KneeDINO: Cross-Plane Attention Fusion for Multi-Label Knee Pathology Detection with Grounded LLM Explainability

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Automated detection of knee pathologies from MRI requires integrating evidence across three orthogonal imaging planes while handling variable scan lengths and class imbalance. ACL tears, meniscus tears, and general structural abnormalities are among the most common orthopedic presentations worldwide, yet radiological interpretation remains time-intensive and subject to inter-reader variability. We present KneeDINO, a multi-label framework that simultaneously detects all three pathologies from axial, coronal, and sagittal volumes. Our central contribution is a per-task cross-plane attention fusion pipeline: three independent RadioDINO ViT-B/16 encoders produce task-specific slice embeddings via per-task attention-pooling heads, and a shared cross-plane Transformer encoder is applied once per pathology, allowing each task to attend across planes using only the slice evidence most relevant to that task. A learnable per-plane embedding distinguishes axial, coronal, and sagittal tokens, which position-free self-attention alone cannot. Attention analysis reveals that all three tasks attend most strongly to the coronal plane, consistent with its clinical importance for joint-space evaluation, with no task exhibiting diagonal-dominant self-attention — confirming genuine cross-plane information exchange. A two-stage explainability pipeline passes volumetric EigenCAM saliency overlays to Gemini, producing plain-English summaries grounded in task-specific visual evidence. KneeDINO achieves AUC of 0.942 (ACL), 0.871 (meniscus), and 0.955 (abnormal) when tested on the MRNet dataset, surpassing prior approaches.

Keywords: Knee MRI · Multi-label classification · Vision Transformer · Cross-plane attention · Explainability · Large language models

1 Introduction

ACL tears, meniscus tears, and broader structural knee injuries are among the most common orthopedic presentations worldwide [10]. MRI is the diagnostic gold standard, yet interpretation is time-intensive and radiologists frequently disagree on borderline findings [9]. Deep learning has reached competitive performance on knee MRI classification [1,6], but two limitations remain before clinical

deployment. First, *late fusion* across imaging planes: prior systems score each plane independently and combine results only at the end [1,6], so no plane can inform how another is interpreted. Second, *opacity*: when a model flags a likely ACL tear, it provides no spatial or linguistic explanation a clinician can evaluate.

We address both limitations with KneeDINO. Rather than pooling slice features with a single shared head and fusing planes once, we assign each pathology task its own slice-selection mechanism. Three per-task attention-pooling heads learn which slices matter most for ACL integrity, meniscal tissue, and general abnormality independently. A shared cross-plane Transformer encoder is then applied once per task, so each pathology can attend across planes using its own slice evidence. A learnable per-plane identity embedding distinguishes axial, coronal, and sagittal tokens, which position-free self-attention cannot. Further contributions include a multi-label RadioDINO-based architecture [8] producing all three predictions in a single forward pass, and an explainability pipeline feeding task-specific EigenCAM maps [7] to Gemini [4], tying generated clinical language directly to the evidence driving each prediction.

2 Related Work

MRNet [1] established the benchmark using per-plane AlexNet models combined via logistic regression. Subsequent work such as Kara and Hardalaç [6] employed progressively operating ResNet50 models with CNN and denoising autoencoder stages for image selection, region localisation, and diagnosis, but still scored each plane independently. Transformer-based approaches [3] improve feature quality but retain this late-fusion structure. RadioDINO [8] is a ViT-B/16 pretrained via self-supervised DINO [2] on multi-modal radiological data; its attention heads identify clinically relevant anatomy without task-specific supervision, providing a strong inductive bias for fine-tuning on MRNet's 1,130 training cases.

Multi-view attention fusion — letting views attend to one another before prediction rather than merging independent outputs — has proven effective in 3D scene understanding. We apply this principle to the three orthogonal planes of a knee MRI, where the token sequence is exactly three plane-level embeddings. For explainability, EigenCAM [7] produces class-discriminative spatial maps; extending such maps to 3D MRI volumes via per-slice aggregation follows established practice [5]. LLM-based report generation [11] typically conditions on raw images or structured findings; our pipeline instead uses pathology-specific saliency overlays, grounding generated text in the evidence the classifier actually used.

