AMDF: Holistic Probing of Vision Foundation Models with Adaptive Multi-Layer Dense Fusion

Turhan Can Kargin[1,2] · Bill Psomas[3,4] · Paul Wagner[5] · Giorgos Tolias[3,4] · Bartosz Zieliński[1,2] · Marcin Przewięźlikowski[1,2]

[1] Jagiellonian University in Kraków [2] GMUM [3] Czech Technical University in Prague [4] VRG [5] RWTH Aachen University

EDS 2026
ELLIS Doctoral Symposium on Trustworthy AI
Lisbon · July 27–30, 2026

TL;DR

The final layer of a Vision Foundation Model is **not** always the best representation. AMDF learns to fuse dense features from **every** frozen backbone layer, independently for each spatial patch and each downstream task.

- We identify an **abstraction–depth correlation**: geometric tasks peak at intermediate layers, semantic tasks peak near the final layer — so no single fixed depth is right for every task.
- We propose **AMDF**: per-layer spatial refinement followed by patch-wise cross-attention fusion across the full VFM depth, with the backbone frozen throughout training.
- AMDF sets the strongest overall probing performance across DINOv3, MAE, SigLIP 2, V-JEPA2.1 and Perception Encoder, on dense geometric, semantic, and shared multi-task benchmarks.
- Its learned depth-attention is interpretable, and recovers the same semantic–geometric split across depth found via exhaustive oracle layer sweeps.

Motivation & Overview

Representations change abstraction level across network depth: low-level geometry, mid-level scene structure, and high-level semantics peak at different layers. AMDF removes the need to hand-pick a layer — it refines every layer, then fuses them **per patch** with a learned depth-attention distribution.

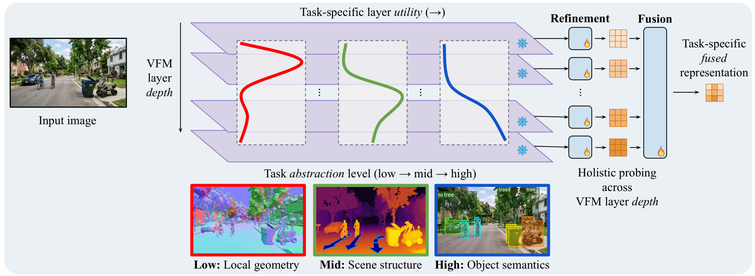


Fig. 1: Task-specific layer utility varies with abstraction level (low/mid/high). AMDF refines each frozen layer, then fuses all layers per patch into a task-specific dense representation.

Abstraction–Depth Correlation: Empirical Evidence

We probe a frozen DINOv3-L/16 independently at every layer $\ell \in \{1,\dots,24\}$ on three NYUv2 tasks of increasing abstraction, training a lightweight head per layer and recording its validation score. All three tasks share the same training images, yet the best-performing layer shifts systematically deeper as task abstraction increases.

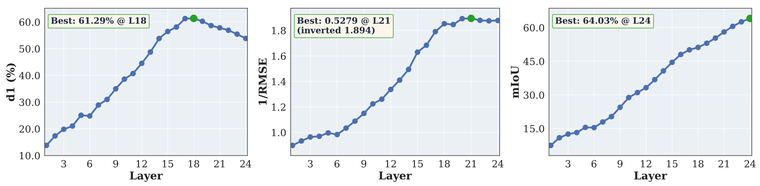


Fig. 2: Layer-wise probing on NYUv2 with DINOv3-L/16: the best layer shifts deeper as tasks become more semantic — surface normals peak at layer 18, depth at layer 21, segmentation at the final layer 24 (the last layer).

Exhaustive per-task layer search is costly and still commits to one *fixed* depth. Worse, a single layer ignores that complementary information may be spread across *multiple* depths at once — while naively concatenating all layers yields prohibitively large dense features. This motivates AMDF's holistic, patch-wise fusion across the full frozen hierarchy.

