HandXFM: Semantic-Structural Distillation for Hand Radiograph Foundation Models

Yuxi Long[1], **Ganlin Feng**[1], **Lianghong Chen**[1], **Liam J. O'Neil**[2], **Carol A. Hitchon**[3], **Pingzhao Hu**[1,4,*]

[1]Department of Computer Science, Western University, London, ON, Canada
[2]Department of Medicine, University of Manitoba, Winnipeg, MB, Canada
[3]Department of Internal Medicine, University of Manitoba, Winnipeg, MB, Canada
[4]Department of Biochemistry, Western University, London, ON, Canada

ylong66@uwo.ca, phu49@uwo.ca

Abstract

Hand radiographs support diverse clinical tasks, including skeletal maturity assessment, rheumatoid arthritis scoring, abnormality screening, and anatomical localization, yet existing models are typically trained separately for each task and dataset. We present HandXFM, a domain-specific foundation model that leverages cross-modality knowledge for comprehensive hand X-ray understanding. HandXFM distills semantic priors from BiomedCLIP and structural cues from a chest X-ray Vision Transformer (ViT) through a unified alignment framework, integrating complementary medical knowledge in an interpretable way. After pretraining, HandXFM can be adapted to a range of musculoskeletal tasks, including bone age estimation, SvH score prediction, abnormality detection, joint localization, bone segmentation, and visual question answering. It consistently outperforms task-specific and single-expert baselines, achieving up to 0.11 improvement in Pearson correlation coefficient (PCC) for bone age prediction and a 12% accuracy gain in abnormality classification. Moreover, HandXFM generalizes to unseen datasets and yields cross-attention maps linking clinical terms to anatomical regions. These results demonstrate that multi-expert distillation effectively unifies semantic and structural supervision within a single pretrained model, establishing HandXFM as an interpretable and generalizable foundation model for hand radiograph analysis.

1 Introduction

Hand radiography is a cornerstone modality in musculoskeletal imaging, used to assess growth and development, diagnose arthritis and trauma, quantify joint damage, and guide treatment. Clinical tasks including bone age estimation, SvH (Sharp/van der Heijde) scoring [42], joint detection, bone segmentation, and abnormality classification rely on hand X-rays as structural evidence. Despite their importance, computational models for hand radiographs are still largely developed separately for each task and dataset, limiting transfer across institutions, supervision levels, and clinical objectives.

Existing approaches are mostly task-specific and annotation-dependent [11, 44], making it difficult to build a unified representation backbone that captures shared anatomical and radiographic patterns. In contrast to the rapid progress of large-scale foundation models in natural and multimodal imagery [29, 19], the hand X-ray domain still lacks a reusable pretrained model tailored to its structure-dominant nature. Prior medical foundation models provide useful semantic priors [47, 43], but often rely on cross-modal supervision or single-expert training, leading to limited structural reasoning and weaker

*Corresponding author. Department of Computer Science, Western University, 1400 Western Road, London, Ontario N6G 2V4, Canada. E-mail: phu49@uwo.ca (P.H.)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

generalization across datasets. They also tend to overlook fine-grained anatomical cues such as bone boundaries, joint gaps, and cortical contours, which are critical for hand radiograph interpretation.

To address these limitations, we present **HandXFM**, a domain-specific foundation model for hand X-rays. HandXFM integrates a transformer-based student encoder with expert teacher models, i.e., domain-specialized pretrained models. It adopts a multi-expert masked latent alignment framework [45] combining alignment and reconstruction objectives, edge-aware masking for structural boundaries, and an Exponential Moving Average (EMA) teacher [14] for stable supervision. These design choices are motivated by key characteristics of hand radiographs: subtle boundary-sensitive findings, heterogeneous downstream supervision, and cross-dataset transfer.

Once pretrained, HandXFM can be adapted to new datasets as a backbone for downstream tasks including regression, classification, detection, segmentation, and VQA. It can be fine-tuned with minimal supervision, making it suitable for multi-center medical pipelines. Extensive experiments show that HandXFM outperforms prior self-supervised and foundation baselines across tasks and generalizes to unseen datasets such as the Digital Hand Atlas. Visualizations indicate that its attention aligns with clinically meaningful regions, supporting anatomical grounding and interpretability.

Our main contributions are summarized as follows:

- We curate a large multi-source hand X-ray dataset and propose HandXFM, a domain-specific foundation model for hand radiographs, capable of learning unified radiographic representations.
- We introduce a multi-expert masked latent alignment framework that fuses semantic and structural priors from experts through alignment, reconstruction, edge-aware masking, and EMA stabilization.
- We validate HandXFM across diverse downstream tasks, including regression, classification, detection, segmentation, and VQA, demonstrating transferability and anatomical interpretability.

2 Related Work

Recent advances in large-scale visual pretraining have produced powerful foundation models such as DINOv3 [31], MAE [19], CLIP [29], and EVA [12]. They learn general-purpose representations from natural images through contrastive or masked objectives, achieving strong transferability. However, assumptions on color textures, open-world semantics, and text-aligned supervision limit their applicability to grayscale, structure-dominant medical domains.

Medical foundation models such as MedCLIP [43], BiomedCLIP [47], and M3AE [9] introduce semantic priors and cross-modal transfer [38, 25]. Yet many rely on textual supervision [3] or single-domain training [28, 46], with limited anatomical or structure-aware reasoning. Recent musculoskeletal and multi-anatomy X-ray foundation models, such as SKELEX [22] and XR-0 [32], improve radiographic transfer but are not tailored to fine-grained hand-specific reasoning.

Hand X-ray studies remain task-specific, covering bone age prediction [18, 2], rheumatoid arthritis scoring [37, 26], and abnormality classification [30, 1]. These CNN- or ViT-based models perform well within individual domains but generalize poorly across datasets and tasks. To the best of our knowledge, no foundation model has been developed specifically for hand X-rays, and no existing approach unifies multi-task hand radiography in a single framework.

Unlike prior methods, HandXFM bridges semantic foundation models and anatomical hand analysis. It combines multi-expert latent alignment, edge-aware masking, and EMA consistency to learn unified, structure-informed representations from unlabeled hand X-rays.

3 Method

HandXFM learns anatomically meaningful representations from unlabeled hand radiographs. Instead of relying on a single self-supervised objective, it combines visual and semantic supervision within a unified pretraining framework. The model aligns its latent space with multiple medical experts, reconstructs occluded features through contextual prediction, and maintains stable representation dynamics over time. An overview of the HandXFM framework is illustrated in Figure 1.

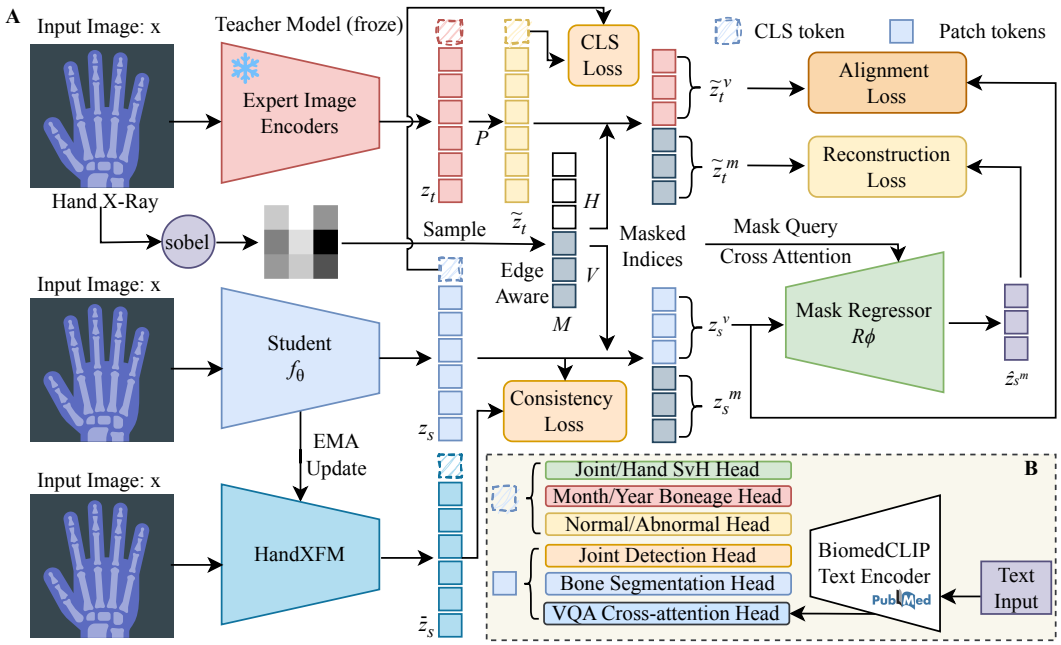

Figure 1: **Overview of the proposed HandXFM framework. (A) Pretraining architecture.** HandXFM jointly aligns the student encoder with multiple frozen medical experts while reconstructing masked latent features through the regressor. Edge-aware masking emphasizes structurally salient regions using Sobel gradients. The student learns from complementary objectives: multi-expert alignment loss, masked latent reconstruction loss, and EMA-based consistency loss, with an additional CLS-level supervision to preserve global semantics. **(B) Downstream adaptation.** After pretraining, the resulting foundation model serves as a universal encoder for diverse radiographic downstream tasks, including bone segmentation, joint detection, normal/abnormal classification, bone age prediction, SvH scoring, and VQA. The pretrained representation provides strong anatomical priors that transfer effectively across tasks and datasets.

3.1 Preliminary

Given an unlabeled hand X-ray image $x \in \mathbb{R}^{3\times224\times224}$, preprocessed as described in the Supp.A.2, the student encoder f_θ produces a set of latent embeddings:

$$z_s = f_\theta(x) = [z_{\text{cls}}; z_1, z_2, \ldots, z_N], \quad z_i \in \mathbb{R}^D, \tag{1}$$

where z_{cls} represents the global token and $\{z_i\}$ denote patch-level embeddings. Following DINOv3 [31], the image is divided into $N = 14 \times 14 = 196$ patches. Frozen medical experts $\{t_j\}_{j=1}^{J}$, each pretrained on different imaging domains, provide complementary latent targets:

$$z_{t_j} = t_j(x) = [z_{\text{cls}}^{(j)}; z_1^{(j)}, \ldots, z_N^{(j)}]. \tag{2}$$

Each frozen expert feature is mapped into the student feature space by a learnable linear projection, anchoring supervision to pretrained medical representations while introducing complementary semantic and structural cues:

$$\tilde{z}^{(j)} = P_j z_{t_j}, \quad P_j : \mathbb{R}^{D_j} \to \mathbb{R}^D. \tag{3}$$

To encourage latent prediction, a binary mask $M \in \{0,1\}^N$ is sampled per image. The visible subset $\mathcal{V} = \{i \mid M_i = 0\}$ is used for alignment, while the masked subset $\mathcal{H} = \{i \mid M_i = 1\}$ is used for reconstruction.

3.2 Multi-Expert Alignment

The alignment objective enables the student to absorb knowledge from multiple medical experts and form a domain-agnostic representation space. Each expert contributes a distinct prior: a text–vision

model such as BiomedCLIP [47] provides semantic abstraction, while a radiographic ViT [8] offers radiology-typical spatial–tonal cues. By jointly aligning to these signals, the student can encode both conceptual and style consistency. Visible patch tokens align with the corresponding expert projections through a mean-squared latent alignment loss:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \frac{1}{J} \sum_{j=1}^{J} \| z_i - \tilde{z}_i^{(j)} \|_2^2. \tag{4}$$

This encourages the model to preserve local relationships that remain semantically consistent across domains. To maintain global coherence, we distill the image-level [CLS] token from all experts:

$$\mathcal{L}_{\text{cls}} = \frac{1}{J} \sum_{j=1}^{J} \| z_{\text{cls}} - \tilde{z}_{\text{cls}}^{(j)} \|_2^2. \tag{5}$$

Together, these objectives transfer both fine-grained spatial correspondences and holistic semantics into the student encoder. By jointly attending to multiple experts, the model learns to represent hand anatomy agnostic to dataset origin or modality differences. The resulting latent space is both semantically organized and structurally smooth, allowing consistent mapping of similar joints, bones, and pathological cues in varying imaging conditions. This multi-expert alignment allows HandXFM to generalize to diverse downstream tasks without task-specific supervision.

3.3 Masked Latent Reconstruction

To complement expert alignment, HandXFM performs token-level masked latent reconstruction by withholding a subset of encoder output tokens and predicting their expert-aligned targets from the remaining tokens. This objective regularizes the student toward spatial continuity and anatomical dependencies, rather than serving as input-level masked image modeling.

Let $z_s^v = \{z_i \mid i \in \mathcal{V}\}$ denote visible patch embeddings and $z_s^m = \{z_i \mid i \in \mathcal{H}\}$ the masked ones. We employ a cross-attention regressor $\mathcal{R}_\phi$ that uses z_s^v as keys and values while querying with learnable mask tokens plus positional embeddings, yielding reconstructed latent representations $\hat{z}_s^m = \mathcal{R}_\phi(z_s^v)$.