3 Method

3.1 Problem Formulation

Each knee examination consists of three MRI volumes acquired in orthogonal planes—axial, coronal, and sagittal—each containing a variable number of slices

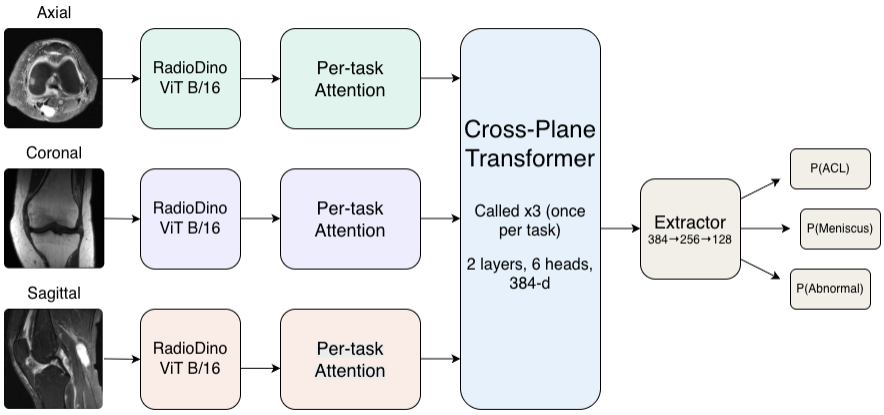

Fig. 1. KneeDINO architecture. Three independent RadioDINO ViT-B/16 encoders process axial, coronal, and sagittal volumes. Each encoder feeds three per-task attention-pooling heads (ACL, meniscus, abnormal), yielding nine task-specific plane embeddings in total. For each task, the corresponding three plane embeddings are passed through a shared cross-plane Transformer encoder, producing a task-specific fused representation that is decoded by its own output head. The optional explainability pipeline (dashed) computes task-specific EigenCAM overlays and passes them to Gemini for plain-English summaries.

at a common spatial resolution. The task is to produce three simultaneous binary pathology probabilities (ACL tear, meniscus tear, general abnormality) from this set of three volumes.

3.2 Plane-Specific RadioDINO Encoders

Each acquisition plane is processed by its own RadioDINO ViT-B/16 encoder, initialised from the same pretrained checkpoint but kept independent during fine-tuning. Slices are resized to 224×224 and replicated to three channels. The `[CLS]` token from each slice serves as a 768-dimensional slice summary. ViT-B/16 has 86M parameters per instance; with three independent copies the total backbone count is 258M, typical for this family. Only the final two transformer blocks and layer normalisation are fine-tuned per encoder ($\approx$15M per encoder, 45M total), keeping updated weights proportionate to MRNet's 1,130 training cases. Independent encoders prevent gradients from one plane contaminating representations learned for another.

3.3 Per-Task Attention-Weighted Slice Pooling

Not every slice contains diagnostically relevant tissue, and crucially, slices informative for ACL tears differ from those informative for meniscal damage. A

shared pooling step is forced to compromise. Therefore, we assign each task its own two-layer MLP (Tanh activation) that maps each slice's 768-d `[CLS]` vector to a scalar score; a softmax over all slices yields per-task slice weights, and the task-specific plane embedding is the resulting weighted sum. The three pooling heads (ACL, meniscus, abnormal) have entirely separate parameters, allowing the model to up-weight posterior-horn slices for meniscal evaluation independently of mid-joint slices for ACL continuity, without slice-level pathology labels.

3.4 Per-Task Cross-Plane Attention Fusion

Prior multi-plane systems combine per-plane scores after each plane has been scored independently, so representations never interact. Even cross-plane attention designs that operate on a single shared pooled vector conflate slice evidence across tasks before inter-plane reasoning occurs. We apply cross-plane attention separately per task, fusing only the plane evidence selected by that task's pooling head.

For each task, the three 768-d task-specific plane embeddings are linearly projected to 384-d and augmented with a learnable per-plane identity vector (axial, coronal, sagittal). Without this embedding, self-attention treats the three tokens as interchangeable. The augmented tokens pass through a compact two-layer Transformer encoder (6 heads, pre-LN, GELU), which is *shared* across tasks to limit parameter count while forcing task-agnostic cross-plane reasoning; task differentiation lives entirely in the upstream pooling heads. The three updated tokens are mean-pooled to a 384-d fused representation, fed to a shared extractor ($384 \to 256 \to 128$, BN-ReLU-dropout) and three independent task heads ($128 \to 64 \to 1$, sigmoid) for ACL, meniscus, and abnormality. Calling the encoder three times — once per task on task-specific evidence — structurally ties inter-plane reasoning to each pathology.

