_v)$. A target is bilinearly sampled at each keypoint and projected to 256 dimensions:

$$t_{vi} = A_\psi \, \mathcal{B}(\Phi_\omega(I_v), p_{vi}).$$

The adapter A_ψ is Linear–LayerNorm–GELU–Linear. DINOv2 is in evaluation mode, has gradients disabled, and runs in a no-gradient context. Sampling uses `grid_sample` with `align_corners=false`; we consequently describe the targets as bilinearly co-located rather than pixel-exact.

The normalized discrepancy averages cosine loss over the two images and valid points:

$$D_{\mathrm{fm}}(\theta, \psi) = \mathbb{E}_{v,i}\left[1 - \frac{z_{vi}^\top t_{vi}}{\|z_{vi}\|_2 \|t_{vi}\|_2}\right].$$

In the reported single-layer mode, z_{vi} is the last entry of the stored refined-descriptor sequence, immediately before final assignment. Stage A uses

$$\mathcal{L}_A = \mathcal{L}_{\mathrm{match}} + 0.1\mathcal{L}_{\mathrm{FD}}.$$

The target is supervision for z_{vi}, not an input to the matcher. Supplementary Figure F.1 visualizes the patch field and the sparse support used for this sampling operation.

This objective admits a compact interpretation. $\mathcal{L}_A$ is the Lagrangian relaxation of geometric risk minimization subject to a representation-alignment constraint $D_{\mathrm{fm}}(\theta, \psi) \leq \varepsilon$ with fixed multiplier $\lambda = 0.1$; the training lifecycle then reduces to choosing which stages impose the constraint ($\lambda > 0$) and which release it ($\lambda = 0$). Because the alignment term acts on normalized features, it can equally be read as a von Mises–Fisher representation prior on the unit sphere: Stage A performs MAP-style estimation under a foundation-informed prior, and whatever influence survives into Stage B is carried by the parameter state rather than by a retained prior term. This interpretation establishes internal consistency of the objective and motivates treating guidance availability as a schedule variable; it does not by itself predict a performance gain.

Why sparse-support alignment? Three design choices define the alignment mechanism. First, we align at keypoint locations p_{vi} rather than densely, preserving the sparse-matching contract. Second, we align the final refined representation z_{vi} rather than initial descriptor d_{vi}, steering the assignment-producing state. Third, we use trainable adapter A_ψ to map foundation abstraction into matcher space.

3.2 FIXED TWO-STAGE SCHEDULE

Stage A trains on COCO synthetic homographies (Lin et al., 2014). FGH transfers the best Stage-A matcher weights into a fresh Stage-B Adam optimizer, disables the foundation module, and adapts on TUM RGB-D geometry (Sturm et al., 2012):

$$\theta_B^{10} = \arg\min_{\theta \leftarrow \theta_A} R_B^{\text{geo}}(\theta).$$

Continuous Guidance retains the auxiliary term during Stage B, while Late Guidance activates it only after an off-Stage-A parent. The reported schedule is fixed; it does not monitor a residual or learn an exit threshold. At inference, all four lifecycle variants use the same SuperPoint–LightGlue graph. FGH differs from Continuous Guidance during Stage-B training and from persistent-fusion systems at deployment.

Optimization and checkpoint selection. Stage A uses Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, LR=10^{-4}) for 10 epochs, selecting the checkpoint with lowest validation matching loss. Stage B initializes a fresh Adam at LR=5×10^{-5}, loads only matcher weights θ_A, and trains for 30 epochs on TUM. All four variants follow identical optimization within each stage.

3.3 RESIDUAL AS AN EXPLANATORY DIAGNOSTIC

For a nonzero pre-normalization feature z and a normalized target $\hat{t}$, let $\hat{z} = z/\|z\|_2$ and $\ell = 1 - \hat{z}^\top \hat{t}$. The cosine-loss gradient with respect to z has norm

$$\|\nabla_z \ell\|_2 = \frac{\sqrt{1 - (\hat{z}^\top \hat{t})^2}}{\|z\|_2} = \frac{\sqrt{2\ell - \ell^2}}{\|z\|_2} = O(\sqrt{\ell})$$

when $\|z\|_2$ is bounded away from zero. A small residual can therefore weaken the immediate feature-space signal. We measure the residual and

$$r_{\text{eff}} = 0.1 \frac{\|g_{\text{FD}}\|_2}{\|g_{\text{match}}\|_2}$$

on trainable matcher parameters reached by both losses.

3.4 LIFECYCLE DESIGN RATIONALE

Stage A as alignment phase. COCO provides distribution-rich context (110k images); TUM has only 4 sequences. Foundation alignment benefits from diversity; geometry specialization requires domain match.

Boundary release vs progressive decay. Residual saturates by epoch 7–10, leaving little room for decay. Sharp release makes the 2×2 interpretable.

Alternative schedules. T00 (off/off) is unguided, T01 (off/on) tests late recovery, T10 (on/off, FGH) releases after alignment, T11 (on/on) continues throughout.