The reconstruction loss compares the predicted masked tokens with the expert-projected targets:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \frac{1}{J} \sum_{j=1}^{J} \| \hat{z}_{s,i}^m - \tilde{z}_i^{(j)} \|_2^2. \tag{6}$$

Unlike pixel-level reconstruction, this formulation constrains the model to infer feature relationships.

Edge-Aware Masking. To ensure that reconstruction focuses on informative regions, we employ an edge-aware masking strategy assigning higher masking probability to patches with strong gradients, typically corresponding to bone and joint boundaries. Let p_i denote the normalized edge strength of patch i. The sampling distribution is:

$$q_i = (1-\alpha)\frac{p_i}{\sum_k p_k} + \alpha\frac{1}{N}, \quad \mathcal{H} \sim \text{SampleWithoutReplacement}(q, \lfloor rN \rfloor), \tag{7}$$

where r is the mask ratio and $\alpha \in [0,1]$ controls the trade-off between edge-based and random masking. By emphasizing structurally salient regions, the model learns geometry-aware features and spatial continuity. The masked latent regression captures inter-bone and joint relationships, improving sensitivity to structural deformation and localization errors for segmentation, landmark detection, and joint-level scoring.

3.4 EMA Consistency

To stabilize optimization and avoid overfitting to instantaneous expert outputs, we use an exponential moving average (EMA) of the student encoder as a smoothed teacher providing stable targets for self-consistency. After each update step, parameters are updated as

$$\theta_m \leftarrow \mu\,\theta_m + (1-\mu)\,\theta, \quad \mu = 0.996, \tag{8}$$

where μ is a constant momentum coefficient controlling update smoothness. Following standard self-distillation frameworks such as DINO and BYOL, we empirically set $\mu = 0.996$ for stable evolution

Algorithm 1 HandXFM pretraining with multi-expert masked latent alignment

Input: Unlabeled hand X-ray dataset $\mathcal{D}$; student encoder f_θ; experts $\{t_j\}_{j=1}^{J}$; regressor $\mathcal{R}_\phi$; projection heads $\{P_j\}$; EMA encoder f_{θ_m}; loss weights $\lambda_{\text{recon}}, \lambda_{\text{ema}}, \lambda_{\text{cls}}$.
Output: Student encoder f_θ and EMA teacher f_{θ_m}.
Initialize $\theta, \phi, \{P_j\}$; set $\theta_m \leftarrow \theta$.
for each minibatch $x \sim \mathcal{D}$ **do**
  Sample binary mask M using edge-aware PatchMasker (Equation (7)).
  Obtain visible index set $\mathcal{V}$ and masked index set $\mathcal{H}$.
  Compute student tokens $z_s = f_\theta(x) = [z_{\text{cls}}; z_1, \ldots, z_N]$.
  for each expert t_j **do**
    Compute expert tokens $z_{t_j} = t_j(x)$ and project $\tilde{z}^{(j)} = P_j z_{t_j}$.
  end for
  Compute alignment loss $\mathcal{L}_{\text{align}}$ (Equation (4)) and CLS loss $\mathcal{L}_{\text{cls}}$ (Equation (5)).
  Predict masked tokens $\hat{z}_s^m = \mathcal{R}_\phi(z_s^v)$ and compute reconstruction loss $\mathcal{L}_{\text{recon}}$ (Equation (6)).
  Compute EMA teacher tokens $\bar{z}_s = f_{\theta_m}(x)$ and loss $\mathcal{L}_{\text{ema}}$ (Equation (9)).
  Compute total loss $\mathcal{L}_{\text{total}}$ (Equation (10)) and update $\theta, \phi, \{P_j\}$ with AdamW.
  Update EMA teacher: $\theta_m \leftarrow \mu\,\theta_m + (1-\mu)\,\theta$.
end for

of the EMA teacher. The EMA encoder produces latent features $\bar{z}_s = f_{\theta_m}(x) = [\bar{z}_{\text{cls}}; \bar{z}_1, \ldots, \bar{z}_N]$. We apply a consistency loss between the current and EMA representations:

$$\mathcal{L}_{\text{ema}} = \frac{1}{|\mathcal{V}| + |\mathcal{H}|} \sum_{i \in \mathcal{V} \cup \mathcal{H}} \| z_{s,i} - \bar{z}_{s,i} \|_2^2. \tag{9}$$

This constraint regularizes feature evolution. The temporal averaging of encoder weights acts as a low-pass filter, preventing instability from stochastic expert signals or local minima. It encourages gradual refinement of high-level concepts and maintains manifold continuity.

The overall pretraining objective combines alignment, reconstruction, and consistency terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{align}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \lambda_{\text{ema}}\mathcal{L}_{\text{ema}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}. \tag{10}$$

The full pretraining loop is summarized in Algorithm 1, and implementation details are provided in Supp. A.

4 Experiments

4.1 Datasets

To study hand radiograph foundation modeling under realistic clinical heterogeneity, we assemble a large corpus spanning development, disease grading, abnormality screening, localization, and structural understanding. The datasets differ in patient population, annotation granularity, label units, and downstream objectives, reflecting the fragmented nature of real-world hand radiograph use. To our knowledge, this is the largest open-access benchmark collection in this domain. **Pretraining uses training splits only; all val/test splits and external-validation datasets are excluded**.

- **Bone Age Prediction.** We evaluate bone age estimation on three datasets: **RSNA** [18] (14,036 radiographs; months), **RHPE** [16] (6,279 images; months + sex), and **Digital Hand Atlas** [15] (1,103 images; years + sex), used for external validation.
- **SvH Score Regression.** To assess rheumatoid arthritis severity, we use **AIRINIIReumHands** [37] (330 radiographs; joint-level erosion and JSN annotations) and the **Anhui** dataset [44] (3,818 bilateral radiographs with image-level averaged scores).
- **Abnormality Classification.** We use the MURA hand subset [30] (6,003 radiographs labeled normal/abnormal), restricting to the hand category.
- **Closed-end VQA.** We use 108 hand X-rays from **MedPix** [27], treating captions as questions and titles as answers.

Table 1: Regression results on SvH scoring and bone age prediction.

SvH Scoring					Bone Age Prediction				
AIRINIIReumHands (Erosion)					**RSNA (Monthly Labels)**				
Method	RMSE↓	MAE↓	R^2 ↑	PCC↑	Method	RMSE↓	MAE↓	R^2 ↑	PCC↑
RheumaVIT [37]	0.445	0.360	0.373	0.766	CCR-BAA [24]	11.911	8.873	0.918	0.959
DINOv3 [31]	0.525	0.300	0.127	0.426	AE-BoNet [34]	13.799	10.190	0.891	0.944
RadFM [46]	0.409	0.231	0.403	0.797	DINOv3 [31]	12.702	9.106	0.907	0.953
RadCLIP [25]	0.402	0.136	0.487	0.739	RadFM [46]	18.395	13.752	0.805	0.898
HandXFM	**0.298**	**0.111**	**0.719**	**0.929**	RadCLIP [25]	12.773	9.379	0.906	0.954
					HandXFM	**9.932**	**7.343**	**0.943**	**0.972**
AIRINIIReumHands (JSN)					**RHPE (Monthly Labels)**				
Method	RMSE↓	MAE↓	R^2 ↑	PCC↑	Method	RMSE↓	MAE↓	R^2 ↑	PCC↑
RheumaVIT [37]	0.758	0.601	0.538	0.847	ConvNeXt [5]	14.288	11.497	0.877	0.965
DINOv3 [31]	0.714	**0.558**	0.591	0.803	BoNet [11]	11.246	8.708	0.926	0.962
RadFM [46]	0.792	0.618	0.496	0.740	DINOv3 [31]	27.828	21.856	0.550	0.790
RadCLIP [25]	0.792	0.635	0.497	0.861	RadFM [46]	23.887	18.233	0.668	0.820
HandXFM	**0.705**	0.608	**0.601**	**0.885**	RadCLIP [25]	12.320	9.656	0.912	0.955
					HandXFM	**11.073**	**8.363**	**0.929**	**0.966**
Anhui (Whole-hands Image)					**DHA (External Validation; Year Labels)**				
Method	RMSE↓	MAE↓	R^2 ↑	PCC↑	Method	RMSE↓	MAE↓	R^2 ↑	PCC↑
ResNet dw-50 [44]	24.150	16.688	0.669	0.835	CCR-BAA [24]	1.528	1.193	0.899	0.952
ResNet-50 [4]	30.967	22.185	0.456	0.678	AE-BoNet [34]	1.636	1.233	0.884	0.941
DINOv3 [31]	31.232	20.301	0.447	0.702	DINOv3 [31]	1.909	1.418	0.842	0.930
RadFM [46]	30.513	20.595	0.472	0.721	RadFM [46]	1.799	1.365	0.859	0.929
RadCLIP [25]	22.671	14.642	0.709	0.869	RadCLIP [25]	1.864	1.420	0.849	0.931
HandXFM	**18.231**	**12.356**	**0.812**	**0.903**	HandXFM	**1.269**	**0.990**	**0.930**	**0.965**

- **Joint Detection and Bone Segmentation (New Annotations).** We curate two RSNA-based datasets: a joint detection set (6,130 images with 17 landmark boxes; Table C33) and a bone segmentation set (549 images with 19-class masks; Table C34). Annotations follow *Terminologia Anatomica* [13] and are aligned with *RSNA RadLex* [23].

Data separation and baseline protocols. Only the designated training partitions contribute images to unlabeled HandXFM pretraining; validation, test, and external-evaluation images are excluded. For RSNA-derived localization benchmarks, detector and segmenter training use RSNA training images, while evaluation uses the disjoint validation partition. Unless otherwise specified, pretrained visual baselines are fully fine-tuned using the same downstream partitions and task-specific evaluation protocols. Detection and segmentation backbones share the same DETR-style downstream architecture, whereas results adopted from published task-specific systems are reported separately. Detailed dataset allocation and overlap controls are provided in Table C21.

4.2 Downstream Tasks

To demonstrate HandXFM's versatility, we apply it as an encoder backbone for diverse downstream tasks (regression, classification, segmentation, detection, and VQA) using lightweight task-specific heads. These heads are intentionally small, ensuring performance reflects representation quality rather than head capacity. The following subsections describe each task and its protocol.

4.2.1 SvH Scoring

We evaluate HandXFM on two rheumatoid arthritis datasets with different supervision levels: AIRINIIReumHands [37] (joint-level) and Anhui [44] (image-level). Because of this granularity difference, we design two lightweight regression heads built on the same pretrained HandXFM encoder.

- **Joint-level head.** For AIRINIIReumHands, we attach two parallel regression branches to predict erosion and JSN scores from the pooled representation. Missing joint annotations are handled with a masked Smooth-ℓ_1 loss so each image contributes supervision only where labels exist.

Table 2: MURA XR_HAND classification (normal vs. abnormal).

Method	AUC↑	AUPRC↑	ACC↑	F1↑	P↑	R↑
ST-SSAD [36]	0.5966	0.4969	0.5978	0.0887	0.6429	0.0476
DenseNet-169 [35]	0.8234	0.7788	0.7109	0.6826	0.6217	0.7566
DINOv3 [31]	0.7405	0.6667	0.6741	0.6589	0.5900	0.7460
RadFM [46]	0.6114	0.5214	0.4891	0.5983	0.4419	**0.9259**
RadCLIP [25]	0.8460	0.8084	0.7500	0.6000	**0.9231**	0.4444
HandXFM (ours)	**0.8599**	**0.8213**	**0.7857**	**0.6903**	0.8843	0.5661

- **Whole-hands head.** For the Anhui dataset, we use a shallow regression head on the CLS token to predict the global SvH score of both hands.

We report RMSE, MAE, R^2, and PCC, which together assess absolute regression error, explained variance, and agreement with radiologist scoring. The results are summarized in Table 1. HandXFM shows consistent improvements across both joint-level and whole-hand SvH scoring tasks.

On AIRINIIReumHands, HandXFM achieves PCCs of 0.929 (erosion) and 0.885 (JSN), outperforming all baselines. Although some baselines achieve lower JSN MAE, HandXFM obtains the highest PCC among the evaluated methods, indicating stronger correlation with the reference damage scores. We report a multi-seed stability analysis for it in Supp. B.8, given its relatively small dataset size.

On the Anhui dataset (single score for both hands), HandXFM reduces RMSE from 24–31 to 18.2 and improves R^2 to 0.812, surpassing RadCLIP (22.67, 0.709) and RadFM (30.51, 0.472). These gains suggest a stronger representation that aggregates bilateral lesions into a global disease estimate.