Implementation & Task Heads

- Every extracted layer is first projected to a shared, low-dimensional bottleneck before the refinement / fusion attention, keeping dense multi-layer attention cheap in memory and compute.
- The VFM backbone stays **frozen** throughout; only the lightweight refinement + fusion probe and the downstream task head are trained end-to-end.
- All available ViT blocks are exposed to AMDF: no manual subset selection or hand-picked layer depth is required.
- AMDF adds only a small fraction of the frozen backbone's parameters — the bottleneck projections, refinement queries, and depth-fusion queries are the sole trainable components besides the task head.

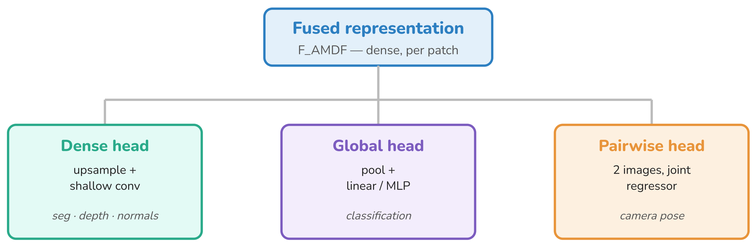
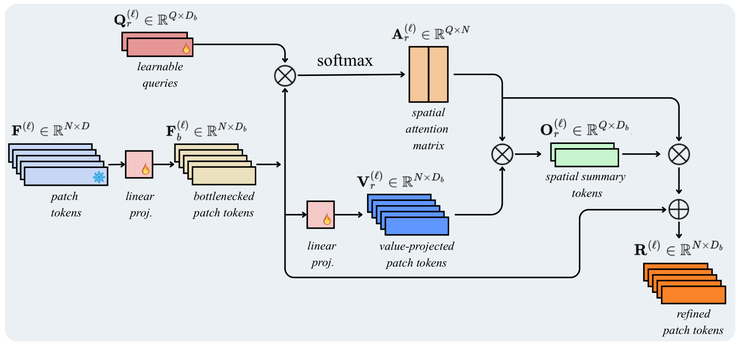


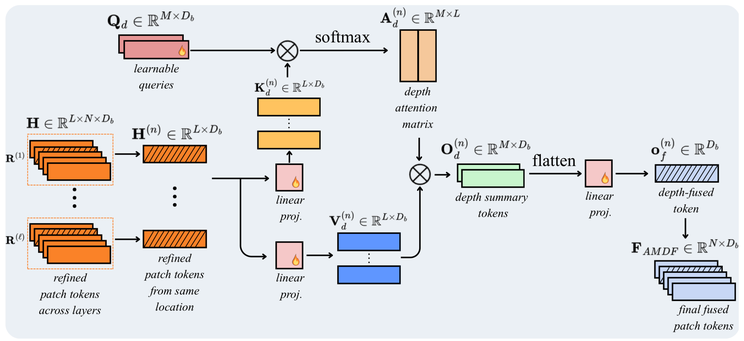
Fig. 3: The same frozen-backbone fusion module feeds three task-specific heads — only the head (and its supervision) changes per task.

Method: Adaptive Multi-Layer Dense Fusion

(1) Layer-specific refinement — learnable per-layer queries cross-attend to each layer's patch tokens and scatter back to the patch grid, preserving dense spatial structure.



(2) Depth-aware fusion — for every spatial patch, learnable depth queries cross-attend over the refined layer stack (with layer position embeddings), yielding a patch-specific mixture over network depth.



Learned depth-attention tracks layer utility

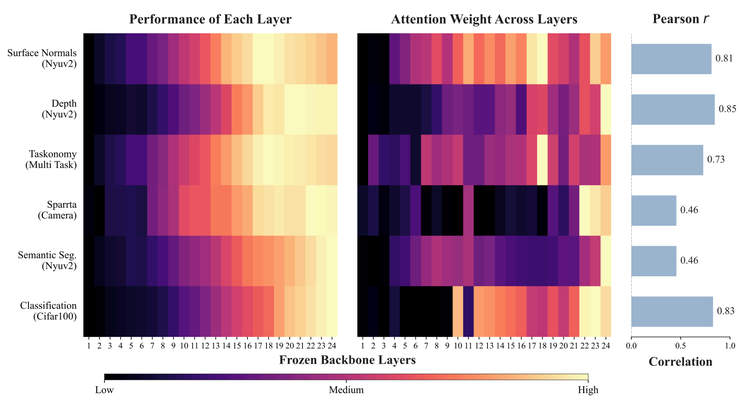


Fig. 4: AMDF's learned depth-attention (middle) tracks per-layer probe performance (left) on DINOv3-L/16. Alignment (Pearson r, right) is strongest for dense geometric tasks (depth r≈0.85, normals r≈0.81, Taskonomy r≈0.73) and shifts toward deep layers for semantic tasks (classification r≈0.83).