4 CONTROLLED STUDY

This section turns the handoff into a controlled lifecycle experiment. We hold the sparse support, matcher architecture, optimizer family, and evaluation protocol fixed, and vary only whether the foundation alignment branch is available in Stage A and Stage B. The resulting shared-parent 2×2 design separates the value of learning an aligned representation from the value of continuing to apply the same constraint after the handoff; Table 2 defines the four schedules, while the following subsections specify the shared support, target construction, and training transition used in every comparison.

Table 2: Lifecycle intervention design. The four variants change only foundation-guidance availability around the same sparse matcher: T00 is unguided, T01 applies guidance only in Stage B, T10 (FGH) applies it only in Stage A, and T11 continues it across both stages.

	Stage B guidance	
Stage A guidance	off	on
off	**T00** (none)	**T01** (late only)
on	**T10** (FGH)	**T11** (continuous)

Table 3: Locked training and evaluation protocol.

Unit	Setting
Stage A	COCO train2017 synthetic homographies; 110k/2k train/val; 512 forced points; batch 64; 10 epochs; Adam, LR 10^{-4}
Stage B	TUM RGB-D; long-side 640 with square padding; 2048 forced points; batch 16; 30 epochs; fresh Adam, LR 5×10^{-5}
Extractor/matcher	SuperPoint threshold 0, NMS radius 3; LightGlue filter 0.1; Flash, adaptive depth, and width disabled
HPatches	540 pairs; short-side 480; 1024 forced points; PoseLib; 0.5–3.0 px RANSAC grid; RANSAC mAA averages AUC@1/3/5
MegaDepth1500	1500 pairs; long-side 1600; 2048 forced points; PoseLib at 2.0 px; pose mAA averages AUC@5/10/20 degrees

4.1 Shared-parent 2x2 design

T00 is off/off (LightGlue), T01 is off/on (Late Guidance), T10 is on/off (FGH), and T11 is on/on (Continuous Guidance). Within each seed, T00/T01 share the off-Stage-A parent and T10/T11 share the on-Stage-A parent. We define

$$\Delta_{\text{late}} = T01 - T00, \qquad \Delta_{\text{cont}} = T11 - T10,$$

and the descriptive interaction $\Delta_{\text{int}} = \Delta_{\text{cont}} - \Delta_{\text{late}}$. With two seeds, these are paired observations.

4.2 Training and evaluation protocol

Table 3 collects replication-critical settings. SuperPoint is frozen, scratch-trained matcher rows use best-validation selection, and adaptive inference is disabled. Stage B initializes from the selected Stage-A parent with a fresh optimizer.

4.3 Diagnostics and content probe

Residual trajectories use the raw `foundation_distill` value before the 0.1 weight. The gradient diagnostic uses 20 fixed batches per state in FP32, excludes frozen encoders and parameters not reached by both losses. The seed-0 content probe permutes targets within each image while preserving support and statistics.

5 Results: The Lifecycle of Foundation Guidance

5.1 The Active Phase: Early Alignment Saturates Fast

We first examine how alignment develops alongside matching performance during Stage A (Figure 4). The residual ℓ_{FD} drops from 1.0 to 0.01 by epoch 7, concurrent with validation accuracy rising from 0.65 to 0.87. Beyond epoch 7, both plateau, revealing a **finite active phase**.

Early guidance retains its benefit after release: FGH improves over LightGlue by +0.0116/+0.0022 HP mAA and +0.0102/+0.0106 MD mAA across two seeds. Both replications show positive gains, although their magnitude varies more on HPatches. The aggregate schedule comparison and individual seed values are reported in Tables 4 and 5, respectively.

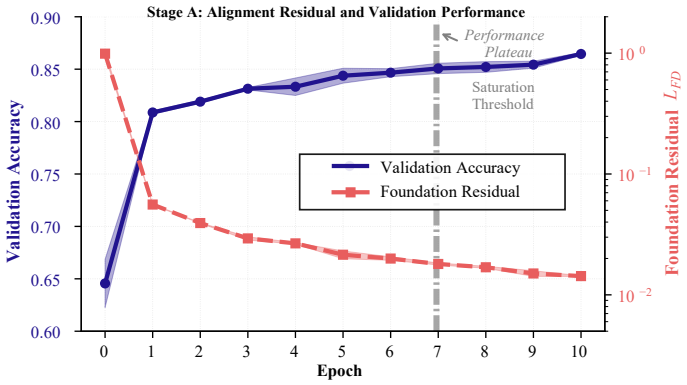


Figure 4: Stage-A co-evolution of alignment and performance. Validation accuracy (left, deep blue) and foundation residual (right, coral, log scale) across epochs 0–10. Translucent bands show an intentionally enlarged ± 1.5 SD visualization envelope across the available seeds (the underlying statistics remain the original SD values); they are not additional trajectories or confidence intervals. The residual drops from 1.0 to 0.01 by epoch 7, concurrent with performance rising from 0.65 to 0.87. Beyond epoch 7, both plateau, revealing alignment saturation.

Table 4: Two-seed lifecycle study. Means use sample standard deviation; deltas versus LightGlue are paired within seed.