4.2.2 Bone Age Prediction

We assess bone age estimation, a benchmark for maturation, on three datasets: RSNA, RHPE, and Digital Hand Atlas (DHA) for external validation. Acknowledging sexual dimorphism, regression heads jointly encode the image and a sex variable, which is projected and concatenated with the image embedding. All models use a Smooth-ℓ_1 loss.

- **RSNA.** As shown in Table 1, HandXFM achieves the lowest MAE (7.34 months) and highest PCC (0.972), outperforming hand-specific pipelines (AE-BoNet, CCR-BAA) and foundation-model baselines (RadFM, RadCLIP). Strong R^2 and PCC indicate preserved maturity ordering beyond absolute error, suggesting multi-expert pretraining captures subtle developmental cues.
- **RHPE.** HandXFM achieves the best overall performance on RHPE, reducing MAE to 8.36 months while reaching $R^2 = 0.929$ and PCC=0.966. It outperforms DINOv3, RadFM, RadCLIP, ConvNeXt, and BoNet, indicating that the pretrained representation effectively captures bone structure and epiphyseal development for pediatric bone age estimation.
- **External validation on DHA.** Training on RSNA (months) and evaluating on DHA (years), HandXFM maintains high correlation (PCC=0.965, $R^2 = 0.930$). As the head is not tuned on DHA, this reflects strong backbone generalization and stable anatomical representations.

4.2.3 Abnormality Classification

We evaluate abnormality classification on the MURA hand subset and summarize the results in Table 2. HandXFM serves as the visual encoder, followed by a CLS-only head with two linear layers applied to the encoder's CLS token. We train with BCE-with-logits and report AUC, AUPRC, ACC, F1, precision (P), and recall (R), using a weighted sampler to address class imbalance. Baselines include a self-taught semi-supervised anomaly detection method (ST-SSAD) [36], a supervised DenseNet-169 pipeline [35], and radiology foundation models (RadFM [46], RadCLIP [25]).

HandXFM achieves the best performance on the MURA hand subset. Compared with the DenseNet-169 pipeline, it improves AUC from 0.823 to 0.860, AUPRC from 0.779 to 0.821, and accuracy from 71.1% to 78.6%, while also achieving the highest F1 (0.690). With a CLS-only head, HandXFM maintains high precision (P = 0.884) and competitive recall (R = 0.566), yielding a strong precision–recall trade-off. In contrast, RadCLIP favors precision (P = 0.923; R = 0.444) and RadFM favors recall (R = 0.926; P = 0.442), leading to weaker discrimination. Relative to DINOv3-only, multi-expert pretraining adds anatomical priors and improves robustness to heterogeneous image quality.

Table 3: Joint detection and bone segmentation results.

Joint detection (17-class)

Method	$mAP_{50:95}$ ↑	mAP_{50} ↑	mAP_{75} ↑
ResNet50 + DETR	0.5303	0.9433	0.5409
ViT + DETR	0.5617	0.9508	0.6057
ConvNeXt + DETR	0.5517	0.9468	0.6023
DINOv3 + DETR (native)	0.5857	0.9630	0.6512
YOLOv12	0.7391	**0.9903**	0.8752
DINOv3 + DETR (adapted)	0.5900	0.9650	0.6487
HandXFM	**0.7636**	0.9802	**0.9026**

Bone segmentation (19-class)

Method	P↑	R↑	$mAP_{50:95}$ ↑
ResNet50 + DETR	0.9426	0.9136	0.4191
ViT + DETR	0.9437	0.9073	0.3482
ConvNeXt + DETR	0.9454	0.9173	0.4356
DINOv3 + DETR (native)	0.9455	0.9191	0.4342
YOLOv12	0.9006	0.9080	0.5051
HandXFM	**0.9481**	**0.9264**	**0.6101**

4.2.4 Joint Detection

We evaluate landmark-level joint detection on a 17-class dataset covering DIP, PIP, MCP joints, and the radius, ulna, and wrist center. A DETR-style decoder [6, 20, 21] is attached to the HandXFM encoder: patch tokens are projected to 256-D embeddings with learnable 2D positional encodings and decoded by a 6-layer Transformer with 100 queries. The decoder predicts class logits and normalized boxes via parallel MLP heads, and is trained with Hungarian matching using a composite loss (classification, ℓ_1, and GIoU) optimized by AdamW with a cosine schedule. Evaluation follows COCO-style mAP metrics (mAP@0.5:0.95, mAP@0.5, mAP@0.75).

As shown in Table 3, HandXFM outperforms all matched DETR-based baselines, including ResNet50, ViT, ConvNeXt, native DINOv3, and domain-adapted DINOv3, and also surpasses the supervised YOLOv12 [40] in mAP@0.5:0.95. It achieves the best high-IoU localization quality (mAP@0.75 = 0.9026), indicating more accurate landmark localization. These results suggest that the gain is attributable to the HandXFM backbone rather than decoder choice alone. Its attention-guided representations capture bone and joint topology, enabling geometric reasoning over part–whole relationships, as illustrated in the left panel of Figure 2, with qualitative examples in Supp. B.

4.2.5 Bone Segmentation

We perform bone segmentation on a 19-class dataset covering phalanges and metacarpals. Following the detection framework, we adopt a DETR architecture: the HandXFM encoder outputs patch tokens projected to a 256-dimensional feature map and processed by a transformer encoder–decoder. Each decoder query predicts boxes and class labels, while a lightweight head refines them into masks. The head upsamples 14×14 feature maps to 224×224 via bilinear interpolation and applies two convolutional layers with group normalization. Training jointly optimizes a set loss (classification, ℓ_1 box, and GIoU) and a segmentation loss (weighted cross-entropy + $0.5\times$ Dice). We report precision/recall and segmentation metrics (mAP@0.5:0.95, mAP@0.5, mAP@0.75).

As summarized in Table 3 and visualized in the left panel of Figure 2, HandXFM achieves the best segmentation performance, outperforming compared backbones and YOLOv12, with mAP@0.5:0.95 improving to 0.6101 and precision rising to 0.9481. These gains are obtained without pixel pretraining supervision and under a matched downstream head, indicating that the improvement comes from the HandXFM backbone. The encoder's latent representations, learned via multi-expert alignment and edge-aware masking, capture bone contours and spatial separations as structural cues, enabling the decoder to delineate complex boundaries with greater geometric fidelity. This suggests that HandXFM internalizes spatial structure during pretraining rather than relying on task-specific supervision.

Comparison with semantic segmentation baselines. We additionally compare HandXFM against U-Net and nnU-Net using pixel-level segmentation metrics on the same evaluation set (Table B5). HandXFM achieves the highest foreground precision (0.4596), recall (0.4984), and F1 score (0.4777), while maintaining pixel accuracy comparable to U-Net and nnU-Net. These measurements complement the COCO-style instance segmentation results, which evaluate the localization and separation of individual anatomical bone instances rather than only pixel-level foreground prediction.

4.2.6 Visual Question Answering

To evaluate cross-modal capability, we conduct a closed-ended visual question answering (VQA) task on the hand subset of MedPix (108 image–question pairs) [33]. We use a fixed train/test split, and all test images are excluded from the unlabeled pretraining corpus. Each sample pairs a radiograph with a caption-derived query, using the diagnostic title as the ground-truth answer. HandXFM combines its

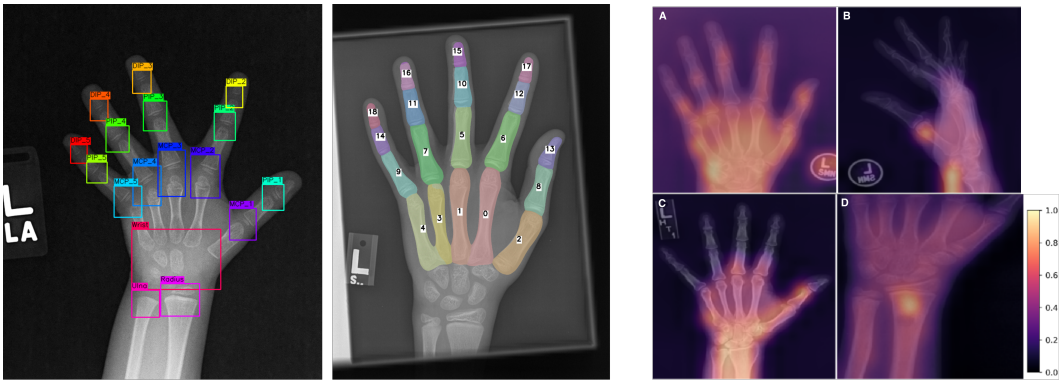

Figure 2: **Qualitative visualizations.** Left: predicted examples of joint detection and bone segmentation, where bounding boxes denote localized joints and colored masks represent segmented bones. Right: text–image cross-attention maps for clinical terms: (A) joint, (B) scaphotrapezial, (C) metacarpals, and (D) radius.

visual encoder with a frozen BiomedCLIP text encoder via cross-attention, followed by a two-layer MLP classifier. Models are trained with cross-entropy and evaluated using Top-1 and Top-5 accuracy. HandXFM achieves 0.6364 Top-1 and 0.9091 Top-5 accuracy, matching RadCLIP in this small-scale setting. Given the limited size of the MedPix hand subset, we interpret this result as evidence of cross-modal capability rather than as a definitive ranking between models, especially since HandXFM uses a visual backbone pretrained without paired image–text supervision.

Visualization. As shown in the right panel of Figure 2, cross-attention maps illustrate clinical terms grounded in radiographic regions. Examples include "joint," "scaphotrapezial," "metacarpals," and "radius." The maps show correspondence between tokens and anatomical structures, demonstrating effective alignment between clinical concepts and spatial regions, and supporting interpretability.

4.3 Ablation Study

We conduct a series of ablation studies to analyze the contributions of design choices in HandXFM. Specifically, we examine: (1) the number of expert teachers used during pretraining, (2) the role of the alignment and reconstruction objectives, and (3) the impact of edge-aware masking.

4.3.1 Teacher Ensemble Ablation

We compare three pretraining configurations: (i) *No teacher* (DINOv3 student only), (ii) *Single teacher* (BiomedCLIP or Chest X-ray ViT), and (iii) *Dual teachers* (HandXFM).

- **Dual-teacher distillation is strongest.** On bone age (Table B6), HandXFM reduces MAE from 9.11 (DINOv3-only) to 7.34 and improves PCC from 0.953 to 0.972; on joint detection, mAP@0.5:0.95 increases from 0.59 to 0.76. On MURA (Table B7), AUC rises from 0.74 to 0.86.
- **Single teachers show task-aligned strengths.** The Chest teacher performs best on structure-sensitive targets; for AIRINIIReumHands JSN it achieves RMSE 0.591 and MAE 0.380. BiomedCLIP improves semantic cues but is weaker on fine-grained structural scoring. Combining both yields the most balanced representation, achieving best or near-best performance across bone age, joint detection (Table B8), and MURA classification.

4.3.2 Ablation on Loss Composition

We compare three settings: (i) *Alignment-only* (teacher–student alignment), (ii) *Reconstruction-only* (masked token prediction), and (iii) *Combined*. Results are summarized in Tables B10 to B13.

- **Combined losses are most reliable.** They achieve best or near-best results across RSNA and RHPE bone age (lowest MAE, highest PCC), MURA (AUC/ACC), and joint detection (mAP@0.5:0.95).
- **Alignment-only.** It excels on some fine-grained regressions (e.g., AIRINIIReumHands erosion MAE) but underperforms on JSN, indicating limited spatial coverage.

- **Reconstruction-only.** It remains competitive, trailing the combined setup but surpassing alignment-only on JSN and detection, showing masked prediction improves spatial completeness.

Overall, alignment sharpens teacher-consistent semantics, reconstruction enhances localization, and combining them yields the most robust cross-task performance.

4.3.3 Ablation on Masking Strategy

We compare the edge-aware masking scheme with a standard random masking baseline during pretraining. Both configurations use a 50% masking ratio and identical training settings. They differ only in how masked patch positions are sampled. Edge-aware masking assigns higher sampling probabilities to patches with strong image gradients, using an edge–uniform mixture with $\alpha = 0.5$.

Across most regression (Table B14) and classification (Table B15) tasks, edge-aware masking consistently outperforms random masking, indicating that emphasizing structurally salient regions benefits both semantic and quantitative prediction. The advantage is more pronounced for localization-heavy tasks. For joint detection (Table B16), mAP@0.5:0.95 improves from 0.7125 to 0.7636 and mAP@0.75 from 0.8485 to 0.9026; for bone segmentation (Table B17), instance accuracy increases from 0.9713 to 0.9751 and mAP@0.5:0.95 from 0.4680 to 0.6101.

Overall, edge-aware masking provides stronger localization performance, while random masking remains competitive on selected regression and pixel-level segmentation metrics.

4.3.4 Additional Representation Analysis

We complement the component ablations with additional experiments examining matched in-domain pretraining and frozen-backbone representation transfer.

