GRAFT: GROWING AGGLOMERATIVE FOUNDATION MODELS VIA CONTINUAL TEACHER DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision foundation models such as DINOv2, SigLIP2, and MASt3R develop complementary capabilities from different pretraining objectives, yet their knowledge remains distributed across separate, specialized models. Multi-teacher knowledge distillation offers a path toward consolidating these capabilities into a single agglomerative backbone, but existing approaches assume a fixed set of teachers, and incorporating a new teacher requires repeating expensive joint distillation over the entire teacher set. We introduce **GRAFT**, a continual multi-teacher distillation framework that enables a unified backbone to progressively acquire capabilities from an open-ended sequence of foundation models. When a new teacher arrives, GRAFT treats the previously distilled model as a teacher for preserving learned capabilities, while the current student jointly learns from both the previous model and the incoming teacher. Furthermore, to reconcile the incompatible representation geometries of heterogeneous teachers, we introduce Teacher Specific Readout Tokens, which grant each teacher an independent read-out of the shared encoder, together with Geometry Agnostic Relational Loss that aligns a vision–language teacher by matching image–text similarity structures rather than raw feature values. We provide GRAFT model, which is a single, continually extensible backbone that unifies five domains, including image understanding, 2D dense prediction, 3D human pose estimation, 3D vision, and vision–language, delivering strong performance across all of them while acquiring each new capability at the cost of a single distillation rather than a full re-distillation.

1 INTRODUCTION

Progress in visual representation learning has produced an increasingly diverse ecosystem of foundation models (Oquab et al., 2023; Radford et al., 2021; Tschannen et al., 2025; Kirillov et al., 2023; Leroy et al., 2024; Baradel et al., 2024). Rather than converging toward a single generalist representation, these models develop complementary capabilities from different pretraining objectives and data sources. DINOv2 provides strong generic visual features, SigLIP2 aligns visual representations with language, MASt3R captures 3D geometry, and specialized models such as SAM target segmentation. Deploying a separate encoder for every capability, however, quickly becomes impractical when memory, computation, or inference latency is constrained. Agglomerative models such as AM-RADIO (Ranzinger et al., 2024b), DUNE (Sarıyıldız et al., 2025), SAK (Lu et al., 2025) and Theia (Shang et al., 2024) demonstrate that representations distilled from heterogeneous foundation models can coexist within a single encoder, typically through teacher specific projection heads and mechanisms for balancing multiple distillation objectives (Ranzinger et al., 2024a).

However, what existing agglomerative models shares is that the teacher set is fixed. Once training is complete, incorporating a newly available foundation model generally requires performing joint distillation again over the enlarged teacher set. This assumption conflicts with how foundation models evolve in practice. A stronger teacher is released, or a task the system did not originally require becomes necessary, and the desirable capability is one that appears after the student was trained. Adding it to an existing agglomerative student is not incremental under current methods. The student must be re-distilled from scratch against the enlarged teacher set, a joint MTKD run that keeps every earlier teacher live. Reaching a target set of capabilities this way costs a full distillation run for every capability added, rather than one teacher's worth of work per step.

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.

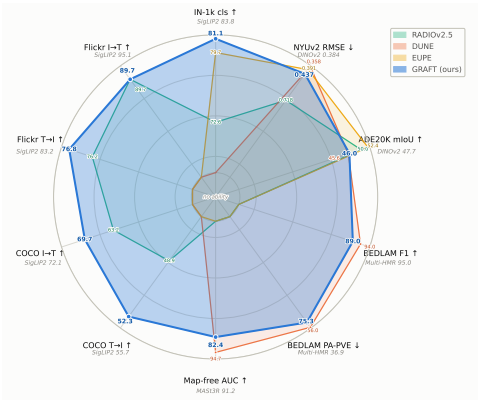


Figure 1: Compared to existing agglomerative models, GRAFT hold a comprehensive capability on various tasks with strong performance on each task.

We study this setting as **continual multi-teacher distillation**. Given a student that has already absorbed several foundation models, and an incoming teacher, our goal is to incorporate the new teacher without accessing any of the original teachers and without restarting distillation from scratch. Meeting this goal raises two challenges. The first is *catastrophic forgetting*: because the earlier teachers are discarded, any drift of the shared trunk toward the incoming teacher silently erodes the tasks the student already supports, and no original teacher remains to repair the damage. The second is *teacher heterogeneity*: foundation models trained with different objectives can organize their representation spaces in substantially different ways. Forcing all of them to match a common representation through the same pointwise feature objective can introduce conflicting supervision. As illustrated in Figure 2, this heterogeneity can emerge early in the network rather than only at the final representation layer.

We introduce **GRAFT**, a continual multi-teacher distillation framework that addresses forgetting through a key observation: the previous student already encodes the capabilities acquired from earlier teachers and can therefore serve as a compact surrogate for them during subsequent distillation. At each round, GRAFT initializes the student from the previously distilled model, freezes a copy of the distilled model as an anchor, and jointly distills from this previous model and the incoming teacher. The anchor model provides an anchor loss that discourages drift from previously acquired representations, while the incoming teacher provides supervision for the new capability. Each graft step therefore requires only two frozen reference models, the previous student and the incoming teacher, regardless of how many capabilities have already been accumulated. Historical knowledge is carried forward inside the student itself rather than through an ever growing buffer of past teachers.

GRAFT further addresses teacher heterogeneity at both the architectural and objective levels. Architecturally, we introduce Teacher Specific Readout Tokens, which provide each teacher with an independent global readout from the shared encoder. Each readout is paired with its own projector and normalization parameters, allowing heterogeneous teachers to extract different global summaries from a common trunk without forcing their representations into a single shared coordinate system. At the objective level, for vision-language teachers, we introduce a Geometry Agnostic Relational Loss that matches the image-text similarity structures rather than match teacher and student feature values directly. This allows the student to preserve the teacher's image text alignment while retaining a feature space compatible with previously acquired capabilities.