3.5 Training Objective

We minimise a class-balanced, task-weighted binary cross-entropy loss with label smoothing ($\epsilon = 0.05$). The meniscus task is weighted 1.8 (others 1.0) to reflect its greater difficulty; within each task, positive-class loss is scaled by the negative-to-positive ratio. Augmentation includes Mixup [12] ($\alpha = 0.2$), horizontal flip ($p = 0.50$), rotation $\pm 20°$ ($p = 0.70$), brightness jitter $[0.8, 1.2]$ ($p = 0.80$), Gaussian noise ($p = 0.40$), and blur ($p = 0.30$). AdamW uses decoupled rates: 5×10^{-5} (backbone, WD 10^{-4}) and 5×10^{-4} (heads, WD 5×10^{-4}), with 1-D parameters excluded from decay. Cosine annealing follows a 3-epoch linear warmup; gradient clipping at L2 norm ≤ 1.0 and early stopping (patience 12) prevent overfitting.

3.6 Explainability Pipeline

EigenCAM [7] is computed at block 11 of each fine-tuned encoder for each task, producing a per-slice spatial activation map. Maps are stacked into a

Table 1. MRNet dataset summary. Prevalences computed on the 1,130 training cases.

Split	Cases	ACL+	Meniscus+	Abnormal+
Train	1,130	20.5%	35.0%	60.7%
Validation	120	19.2%	33.3%	58.3%

saliency volume, normalised to $[0, 1]$, and blended with MRI slices (weight 0.4, jet colormap). The three highest-activation overlays, the predicted probability vector, and a risk category are passed to Gemini [4] via its vision-language API with a fixed prompt instructing it to: (1) summarise findings in plain English; (2) relate highlighted anatomy to the predicted pathology; (3) recommend next clinical steps; and (4) append a standardised AI disclaimer. Grounding the language model on saliency overlays ensures generated text describes the evidence the classifier used, not incidental image features.

4 Experiments

4.1 Dataset

MRNet [1] contains 1,370 knee MRI examinations, each acquired in three orthogonal planes (Table 1). Following the standard split, we train on 1,130 cases and evaluate on 120; held-out test labels are not publicly available. All splits are performed at the subject level to prevent information leakage between train and validation.

4.2 Implementation

All experiments run in PyTorch with mixed-precision training (AMP) on a single NVIDIA A100 GPU. Slices are resized to 224×224 and normalised with ImageNet channel statistics. At inference, we average predictions over four test-time augmentation variants (the original scan plus three horizontal-flip combinations). The effective batch size is 8, achieved via a physical batch size of 4 with two-step gradient accumulation.

4.3 Quantitative Results

Table 2 reports AUC on the 120-case validation set. Two ablations isolate the contribution of each architectural decision. The "w/o cross-plane attention" row replaces the Transformer fusion block with concatenation of the three plane embeddings followed by an MLP, keeping per-task pooling intact; this isolates the benefit of cross-plane attention specifically. The "shared slice pooling" row restores cross-plane attention but replaces the three per-task pooling heads with a single shared pooling head, isolating the benefit of task-specific slice selection. Together these rows show that both components contribute independently: cross-plane attention improves inter-plane evidence integration, while per-task pooling ensures each task selects the slices most anatomically relevant to its own pathology.

Table 2. AUC on MRNet validation set (120 cases).

Method	ACL	Meniscus	Abnormal
MRNet (AlexNet, multi-plane) [1]†	0.965	0.847	0.937
Kara & Hardalaç (ResNet50) [6]†	0.894	0.799	0.932
ViT-B/16 (ImageNet, shared fusion)*	0.901	0.838	0.921
KneeDINO (ours)	0.942	0.871	0.955
w/o cross-plane attention (concat.+MLP)	0.920	0.814	0.931
shared slice pooling (no per-task heads)	0.925	0.839	0.950
w/o triple plane (axial only)	0.854	0.794	0.892

†Results from original paper. *Our reimplementation.