Matched pretraining control. To distinguish the effect of the HandXFM pretraining objective from additional exposure to hand radiographs, we compare native DINOv3, DINOv3 adapted on the same hand-radiograph training corpus, and HandXFM using the same joint-detection protocol. As shown in Table 3, the domain-adapted DINOv3 achieves mAP@0.5:0.95 of 0.5900, compared with 0.5857 for native DINOv3 and 0.7636 for HandXFM. The improvement obtained by HandXFM suggests that its pretraining formulation contributes beyond additional in-domain data exposure.

Frozen-backbone evaluation. We also evaluate the pretrained encoders without updating their backbone parameters. On Anhui whole-hand SvH scoring, frozen HandXFM achieves PCC 0.684 and $R^2 = 0.333$, compared with PCC 0.654 and $R^2 = 0.241$ for RadCLIP (Table B19). These results indicate that the HandXFM representation retains task-relevant information accessible to a lightweight regression head, even without end-to-end backbone fine-tuning.

5 Conclusion

We presented HandXFM, a multi-expert foundation model for comprehensive hand X-ray understanding. By aligning a DINOv3 student with two complementary teachers, BiomedCLIP for semantic priors and a Chest X-ray ViT for structural cues, HandXFM learns transferable representations that generalize across diverse downstream tasks, including bone-age and SvH regression, abnormality classification, joint detection, bone segmentation, and visual question answering. Extensive experiments show consistent improvements over prior baselines and strong generalization to unseen datasets such as the Digital Hand Atlas. Ablation studies highlight the effectiveness of multi-expert distillation, the synergy between alignment and reconstruction losses, and the benefit of edge-aware masking for structural localization. Future work may extend HandXFM to integrate a broader spectrum of complementary expert knowledge beyond the current modalities. Overall, HandXFM provides a unified and interpretable foundation for musculoskeletal imaging, illustrating how multi-expert distillation bridges structural precision with semantic understanding and offers a scalable blueprint for domain-specialized medical foundation models.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Canada Research Chairs Tier II Program (CRC-2021-00482), the Canadian Institutes of Health Research (PJT-212993), and the Canada Foundation for Innovation (CFI) John R. Evans Leaders Fund (JELF, project no. 43481).

Competing interests. The authors declare no competing financial interests.

References

[1] Laith Alzubaidi, Asma Salhi, Mohammed A. Fadhel, Jinshuai Bai, Freek Hollman, Kristine Italia, Roberto Pareyon, A. S. Albahri, Chun Ouyang, Jose Santamaría, Kenneth Cutbush, Ashish Gupta, Amin Abbosh, and Yuantong Gu. Trustworthy deep learning framework for the detection of abnormalities in x-ray shoulder images. *PLOS ONE*, 19(3):e0299545, March 2024. doi: 10.1371/journal.pone.0299545.

[2] Ifrah Andleeb, Bilal Zahid Hussain, Julie Joncas, Soraya Barchi, Marjolaine Roy-Beaudry, Stefan Parent, Guy Grimard, Hubert Labelle, and Luc Duong. Automatic evaluation of bone age using hand radiographs and pancorporal radiographs in adolescent idiopathic scoliosis. *Diagnostics (Basel)*, 15(4):452, February 2025. doi: 10.3390/diagnostics15040452. URL `https://doi.org/10.3390/diagnostics15040452`.

[3] Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, Rogier van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay S. Chaudhari. A vision–language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering*, 9(4):494–506, April 2025. ISSN 2157-846X. doi: 10.1038/s41551-024-01246-y. URL `https://doi.org/10.1038/s41551-024-01246-y`.

[4] Zhiyan Bo, Laura C. Coates, and Bartłomiej W. Papież. Deep learning models to automate the scoring of hand radiographs for rheumatoid arthritis. In Moi Hoon Yap, Connah Kendrick, Ardhendu Behera, Timothy Cootes, and Reyer Zwiggelaar, editors, *Medical Image Understanding and Analysis*, pages 398–413, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-66958-3.

[5] Joshua T. Bram, Ayoosh Pareek, Samuel A. Beber, Ruth H. Jones, M. Moein Shariatnia, Amir Daliliyazdi, Olivia C. Tracey, Daniel W. Green, and Peter D. Fabricant. Determination of skeletal age from hand radiographs using deep learning. *The American Journal of Sports Medicine*, 53(11):2715–2725, September 2025. doi: 10.1177/03635465251359618. Epub 2025 Aug 15.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58451-1. doi: 10.1007/978-3-030-58452-8_13. URL `https://doi.org/10.1007/978-3-030-58452-8_13`.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[8] chanelcolgate. vit-base-patch16-224-chest-x-ray: fine-tuned vit-base on a chest-x-ray classification dataset. `https://huggingface.co/chanelcolgate/vit-base-patch16-224-chest-x-ray`, 2025.

[9] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 679–689, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16443-9.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] María Escobar, Cristina González, Felipe Torres, Laura Daza, Gustavo Triana, and Pablo Arbeláez. Hand pose estimation for pediatric bone age assessment. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019: Proceedings, Part VI*, pages 531–539, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-32225-0. doi: 10.1007/978-3-030-32226-7_59. URL https://doi.org/10.1007/978-3-030-32226-7_59.

[12] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

[13] Federative International Programme for Anatomical Terminology (FIPAT). *Terminologia Anatomica: International Anatomical Terminology*. Thieme Medical Publishers, Stuttgart, Germany, 2nd edition, 2019.

[14] Ning Gao, Sanping Zhou, Le Wang, and Nanning Zheng. Pmt: Progressive mean teacher via exploring temporal consistency for semi-supervised medical image segmentation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 144–160, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73113-6.

[15] Arkadiusz Gertych, Aifeng Zhang, James Sayre, Sylwia Pospiech-Kurkowska, and H. K. Huang. Bone age assessment of children using a digital hand atlas. *Computerized Medical Imaging and Graphics*, 31(4–5):322–331, June 2007. doi: 10.1016/j.compmedimag.2007.02.012.

[16] Cristina González, María Escobar, Laura Daza, Felipe Torres, Gustavo Triana, and Pablo Arbeláez. Simba: Specific identity markers for bone age assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 753–763. Springer, 2020.

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[18] Safwan S. Halabi, Luciano M. Prevedello, Jayashree Kalpathy-Cramer, Artem B. Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, Felipe Campos Kitamura, Hans H. Thodberg, Leon Chen, George Shih, Katherine Andriole, Marc D. Kohli, Bradley J. Erickson, and Adam E. Flanders. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019. doi: 10.1148/radiol.2018180736.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

[20] Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. DEIM: DETR with Improved Matching for Fast Convergence. *arXiv e-prints*, art. arXiv:2412.04234, December 2024. doi: 10.48550/arXiv.2412.04234.

[21] Shihua Huang, Yongjie Hou, Longfei Liu, Xuanlong Yu, and Xi Shen. Real-Time Object Detection Meets DINOv3. *arXiv e-prints*, art. arXiv:2509.20787, September 2025. doi: 10.48550/arXiv.2509.20787.

[22] Shinn Kim, Soobin Lee, Kyoungseob Shin, Han-Soo Kim, Yongsung Kim, Minsu Kim, Juhong Nam, Somang Ko, Daeheon Kwon, Wook Huh, et al. A generalizable large-scale foundation model for musculoskeletal radiographs. *arXiv preprint arXiv:2602.03076*, 2026.

[23] Curtis P. Langlotz. Radlex: A new method for indexing online educational materials. *Radiographics*, 26(6):1595–1597, November 2006. doi: 10.1148/rg.266065168.

[24] Zhangyong Li, Wang Chen, Yang Ju, Yong Chen, Zhengjun Hou, Xinwei Li, and Yuhao Jiang. Bone age assessment based on deep neural networks with annotation-free cascaded critical bone region extraction. *Frontiers in Artificial Intelligence*, 6:1142895, March 2023. doi: 10.3389/frai.2023.1142895. eCollection 2023.

[25] Zhixiu Lu, Hailong Li, Nehal A. Parikh, Jonathan R. Dillman, and Lili He. RadCLIP: Enhancing radiologic image analysis through contrastive language-image pretraining. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):17613–17622, October 2025. doi: 10.1109/TNNLS.2025.3568036.

[26] Hajar Moradmand and Lei Ren. Multistage deep learning methods for automating radiographic sharp score prediction in rheumatoid arthritis. *Scientific Reports*, 15(1):3391, January 2025. doi: 10.1038/s41598-025-86073-0. URL https://doi.org/10.1038/s41598-025-86073-0.

[27] National Library of Medicine. MedPix®: Database of Medical Images, Teaching Cases, and Clinical Topics. https://medpix.nlm.nih.gov/. Accessed: 2026-03-06.

[28] Magdalini Paschali, Zhihong Chen, Louis Blankemeier, Maya Varma, Alaa Youssef, Christian Bluethgen, Curtis Langlotz, Sergios Gatidis, and Akshay Chaudhari. Foundation models in radiology: what, how, why, and why not. *Radiology*, 314(2):e240597, 2025.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[30] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. *arXiv e-prints*, art. arXiv:1712.06957, December 2017. doi: 10.48550/arXiv.1712.06957.

[31] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv e-prints*, art. arXiv:2508.10104, August 2025. doi: 10.48550/arXiv.2508.10104.

[32] Nishank Singla, Krisztian Koos, Farzin Haddadpour, Amin Honarmandi Shandiz, Lovish Chum, Xiaojian Xu, Qing Jin, and Erhan Bas. Multi anatomy x-ray foundation model. *arXiv preprint arXiv:2509.12146*, 2025.

[33] Irene Siragusa, Salvatore Contino, Massimo La Ciura, Rosario Alicata, and Roberto Pirrone. Medpix 2.0: A comprehensive multimodal biomedical data set for advanced ai applications with retrieval augmented generation and knowledge graphs. *Data Science and Engineering*, July 2025. ISSN 2364-1541. doi: 10.1007/s41019-025-00297-8. URL http://dx.doi.org/10.1007/s41019-025-00297-8.

[34] Mojtaba Sirati-Amsheh, Elham Shabaninia, and Ali Chaparian. Ae-bonet: A deep learning method for pediatric bone age estimation using an unsupervised pre-trained model. *Journal of Biomedical Physics and Engineering*, 15(3):271–280, June 2025. doi: 10.31661/jbpe.v0i0.2304-1609. eCollection 2025 Jun.

[35] Anna Solovyova and Igor Solovyov. X-ray bone abnormalities detection using mura dataset. *arXiv preprint arXiv:2008.03356*, 2020. URL https://arxiv.org/abs/2008.03356.

[36] Antoine Spahr, Behzad Bozorgtabar, and Jean-Philippe Thiran. Self-taught semi-supervised anomaly detection on upper limb x-rays. In *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1632–1636. IEEE, 2021. doi: 10.1109/ISBI48211.2021.9433771.

[37] Alexander Stolpovsky, Elizaveta Dakhova, Polina Druzhinina, Polina Postnikova, Daniil Kudinsky, Alexander Smirnov, Anastasia Sukhinina, Alexander Lila, and Anvar Kurmukov. Rheumavit: Transformer-based model for automated scoring of hand joints in rheumatoid arthritis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2522–2531, October 2023.

[38] Yiyao Sun, Xinran Wen, Yan Zhang, Lijun Jin, Chunna Yang, Qianhui Zhang, Mingchen Jiang, Zhaoyang Xu, Wei Guo, Juan Su, and Xiran Jiang. Visual-language foundation models in medical imaging: A systematic review and meta-analysis of diagnostic and analytical applications. *Computer Methods and Programs in Biomedicine*, 268:108870, August 2025. ISSN 0169-2607. doi: 10.1016/j.cmpb.2025.108870. URL https://www.sciencedirect.com/science/article/pii/S0169260725002871.

[39] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[40] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[42] D. M. F. M. Van der Heijde. How to read radiographs according to the sharp/van der heijde method. *The Journal of Rheumatology*, 27(1):261–263, January 2000.

[43] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3876–3887, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.256. URL https://pubmed.ncbi.nlm.nih.gov/39144675/.

[44] Zijian Wang, Jian Liu, Zongyun Gu, and Chuanfu Li. An efficient cnn for hand x-ray overall scoring of rheumatoid arthritis. *Complexity*, 2022(1):5485606, 2022. doi: 10.1155/2022/5485606. URL https://onlinelibrary.wiley.com/doi/10.1155/2022/5485606.

[45] Yibing Wei, Abhinav Gupta, and Pedro Morgado. Towards latent masked image modeling for self-supervised visual representation learning. In *Computer Vision – ECCV 2024: Proceedings, Part XXXIX*, pages 1–17, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72932-4. doi: 10.1007/978-3-031-72933-1_1. URL https://doi.org/10.1007/978-3-031-72933-1_1.

[46] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, August 2025. doi: 10.1038/s41467-025-62385-7. URL https://doi.org/10.1038/s41467-025-62385-7.

[47] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2 (1), 2024. doi: 10.1056/AIoa2400640. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400640.