Together, we turn capability consolidation from a one time operation into a continual process. GRAFT progressively grows a single shared backbone by absorbing one foundation model at a time, requiring only a single incremental distillation stage for each incoming teacher. We evaluate the resulting model across five capability families spanning image understanding, 2D dense prediction, 3D human pose estimation, 3D vision, and vision language understanding. Compared with existing agglomerative models, GRAFT supports a comprehensive range of heterogeneous capabilities while maintaining strong performance across individual tasks, with the additional ability to continually incorporate new capabilities without full joint re-distillation.

Our contributions are summarized as follows:

- We formalize continual multi-teacher distillation for vision foundation models, a setting in which heterogeneous teachers arrive over time with no access to prior teachers.
- We introduce **GRAFT**, a continual distillation framework that addresses both forgetting and teacher heterogeneity. GRAFT uses the previously distilled student as a frozen retention teacher to preserve accumulated capabilities while learning from each incoming teacher. To accommodate heterogeneous representations within a shared encoder, GRAFT further introduces Teacher Specific Readout Tokens for independent teacher specific readouts and a Geometry Agnostic Relational Loss that transfers vision-language knowledge by matching image text similarity structures rather than raw feature values.

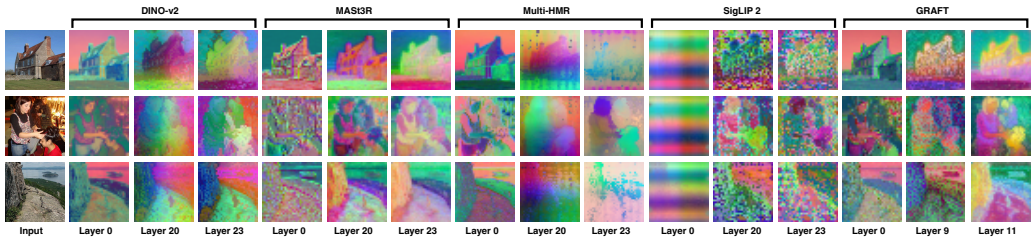


Figure 2: **Per-patch feature geometry across depth.** Each cell visualizes the top-3 PCA components of patch tokens (mapped to RGB, fit per cell) at a shallow, deep, and final block for four ViT-L teachers and GRAFT (ViT-B) on three ADE20K images. The teachers diverge from the shallowest blocks: DINOv2 resolves object structure, MASt3R shows geometric texture, SigLIP2 carries only positional patterns, and Multi-HMR progressively organizes around people. GRAFT unifies these in a single trunk—its patch tokens retain sharp object/part structure from its dense teachers (DINOv2, MASt3R, Multi-HMR), while SigLIP2's vision–language capability is grafted relationally onto a dedicated summary token (§3.3) and thus does not appear in this patch view.

- We demonstrate a distilled model spanning five capability families, including image understanding, 2D dense prediction, 3D human pose estimation, 3D vision, and vision language understanding, and show that GRAFT can acquire new capabilities while retaining previously distilled ones without full joint re-distillation over the accumulated teacher set.

2 Related Work

2.1 Vision foundation models

Vision foundation models acquire markedly different representational biases from their pretraining objectives and data. Self-supervised visual pretraining produces broadly transferable representations with strong spatial structure, supporting tasks such as semantic segmentation and monocular depth estimation (Oquab et al., 2023; Siméoni et al., 2025). Image–text contrastive pretraining instead emphasizes semantic alignment across modalities, enabling zero-shot recognition and cross-modal retrieval (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025; Fang et al., 2024). Other models specialize more explicitly: SAM learns promptable segmentation (Kirillov et al., 2023), DUSt3R and MASt3R target multi-view geometry and 3D reconstruction (Wang et al., 2024; Leroy et al., 2024), and MultiHMR targets multi-person human mesh recovery (Baradel et al., 2024). Despite their broad individual transferability, these models exhibit complementary strengths across task families rather than a single representation uniformly dominating across all capabilities. This complementarity motivates methods that consolidate multiple foundation models into a shared encoder.

2.2 Agglomerative models and multi-teacher distillation

Multi-teacher knowledge distillation provides a natural route for consolidating complementary foundation models into a shared backbone. AM-RADIO (Ranzinger et al., 2024b) introduced the agglomerative model paradigm, using teacher-specific heads to distill global and spatial representations from multiple teachers into a common ViT. Subsequent methods (Sarıyıldız et al., 2024; 2025; Lu et al., 2025; Ranzinger et al., 2024a; Shang et al., 2024; Heinrich et al., 2025; Chaybouti et al., 2025) expand the teacher set and improve how heterogeneous representations are reconciled. PHI-S (Ranzinger et al., 2024a) addresses mismatched teacher feature statistics through isotropic standardization. DUNE (Sarıyıldız et al., 2025) distills 2D, 3D, and human-perception teachers using teacher-specific transformer projectors. SAK (Lu et al., 2025) preserves teacher-specific biases through separate adaptation paths.

Multi-teacher distillation has also been explored beyond general-purpose agglomerative encoders. Related work integrates multiple visual teachers into vision–language models through lightweight adaptation modules (Cao et al., 2025; Wang et al., 2026a), distills foundation models using compact proxy datasets for greater data efficiency (Zhang et al., 2025; Lee et al., 2025; Frank & Davis, 2025), studies heterogeneous and task-agnostic transfer objectives (Li et al., 2025; Chen et al., 2025; Busbridge et al., 2025; Formont et al., 2026), and consolidates domain-specific teachers for med-

ical imaging, 3D perception, and robotics (Grashei et al., 2025; Zeng et al., 2025; Li et al., 2026; Wang et al., 2026b). These approaches broaden the scope of multi-teacher distillation, but existing agglomerative formulations predominantly optimize over a teacher collection available during the corresponding distillation procedure.

Despite these advances, existing methods are designed for a fixed teacher set. Adding a teacher after training therefore requires revisiting previous teachers or performing another joint consolidation stage. GRAFT instead treats the agglomerative backbone as continually extensible.