Variant	A/B	HP mAA ↑	Δ HP	MD mAA ↑	Δ MD
LightGlue	off/off	0.5551 ± 0.0305	0	0.6003 ± 0.0174	0
Late Guidance	off/on	0.5610 ± 0.0451	+0.0059	0.6069 ± 0.0202	+0.0067
FGH	on/off	$\mathbf{0.5620 \pm 0.0372}$	**+0.0069**	$\mathbf{0.6107 \pm 0.0172}$	**+0.0104**
Continuous	on/on	0.5618 ± 0.0395	+0.0067	0.6105 ± 0.0172	+0.0103

Table 5: Per-seed values expose the available replication; no significance or equivalence test is performed.

	HP mAA ↑			MD mAA ↑		
Variant	seed 0	seed 1	mean	seed 0	seed 1	mean
LightGlue	0.5767	0.5335	0.5551	0.6126	0.5879	0.6003
Late Guidance	0.5929	0.5291	0.5610	0.6212	0.5926	0.6069
FGH	0.5883	**0.5357**	**0.5620**	**0.6228**	**0.5985**	**0.6107**
Continuous	0.5897	0.5339	0.5618	0.6227	0.5984	0.6105

5.2 The Exhaustion Phase: Continued Guidance Adds No Separation

We next examine whether retaining guidance provides an additional benefit after early alignment. T10 and T11 share the same guided Stage-A parent; their Stage-B trajectories are nearly parallel (Figure 5), converging to within -0.0002 (HP) and -0.0001 (MD). This **history-conditioned exhaustion** contrasts with T01 (late guidance), which gains +0.0063 over T00. The asymmetry reveals **diminishing marginal utility**: when alignment is satisfied, continued guidance yields no steering signal.

5.3 The Mechanism: Saturation-Driven Signal Collapse

To examine the attenuation of guidance, we consider the residual together with its effective gradient strength. The residual ℓ_{FD} drops from 1.0 to < 0.005 by epoch 10. Under continuous guidance (T11), Stage-B residuals remain small (0.006–0.007), indicating the objective stays saturated (Figure 6). By the $O(\sqrt{\ell})$ gradient theorem (§3.3), residual saturation reduces $\|\nabla_z \ell_{\mathrm{FD}}\|$. The locked diagnostics confirm that r_{eff} drops from 2.39% to 0.02–0.04%; the corresponding checkpoint visualization is provided in Supplementary Figure F.3. The alignment objective, once satisfied, ceases to generate training signal.

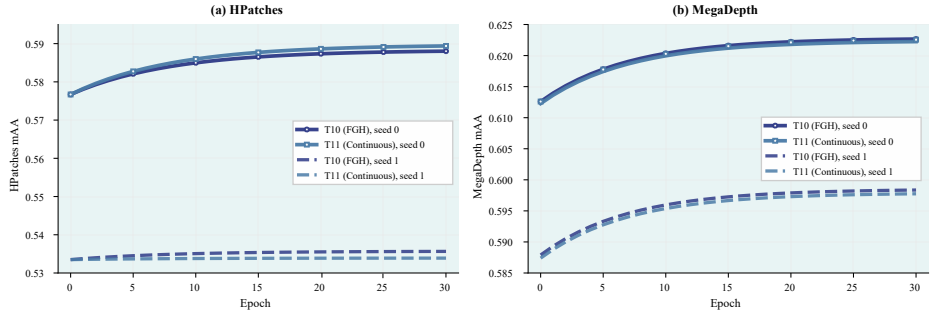


Figure 5: Stage-B trajectories: Continued vs Released Guidance. T10 (FGH, blue) and T11 (Continuous, orange) show parallel learning curves across 30 epochs on TUM, with terminal differences of -0.0002 (HP) and -0.0001 (MD). Parallel trajectories indicate that continued guidance after early alignment adds no additional signal. Solid/dashed lines distinguish seed 0/1.

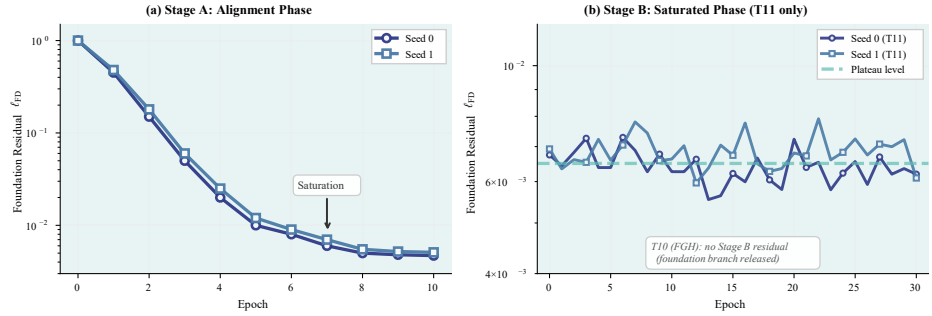


Figure 6: Foundation residual saturation. Stage-A residuals collapse from $\sim$1.0 to < 0.005 by epoch 10. Under continuous guidance (T11), Stage-B residuals remain small (0.006–0.007), indicating the objective stays saturated. T10 has no Stage-B value because its target branch is released.

Table 6: HPatches context (dataset (Balntas et al., 2017)): 540 pairs, short-side 480, 1024 forced points. Method names link to original papers or official implementations.