Supplementary Material

Table of Contents

A. Pretrain Configuration and Hyperparameters 16
A.1. Student and Teacher Backbones 16
Student encoder 16
Expert teachers 16
A.2. Pretraining Data and Augmentation 16
Data sources 16
Image preprocessing 17
A.3. Masking and Latent Regressor 17
Patch masking 17
Masked token regressor 17
A.4. Losses, EMA, and Optimization 17
Latent alignment and reconstruction 17
EMA teacher 17
A.5. Ablation Pretraining Variants 18

B. Results Details 18
B.1. Joint Detection Results 18
B.2. Bone Segmentation Results 18
B.3. Ablation on Expert Composition 19
B.4. Ablation on Loss Composition 19
B.5. Ablation on Masking Strategy 21
B.6. Effect of Higher-Resolution Downstream Fine-Tuning 22
B.7. Full Fine-Tuning Protocol and Frozen-Backbone Analysis 23
B.8. Multi-Seed Stability Analysis 23
B.9. VQA Case Analysis 24

C. Dataset Details 25
C.1. Public Datasets 26
C.2. Newly Annotated Datasets 30

D. Limitations and Future Work 31

E. Reproducibility Check 32
E.1. Compute 32
E.2. Data organization and preprocessing 33
E.3. Training configuration 34
E.4. Code entry points and artifacts 34

F. Broader Impact 35

A Pretrain Configuration and Hyperparameters

This section records the implementation settings needed to reproduce HandXFM pretraining. We focus on the components and configuration of the proposed method; the learning objectives themselves are introduced in Section 3.

A.1 Student and Teacher Backbones

Student encoder. HandXFM initializes its student from a DINOv3 ViT-B/16 checkpoint using the Hugging Face `AutoModel` interface. The original ViT architecture is retained: 16×16 patches, 12 transformer layers, 12 attention heads, and a hidden width of 768 (MLP ratio 4). At the pretraining resolution of 224×224, the model produces 196 patch tokens and one `[CLS]` token. When the checkpoint contains register tokens, these are excluded from the tokens used in the distillation objectives. The resulting encoder has approximately 86M parameters. The final exponential-moving-average (EMA) encoder, rather than the instantaneous student, supplies the default weights for downstream experiments.

The 224×224 pretraining resolution limits the computational cost of multi-expert supervision; downstream tasks may use higher input resolutions as specified in their respective experimental protocols. The student architecture is not modified: HandXFM learns a hand-radiograph representation through its pretraining objectives.

Expert teachers. Two pretrained visual encoders supply complementary latent supervision. The BiomedCLIP ViT-B/16 image encoder is loaded with `open_clip.create_model_and_transforms` and its associated preprocessing. Its visual representations were learned through biomedical image–text contrastive pretraining, providing language-aligned visual features without requiring text inputs during HandXFM pretraining. A chest-radiograph ViT-B/16, loaded with `AutoModel` and its `AutoImageProcessor`, contributes a distinct radiography-domain appearance prior.

Here, *frozen* means that neither expert receives gradient updates during HandXFM pretraining. Both operate only as feature-target generators; the student, the expert-specific linear projection layers P_j, and the masked latent regressor are optimized. Each P_j maps the corresponding expert representation to the 768-dimensional student space and is instantiated as a `LazyLinear` layer. Teacher targets are organized into a global `[CLS]` representation and a spatial token sequence. For patch-wise supervision, a teacher sequence with a different number of spatial positions is truncated or padded to the student's 196-token grid.

Radiography-domain teacher provenance. The radiography-domain teacher uses the following Hugging Face checkpoint:

```
chanelcolgate/vit-base-patch16-224-chest-x-ray
```

Starting from a standard ViT-B/16 checkpoint, this model was fine-tuned for binary normal-versus-pneumonia classification on the public trpakov chest X-ray classification collection. According to its model card, this source dataset has 12,230 training, 1,165 validation, and 582 test images. This teacher was not trained to identify hand joints or bone boundaries. Its role in HandXFM is to provide features adapted to projection radiographs, whereas BiomedCLIP supplies representations shaped by biomedical image–text learning. We do not assume an exclusive semantic/structural division between the teachers; the corresponding single- and dual-teacher comparisons are reported in Section 4.3.

A.2 Pretraining Data and Augmentation

Data sources. We curated approximately 30K hand radiographs from six publicly available datasets: RSNA Bone Age, RHPE, MURA `XR_HAND`, AIRINIIReumHands, Anhui RA, and MedPix Hand. HandXFM uses images from the designated training partitions for unlabeled pretraining, with validation and test images excluded. Radiographs are processed as whole-hand inputs rather than cropped into individual joint or bone regions. The dataset partitions used for downstream training and evaluation are detailed in Table C21.

Image preprocessing. The notation $x \in \mathbb{R}^{3\times224\times224}$ in Section 3 denotes the standardized network input, not three independently acquired X-ray channels. Images from Anhui RA and MedPix may be encoded as RGB; grayscale radiographs from RSNA, RHPE, MURA, and AIRINIIReumHands are represented by repeating their intensity channel across the three RGB channels. This permits use of the pretrained three-channel ViT encoders and does not introduce additional color information.

Augmentations are applied before feature extraction. Following a resize to 256×256, each input undergoes a randomly applied horizontal flip ($p = 0.5$) and a rotation bounded by $\pm10^\circ$ with zero padding; the central 224×224 region is retained afterward. `ToTensor` converts pixel intensities to $[0, 1]$. Backbone-specific input normalization is subsequently handled by the relevant model processor; for the DINOv3 processor, `do_rescale=False` avoids rescaling tensors a second time. BiomedCLIP uses the preprocessing associated with its `open_clip` configuration.

A.3 Masking and Latent Regressor

Patch masking. HandXFM masks 50% of the 14×14 patch locations. For each image, we aggregate grayscale Sobel-gradient magnitudes into 196 patch-level values via adaptive average pooling. The resulting grid biases the masking distribution toward edge-rich locations; it is combined with uniform sampling according to Section 3. In the reported HandXFM configuration, `mask_type=edge`, `mask_ratio=0.5`, `hw=14`, and `mask_edge_alpha=0.5`. The coefficient $\alpha = 0.5$ is the weight of the uniform component; patches with stronger edges are more likely to be *masked* and predicted. Patches are selected without replacement.

Masked token regressor. The latent regressor $\mathcal{R}_\phi$ uses learned mask queries to attend to visible student patch features and estimate representations at masked locations. Its architecture has six transformer blocks, hidden width 768, eight attention heads, MLP ratio 4, GELU activations, LayerNorm, and linearly increasing stochastic depth up to 0.1. Patch locations are specified using fixed 2D sine–cosine embeddings of size 196×768, assigned to both the context features and the mask queries. The predictions are trained against expert-projected latent targets, not reconstructed image pixels.

A.4 Losses, EMA, and Optimization

Latent alignment and reconstruction. The visible-token alignment and masked-token reconstruction terms use mean-squared differences between student representations (or regressor predictions) and the corresponding projected expert targets. The contributions from the available experts are averaged, and an additional `[CLS]` alignment term supervises the global representation. In the objective defined in Section 3, the main-run coefficients are $\lambda_{\text{align}} = 1$, $\lambda_{\text{recon}} = 1$, $\lambda_{\text{ema}} = 0.01$, and $\lambda_{\text{cls}} = 0.01$.

The loss coefficients were selected by examining the relative magnitudes of the individual terms during preliminary pretraining. Alignment and reconstruction had comparable scales and were assigned unit weights, whereas the auxiliary EMA and CLS terms were assigned smaller coefficients to prevent them from dominating the objective. These weights were kept fixed for pretraining runs using the complete objective.

EMA teacher. An EMA copy of the student is maintained with momentum $\mu = 0.996$ and updated after each optimization step as $\theta_m \leftarrow \mu\theta_m + (1 - \mu)\theta$. Its parameters are never updated by backpropagation. The EMA features supply a temporal consistency target for the current student, with the corresponding contribution controlled by $\lambda_{\text{ema}} = 0.01$. Weight-averaged teachers are established techniques in self-training and self-distillation, including Mean Teacher [39], BYOL [17], and DINO [7]; EMA is used here for stabilization and is not claimed as an independently novel algorithmic component.

Optimizer and learning rates. We use `AdamW` to optimize the student encoder and the regressor/projection modules in separate parameter groups, while the frozen experts and EMA encoder are excluded from gradient-based updates. The encoder and head learning rates are 3×10^{-5} and 1×10^{-4}, respectively. The AdamW coefficients are $(\beta_1, \beta_2) = (0.9, 0.999)$, with weight decay set to zero. The complete main-run settings are listed in Table A4.

Table A4: HandXFM pretraining configuration. All values describe the main setting; ablations modify only the indicated component.

Component	Configuration
Student	DINOv3 ViT-B/16; 12 blocks; 12 heads; $D = 768$
Input resolution / grid	224×224 / 14×14 ($N = 196$)
Masking	Edge-aware; ratio 0.5; $\alpha = 0.5$
Latent regressor	6 blocks; $D = 768$; 8 heads; MLP ratio 4
Regressor regularization	GELU; LayerNorm; drop path up to 0.1
Loss weights	$(\lambda_{\text{align}}, \lambda_{\text{recon}}, \lambda_{\text{ema}}, \lambda_{\text{cls}}) = (1, 1, 0.01, 0.01)$
EMA momentum	0.996
Optimizer	`AdamW`
Learning rates (student / heads)	3×10^{-5} / 1×10^{-4}
Weight decay	0
AdamW betas (β_1, β_2)	$(0.9, 0.999)$
Epochs / batch size	10 / 512
Student warmup freeze	2 epochs
Workers / logging / checkpointing	8 / 100 iterations / 5 epochs

Training schedule. Pretraining runs for 10 epochs with batch size 512. The student encoder is frozen during the first two epochs, then unfrozen; the trainable projection and regression modules are optimized throughout. The final EMA encoder is exported for downstream adaptation. For reproducibility, the saved training state records both encoder copies, the latent-prediction modules, optimizer variables, epoch numbers, and aggregate losses.

A.5 Ablation Pretraining Variants

The teacher-composition, objective, and masking ablations use the same student initialization, pretraining data, optimization settings, and downstream evaluation protocols unless otherwise specified. The single-teacher configurations receive latent supervision from either BiomedCLIP or the chest-radiograph encoder, while the dual-teacher configuration uses both. The objective ablations isolate visible-token alignment and masked latent reconstruction. For the masking ablation, uniformly sampled patches replace edge-biased masking while maintaining the same masking ratio of 0.5.

B Results Details

B.1 Joint detection result

Figure B3 illustrates predictions for the 17 anatomical targets across representative hand radiographs. The examples include variation in positioning, finger overlap, and local anatomy. The detector operates on full radiographs without requiring a separate image-registration step. These visual examples supplement the quantitative localization metrics in the main paper; they are not, on their own, a measurement of robustness under every acquisition condition.

B.2 Bone segmentation result

The examples in Figure B4 visualize the correspondence between the radiographs and their predicted anatomical masks. These results provide examples of instance-aware bone delineation across subjects; the quantitative instance-localization and pixel-level comparisons are reported separately. The proposed method is intended to identify individual bone categories, rather than merely segmenting the aggregate bone foreground.

Table B5: Pixel-level bone-segmentation comparison with conventional medical semantic-segmentation baselines on the same evaluation set. Results complement the instance-level COCO-style evaluation; pixel accuracy alone can be influenced by background prevalence. Higher is better for all metrics.

Method	Pixel ACC	Precision	Recall	F1
nnU-Net	0.9587	0.4384	0.4762	0.4538
U-Net	**0.9597**	0.4549	0.4803	0.4645
HandXFM	0.9586	**0.4596**	**0.4984**	**0.4777**

To complement the instance-level baselines in the main paper, Table B5 compares HandXFM with U-Net and nnU-Net using pixel-wise measures. HandXFM obtains the highest precision, recall, and F1 in this comparison, whereas U-Net achieves slightly higher pixel accuracy. These results address a different aspect of performance from the reported COCO-style instance mAP; the pixel metrics do not by themselves evaluate the correct separation of neighboring bone instances.

B.3 Ablation on Expert Composition

Effect of expert on regression tasks. Table B6 compares the DINOv3-only control with BiomedCLIP-only, chest-radiograph-only, and dual-teacher supervision under the same downstream protocols. The effects depend on the target: The chest-only variant is stronger than BiomedCLIP-only on the RSNA and RHPE bone-age metrics and on several AIRINIIReumHands measures. The dual-teacher model leads on RSNA, RHPE, Anhui, and erosion metrics, as well as AIRINIIReumHands JSN PCC; for JSN RMSE, MAE, and R^2, chest-only performs better. Thus, the teacher contributions are complementary but cannot be assigned exclusively semantic or anatomical roles.