2.3 Continual learning

Preventing degradation of earlier capabilities is the central problem of continual learning. Regularization methods penalize movement of parameters important to previous tasks (Kirkpatrick et al., 2017); distillation-based methods constrain the model's outputs toward those of its former self (Li & Hoiem, 2017); replay methods retain or synthesize data from earlier tasks (Rebuffi et al., 2017). SE2D (Michel et al., 2026) study teachers trained on different domains and identify both the transfer of knowledge absent from the student's current data and its subsequent forgetting as central challenges. This establishes that sequential teacher distillation can be performed without retaining the full history of teacher models. While SE2D preserves predictive knowledge by matching logits from sequential teachers trained on different domains, GRAFT continually grows an agglomerative model whose teachers contribute fundamentally different representation-level capabilities, which span image understanding, dense prediction, 3D human pose, 3D geometry, and vision–language understanding, and therefore cannot generally be represented through a common logit space.

3 Method

3.1 Problem definition: Continual multi-teacher distillation

We formalize **continual multi-teacher knowledge distillation** (continual MTKD). Heterogeneous vision teachers $T_1, T_2, \ldots$ arrive one at a time, starting from an initial student S_1, an arbitrary entry point such as a randomly initialized network or an off-the-shelf pretrained checkpoint. When teacher T_k arrives, the learner must extend the previous student S_{k-1} into a student S_k that supports every task seen so far, $1, \ldots, k$, at the cost of distilling that single teacher.

At round k, the learner is given exactly three inputs:

- the prior student S_{k-1}, a shared-trunk encoder that already supports tasks $1, \ldots, k-1$;
- an image corpus $\mathcal{D}_k$;
- the incoming teacher T_k, accessible only as a frozen feature extractor.

The defining constraint is that the past teachers $T_1, \ldots, T_{k-1}$ are unavailable. Each student is evaluated by acquisition and retention. Acquisition measures how effectively the new task k is assimilated, and retention measures how robustly the earlier tasks $1, \ldots, k-1$ are preserved. In the following sections, we first articulate the inherent challenges of continual MTKD, and subsequently present our proposed framework, GRAFT, to overcome them.

3.2 The GRAFT pipeline

Continual multi-teacher knowledge distillation (MTKD) inherently suffers from catastrophic forgetting. As the model distills knowledge from a new teacher, the absence of supervision from discarded previous teachers causes the representation to drift toward the current task. We address this problem with GRAFT, a continual MTKD framework that uses the previously distilled student as a compact functional proxy for all past teachers. Because the previous student consolidates the functions acquired from earlier teachers, it can provide retention supervision.

Consider continual distillation over a sequence of teachers $T_1, \ldots, T_k$. At round k, GRAFT maintains only the incoming teacher T_k and the anchor model S_{k-1}, which is a frozen student obtained from round $k-1$. The active student S_k is warm started from S_{k-1}. S_k is then optimized to simultaneously acquire the new teacher while preserving the learned task functions. Given an input image

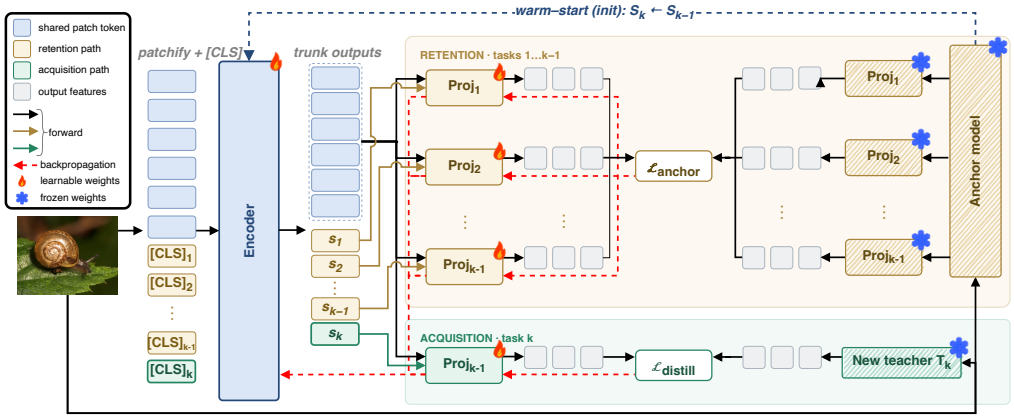


Figure 3: **Overview of GRAFT at round k.** The active student S_k is warm-started from S_{k-1}. A learnable Teacher Specific Readout Token (TSRT) is assigned to each task and jointly encoded with the shared patch tokens, producing task specific summary representations that are mapped by task-specific projectors. Training follows two concurrent paths. In the retention path, outputs for previously acquired tasks $1, \ldots, k-1$ are constrained by $\mathcal{L}_{\text{anchor}}$ to remain consistent with those of the frozen anchor model S_{k-1}, which preserves previously acquired task functions. In the acquisition path, the new task readout is optimized through $\mathcal{L}_{\text{distill}}$ to acquire knowledge from the incoming teacher T_k. Only the active student S_k is updated, while both S_{k-1} and T_k remain frozen.

$\mathbf{x}$, we denote the student output associated with task j at round k as

$$\mathbf{s}_j^{(k)} = \mathcal{H}_j^{(k)}\left(\mathcal{E}^{(k)}(\mathbf{x})\right), \tag{1}$$

where $\mathcal{E}$ denotes the shared visual backbone and $\mathcal{H}_j$ denotes the task specific projector associated with task j. This formulation allows different tasks to employ distinct readouts and distillation objectives while sharing the same visual backbone.