Conclusion

- VFM capabilities are **distributed across depth**, not concentrated in the final layer — final-layer probing systematically underestimates them.
- AMDF's dense, patch-wise depth fusion beats final-layer, oracle single-layer, and static multi-layer (mean / weighted / concat) baselines across geometric, semantic and multi-task benchmarks.
- Learned depth-attention is interpretable: it recovers the same semantic–geometric split found via exhaustive oracle layer sweeps.
- This encourages moving beyond final-layer probing — frozen VFMs should be studied as full representational hierarchies, not single-output feature extractors.
- **No manual layer search needed:** AMDF replaces exhaustive per-task, per-layer sweeps with one learned, adaptive probe.

Results on NYUv2 (increasing abstraction)

Surface normals — d_1 (↑)

Method	DINOv3-S	DINOv3-B	DINOv3-L	MAE-L	SigLIP2-L	V-JEPA2.1-L	PESpatial-L	Rank
BestLayer (oracle)	52.47	57.67	61.32	53.12	41.95	61.89	54.43	5.00
LastLayer	48.75	51.90	53.01	49.13	30.58	60.01	40.04	6.00
AvgLayer	52.88	58.87	62.86	55.42	47.16	62.03	56.12	3.86
WeightedLayers	53.03	59.08	62.90	55.45	49.28	61.95	56.99	3.14
ConcatLayers	54.00	59.61	63.27	55.97	50.14	62.39	58.33	2.00
AMDF (ours)	54.65	59.63	63.40	56.26	52.57	62.97	59.66	1.00

Depth estimation — d_1 (↑)

Method	DINOv3-S	DINOv3-B	DINOv3-L	MAE-L	SigLIP2-L	V-JEPA2.1-L	PESpatial-L	Rank
BestLayer (oracle)	80.01	83.99	85.02	74.17	76.46	86.91	85.98	4.71
LastLayer	79.28	82.33	83.23	68.43	72.23	86.23	83.55	6.00
AvgLayer	79.59	84.73	87.20	76.56	79.90	87.51	86.08	3.00
WeightedLayers	80.08	84.49	87.32	76.55	80.10	87.27	85.87	3.29
ConcatLayers	80.40	84.48	87.52	76.95	80.54	87.31	87.13	2.14
AMDF (ours)	80.47	**84.47**	87.56	**76.43**	81.42	87.95	87.40	1.86

Semantic segmentation — mIoU (↑)

Method	DINOv3-S	DINOv3-B	DINOv3-L	MAE-L	SigLIP2-L	V-JEPA2.1-L	PESpatial-L	Rank
BestLayer (oracle)	52.29	58.75	64.03	33.21	44.01	56.77	59.24	4.93
LastLayer	52.29	58.75	64.03	33.21	44.01	56.77	59.24	4.93
AvgLayer	52.40	58.88	63.96	38.69	48.09	56.26	59.84	3.57
WeightedLayers	52.99	59.34	64.07	38.74	48.25	56.22	59.58	2.64
ConcatLayers	52.75	58.98	64.07	38.51	47.21	55.93	59.64	3.64
AMDF (ours)	53.22	60.12	64.86	38.83	**47.98**	57.48	60.00	1.29

Shared multi-task representation (Taskonomy)

Aggregate L_1 ×100 (↓)