Effect of expert on classification task. On MURA (Table B7), the dual-teacher model records the highest AUC, accuracy, and precision. The chest-only variant has the highest F1, while DINOv3-only has the highest recall. These differences show a trade-off between precision and sensitivity; an overall improvement in AUC should not be interpreted as superiority on every operating-point metric.

Effect of expert on joint detection. The matched joint-localization comparison in Table B8 is particularly informative: BiomedCLIP-only falls below the DINOv3-only baseline on all three reported metrics, whereas chest-only yields modest gains on the aggregate and stricter-IoU metrics. Dual-teacher HandXFM substantially improves mAP@0.5:0.95 and mAP@0.75. These results establish that the benefit arises from the combination of teachers, rather than assuming that every individual teacher improves every task.

Effect of expert on bone segmentation. The segmentation ablation (Table B9) separates instance classification, instance localization, and pixel-level segmentation. Dual-teacher supervision leads on instance classification and on the two stricter localization summaries (mAP@0.5:0.95 and mAP@0.75). However, chest-only has the highest mAP@0.5, and BiomedCLIP-only attains higher pixel accuracy and recall than the dual model. This pattern supports the value of combined supervision for bone-level localization without implying a uniform advantage on all pixel metrics.

B.4 Ablation on Loss Composition

Effect of loss on regression tasks. The regression results in Table B10 indicate that the preferred objective depends on the benchmark and metric. The combined formulation performs best on RSNA, RHPE, and Anhui. For AIRINIIReumHands erosion, alignment-only has the lowest RMSE and MAE and the highest R^2, whereas the combined objective has the highest PCC. For JSN, reconstruction-only has the lowest errors while the combined formulation again has the highest PCC. This is evidence for cross-task complementarity, not a claim that the combined loss dominates every measurement.

Effect of loss on classification task. In MURA classification (Table B11), the joint objective produces the highest AUC, accuracy, and precision, whereas reconstruction-only yields the highest

Table B6: Ablation on expert composition—regression. Lower is better for RMSE/MAE; higher is better for R^2/PCC.

Dataset	Metric	DINOv3	BiomedCLIP	Chest X-ray	HandXFM
			Bone Age		
RSNA	RMSE ↓	12.702	10.497	10.178	**9.932**
	MAE ↓	9.106	7.683	7.584	**7.343**
	R^2 ↑	0.907	0.937	0.940	**0.943**
	PCC ↑	0.953	0.968	0.970	**0.972**
RHPE	RMSE ↓	27.828	18.487	12.462	**11.073**
	MAE ↓	21.856	13.484	9.388	**8.363**
	R^2 ↑	0.550	0.801	0.910	**0.929**
	PCC ↑	0.790	0.911	0.956	**0.966**
			SvH Scoring		
AIRINII (Erosion)	RMSE ↓	0.525	0.420	0.306	**0.298**
	MAE ↓	0.300	0.230	0.144	**0.111**
	R^2 ↑	0.127	0.440	0.704	**0.719**
	PCC ↑	0.426	0.758	0.894	**0.929**
AIRINII (JSN)	RMSE ↓	0.714	0.738	**0.591**	0.705
	MAE ↓	0.558	0.575	**0.380**	0.608
	R^2 ↑	0.591	0.562	**0.720**	0.601
	PCC ↑	0.803	0.782	0.870	**0.885**
Anhui (SvH)	RMSE ↓	31.232	26.115	20.567	**18.231**
	MAE ↓	20.301	17.094	13.529	**12.356**
	R^2 ↑	0.447	0.613	0.760	**0.812**
	PCC ↑	0.702	0.796	0.884	**0.903**

Table B7: Ablation on expert composition—MURA normal/abnormal classification. Higher is better for all metrics.

Metric	DINOv3 Only	BiomedCLIP Only	Chest X-ray Only	Dual (HandXFM)
AUC ↑	0.7405	0.7771	0.8485	**0.8599**
ACC ↑	0.6741	0.6987	0.7768	**0.7857**
F1 ↑	0.6589	0.6650	**0.6933**	0.6903
P ↑	0.5900	0.6262	0.8248	**0.8843**
R ↑	**0.7460**	0.7090	0.5979	0.5661

F1 and recall. The distinction matters because threshold-dependent metrics can respond differently to the same learned representation.

Effect of loss on joint detection. For detection, reconstruction-only exceeds alignment-only on each reported mAP measure, while the combined loss provides a larger improvement at the stricter IoU threshold (Table B12). This suggests that the two terms interact constructively for localization, but the experiment does not isolate the individual contribution of the auxiliary EMA or CLS terms.

Effect of loss on bone segmentation. The bone-segmentation comparison in Table B13 is likewise metric dependent. The combined objective leads for instance-level accuracy, recall, F1, and all reported instance-localization measures. Alignment-only provides higher pixel accuracy, precision, recall, and F1 than the combined model, while reconstruction-only ties the combined model on

Table B8: Ablation on expert composition—17-class joint detection. Higher is better for all metrics.

Metric	DINOv3 Only	BiomedCLIP Only	Chest X-ray Only	Dual (Ours)
mAP@0.5:0.95 ↑	0.5900	0.5550	0.6017	**0.7636**
mAP@0.5 ↑	0.9650	0.9515	0.9621	**0.9802**
mAP@0.75 ↑	0.6487	0.5810	0.6771	**0.9026**

Table B9: Ablation on expert composition—bone segmentation, including instance-level classification, instance localization, and pixel-level metrics. Higher is better. The DINOv3-only configuration is the no-expert-teacher pretraining ablation, rather than the native DINOv3 checkpoint evaluated in the main baseline comparison.

Metric	DINOv3 Only	BiomedCLIP Only	Chest X-ray Only	Dual (HandXFM)
Instance-level Classification				
ACC ↑	0.9474	0.9732	0.9675	**0.9751**
P ↑	0.9435	0.9445	0.9428	**0.9481**
R ↑	0.9000	0.9245	0.9191	**0.9264**
F1 ↑	0.9210	0.9344	0.9307	**0.9371**
Instance-level Localization				
mAP@0.5:0.95 ↑	0.3121	0.4800	0.5187	**0.6101**
mAP@0.5 ↑	0.6998	0.9165	**0.9632**	0.9466
mAP@0.75 ↑	0.2092	0.4397	0.5349	**0.7030**
Pixel-level Segmentation				
ACC ↑	0.9512	**0.9593**	0.9589	0.9586
P ↑	0.4174	0.4582	0.4528	**0.4596**
R ↑	0.4850	**0.5026**	0.5024	0.4984
F1 ↑	0.4436	**0.4777**	0.4753	**0.4777**

instance precision. Accordingly, the main benefit of combining these objectives is observed in instance discrimination and localization rather than in every pixel-wise score.

B.5 Ablation on Masking Strategy

Effect of masking strategy on regression tasks. As shown in Table B14, edge-aware masking improves the main RSNA, RHPE, and Anhui bone-age or SvH regression measures. On the smaller AIRINIIReumHands dataset, the trend is mixed: random masking has slightly higher erosion PCC and lower JSN RMSE/MAE, whereas edge-aware masking leads on erosion RMSE/R^2 and JSN PCC. The edge-biased sampling policy therefore offers a generally useful inductive bias, with exceptions that motivate the multi-seed analysis in Section B.8.

Effect of masking strategy on classification tasks. On MURA, edge-aware masking improves all five reported classification measures, including AUC (0.7739 to 0.8599) and F1 (0.5860 to 0.6903), relative to random masking (Table B15). This result is consistent with a benefit from emphasizing radiographically informative boundaries during pretraining, although the evaluation does not identify which specific image regions cause the improvement.

Table B10: Ablation on loss composition—regression. Lower is better for RMSE/MAE; higher is better for R^2/PCC.

Dataset	Metric	Alignment Only	Reconstruction Only	Combined
		Bone Age Prediction		
RSNA	RMSE ↓	10.200	10.415	**9.932**
	MAE ↓	7.600	7.676	**7.343**
	R^2 ↑	0.940	0.938	**0.943**
	PCC ↑	0.970	0.968	**0.972**
RHPE	RMSE ↓	11.622	12.329	**11.073**
	MAE ↓	8.907	9.372	**8.363**
	R^2 ↑	0.921	0.912	**0.929**
	PCC ↑	0.964	0.957	**0.966**
		SvH Scoring		
AIRINII (Erosion)	RMSE ↓	**0.223**	0.240	0.298
	MAE ↓	**0.104**	0.141	0.111
	R^2 ↑	**0.843**	0.818	0.719
	PCC ↑	0.919	0.906	**0.929**
AIRINII (JSN)	RMSE ↓	0.804	**0.649**	0.705
	MAE ↓	0.601	**0.416**	0.608
	R^2 ↑	0.481	**0.662**	0.601
	PCC ↑	0.777	0.859	**0.885**
Anhui (SvH)	RMSE ↓	20.987	21.194	**18.231**
	MAE ↓	13.453	13.939	**12.356**
	R^2 ↑	0.750	0.745	**0.812**
	PCC ↑	0.887	0.880	**0.903**

Effect of masking strategy on joint detection. For joint detection (Table B16), the greatest improvement appears under stricter localization criteria: mAP@0.75 rises from 0.8485 to 0.9026 and average mAP@0.5:0.95 from 0.7125 to 0.7636. The random-mask model is marginally higher at mAP@0.5 (0.9844 versus 0.9802), indicating that the advantage of edge-aware masking is most visible for tighter overlap requirements.

Effect of masking strategy on bone segmentation. For the segmentation task, edge-aware masking improves all reported instance classification and localization scores (Table B17); the mAP@0.75 increase is especially pronounced (0.4617 to 0.7030). Pixel-level results show a different balance: the random-mask variant has higher pixel accuracy, recall, and F1, whereas edge-aware masking has higher pixel precision. These measurements support the localization-oriented role of the proposed masking policy without implying that it optimizes every foreground-pixel metric.

B.6 Effect of Higher-Resolution Downstream Fine-Tuning

HandXFM is pretrained with 224×224 inputs. Downstream tasks can use a larger input grid by interpolating the pretrained positional embeddings with two-dimensional bicubic interpolation, a standard ViT adaptation practice [10, 41]. This strategy permits evaluation of resolution-dependent features without changing the pretraining corpus or retraining the foundation model.

Table B11: Ablation on loss composition—MURA normal/abnormal classification. Higher is better for all metrics.

Metric	Alignment Only	Reconstruction Only	Combined
AUC ↑	0.8413	0.8430	**0.8599**
ACC ↑	0.7790	0.7723	**0.7857**
F1 ↑	0.6937	**0.7243**	0.6903
P ↑	0.8629	0.7403	**0.8843**
R ↑	0.5661	**0.7090**	0.5661

Table B12: Ablation on loss composition—17-class joint detection. Higher is better for all metrics.

Metric	Alignment Only	Reconstruction Only	Combined
mAP@0.5:0.95 ↑	0.5765	0.5995	**0.7636**
mAP@0.5 ↑	0.9624	0.9702	**0.9802**
mAP@0.75 ↑	0.6230	0.6699	**0.9026**

In the Anhui whole-hand SvH experiment, only the downstream fine-tuning resolution is changed (Table B18). Performance improves from 224×224 to 384×384, reducing RMSE from 18.231 to 15.195 and increasing PCC from 0.903 to 0.933. At 448×448, performance remains higher than the 224×224 configuration but is lower than the 384×384 result. We therefore report a task-specific benefit of moderately higher downstream resolution, not a general claim that increasing resolution always improves performance.

B.7 Full Fine-Tuning Protocol and Frozen-Backbone Analysis

Unless explicitly noted, the principal foundation-model comparisons use end-to-end fine-tuning with each method initialized from its own released checkpoint and trained on the same downstream data partition and task-specific evaluation protocol. Thus, the main results reflect adapted models, not frozen-backbone linear probes. The exception analyzed here holds the visual encoder fixed while fitting the same lightweight regression head on Anhui whole-hand SvH scoring.

As reported in Table B19, HandXFM achieves the highest frozen-backbone PCC (0.684) and R^2 (0.333), together with the lowest RMSE and MAE of the evaluated frozen encoders. This provides complementary evidence that the pretrained representation makes task-relevant features accessible without encoder updates. The frozen HandXFM result is nevertheless substantially weaker than its fully fine-tuned performance in the main paper; consequently, the probe should not be interpreted as a replacement for downstream adaptation.

B.8 Multi-Seed Stability on AIRINIIReumHands

Because AIRINIIReumHands contains relatively few labeled radiographs, single-run ablations can be sensitive to the downstream random seed. We therefore repeated the fixed-split evaluation using five seeds $\{0, 1, 2, 42, 3407\}$ and report the mean and standard deviation for each metric in Table B20. These experiments quantify variability over the selected random seeds; they do not establish uncertainty with respect to patient resampling or changes in clinical site.