Learning is decomposed into two complementary objectives. The acquisition objective transfers knowledge from the incoming teacher T_k into the newly instantiated task projector:

$$\mathcal{L}_{\text{distill}}^{(k)} = d\left(\mathbf{s}_k^{(k)}, T_k(\mathbf{x})\right), \tag{2}$$

where $d(\cdot,\cdot)$ denotes the distillation objective. In parallel, the retention objective preserves the functions learned from all previous teachers. For every previous task $j < k$, the current student is constrained to reproduce the response of the corresponding readout in the frozen anchor model:

$$\mathcal{L}_{\text{anchor}}^{(k)} = \sum_{j=1}^{k-1} \lambda_j d\left(\mathbf{s}_j^{(k)}, \mathcal{H}_j^{(k-1)}\left(\mathcal{E}^{(k-1)}(\mathbf{x})\right)\right), \tag{3}$$

where $\mathcal{E}^{(k-1)}$ and $\mathcal{H}_j^{(k-1)}$ denote the backbone and projector for task j of anchor model S_{k-1}, and all parameters of S_{k-1} are frozen. This objective regularizes the student in function space by maintaining the outputs of all previously instantiated task readouts. The complete objective at continual round k is

$$\mathcal{L}^{(k)} = \underbrace{d\left(\mathbf{s}_k^{(k)}, T_k(\mathbf{x})\right)}_{\text{acquisition}} + \underbrace{\sum_{j=1}^{k-1} \lambda_j d\left(\mathbf{s}_j^{(k)}, \mathcal{H}_j^{(k-1)}\left(\mathcal{E}^{(k-1)}(\mathbf{x})\right)\right)}_{\text{retention}}, \tag{4}$$

where λ controls the strength of historical preservation. The first term expands the student's capability toward the incoming teacher, while the second explicitly constrains previously learned task functions from drifting during this update. GRAFT therefore casts continual MTKD as an additive learning objective in which acquisition of new knowledge is regularized by preservation of previously consolidated functions.

A key property of this formulation is that historical supervision is compressed into a single anchor model. At round k, GRAFT requires the incoming teacher T_k, and the frozen historical student

S_{k-1}, rather than the complete collection of past teachers $T_1, \ldots, T_{k-1}$. Consequently, the number of large backbone models required for retention is independent of the number of previous teachers. The following subsection addresses a complementary challenge: heterogeneous teachers can impose conflicting representational requirements on the shared backbone. We therefore introduce task specific readout tokens and geometry agnostic objectives to isolate these heterogeneous interfaces.

3.3 Addressing representation heterogeneity

Teacher Specific Readout Tokens. A straightforward multi teacher architecture attaches k task specific projection heads to a single global student representation. Although the projection heads are independent, all tasks must read from the same global summary vector $\mathbf{z}$. This creates a representational bottleneck when the corresponding teachers encode substantially different semantics or organize their representation spaces differently.

To disentangle these heterogeneous readouts while retaining a shared visual representation, we introduce Teacher Specific Readout Tokens (TSRT). Instead of using a single global summary token, we associate each task j with an independent learnable readout token $\mathbf{c}_j$. For an input image $\mathbf{x}$, all k readout tokens are prepended jointly to the shared visual patch tokens $v(\mathbf{x})$:

$$\mathbf{Z}_{\text{in}} = \big[\mathbf{c}_1, \ldots, \mathbf{c}_k, v(\mathbf{x})\big]. \tag{5}$$

Then $\mathbf{Z}_{\text{in}}$ is processed by the shared student backbone in the forward pass:

$$\big[\tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_k, \mathbf{Z}_{\text{patch}}\big] = \mathcal{E}^{(k)}\big(\mathbf{Z}_{\text{in}}\big), \tag{6}$$

where $\tilde{\mathbf{c}}_j$ denotes the contextualized readout token associated with task j, and $\mathbf{Z}_{\text{patch}}$ denotes the shared contextualized patch representation. Each task then reads from the shared representation through its corresponding projector. Specifically, the projector receives its dedicated readout token together with the common patch representation:

$$\mathbf{s}_j^{(k)} = \mathcal{H}_j^{(k)}\big(\tilde{\mathbf{c}}_j \oplus \mathbf{Z}_{\text{patch}}\big). \tag{7}$$

This architecture therefore shares the visual computation while factorizing the global readout across tasks. A single global readout forces these heterogeneous objectives to compete through the same summary representation, even when the underlying visual features useful to them can be shared. TSRT instead separates the shared visual representation from the task dependent readout, allowing each task to summarize the common patch features according to its own distillation objective.

Geometry Agnostic Relational Loss. Vision-language teachers introduce a distinct challenge in heterogeneous multi teacher distillation because their representation geometry can differ substantially from that of dense or geometric vision teachers. As illustrated by the PCA visualization in Figure 2, the representations produced by these encoders diverge from the shallowest layers onward. A conventional feature distillation objective directly aligns the student and teacher image embeddings. Such an objective implicitly assumes that direct correspondence between the student and teacher feature spaces is desirable. However, enforcing this correspondence can perturb features required by previously distilled tasks.

We therefore avoid directly matching the image embeddings of the vision language teacher. Instead, we exploit the property for which such a teacher is particularly valuable: its learned alignment between images and language. Give student and supervising feature outputs $\mathbf{s}_j$ and $\hat{\mathbf{s}}_j$ we require the student and teacher to induce similar relationships between image and text features:

$$\text{sim}\left(\mathbf{s}_j, \mathbf{t}\right) \approx \text{sim}\left(\hat{\mathbf{s}}_j, \mathbf{t}\right), \tag{8}$$

where $\mathbf{t}$ is the text features. where $\mathbf{t}$ denotes the text features produced by the frozen text encoder. We convert these image text similarities into probability distributions:

$$p_j^t = \text{softmax}\left(\frac{\hat{\mathbf{s}_j}^\top \mathbf{t}}{\tau_t}\right), \qquad p_j^s = \text{softmax}\left(\frac{\mathbf{s}_j^\top \mathbf{t}}{\tau_s}\right), \tag{9}$$

where τ_t and τ_s are the teacher and student temperatures, respectively. We then transfer the relational structure by defining

$$d\left(\mathbf{s}_j, \hat{\mathbf{s}}_j\right) = \text{KL}\left(p_j^t, \|, p_j^s\right). \tag{10}$$

Table 1: **Main results across five task families.** All models use a ViT-B backbone; teacher rows (ViT-L) are shown as per-task ceilings. Metric directions are marked in the header, best values are in bold and second-best underlined, and "-" indicates the model lacks that capability. ‡ Segmentation for RADIOv2.5 is distilled from SAM, and for SigLino and EUPE from DINOv3. § DUNE result obtained by fine-tuning on non-public data.