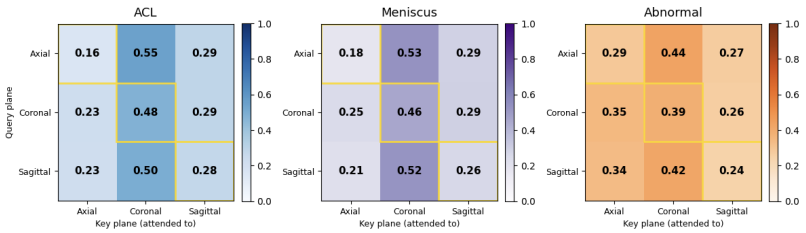

Fig. 2. Validation-set-averaged cross-plane attention weights (final encoder layer, mean over all heads and 120 cases), shown separately per task. Rows are query planes; columns are key planes. All three tasks attend most strongly to the coronal key plane; no matrix is diagonally dominant, confirming genuine cross-plane information exchange. The abnormal task (right) is most distributed, consistent with general structural abnormality manifesting across all views.

4.4 Cross-Plane Attention Patterns

We average attention weights from the final fusion encoder layer across all 120 validation cases, separately per task, yielding three 3×3 matrices (Fig. 2). Unlike scalar late-fusion weights, these record how much each plane's representation incorporated evidence from the others during feature formation. All three tasks show consistently stronger attention toward the coronal key plane (ACL: 0.55; meniscus: 0.53; abnormal: 0.44 for axial query rows), consistent with the coronal view's clinical primacy for joint-space and ligament evaluation. Notably, no matrix is diagonally dominant: off-diagonal weights are substantial in all tasks, confirming that genuine cross-plane information exchange occurs rather than each plane simply attending to itself. The abnormal task produces the most distributed pattern (range 0.24–0.44), reflecting that general structural abnormality can manifest across all planes equally.

Fig. 3 shows EigenCAM overlays for a representative validation case (Case 105) with predicted probabilities of 56.8% (ACL), 91.4% (meniscus), and 99.2% (abnormal)—a high-confidence meniscus and general-abnormality case rather than a clear-cut ACL-positive one. Importantly, EigenCAM is computed across

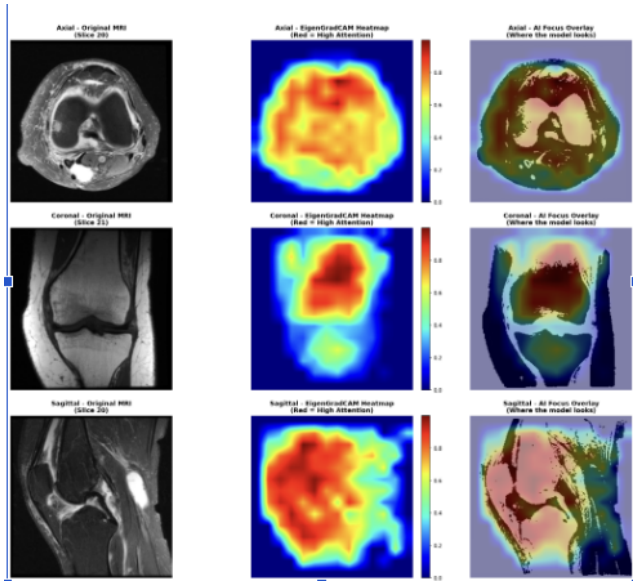

Fig. 3. EigenCAM saliency overlays for validation Case 105 (predicted probabilities: ACL 56.8%, meniscus 91.4%, abnormal 99.2%). EigenCAM is computed volumetrically across all slices per plane; each row shows the highest-activation slice for that plane (axial, coronal, sagittal). The left column shows the original MRI slice, the centre column shows the raw EigenCAM activation map (colour scale 0–1, red indicating high attention), and the right column shows the blended overlay. Coronal and sagittal activation is concentrated near the central joint space, consistent with high meniscus and abnormality confidence; axial activation is more diffuse, consistent with the weaker near-threshold ACL prediction.

all slices in each volume rather than a single representative slice, producing a full volumetric saliency map; the slices shown are the highest-activation examples from each plane, avoiding cherry-picking while preserving readability. Activation is broadly distributed across all three planes rather than sharply localised to a single anatomical landmark, with the coronal and sagittal views showing the most concentrated high-attention regions (dark red) near the central joint space, consistent with the model's high meniscus and abnormality confidence. The axial heatmap is more diffuse, in line with its weaker, near-threshold contribution to the ACL prediction for this case. The corresponding Gemini-generated summary localises its language to the same regions: "Elevated signal is concentrated around the central joint space in the coronal and sagittal views, consistent with meniscal and general structural abnormality; ACL-region signal is comparatively weaker." All generated summaries carry a standardised disclaimer that outputs are AI-generated and require clinical review before use.