HandXFM has the lowest mean erosion RMSE (0.3201) and highest erosion PCC (0.9568). For JSN, it also has the best mean RMSE (0.7067), R^2 (0.5810), and PCC (0.8844), although the random-masking variant has a lower mean MAE (0.4692 versus 0.5470). Residual standard deviations are non-negligible—for example, HandXFM JSN PCC varies by 0.0216 across seeds—so the complete design is best characterized as having stronger mean performance on most reported measures, rather than being uniformly superior or insensitive to initialization.

Table B13: Ablation on loss composition—bone segmentation, including instance and pixel-level metrics. Higher is better.

Metric	Alignment Only	Reconstruction Only	Combined
Instance-level Classification			
ACC ↑	0.9722	0.9512	**0.9751**
P ↑	0.9436	**0.9481**	**0.9481**
R ↑	0.9236	0.9036	**0.9264**
F1 ↑	0.9334	0.9251	**0.9371**
Instance-level Localization			
mAP@0.5:0.95 ↑	0.5071	0.2444	**0.6101**
mAP@0.5 ↑	0.9107	0.6872	**0.9466**
mAP@0.75 ↑	0.5056	0.0984	**0.7030**
Pixel-level Segmentation			
ACC ↑	**0.9602**	0.9600	0.9586
P ↑	**0.4628**	0.4616	0.4596
R ↑	**0.5016**	0.4974	0.4984
F1 ↑	**0.4804**	0.4769	0.4777

B.9 VQA case analysis

The MedPix-derived closed-set VQA task serves as an exploratory assessment of cross-modal transfer rather than a large-scale clinical VQA benchmark. HandXFM pretraining does not use paired image–text supervision directly. Instead, the student learns from BiomedCLIP's frozen, language-aligned visual encoder. During downstream adaptation, HandXFM image features interact with a frozen BiomedCLIP text encoder through cross-attention, followed by an MLP classifier over a fixed answer vocabulary.

Figure B5 presents four radiographs with correct Top-1 predictions but different confidence levels. The first three radiographs share the same caption-derived question and ground-truth diagnosis of chondrocalcinosis, with Top-1 probabilities of 78.3%, 84.0%, and 84.7%, respectively. The fourth example, involving traumatic phalanx dislocation, is also correctly classified but has a lower Top-1 probability of 23.5%. These examples illustrate variation in prediction confidence within the closed-set task rather than distinct levels of diagnostic reasoning.

This evaluation remains limited by the small dataset and the use of existing case descriptions and diagnostic titles as question–answer pairs. Furthermore, the VQA adapter operates on visual patch tokens without explicit joint or bone localization. The predicted probabilities and cross-attention maps do not establish whether the model relies on anatomically appropriate evidence. Accordingly, these qualitative results illustrate the behavior of the multimodal adapter but do not demonstrate open-ended clinical reasoning or provide causal explanations of its predictions.

Table B14: Ablation on masking strategy—regression. Lower is better for RMSE/MAE; higher is better for R^2/PCC.

Dataset	Metric	Random Masking	Edge-aware Masking
Bone Age Prediction			
RSNA	RMSE ↓	10.729	**9.932**
	MAE ↓	7.870	**7.343**
	R^2 ↑	0.934	**0.943**
	PCC ↑	0.967	**0.972**
RHPE	RMSE ↓	11.650	**11.073**
	MAE ↓	8.789	**8.363**
	R^2 ↑	0.921	**0.929**
	PCC ↑	0.962	**0.966**
SvH Scoring			
AIRINII (Erosion)	RMSE ↓	0.438	**0.298**
	MAE ↓	0.127	**0.111**
	R^2 ↑	0.391	**0.719**
	PCC ↑	**0.941**	0.929
AIRINII (JSN)	RMSE ↓	**0.689**	0.705
	MAE ↓	**0.550**	0.608
	R^2 ↑	**0.618**	0.601
	PCC ↑	0.812	**0.885**
Anhui (SvH)	RMSE ↓	21.402	**18.231**
	MAE ↓	13.911	**12.356**
	R^2 ↑	0.740	**0.812**
	PCC ↑	0.884	**0.903**

C Dataset Details

Publicly accessible hand-radiograph datasets remain relatively limited compared with large-scale natural-image and chest-radiograph resources. We curated approximately 30K hand radiographs from six publicly available datasets, encompassing skeletal development, rheumatoid arthritis, musculoskeletal abnormalities, and image-associated clinical descriptions. These resources cover nearly all major public hand-radiograph datasets identified as relevant to the tasks investigated here. Bringing them together broadens the variation in anatomy, patient populations, and imaging conditions available for representation learning. HandXFM leverages this curated collection to learn transferable visual representations without requiring task-specific labels during pretraining.

Table C21 distinguishes images used without labels for HandXFM pretraining from the labeled examples used to fit each downstream task and those reserved for evaluation. The unlabeled pretraining set is constructed only from the designated training partitions of RSNA, RHPE, MURA Hand, AIRINIIReumHands, Anhui RA, and MedPix Hand. Reusing a training image during downstream fine-tuning is intentional; *held-out evaluation images are not used in either pretraining or downstream fitting*. The Digital Hand Atlas (DHA) is excluded entirely until external evaluation.

The allocation is defined before downstream training. We preserve the source study- or patient-level grouping when the corresponding identifiers or official grouping information are available, notably for MURA. For collections without reliable patient identifiers, such as the supplied Anhui image

Table B15: Ablation on masking strategy—MURA normal/abnormal classification. Higher is better for all metrics.

Metric	Random Masking	Edge-aware Masking
AUC ↑	0.7739	**0.8599**
ACC ↑	0.7098	**0.7857**
F1 ↑	0.5860	**0.6903**
P ↑	0.7360	**0.8843**
R ↑	0.4868	**0.5661**

Table B16: Ablation on masking strategy—17-class joint detection. Higher is better for all metrics.

Metric	Random Masking	Edge-aware Masking
mAP@0.5:0.95 ↑	0.7125	**0.7636**
mAP@0.5 ↑	**0.9844**	0.9802
mAP@0.75 ↑	0.8485	**0.9026**

split, patient-level duplication cannot be fully ruled out by split membership alone. Checksum-based comparison of image manifests is used to screen for identical files across pretraining and held-out evaluation partitions; this does not establish the absence of all near-duplicates or unidentified repeated patients. The reported dataset sizes and separation protocol are therefore interpreted at the level supported by the available metadata.

C.1 Public Datasets

C.1.1 RSNA Bone Age Challenge.

The RSNA Pediatric Bone Age dataset consists of left-hand radiographs with skeletal-age labels in months and recorded sex information. The reference age is based on multiple radiologist readings, including repeat assessment, combined through a bias-adjusted weighting procedure. We use its labeled training partition for unlabeled pretraining and for downstream bone-age fitting. The labeled validation partition is reserved for model evaluation; the challenge test images do not provide public reference labels and are not scored in our experiments. The same source partitions also determine the separation of our newly annotated detection and segmentation examples (Section C.2).

The available demographic summary for the training images is reproduced in Table C23. These counts describe radiographs in the provided release; they should not be interpreted as an independently verified count of unique patients.

C.1.2 Radiological Hand Pose Estimation (RHPE).

RHPE contains pediatric hand radiographs acquired in Colombia, with sex and bone-age annotations. The release includes both left- and right-hand images and additionally provides anatomical hand-pose landmarks. For HandXFM, the image-level bone-age labels support regression transfer, while only the training radiographs enter the unlabeled pretraining pool. Validation and test images remain excluded from pretraining. The demographic and partition summaries appear in Tables C24 and C25.

The RHPE hand-pose annotations were generated through a multi-annotator protocol for 17 landmarks and are not used as supervision in HandXFM pretraining. This distinction matters because the model is pretrained on image pixels without the bone-age or landmark labels, even when the original source dataset makes those annotations available.

Table B17: Ablation on masking strategy—bone segmentation, including instance and pixel-level measures. Higher is better.

Metric	Random Masking	Edge-aware Masking
Instance-level Classification		
ACC ↑	0.9713	**0.9751**
P ↑	0.9419	**0.9481**
R ↑	0.9227	**0.9264**
F1 ↑	0.9321	**0.9371**
Instance-level Localization		
mAP@0.5:0.95 ↑	0.4680	**0.6101**
mAP@0.5 ↑	0.8914	**0.9466**
mAP@0.75 ↑	0.4617	**0.7030**
Pixel-level Segmentation		
ACC ↑	**0.9596**	0.9586
P ↑	0.4555	**0.4596**
R ↑	**0.5178**	0.4984
F1 ↑	**0.4836**	0.4777

Table B18: Effect of downstream input resolution on Anhui whole-hand SvH scoring. All variants share the 224×224 pretrained backbone and differ only in fine-tuning resolution.

Input Resolution	RMSE ↓	MAE ↓	R^2 ↑	PCC ↑
224×224	18.231	12.356	0.812	0.903
384×384	**15.195**	**10.513**	**0.869**	**0.933**
448×448	16.404	11.601	0.847	0.928

C.1.3 Digital Hand Atlas (DHA).

DHA provides pediatric left-hand radiographs with skeletal-age assessments and demographic information collected to represent different population groups. We use **only the first collection cycle of 1,103 images** for external bone-age evaluation. DHA contributes neither unlabeled pretraining images nor downstream training samples: the bone-age model trained on RSNA is applied to DHA without DHA-specific adaptation. The original atlas also includes a subsequent 287-image collection cycle, which is listed for dataset context in Table C26 but is not part of the reported external benchmark.

The population and acquisition differences between DHA and RSNA make DHA a useful test of transfer beyond the distributions contributing to training. Its data are strictly external to all HandXFM model fitting.

C.1.4 AIRINIIReumHands.

AIRINIIReumHands contains 330 bilateral hand radiographs, including 42 healthy controls and 288 rheumatoid-arthritis cases. The dataset reports joint-level erosion and joint-space narrowing (JSN) scores derived from multiple expert readers, with an average patient age of 48 ± 14 years. The modified Sharp-style annotation scheme considers 17 erosion sites and 19 JSN sites per hand

Table B19: Frozen-backbone comparison on Anhui whole-hand SvH scoring; only the regression head is fitted.

Method	RMSE ↓	MAE ↓	R^2 ↑	PCC ↑
ResNet-50	35.288	23.869	0.290	0.602
RadFM	36.717	26.092	0.236	0.521
RadCLIP	36.579	22.830	0.241	0.654
HandXFM	**34.284**	**21.534**	**0.333**	**0.684**

Table B20: Multi-seed evaluation on AIRINIIReumHands over five runs with a fixed split and seeds $\{0, 1, 2, 42, 3407\}$. Mean $\pm$ standard deviation.

Task	Method	RMSE ↓	MAE ↓	R^2 ↑	PCC ↑
Erosion	RheumaVIT	0.4166 ± 0.0426	0.2998 ± 0.0537	0.4443 ± 0.1109	0.7487 ± 0.0355
	RadFM	0.3790 ± 0.0148	0.2148 ± 0.0204	0.5441 ± 0.0355	0.7572 ± 0.0187
	RadCLIP	0.3729 ± 0.0985	0.1308 ± 0.0299	0.5287 ± 0.2282	0.8969 ± 0.0755
	HandXFM (random masking)	0.3256 ± 0.1064	0.1248 ± 0.0176	0.6281 ± 0.2114	0.9207 ± 0.0439
	HandXFM (ours)	$\mathbf{0.3201 \pm 0.0507}$	$\mathbf{0.1236 \pm 0.0139}$	$\mathbf{0.6672 \pm 0.0914}$	$\mathbf{0.9568 \pm 0.0137}$
JSN	RheumaVIT	0.7584 ± 0.0563	0.5156 ± 0.0484	0.5357 ± 0.0718	0.8237 ± 0.0244
	RadFM	0.7642 ± 0.0306	0.5760 ± 0.0180	0.5304 ± 0.0378	0.7793 ± 0.0223
	RadCLIP	0.8904 ± 0.0783	0.7240 ± 0.0735	0.3585 ± 0.1104	0.7852 ± 0.0865
	HandXFM (random masking)	0.7156 ± 0.1287	$\mathbf{0.4692 \pm 0.1463}$	0.5756 ± 0.1635	0.8167 ± 0.0276
	HandXFM (ours)	$\mathbf{0.7067 \pm 0.1497}$	0.5470 ± 0.1375	$\mathbf{0.5810 \pm 0.1738}$	$\mathbf{0.8844 \pm 0.0216}$

(Table C28). HandXFM uses 240 training images without their labels in pretraining; the same partition supplies downstream training labels, with 60 validation and 30 test images kept separate. Its small supervised training size motivates the five-seed stability analysis in the Results Details appendix.

The released scores use a Dawid–Skene aggregation approach to combine annotator responses. Joint-region bounding boxes are available for a subset of 268 radiographs; these source annotations are distinct from the RSNA-based 17-class localization dataset created in this work. We retain the source-dataset characteristics in Table C29 to distinguish its annotation protocol from our own.

C.1.5 Anhui RA Dataset.