		Image Understanding	2D Dense Prediction		3D Human Perception		3D Vision	Vision-Language	
Model	**Arch.**	**IN-1k cls ↑**	**Depth ↓**	**mIoU ↑**	**F1 ↑**	**PA-PVE ↓**	**AUC ↑**	**COCO (T→I/I→T) ↑**	**Flickr (T→I/I→T) ↑**
Teachers (ceilings)									
DINOv2	ViT-L	–	0.384	47.70	–	–	–	–	–
Multi-HMR	ViT-L	–	–	–	95.0	36.9	–	–	–
MASt3R	ViT-L	–	–	–	–	–	91.2	–	–
SigLIP2	ViT-L	83.8	–	–	–	–	–	55.7 / 72.1	83.2 / 95.1
Agglomerative baselines									
RADIOv2.5	ViT-B	72.6	0.718	50.0§	–	–	–	48.9 / 63.2	76.2 / 89.7
DUNE	ViT-B	–	**0.358**	45.6	**94**	**56.0**	**94.7**‡	–	–
SigLino	ViT-B	80.5	0.750§	**52.9**§	–	–	–	**55.6 / 72.9**	**81.9 / 94.2**
EUPE	ViT-B	79.7	<u>0.391</u>§	<u>52.4</u>§	–	–	–	–	–
GRAFT models									
S_1 DINOv2	ViT-B	–	0.420	45.5	–	–	–	–	–
$S_2 = S_1+$ Multi-HMR	ViT-B	–	0.429	45.5	89	70.8	–	–	–
$S_3 = S_2+$ MASt3R	ViT-B	–	0.433	45.7	89	75.6	84.2	–	–
$S_4 = S_3+$ SigLIP2	ViT-B	81.2	0.441	46.0	89	78.1	81.8	51.9 / 69.0	76.2 / 88.4
GRAFT (ours)	ViT-B	**81.1**	0.437	46.0	<u>89</u>	<u>75.3</u>	<u>82.4</u>	<u>52.3</u> / <u>69.7</u>	<u>76.8</u> / <u>89.7</u>

Thus, the relational objective transfers the teacher's relative semantic relationships with the text features rather than enforcing pointwise equality between $\mathbf{s}_j^{(k)}$ and $\hat{\mathbf{s}}_j^{(k)}$. The student is therefore not required to reproduce the absolute coordinates of the teacher representation. In practice, we compute the relational objective using batchwise image-text similarity matrices for each training batch. We form the matrices over the global batch to provide each visual representation with a larger set of text features for relational comparison.

4 Experiments

4.1 Datasets and metrics

Teachers models. We distill four heterogeneous teachers , each a ViT-L encoder. DINOv2 (Oquab et al., 2023) is a self-supervised dense teacher whose features transfer to segmentation and depth. Multi-HMR (Baradel et al., 2024) is for 3D human pose and mesh recovery. MASt3R (Leroy et al., 2024) is for two-view 3D geometry and matching. SigLIP2 (Tschannen et al., 2025), an image–text model that aligns visual and language representations. The default vision–language teacher is SigLIP2, with DFN5B (Fang et al., 2024) compared head to head in §4.3.

Datasets. Distillation uses datasets drawn from a mixture that spans all four teacher domains. General 2D understanding and vision–language draw on DataComp-12M (Gadre et al., 2023), ImageNet-21k (Ridnik et al., 2021), Mapillary Street-level Sequences (Warburg et al., 2020), and Google Landmarks Dataset v2 (Weyand et al., 2020). 3D human perception draws on BEDLAM (Black et al., 2023), AGORA (Patel et al., 2021), and CUFFS (Baradel et al., 2024). 3D geometry draws on ARKitScenesV2 (Baruch et al., 2021), BlendedMVS (Yao et al., 2020), CO3Dv2 (Reizenstein et al., 2021), DL3DV (Ling et al., 2024), Habitat512 (Savva et al., 2019), MegaDepthDense (Li & Snavely, 2018), Niantic (Arnold et al., 2022), TartanAir (Wang et al., 2020), Unreal4K (Tosi et al., 2021), VirtualKITTI (Cabon et al., 2020), and WildRGB (Xia et al., 2024).

Implementation details. The student model is a ViT-B/14 encoder at 448px. We evaluate every task family with its standard protocol, spanning the five domains GRAFT unifies. Following previous method (Sarıyıldız et al., 2025), we drop the projectors and finetune a head for each downstream task. Image understanding is measured by ImageNet-1k classification top-1 accuracy. Dense prediction is measured by ADE20K linear-probe segmentation (mIoU) and monocular depth (RMSE). 3D human pose is measured on BEDLAM by F1 and PA-PVE. 3D geometry uses the MASt3R-based suite, covering map-free relocalization, multi-view depth, and multi-view pose. For the 3D-geometry evaluations we substitute the GRAFT encoder for the MASt3R encoder, re-initialize the encoder–decoder linear layer, and fine-tune the decoder and prediction heads. Vision–language is measured by COCO-5K and Flickr-30K retrieval in both directions, at R@1. We substitute the GRAFT encoder for the visual encoder and keep the same text encoder.

4.2 Main results

Performance across different domains. GRAFT is the only model in Table 1 that reports a result in every column: a single ViT-B backbone spanning image understanding, 2D dense prediction,

Table 2: **Ablation studies for vision language distillation.** We study the distillation strategy, vision language projector architecture, and choice of vision language teacher. All experiments start from a student distilled from DINOv2 and further distill an incoming vision language teacher.