Method	DINOv3-S	DINOv3-B	DINOv3-L	MAE-L	SigLIP2-L	V-JEPA2.1-L	PESpatial-L	Rank
BestLayer (oracle)	31.86	28.94	27.94	28.93	33.19	26.99	30.81	5.00
LastLayer	33.02	30.41	29.75	30.41	36.83	27.78	33.68	6.00
AvgLayer	30.50	28.11	26.46	28.04	30.94	26.60	30.24	3.57
WeightedLayers	30.43	28.07	26.45	28.09	30.69	26.61	30.16	3.14
ConcatLayers	30.35	28.13	26.38	28.03	30.57	26.39	30.12	2.29
AMDF (ours)	29.53	27.56	26.01	27.98	29.97	26.14	29.62	1.00

Fine-Grained & Wildlife Classification (Top-1)

Stanford Cars — fine-grained vehicle recognition

Method	DINOv3-S	DINOv3-B	DINOv3-L	MAE-L	SigLIP2-L	V-JEPA2.1-L	PESpatial-L	Rank
BestLayer (oracle)	72.76	86.77	89.81	47.93	94.11	50.89	64.33	6.14
LastLayer	72.76	86.77	89.73	47.89	93.89	50.89	64.33	6.71
AvgLayer	67.88	83.65	87.16	56.92	94.34	59.89	63.52	7.43
WeightedLayers	73.92	86.49	90.55	59.45	94.43	60.74	64.95	4.43
ConcatLayers	72.51	85.99	89.95	63.29	94.18	64.18	63.93	5.57
AvgPool	68.41	86.61	89.05	35.55	93.93	42.01	57.46	8.14
EfficientProbe	89.32	92.03	93.78	88.07	94.83	84.33	87.46	1.57
AttentiveFusionAP	74.79	87.69	90.85	58.99	94.63	59.94	65.19	3.57
AMDF (ours)	88.98	92.56	94.69	87.78	95.15	89.11	86.22	1.43

CzechLynx — individual wildlife re-identification

Method	DINOv3-S	DINOv3-B	DINOv3-L	MAE-L	SigLIP2-L	V-JEPA2.1-L	PESpatial-L	Rank
BestLayer (oracle)	27.57	35.53	40.04	27.61	20.19	31.45	23.79	7.57
LastLayer	27.57	35.53	39.18	27.26	28.92	31.45	23.79	7.57
AvgLayer	33.72	38.24	42.72	35.01	35.92	37.65	29.18	4.43
WeightedLayers	34.12	39.02	44.12	35.19	36.65	37.55	29.88	3.00
ConcatLayers	34.01	38.68	43.13	35.02	35.82	36.75	29.65	4.14
AvgPool	26.07	34.28	37.22	24.11	28.46	29.91	22.83	8.86
EfficientProbe	37.12	47.49	52.35	46.47	35.11	42.72	32.26	2.00
AttentiveFusionAP	31.28	37.12	40.40	31.09	34.89	34.33	27.92	6.00
AMDF (ours)	37.50	44.73	52.10	41.34	41.75	47.82	39.92	1.43

NAVI Camera Pose (Pairwise Head)

Method	DINOv3-S		DINOv3-B		DINOv3-L		MAE-L		SigLIP2-L		V-JEPA2.1-L		PESpatial-L		Rank
	Rot.	Trans.	Rot.	Trans.	Rot.	Trans.	Rot.	Trans.	Rot.	Trans.	Rot.	Trans.	Rot.	Trans.	
BestLayer (oracle)	46.61	0.18	42.98	0.16	36.24	0.14	39.71	0.16	51.42	0.18	38.95	0.16	61.14	0.21	4.43
LastLayer	46.86	0.18	44.18	0.17	40.17	0.16	41.01	0.16	63.69	0.23	39.13	0.16	65.97	0.23	5.68
AvgLayer	42.89	0.17	32.81	0.14	33.50	0.13	35.80	0.15	56.44	0.21	41.14	0.16	62.69	0.22	4.14
WeightedLayers	42.31	0.16	32.63	0.14	32.22	0.13	33.69	0.14	55.05	0.20	40.01	0.16	56.61	0.20	3.14
ConcatLayers	39.22	0.16	31.12	0.13	28.82	0.12	31.58	0.13	43.48	0.17	38.79	0.15	51.17	0.19	1.68
AMDF (ours)	33.69	0.14	27.49	0.12	35.79	0.14	36.01	0.15	27.53	0.13	35.31	0.15	48.56	0.19	1.93