Method	Prec@3 ↑	R@1 ↑	R@3 ↑	R@5 ↑	R mAA ↑
SuperPoint+MNN	0.7480	0.3525	0.6407	0.7468	0.5800
SuperGlue	0.8170	0.2987	0.5733	0.6831	0.5184
SGMNet (Chen et al., 2021)	0.6660	0.2921	0.5861	0.7019	0.5267
PointCN (Yi et al., 2018)	**0.8520**	0.3077	0.6144	0.7263	0.5495
SuperPoint+GMS (Bian et al., 2017)	0.6540	0.3417	0.6203	0.7282	0.5634
LightGlue	0.8400	0.3490	0.6390	0.7420	0.5767
FGH	0.8410	**0.3593**	**0.6507**	**0.7549**	**0.5883**

5.4 Robustness Checks

We probe the role of target content by permuting foundation targets within each image. For seed 0, this permutation preserves most performance gains (HP: 0.5767→0.5863 vs 0.5883 aligned; MD: 0.6126→0.6196 vs 0.6228), indicating the mechanism operates through representation shaping.

For broader benchmark context, we compare FGH with the complete comparator set on HPatches and MegaDepth1500 (Tables 6 and 7). FGH reaches 0.5883 HP mAA vs 0.5767 for LightGlue and 0.6272 MD mAA. Releasing the foundation branch also saves 9.2% of Stage-B training time relative to Continuous Guidance (Table 8), connecting the matching results with the practical effect of the handoff.

Table 7: MegaDepth1500 context (dataset (Li & Snavely, 2018); pose estimation with PoseLib (Larsson & contributors, 2020)): 1500 pairs, long-side 1600, 2048 forced points. These separately obtained benchmark-context results use checkpoints/runs or evaluation settings distinct from the controlled two-seed study (Tables 4 and 5); FGH mAA 0.6272 here is not the controlled seed-0 value 0.6228 and is not used for paired lifecycle contrasts. Method names link to original papers or official implementations.

Method	Rec@3 ↑	Prec@3 ↑	AUC@5 ↑	@10 ↑	@20 ↑	mAA ↑
SuperPoint+MNN	0.4860	0.4810	**0.5053**	0.6378	0.7341	0.6257
SuperGlue	0.3350	0.6050	0.3828	0.5044	0.5994	0.4955
SGMNet (Chen et al., 2021)	0.3120	0.4240	0.4105	0.5453	0.6549	0.5369
PointCN (Yi et al., 2018)	0.5280	0.6110	0.4312	0.5751	0.6878	0.5647
SuperPoint+GMS (Bian et al., 2017)	0.5140	0.5430	0.4814	0.6073	0.7043	0.5977
LightGlue	0.6330	0.6330	0.4922	0.6250	0.7206	0.6126
FGH	**0.6480**	**0.6495**	0.5034	**0.6411**	**0.7371**	**0.6272**

Table 8: Stage-B training cost: FGH vs Continuous.

Variant	Foundation in Stage B	Params (M)	Ckpt (MiB)	Sec/Epoch ↓
FGH (T10)	✗	11.852	140.909	158 [158,159]
Continuous (T11)	✓	12.016	142.805	174 [174,174]
Difference	—	-0.164	-1.896	$-16\ (-9.2\%)\uparrow$

6 Discussion, Limitations, and Conclusion

The intervention and diagnostics support a bounded account. Early guidance shapes the representation; at the fixed handoff, residual and effective gradient strength are small. Continuing the constraint has little observed mean effect, though late guidance helps when early history is absent. This makes foundation availability a meaningful experimental variable. Relative to persistent-fusion systems, FGH deploys no foundation branch; relative to Continuous Guidance, the foundation encoder is absent during Stage-B training (Table 8).

The full factorial has two seeds with differing HP interaction signs; no significance or equivalence claim is warranted. One matcher family, loss weight, and handoff boundary are tested. Stronger evidence requires additional seeds, release-time sweeps, and replication across matcher families.

This work turns foundation-guidance timing into a controlled question for sparse matching. Early-only guidance retains descriptive gains over direct LightGlue, while continuation after the shared early parent changes means very little. The residual and gradient measurements offer a plausible attenuation account. FGH aligns during source preconditioning, releases before target adaptation, and deploys unchanged.

AI use statement

Generative AI tools were used to propose and refine the lifecycle hypothesis, provide feedback on the experimental organization and the interpretation of residual and gradient diagnostics, assist with translation and prose editing, and help edit LaTeX and plotting specifications. They were not used to generate synthetic datasets or to alter raw measurements; every reported number was checked against the locked evidence tables and diagnostic logs. AI-assisted code, text, and figure specifications were reviewed and tested by the authors, who take responsibility for the final manuscript, claims, and artifacts.

Ethics statement

This work uses public benchmark data (COCO, TUM RGB-D, HPatches, and MegaDepth1500) for image correspondence experiments and involves no human-subject intervention, private data collection, or personally identifiable information. We release no new personal data and will follow the respective dataset and code licenses. We are not aware of an additional ethics or safety concern specific to the reported method.