5 Discussion

The results confirm that both per-task slice pooling and cross-plane attention contribute independently to performance. The anatomical motivation for per-task pooling is direct: ACL-informative slices differ from meniscus-informative ones, and a shared head cannot optimise for both simultaneously. Increasing the meniscus task weight to 1.8 and adding label smoothing further address the imbalanced difficulty of that class. Keeping backbone encoders independent prevents gradient contamination across planes, while sharing the fusion encoder across tasks keeps parameter count suitable for MRNet's 1,130 training cases. The attention analysis (Sect. 4.4) reveals consistent coronal dominance across all tasks, with no diagonal-dominant self-attention pattern, providing mechanistic evidence that the model genuinely integrates cross-plane information rather than routing each plane to its own head.

KneeDINO targets triage and screening, not replacement of radiologist review. The EigenCAM-to-Gemini pipeline provides spatially grounded explanations supporting clinical interpretability requirements. Limitations include: slice-level `[CLS]` aggregation may miss diffuse pathologies; with only 120 validation cases, AUC estimates carry substantial uncertainty and bootstrap CIs are needed before drawing strong conclusions from small inter-method gaps; and external validation across scanner vendors is required before clinical deployment.

6 Conclusion

We presented KneeDINO, a multi-label knee MRI classification system built around per-task cross-plane attention fusion. Three independent RadioDINO encoders feed per-task attention-pooling heads; a shared cross-plane Transformer encoder is applied once per task, letting each pathology attend across planes using its own slice evidence. A learnable per-plane identity embedding and pre-LN stabilise the shallow fusion Transformer. Attention analysis reveals consistent coronal dominance and genuine cross-plane information exchange across all three tasks. An EigenCAM-to-Gemini pipeline grounds plain-English clinical summaries in task-specific visual evidence. Tested on MRNet, KneeDINO achieves a mean AUC of 0.923 across three pathologies, enabling automated accurate detection of knee pathologies.

Reproducibility Statement. The MRNet dataset [1] and the pretrained RadioDINO [8] checkpoint used in this work are both publicly available. All architectural details, hyperparameters, and training procedures are fully specified in Sect. 3.

References

1. Bien, N., Rajpurkar, P., Ball, R.L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B.N., Yeom, K.W., Shpanskaya, K., Halabi, S., Zucker, E., Fanton, G., Amanatullah, D.F., Beaulieu, C.F., Riley, G.M., Stewart, R.J., Blankenberg, F.G., Larson,

D.B., Jones, R.H., Langlotz, C.P., Ng, A.Y., Lungren, M.P.: Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLOS Medicine **15**(11), e1002699 (2018)

2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (2021)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16×16 words: Transformers for image recognition at scale. In: Proc. International Conference on Learning Representations (ICLR) (2021)
4. Gemini Team, Google: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
5. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proc. IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017)
6. Kara, A.C., Hardalaç, F.: Detection and classification of knee injuries from MR images using the MRNet dataset with progressively operating deep learning methods. Machine Learning and Knowledge Extraction **3**(4), 1009–1029 (2021). `https://doi.org/10.3390/make3040050`
7. Muhammad, M.B., Yeasin, M.: Eigen-CAM: Class activation map using principal components. In: Proc. International Joint Conference on Neural Networks (IJCNN) (2020)
8. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., Wetscherek, M., Codella, N., Hyland, S.L., Alvarez-Valle, J., Oktay, O.: RAD-DINO: Exploring scalable medical image encoders beyond text supervision. arXiv preprint arXiv:2401.10815 (2024)
9. Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. BMJ **359**, j4683 (2017)
10. Sanders, T.L., Maradit Kremers, H., Bryan, A.J., Larson, D.R., Dahm, D.L., Levy, B.A., Stuart, M.J., Krych, A.J.: Incidence of anterior cruciate ligament tears and reconstruction: A 21-year population-based study. The American Journal of Sports Medicine **44**(6), 1502–1507 (2016)
11. Nicolson, A., Dowling, J., Koopman, B.: Improving chest X-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023)
12. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: Proc. International Conference on Learning Representations (ICLR) (2018)