	Acquisition	Retention			Acquisition	Retention	
Variant	**IN-1k cls ↑**	**mIoU ↑**	**Depth RMSE ↓**	**Variant**	**IN-1k cls ↑**	**mIoU ↑**	**Depth RMSE ↓**
Distillation strategy				*Vision–language projector*			
Direct distillation	**82.6**	35.4	0.584	Frozen encoder + Transformer	75.5	45.5	**0.425**
LoRA finetuning	76.2	45.7	**0.425**	Transformer, 1 layer	80.4	46.0	0.448
LPFT	78.0	40.5	0.442	Transformer, 3 layers	75.9	45.4	0.451
EWC	77.1	40.2	0.508	**MLP + residual (ours)**	**81.2**	**46.0**	0.450
GRAFT w/o TSRT & Relational Loss	77.4	40.6	0.489	*Vision–language teacher*			
GRAFT w/o Relational Loss	80.1	43.6	0.455	DFN5B (softmax InfoNCE)	80.0	45.7	0.454
GRAFT (ours)	81.2	**46.0**	0.450	**SigLIP2 (pairwise sigmoid, ours)**	**81.2**	**46.0**	**0.450**

3D human perception, 3D vision, and vision–language. Each baseline covers only the families its teacher set targets and is blank elsewhere. This breadth does not cost strength. On image understanding GRAFT attains the strongest classification of any agglomerative baseline. On 2D dense prediction it stays competitive with the segmentation- and depth-specialised baselines. On 3D human perception it trails only DUNE, the sole baseline that targets pose. On 3D vision it acquires geometric capability that most baselines lack outright. On vision–language it is second only to the vision–language-specialised baseline and ahead of every other agglomerative model that attempts retrieval. GRAFT therefore leads on image understanding and matches or approaches the strongest specialised baseline on every remaining family, while being the only model to cover all of them.

Table 1 places GRAFT beside baselines trained under materially different conditions, so it should not be read as a controlled ranking. Two differences dominate, detailed in Appendix A. First, the teacher sets differ in number and scale: several baselines distil from far larger teachers, up to $\sim$1.1B parameters for RADIOv2.5 and 1.9B for EUPE, whereas every GRAFT teacher is a ViT-L of $\sim$304M. Second, the training corpora differ in size and composition; several baseline corpora are undisclosed, and SigLino reports image tokens rather than images. Because a stronger teacher or a larger, better-matched corpus directly lifts the metrics Table 1 tabulates, a baseline's lead on a given family reflects its setup as much as its method. GRAFT also carries the leanest budget of the five, never training its trunk from scratch: it warm-starts from the DUNE base and re-pays no base-encoder cost per graft. The controlled comparisons that hold teacher, data, and schedule fixed are reported in §4.3; Table 1 instead makes the coverage point, that one ViT-B under a single fixed teacher and data budget answers in every column.

Continually growing the model from a single teacher. The main-table GRAFT warm-starts from DUNE and performs a single vision–language graft (§4.1). This ablation removes that convenience and asks the harder question the setting poses: can GRAFT *build* the agglomerative model itself, one teacher at a time, with no pre-assembled base? Starting from a student distilled from a single dense teacher (DINOv2), we graft Multi-HMR, then MASt3R, then SigLIP2, and evaluate the student S_k after every graft in Table 1. The table reads two ways. Across a row, each freshly grafted capability switches on at a level comparable to the warm-started model. Down a column, a capability introduced at one step persists through every later graft rather than collapsing: dense segmentation is retained in full and even edges up, depth drifts only modestly, pose detection holds while vertex error drifts within a bounded range, and 3D-vision accuracy gives back only a little after the vision–language graft. The four-capability model built entirely by continual grafting reaches performance comparable to the DUNE-warm-started GRAFT, trailing it only slightly on 3D pose and geometry.

4.3 Ablations

Distillation strategy. In Table 2, we compare GRAFT with alternative strategies for incorporating a new vision-language teacher into an already distilled student from DINOv2. Direct distillation directly optimizes the student toward the new teacher and achieves strong acquisition of the new capability, but substantially degrades the previously distilled dense and geometric capabilities, illustrating the stability plasticity conflict in sequential teacher integration. LoRA (Hu et al., 2021) finetuning instead introduces a portable task specific adaptation. We evaluate retention with the LoRA parameters removed, which recovers the original student behavior, but the resulting acquisition of the new capability remains limited. We further consider linear probe then finetune (LPFT) (Kumar et al., 2022) and EWC based (Kirkpatrick et al., 2017) continual learning, both of which improve the balance between acquisition and retention but still exhibit noticeable interference with previously learned capabilities. Starting from the basic GRAFT pipeline without TSRT or the relational objective, we observe a clear stepwise improvement as each proposed component is introduced.

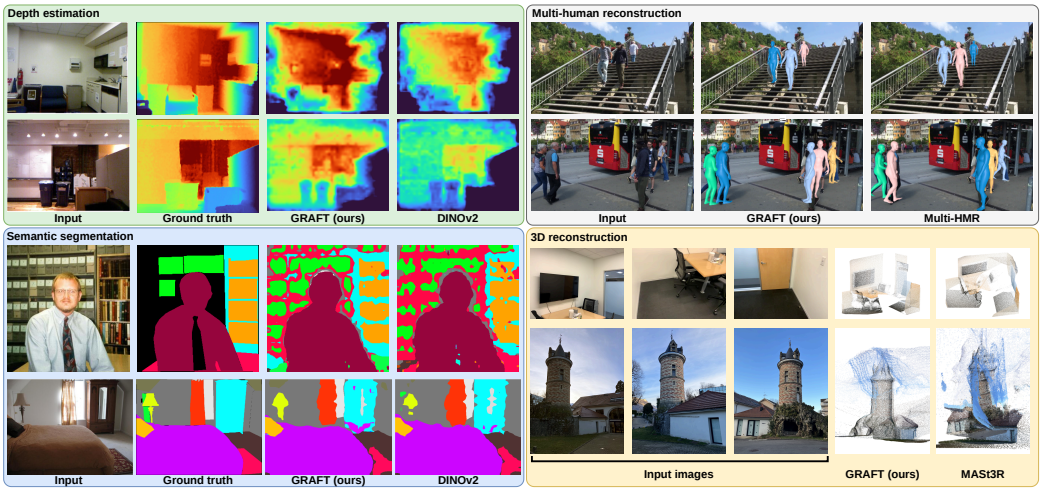


Figure 4: **Qualitative comparison** on depth estimation, semantic segmentation, multi-human reconstruction, and 3D reconstruction between GRAFT and corresponding teachers.