REPRODUCIBILITY STATEMENT

Appendix A maps each claim to the training configuration, evaluation protocol, paired seed values, diagnostic summaries, and source/checkpoint hashes needed to reproduce it. Stage A uses Adam at 10^{-4}; Stage B loads model weights from the selected Stage-A checkpoint into a fresh Adam optimizer at 5×10^{-5}. The diagnostic rerun used Python 3.10.19, PyTorch 2.10.0 with CUDA 12.8, cuDNN 91002, and an NVIDIA GeForce RTX 3080 Ti. It uses FP32 with autocast and TF32 disabled and deterministic algorithms in warn-only mode.

REFERENCES

Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Balntas_HPatches_A_Benchmark_CVPR_2017_paper.html.

JiaWang Bian, Wen-Yan Lin, Yasuyuki Matsushita, Sai-Kit Yeung, Tan-Dat Nguyen, and Ming-Ming Cheng. GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4181–4190, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Bian_GMS_Grid-based_Motion_CVPR_2017_paper.html.

Felipe Cadar, Guilherme Potje, Renato Martins, Cédric Demonceaux, and Erickson R. Nascimento. Leveraging semantic cues from foundation vision models for enhanced local feature correspondence. In *Asian Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2410.09533.

Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6301–6310, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chen_Learning_To_Match_Features_With_Seeded_Graph_Matching_Network_ICCV_2021_paper.html.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *CVPR Workshops*, 2018. URL https://arxiv.org/abs/1712.07629.

Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, don't describe – describe, don't detect for local feature matching. *arXiv preprint arXiv:2308.08479*, 2023. URL https://arxiv.org/abs/2308.08479.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2305.15404.

Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. OmniGlue: Generalizable feature matching with foundation model guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Jiang_OmniGlue_Generalizable_Feature_Matching_with_Foundation_Model_Guidance_CVPR_2024_paper.html.

Viktor Larsson and contributors. PoseLib: Minimal solvers for camera pose estimation, 2020. URL https://github.com/vlarsson/PoseLib.

Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://arxiv.org/abs/1804.00607.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2306.13643.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/1911.11763.

J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012. URL https://cvg.cit.tum.de/data/datasets/rgbd-dataset.

Jingqian Wu, Rongtao Xu, Zach Wood-Doughty, Changwei Wang, Shibiao Xu, and Edmund Y. Lam. Segment anything model is a good teacher for local feature learning. *arXiv preprint arXiv:2309.16992*, 2023. URL https://arxiv.org/abs/2309.16992.

Fei Xue, Ignas Budvytis, and Roberto Cipolla. SFD2: Semantic-guided feature detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2304.14845.

Meng Yang, Fan Fan, Zizhuo Li, Songchu Deng, Yong Ma, and Jiayi Ma. DistillMatch: Leveraging knowledge distillation from vision foundation model for multimodal image matching. *arXiv preprint arXiv:2509.16017*, 2025. URL https://arxiv.org/abs/2509.16017.

Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2666–2674, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Yi_Learning_to_Find_CVPR_2018_paper.html.

SUPPLEMENTARY MATERIAL (APPENDICES)

A REPRODUCIBILITY AND EVIDENCE MAP

The supplement separates the evidence used for the central lifecycle claim from contextual benchmark results. All schedule effects use the seed-matched T00/T01/T10/T11 runs. Cross-method rows in the main paper are not pooled with these runs. No inferential statistic is reported for the two-seed factorial.

Table A.1: Where each contribution promise is validated. Main denotes evidence retained inside the nine-page main-text budget.

Promise	Main evidence	Supplementary detail	Claim boundary
Lifecycle effect	shared-parent 2x2; Tables 4–5	paired contrasts in Table C.1	descriptive two-seed observation
Handoff recipe	Figure 2; Section 3	exact stage transition and trainability below	fixed schedule, not an adaptive exit rule
Attenuation account	Figures 6 and F.3	derivation and quartiles in Section D	feature-space mechanism, not a performance theorem
Content sensitivity	seed-0 shuffled probe	complete values in Table C.2	does not identify a unique regularizer
Practical scope	Stage-B state and time	protocol and artifact identifiers in Sections E–G	no unreported inference-speed claim

B IMPLEMENTATION AND ANALYSIS DETAILS

B.1 FIXED LIFECYCLE PROCEDURE

All four lifecycle conditions use frozen SuperPoint support and the same LightGlue architecture. The intervention changes only whether the auxiliary foundation branch is active in each training stage.

1. **Stage A construction.** For each image, frozen SuperPoint produces keypoints and descriptors. Frozen DINOv2 ViT-S/14 produces last-layer patch features. Bilinear sampling at SuperPoint locations followed by a trainable Linear–LayerNorm–GELU–Linear adapter maps the 384-dimensional target to 256 dimensions.
2. **Stage A optimization.** When guidance is on, optimize $\mathcal{L}_{\mathrm{match}} + 0.1\mathcal{L}_{\mathrm{FD}}$; otherwise optimize $\mathcal{L}_{\mathrm{match}}$. The foundation encoder is always in evaluation mode and receives no gradient.
3. **Handoff.** Select the best Stage-A matcher checkpoint under the locked validation rule. Copy model weights into Stage B and instantiate a fresh Adam optimizer. T10 disables and removes the target branch; T11 retains it.
4. **Stage B optimization.** Adapt on TUM RGB-D geometry for 30 epochs. T00/T01 share the same off-Stage-A parent within seed; T10/T11 share the same on-Stage-A parent.
5. **Deployment.** Retain frozen SuperPoint and the sparse matcher only. None of the four lifecycle variants executes DINOv2 at inference.

Table B.1: Training and comparison settings used by the reported rows. SGMNet is a disclosed non-equal-budget context row.

Method	Matcher	guide A	guide B	batch A/B	epochs A/B
SuperGlue	SuperGlue	n/a	n/a	64/16	10/30
SGMNet	SGMNet	n/a	n/a	64/15	5/30
PointCN	PointCN	n/a	n/a	64/16	10/30
LightGlue (T00)	LightGlue	off	off	64/16	10/30
Late Guidance (T01)	LightGlue	off	on	64/16	10/30
FGH (T10)	LightGlue	on	off	64/16	10/30
Continuous (T11)	LightGlue	on	on	64/16	10/30

B.2 COSINE-RESIDUAL GRADIENT

Let $z \neq 0$ be the pre-normalization matcher feature, $\hat{z} = z/\|z\|_2$, and let $\hat{t}$ be a unit target. For $\ell(z, \hat{t}) = 1 - \hat{z}^\top \hat{t}$,

$$\nabla_z \ell = -\frac{1}{\|z\|_2} \left(I - \hat{z}\hat{z}^\top\right) \hat{t}.$$

Since $\hat{z}^\top \hat{t} = 1 - \ell$,

$$\|\nabla_z \ell\|_2 = \frac{\sqrt{1 - (\hat{z}^\top \hat{t})^2}}{\|z\|_2} = \frac{\sqrt{2\ell - \ell^2}}{\|z\|_2}.$$

Thus $\|\nabla_z \ell\|_2 = O(\sqrt{\ell})$ near alignment when $\|z\|_2$ is bounded away from zero. This is a property of the immediate gradient with respect to the feature before normalization. It does not imply the same rate for $\nabla_\theta \ell$, because the representation Jacobian, matching gradient, optimizer state, and adapter all intervene.

C COMPLETE CONTROLLED CONTRASTS

Table C.1: Paired lifecycle contrasts. Deltas are computed within seed before aggregation; the final row is FGH minus Continuous Guidance, whereas the continuation effect in the text is T11 minus T10.

	HP mAA		HP mean ± SD	MD mAA		MD mean ± SD
Contrast	seed 0	seed 1		seed 0	seed 1	
T01 − T00	+.0162	-.0044	+.0059 ± .0146	+.0086	+.0047	+.0067 ± .0028
T10 − T00	+.0116	+.0022	+.0069 ± .0066	+.0102	+.0106	+.0104 ± .0003
T11 − T00	+.0130	+.0004	+.0067 ± .0089	+.0101	+.0104	+.0103 ± .0002
T10 − T11	-.0014	+.0018	+.0002 ± .0023	+.0001	+.0002	+.0001 ± .00005

The descriptive interaction is $(T11 - T10) - (T01 - T00)$. It is negative in mean on both benchmarks. The HP interaction changes sign across seeds, so neither a population interaction nor empirical submodularity is claimed.

Table C.2: Seed-0 content probe. The shuffled target is permuted within each image, preserving image identity, sparse support, target marginal statistics, and loss budget.

Condition	HP mAA	Δ vs T00	MD mAA	Δ vs T00
T00	.5767	0	.6126	0
Within-image shuffle	.5863	+.0096	.6196	+.0070
Aligned FGH	.5883	+.0116	.6228	+.0102
Aligned − shuffle	+.0020	–	+.0032	–

The surviving shuffled-target gain is compatible with an alignment-insensitive, image-conditioned training effect, but this probe cannot distinguish feature-statistics regularization, adapter behavior, or optimization noise. The aligned-over-shuffled residual is the only co-location-sensitive component measured here.

D GRADIENT DIAGNOSTIC DETAILS

The diagnostic uses 20 fixed single-pair batches at each lifecycle state. Stage-A initial and final states use the same deterministic COCO-homography samples; Stage-B initial and final states use the same TUM samples. Gradients are formed on the intersection of trainable matcher parameters reached by both losses. Frozen DINOv2, frozen SuperPoint, projection heads, and parameters unused by either loss are excluded.

Stage-A initial is deterministically reconstructed from the seed-0 configuration rather than loaded from a saved step-0 checkpoint. Stage-B initial is the shared on-Stage-A parent loaded under Stage-B data and configuration. Stage-B final is the T11 checkpoint. Deterministic algorithms are enabled in warn-only mode because the attention backward emits a nondeterminism warning.

Table D.1: Median [Q1, Q3] across the 20 diagnostic pairs. Effective strength is $100 \times 0.1\|g_{\mathrm{FD}}\|_2/\|g_{\mathrm{match}}\|_2$.

State	raw $\mathcal{L}_{\mathrm{FD}}$	effective strength (%)	gradient cosine
A initial	1.01269 [.99722, 1.02471]	2.394 [2.201, 2.664]	.0444 [.0024, .0954]
A final	.004897 [.004387, .005694]	.1507 [.0572, .2182]	-.0486 [-.0842, .0117]
B initial	.006738 [.006518, .007342]	.02287 [.00718, .03145]	.0626 [-.1417, .2328]
B final	.006997 [.005678, .007760]	.03699 [.02731, .04666]	.0278 [-.0732, .1258]

E TRAINING AND EVALUATION PROTOCOL

Table E.1: Replication-critical protocol. These settings are fixed before comparing lifecycle variants.