These results support the complementary roles of task specific readouts and relational distillation in integrating heterogeneous teachers without disrupting previously acquired representations.

Choice of projector. We study the architecture of the projector. One option is to attach a Transformer projector and optimize only the newly introduced head while keeping the distilled encoder frozen. This largely avoids modifying the previously distilled backbone and therefore preserves part of its original capabilities, but provides insufficient flexibility for acquiring the new vision-language representation. Increasing the Transformer depth does not consistently improve performance, suggesting that a more expressive projector alone does not resolve the representation mismatch between the student and the vision language teacher. Our final design instead uses a lightweight MLP projector with a residual connection. Despite its substantially simpler architecture, it achieves acquisition and retention comparable to or better than the Transformer based alternatives across the evaluated tasks. This indicates that the benefits of GRAFT primarily arise from the proposed representation and optimization design rather than from increasing projector capacity.

Vision language teacher. We examine whether the effectiveness of GRAFT depends strongly on the choice of vision-language teacher. In addition to SigLIP2, we replace the supervising model with DFN5B, which has a comparable model scale and capability GRAFT exhibits similar acquisition and retention behavior with both teachers. This result suggests that the proposed grafting mechanism is robust to the specific choice of vision language teacher and can transfer vision-language knowledge without relying on properties unique to SigLIP2.

4.4 Qualitative results

We provide qualitative comparisons between GRAFT and its corresponding teacher models across four downstream tasks: depth estimation, semantic segmentation, multi-human reconstruction, and 3D reconstruction. As shown in Figure 4, our method achieves competitive or superior performance.

5 Conclusion

We introduced GRAFT, a continual multi-teacher distillation framework that enables agglomerative vision foundation models to grow beyond a fixed teacher set. Instead of revisiting all previously integrated teachers, GRAFT uses the previous student as a frozen retention teacher and distills jointly from it and the incoming foundation model, allowing new capabilities to be acquired while preserving prior ones at the cost of a single incremental distillation stage. To accommodate heterogeneous representation geometries, GRAFT further introduces Teacher Specific Readout Tokens and a Geometry Agnostic Relational Loss that transfers vision–language knowledge through image–text similarity structure rather than direct feature matching. Using this formulation, a single ViT-B backbone progressively integrates image understanding, 2D dense prediction, 3D human pose estimation, 3D vision, and vision–language capabilities.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, find related work, and implement methods. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Generating synthetic datasets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, and assisting in the writing of proofs are not applicable to this work. We have reviewed all AI-assisted work. We proofread the LLM-polished writing, reviewed the literature, and checked the code implementation. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision (ECCV)*, 2022.

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision (ECCV)*, 2024.

Axel Barroso-Laguna, Sowmya Munukutla, Victor Adrian Prisacariu, and Eric Brachmann. Matching 2d images in 3d: Metric relative pose from metric correspondences. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

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.

Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 2023.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. Distillation scaling laws. *arXiv preprint arXiv:2502.08606*, 2025.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Sofian Chaybouti, Sanath Narayan, Yasser Dahou, Phúc H Lê Khac, Ankit Singh, Ngoc Dung Huynh, Wamiq Reyaz Para, Hilde Kuehne, and Hakim Hacid. Siglino: Efficient multi-teacher distillation for agglomerative vision foundation models. *arXiv preprint arXiv:2512.20157*, 2025.

Zhaowei Chen, Borui Zhao, Yuchen Ge, Yuhao Chen, Renjie Song, and Jiajun Liang. Asymmetric decision-making in online knowledge distillation: Unifying consensus and divergence. *arXiv preprint arXiv:2503.06685*, 2025.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *International Conference on Learning Representations*, volume 2024, 2024.

Philippe Formont, Maxime Darrin, Banafsheh Karimian, Eric Granger, Jackie CK Cheung, Ismail Ayed, Mohammadhadi Shateri, and Pablo Piantanida. Learning task-agnostic representations through multi-teacher distillation. In *Advances in Neural Information Processing Systems*, 2026.

Logan Frank and Jim Davis. What makes a good dataset for knowledge distillation$f$. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, 2023.

Christian Grashei, Christian Brechenmacher, Rao Muhammad Umer, Jingsong Liu, Carsten Marr, Ewa Szczurek, and Peter J Schüffler. Pathryoshka: Compressing pathology foundation models via multi-teacher knowledge distillation with nested embeddings. *arXiv preprint arXiv:2511.23204*, 2025.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*. IEEE, 2023.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 2017.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

Jungsoo Lee, Debasmit Das, Munawar Hayat, Sungha Choi, Kyuwoong Hwang, and Fatih Porikli. Customkd: Customizing large vision foundation for edge model improvement via knowledge distillation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision (ECCV)*, 2024.

Guopeng Li, Qiang Wang, Ke Yan, Shouhong Ding, Yuan Gao, and Gui-Song Xia. Fuse before transfer: Knowledge fusion for heterogeneous distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2025.

Yue Li, Qi Ma, Runyi Yang, Mengjiao Ma, Bin Ren, Nikola Popovic, Nicu Sebe, Theo Gevers, Luc Van Gool, Danda Pani Paudel, et al. Chorus: Multi-teacher pretraining for holistic 3d gaussian scene encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Yuxiang Lu, Shengcao Cao, and Yu-Xiong Wang. Swiss army knife: Synergizing biases in knowledge from vision foundation models for multi-task learning. In *International Conference on Learning Representations*, volume 2025, 2025.