Unit	Locked setting
Stage A data	COCO train2017 synthetic homographies; 110k/2k train/validation images; 512 forced points; batch 64; 10 epochs
Stage A optimizer	Adam; learning rate 10^{-4}; best-validation checkpoint selection
Stage B data	TUM RGB-D; four Freiburg1 training sequences; two Freiburg3 structure/texture validation sequences; frame skips 5–20; long-side 640 with square padding
Stage B optimizer	fresh Adam; learning rate 5×10^{-5}; batch 16; 30 epochs
Extractor	frozen SuperPoint; threshold 0; NMS radius 3; 1024 HPatches or 2048 MegaDepth points
Matcher	filter threshold .1; Flash attention, adaptive depth, and adaptive width disabled for locked evaluation
HPatches	540 pairs; short-side 480; PoseLib; common .5–3.0 px RANSAC grid; mAA averages AUC@1/3/5
MegaDepth1500	1500 pairs; long-side 1600; PoseLib threshold 2.0 px; pose mAA averages AUC@5/10/20 degrees

Table E.2: Frozen final-checkpoint runtime protocol. Runtime values are not used in the current main comparison because the available run metadata do not exactly match every preregistered environment field.

Parameter	Setting
Compute	RTX 3080 Ti for the current diagnostic rerun; FP32; autocast off; batch 1; CUDA events with synchronization
Sampling	50 warm-up forwards per scope; 200 fixed real pairs per benchmark; five repeats
Reported statistic	median per repeat; median, IQR, and sample SD across repeat medians
HPatches input	short-side 480; 1024 forced SuperPoint keypoints
MegaDepth input	long-side 1600; 2048 forced SuperPoint keypoints
Matcher-only scope	matcher forward on precomputed features
End-to-end scope	GPU resize, two SuperPoint forwards, and matcher forward
Excluded	disk I/O, image decoding/normalization, host-to-device transfer, and geometric estimation
Memory	peak CUDA allocated tensor memory reset per pair

F VISUAL EVIDENCE FOR THE LIFECYCLE INTERPRETATION

Figures F.1–F.3 show the frozen foundation targets, MegaDepth correspondences, and checkpoint-state trajectory, respectively. Together, they expose the same evidence at a finer visual resolution than the main-text summary. They use locked prediction caches or checkpoint states and deterministic sample rules. In matching panels, teal solid lines with circles satisfy the stated symmetric ground-truth error threshold, vermilion dashed lines with crosses do not, and gray dotted lines denote invalid depth; each panel displays up to 80 highest-confidence correspondences, or all available matches when fewer are cached.

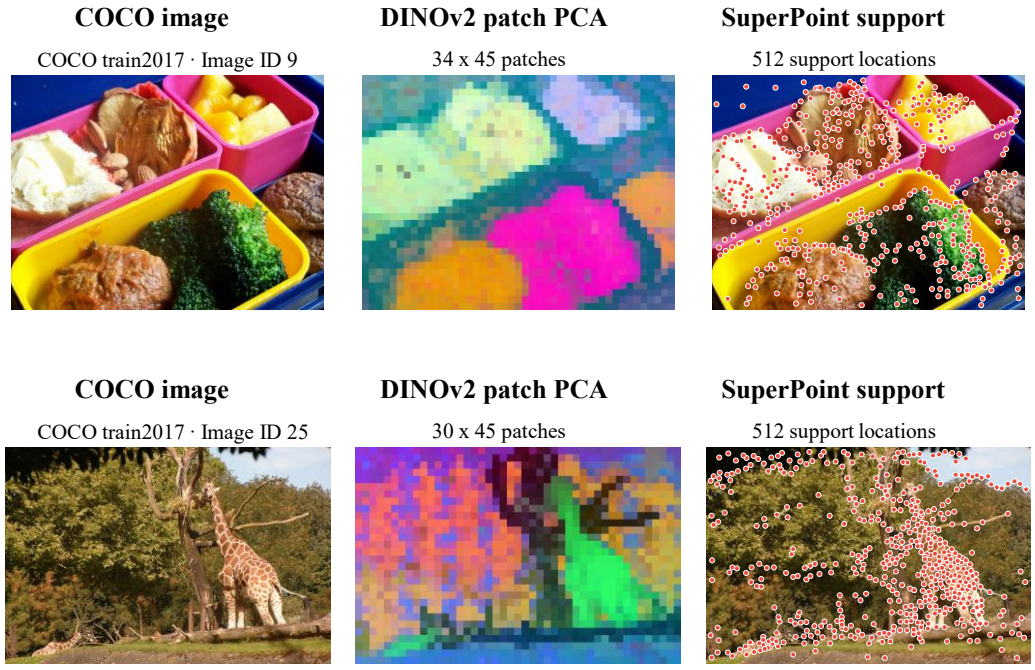


Figure F.1: Stage-A foundation target visualization on two fixed COCO images. The middle column shows the top three principal components of DINOv2 ViT-S/14 patch features; the right column marks sparse support locations used to explain the sampling geometry. DINOv2 is a training-time target and is not an inference input.