Nicolas Michel, Maorong Wang, Jiangpeng He, and Toshihiko Yamasaki. Continual distillation of teachers from different domains. *arXiv preprint arXiv:2605.04059*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021.

Mike Ranzinger, Jon Barker, Greg Heinrich, Pavlo Molchanov, Bryan Catanzaro, and Andrew Tao. Phi-s: Distribution balancing for label-free multi-teacher distillation. *arXiv preprint arXiv:2410.01680*, 2024a.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024b.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2017.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Unic: Universal classification models via multi-teacher distillation. In *European Conference on Computer Vision (ECCV)*, 2024.

Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Pau De Jorge, Diane Larlus, and Yannis Kalantidis. Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019.

Jinghuan Shang, Karl Schmeckpeper, Brandon B May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. *arXiv preprint arXiv:2407.20179*, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *arXiv preprint arXiv:2104.00680*, 2021.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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)*. IEEE, 2024.

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

Yimu Wang, Mozhgan Nasr Azadani, Sean Sedwards, and Krzysztof Czarnecki. Hawaii: Hierarchical visual knowledge transfer for efficient vision-language models. In *Advances in Neural Information Processing Systems*, 2026a.

Yixiao Wang, Mingxiao Huo, Zhixuan Liang, Yushi Du, Lingfeng Sun, Haotian Lin, Jinghuan Shang, Chensheng Peng, Mohit Bansal, Mingyu Ding, et al. Ver: Vision expert transformer for robot learning via foundation distillation and dynamic routing. In *International Conference on Learning Representations*, 2026b.

Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2–a large-scale benchmark for instance-level recognition and retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

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

Chengxi Zeng, Yuxuan Jiang, Fan Zhang, Alberto Gambaruto, and Tilo Burghardt. Agglomerating large vision encoders via distillation for vfss segmentation. *arXiv preprint arXiv:2504.02351*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023.

Yitian Zhang, Xu Ma, Yue Bai, Huan Wang, and Yun Fu. Accessing vision foundation models via imagenet-1k. In *International Conference on Learning Representations*, 2025.

A COMPARISON CONDITIONS

The main comparison in Table 1 is not strictly like-for-like: the methods differ in the teachers they distill and in the corpora they train on, both of which directly move the reported metrics. Table 3 collects these conditions.

Table 3: **Teacher sets and training data differ across methods**, so the head-to-head in Table 1 is not like-for-like. Teacher scale is per model and counts are the number of distillation teachers. Entries are stated by each source where available and marked unavail when a source does not disclose them. SigLino reports image tokens rather than images, and EUPE additionally trains a from-scratch 1.9B proxy before distillation. GRAFT's teachers are DUNE's three, inherited through the warm start, plus the grafted vision–language teacher; its data subtotal counts DataComp-12M and ImageNet-19k and excludes Mapillary and GLDv2, whose sizes are undisclosed.

Method	**Teachers (count × scale)**	**Training data**
RADIOv2.5	4× up to ∼1.1B (DINOv2-g, SAM-H, DFN-CLIP-H)	unavail
DUNE	3× ViT-L/14 (∼304M)	∼20.7M images
SigLino	2× ViT-L	230B image tokens
EUPE	PEcore-G 1.9B, PElang-G 1.7B, DINOv3-H+ 840M (+1.9B proxy)	unavail
GRAFT (ours)	3× ViT-L/14 (∼304M) + SigLIP2 (ViT-L)	∼21.7M images

B EXTENDED RESULTS

3D geometry: map-free relocalization. Table 4 examines GRAFT's 3D-geometry capability in detail on the map-free relocalization leaderboard, with the GRAFT encoder substituted into MASt3R and only the decoder fine-tuned. GRAFT ranks second only to the MASt3R teacher it distills from, ahead of the dedicated matching and two-view baselines on virtual-correspondence AUC, reprojection error, and relative-pose AUC. The residual gap to the teacher concentrates in fine-grained pose precision and reflects the setting: GRAFT compresses a ViT-L geometry specialist into a ViT-B trunk shared with four other capabilities and adapts no backbone weights at evaluation. That a shared, unadapted trunk still recovers most of a dedicated matcher's relocalization accuracy shows the geometry graft leaves genuinely usable spatial structure in the trunk.

Table 4: **Map-free visual relocalization**, official leaderboard (test split). The GRAFT encoder is substituted for the MASt3R encoder and the MASt3R decoder is fine-tuned, with the backbone frozen. AUC and precision (Prec.) are percentages, and metric directions are marked in the header. Baseline rows are quoted from Sarıyıldız et al. (2025), Table 4. This is the leaderboard split, whereas the 3D-Vision column of Table 1 reports the validation-split VCRE < 90 px AUC, so the two are not directly comparable. The MASt3R leaderboard entry is a private checkpoint that scores slightly above the released model we use as teacher.

		VCRE < 45 **px**		**VCRE** < 90 **px**		**Median reproj.**	**Relative pose error**			
Method	**Encoder**	**AUC** ↑	**Prec.** ↑	**AUC** ↑	**Prec.** ↑	**error (px)** ↓	**Trans. (m)** ↓	**Rot. (°)** ↓	**Prec.** ↑	**AUC** ↑
LoFTR (Sun et al., 2021)	CNN	39.7	18.2	61.8	33.5	166.8	2.31	39.4	26.9	9.8
DUSt3R (Wang et al., 2024)	ViT-L	45.9	28.7	69.8	50.4	115.8	0.99	7.1	39.4	21.4
MicKey (Barroso-Laguna et al., 2024)	ViT-L	57.2	31.2	74.8	49.3	129.5	1.59	26.0	28.3	12.0
MASt3R (Leroy et al., 2024)	ViT-L	**81.7**	**63.0**	**93.3**	**79.3**	**48.8**	**0.37**	**2.2**	**74.0**	**54.7**
GRAFT (S_4, ours)	ViT-B	60.9	48.0	82.2	69.1	73.1	0.64	7.1	27.9	38.8