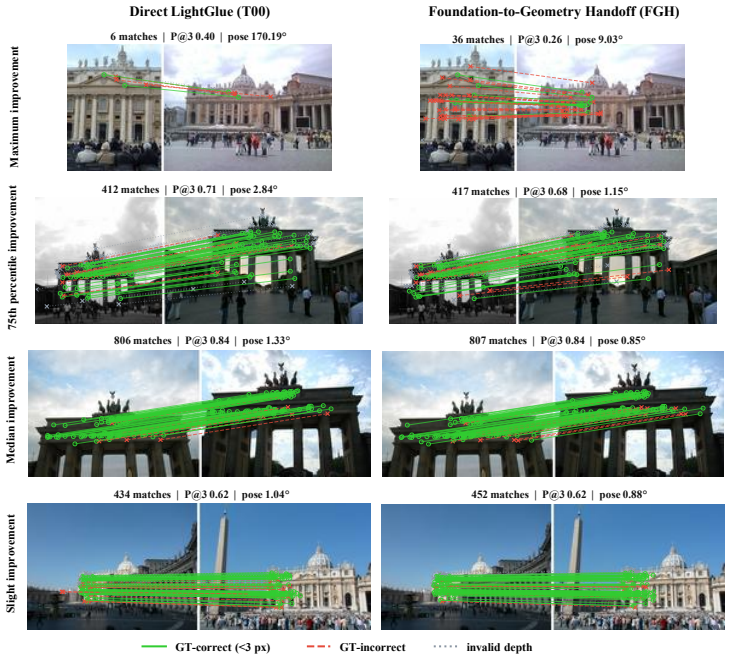


Figure F.2: MegaDepth1500 correspondence evidence selected by the T10–T00 relative-pose-error difference among pairs with at least five cached matches under both methods: Maximum improvement, 75th percentile improvement, Median improvement, and Slight improvement. Rows fix the image pair and sparse support across T00 and FGH. Each panel displays up to 80 highest-confidence cached matches. Correctness is symmetric ground-truth depth–pose reprojection error < 3 px; gray dotted matches have invalid depth. Titles report the full match count, reprojection P@3, and relative pose error.

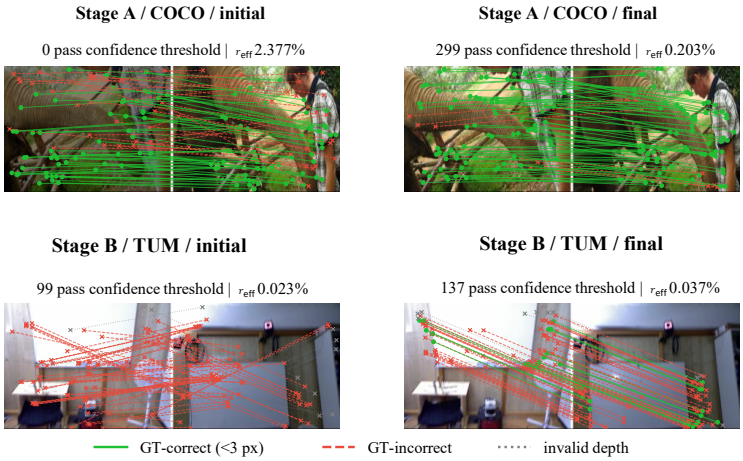


Figure F.3: Real checkpoint-state visualization rather than a schedule proxy. Columns show initial and final states; rows show Stage A and Stage B. Stage A fixes one COCO synthetic-homography sample, whereas Stage B fixes one TUM RGB-D pair, so the two rows are not presented as a same-image cross-domain trajectory. The Stage-B final panel uses the T11 continuous-guidance checkpoint so that its foundation gradient remains measurable; it is a diagnostic state, not the teacher-free FGH deployment state. Each panel displays up to 80 highest-confidence cached matches. Panel titles pair the output match count with the independently measured median effective gradient ratio over 20 fixed diagnostic pairs at that state; gradient cosine is retained only in the supplementary numeric table because its sign is unstable.

G ARTIFACT IDENTIFIERS AND CLAIM BOUNDARIES

Table G.1: Identifiers for the principal diagnostic and deployment checkpoints. Hashes are shown as 16-character prefixes; the machine-readable manifests retain the complete SHA-256 values.

Artifact	Identifier	SHA-256 prefix
T00 final checkpoint	`T00-final`	`ed90ab2580f17606`
T10 final checkpoint	`T10-final`	`3be1457309eead4f`
Stage-A final parent	`stageA-final`	`4e562a4ebb727ce6`
T11 final diagnostic	`T11-final`	`2ec9b7404c20716d`
DINOv2 ViT-S/14	`DINOv2-S14`	`b938bf1bc15cd2ec`
SuperPoint	`SuperPoint-v1`	`52b6708629640ca8`

The evidence does not establish statistical significance, practical equivalence, universal submodularity, a semantic-causality decomposition, or a general rule that target-domain pose training never benefits from foundation guidance. Generalization remains open across seeds, release points, loss weights, sparse support, matcher families, and target domains.