The Anhui RA collection contains 3,818 bilateral-hand examinations annotated for rheumatoid-arthritis damage severity. Radiologist-derived erosion, JSN, and overall Sharp–van der Heijde (SvH) scores provide supervision for the whole-hand regression task. Only the **2,700 training images** are included in HandXFM pretraining. We use the corresponding image-level split of 760 validation and 358 test images for model assessment; the three partitions sum to 3,818. The provided release does not support an independent audit of patient-level overlap, which is a limitation of this split.

C.1.6 MURA Hand Subset.

We select the `XR_HAND` category from the MURA musculoskeletal radiography release for normal-versus-abnormal classification. The selected subset contains 6,003 radiographs, including 5,543 training images and 460 validation images (Table C32). A study may contain multiple radiographic

Table C21: Allocation of source datasets to unlabeled HandXFM pretraining, downstream model fitting, and held-out evaluation. Newly annotated localization images follow the official RSNA partitions.

Dataset	Downstream task	Unlabeled pretraining	Downstream fitting	Held-out evaluation / separation
RSNA Bone Age	Bone-age regression	12,611 training images	Official training partition	1,425 labeled validation images; 200 unlabeled challenge test images not evaluated. Official partitions.
RHPE	Bone-age regression	5,487 training images	Training partition	713 validation; 79 test images. Disjoint splits.
MURA Hand	Abnormality classification	5,543 training images	Training partition	460 validation images. Official study-disjoint partition.
AIRINIIReumHands	Joint-level SvH regression	240 training images	240 training images	60 validation; 30 test images. Disjoint splits.
Anhui RA	Whole-hand SvH regression	2,700 training images	2,700 training images	760 validation; 358 test images. Image-level split.
MedPix Hand	Closed-set VQA	97 training images	97 training image–text pairs	11 held-out test pairs. Test images excluded.
DHA	External bone-age transfer	**Not used**	**Not used**; RSNA-trained predictor transferred directly	1,103 external images. Entirely excluded.
Joint detection (ours)	17-class localization	Already included through RSNA training images	RSNA-train annotations	RSNA-validation annotations; 6,130 images annotated in total. RSNA partitions.
Bone segmentation (ours)	19-class segmentation	Already included through RSNA training images	RSNA-train annotations	RSNA-validation annotations; 549 images annotated in total. RSNA partitions.

Table C22: RSNA bone-age partition sizes and reported mean skeletal age. The challenge test labels are unavailable for the evaluations in this paper.

Split	#Images	Mean Age (mo)	Labels
Training	12,611	127	Yes
Validation	1,425	127	Yes
Test	200	132	No

views. HandXFM pretraining uses only images from the designated training partition, while the held-out validation partition is reserved for abnormality classification.

C.1.7 MedPix Dataset.

The MedPix teaching-case archive provides radiographic images linked to diagnostic titles and descriptive text [27]. We select 108 hand X-ray image–text examples and create a closed-set evaluation by treating the descriptive caption as a question-like textual input and the diagnostic title as the answer label. The resulting split contains **97 training pairs and 11 held-out test pairs**. Only training images participate in HandXFM's unlabeled pretraining, and no test image is reused in model

Table C23: RSNA bone-age training-image demographics, using the reported sex and age categories.

Metric	Value
Training images	12,611
Male-labeled images	6,833
Female-labeled images	5,778
Mean Age (years)	10.61
0–5 years	801
6–10 years	3,487
11–15 years	7,224
16–20 years	1,099

Table C24: RHPE partition sizes and reported mean bone ages.

Split	#Images	Mean Age (mo)	Labels
Training	5,487	126	Yes
Validation	713	126	Yes
Test	79	126	Yes

fitting. Since the questions are derived from existing case descriptions rather than independently authored clinical queries, this benchmark measures constrained multimodal transfer and should not be regarded as open-ended clinical VQA.

C.2 Newly Annotated Datasets

Joint Detection Dataset. We construct a 17-category joint/landmark detection benchmark by annotating bounding boxes in 6,130 RSNA radiographs. The classes cover DIP joints of digits 2–5, proximal interphalangeal joints of digits 2–5 and the thumb interphalangeal joint (`PIP_1` in the dataset label scheme), MCP joints of digits 1–5, two distal forearm landmarks, and the wrist center. The 17 target categories comprise the DIP, PIP/IP, and MCP joints, together with distal radius, distal ulna, and wrist-center landmarks. The corresponding class names and anatomical definitions are listed in Table C33.

Annotation images were sourced from the official RSNA train and validation partitions, at approximately a 9:1 ratio. Detector fitting uses only labels from the RSNA training partition, and evaluation uses annotations on RSNA validation images. Since unlabeled pretraining likewise excludes RSNA validation images, no detection-evaluation image is included in HandXFM pretraining or detector training.

Bone Segmentation Dataset. We annotate 549 RSNA radiographs with pixel-level masks for 19 individual metacarpal and phalangeal categories. Annotations from the RSNA training partition are used for downstream model fitting, whereas those from the validation partition are reserved for evaluation. The complete class definitions and output-label mapping are provided in Table C34.

Annotation Standards. Our anatomical class definitions follow *Terminologia Anatomica* (2nd edition) and are cross-checked against RSNA RadLex terminology. Digits are indexed 1 through 5 from thumb to little finger. The anatomical naming convention requires a small distinction from the stored detection class name: `PIP_1` refers to the thumb's single interphalangeal joint, not to a separate proximal interphalangeal joint. The distal-radius, distal-ulna, and wrist-center classes are localization landmarks rather than additional finger joints. For segmentation, each integer class ID from 0 to 18 corresponds to one specific bone, as defined in Table C34.

Table C25: Reported demographics of RHPE training radiographs.

Metric	Value
Training images	5,487
Male-labeled images	2,524
Female-labeled images	2,963
Mean Age (years)	10.5
0–5 years	340
6–10 years	1,590
11–15 years	2,720
16–20 years	837

Table C26: DHA collection cycles. Only the first cycle is used for the reported external-transfer experiment.

Cycle	#Images	Age range (years)	Readings per image
First Cycle	1,103	1–18	2
Second Cycle	287	5–14	4

Two annotators with musculoskeletal training participated in constructing and reviewing the new annotations. Initial bounding boxes and masks were produced by one annotator using Roboflow, with subsequent joint review and consensus resolution. An automated quality-assurance screen flagged incomplete cases, missing or repeated class labels, unexpected category counts, questionable hand orientation, and anatomically implausible spatial arrangements; flagged annotations were reviewed before finalization. A further audit comprised 1,000 randomly sampled detection images and 100 segmentation masks, reviewed by both annotators for landmark placement, mask alignment, and correct category assignment. Because the annotation process used consensus review rather than two independent full annotation sets, inter-annotator IoU or Cohen's κ cannot be reported as independent agreement measures.

D Limitations and Future Work

Despite its performance across heterogeneous downstream tasks, HandXFM has several limitations related to expert supervision, data coverage, and downstream adaptation.

Expert fusion formulation. HandXFM combines the latent targets of two frozen visual experts through a shared alignment and reconstruction framework. Although the teacher-composition ablations demonstrate complementary transfer across tasks, the current formulation assigns equal importance to the two experts rather than adapting their contributions to individual images or anatomical regions. Moreover, the relative loss weights were selected based on preliminary loss-scale observations rather than a comprehensive hyperparameter sensitivity study. The objective ablations establish the contribution of alignment and reconstruction, but do not characterize sensitivity to all possible weighting configurations. Future work could investigate adaptive expert weighting and alternative fusion mechanisms, particularly for tasks with different requirements for global semantic information and localized anatomical detail.

Scope of pretraining data. We curated approximately 30K hand radiographs from six publicly available datasets, covering multiple clinical tasks and imaging conditions. Nevertheless, the available data remain limited in comparison with larger medical imaging domains. In particular, uncommon abnormalities, advanced disease manifestations, and variations in acquisition protocols may not be

Table C27: Demographic and image-format information for the evaluated DHA first-cycle collection.

Metric	Value
Total Patients	1,103
Male Patients	553
Female Patients	550
Ethnic Groups	CA, AA, HI, AS
Mean Age (years)	9.5
Age Range	1–18
Image Resolution	2K $\times$ 2K
Format	DICOM

Table C28: Scored anatomical sites in AIRINIIReumHands (per hand).

Regions	#Sites
Erosion: Interphalangeal, 4 DIP, 4 PIP, 5 MCP, 1st metacarpal base, wrist, ulnar, radiocarpal	17
JSN: Interphalangeal, 4 DIP, 4 PIP, 5 MCP, CMC (3–5), wrist, radiocarpal	19

sufficiently represented. The diversity of the training sources also does not guarantee equivalent performance across institutions or patient subgroups. Additional datasets encompassing broader patient populations, disease severity, and imaging conditions would enable more extensive assessment of representation transfer and robustness.

Downstream adapter design. HandXFM is evaluated using relatively lightweight task-specific heads to assess the transferability of its pretrained representations. However, these adapters may not fully exploit the anatomical information encoded by the backbone. For example, the VQA adapter applies cross-attention over visual patch tokens without explicitly incorporating detected joints or segmented bones. This can limit its ability to distinguish abnormalities requiring precise assessment of local geometry, cortical continuity, or relationships between anatomical structures.

The MedPix-derived VQA evaluation is also limited by its small sample size and closed-set answer formulation. Consequently, performance on this benchmark does not establish open-ended clinical reasoning capability. Similarly, the cross-attention visualizations provide qualitative examples of text–image correspondence rather than validated causal explanations of model predictions. Future work could investigate anatomy-conditioned multimodal adapters and larger benchmarks with region-level text annotations to evaluate visual grounding more rigorously.

E Reproducibility Check

We summarize the computational environment, data processing procedures, and experimental configurations used for HandXFM pretraining and downstream evaluation. Additional architecture and optimization details are provided in Section A, while dataset-specific partitions are described in Section C.

E.1 Compute

Hardware. Experiments were conducted on a Linux server equipped with an NVIDIA A100 GPU and an Intel Xeon Platinum 8480+ CPU.

Table C29: Clinical and labeling characteristics reported for AIRINIIReumHands.

Metric	Value
Total Patients	330
Healthy Controls	42
RA Patients	288
Average Age (years)	48 ± 14
Acquisition Period	2019–2022
Hand Views	Bilateral (1 radiograph per subject)
Annotation Experts	3 radiologists (5, 7, and 30 years of experience)
Scoring Method	Modified Sharp (erosion: 0–5, JSN: 0–4)
Consensus Algorithm	Dawid–Skene aggregation
Joint Locations Annotated	21 per hand (bounding boxes)

Table C30: Anhui RA image-level partitions used for pretraining and SvH regression.

Split	#Images	Annotations	Method
Training	2,700	JSN, Erosion, Sharp score	SvH
Validation	760	JSN, Erosion, Sharp score	SvH
Test	358	JSN, Erosion, Sharp score	SvH

Software. HandXFM is implemented in PyTorch. The implementation uses Hugging Face Transformers for the student backbone and radiography-domain teacher, `open_clip` for BiomedCLIP, and `timm` and `torchvision` for model components and image processing.

Reproducibility controls. For each downstream benchmark, all compared models use the designated training and evaluation partitions and the same task-specific metric definitions. Model initialization, data processing, optimization, and masking configurations are specified independently of the evaluation data. The multi-seed experiments on AIRINIIReumHands additionally quantify variability across different random initializations, as detailed in Section B.8.

E.2 Data organization and preprocessing

Data roots. Pretraining images are organized by dataset source and loaded from configurable root directories. The image loader supports commonly used radiographic image formats, including PNG, JPEG, and TIFF. Dataset-specific training partitions determine which images are eligible for pretraining. The allocation of images across pretraining, downstream training, and evaluation is summarized in Table C21.

Transforms. Pretraining applies resizing, random horizontal flipping, rotation, and center cropping to obtain 224×224 inputs. Pixel intensities are converted to $[0, 1]$, followed by the preprocessing required by each pretrained encoder. To avoid repeated input scaling, the DINOv3 image processor uses `do_rescale=False`. The complete augmentation settings are reported in Section A.2.

Pretraining split policy. HandXFM uses only the designated training partitions of the six source datasets for unlabeled pretraining. Validation and test images are excluded, and the Digital Hand Atlas is held out for external bone-age evaluation. For the RSNA-derived joint-detection and bone-segmentation benchmarks, downstream training annotations originate from the RSNA training

Table C31: Anhui RA source-dataset characteristics and the train/validation/test partition counts.

Metric	Value
Total Radiographs	3,818
Institution	Anhui University of Traditional Chinese Medicine
Acquisition Period	2020–2022
Average Age (years)	Not reported
Annotation Experts	2 radiologists
Scoring System	Sharp–van der Heijde (SvH)
Labeled Attributes	JSN, Erosion, Sharp Score
Split Ratio (Train/Val/Test)	2700 / 760 / 358
Image Format	X-ray (DICOM)
Diagnosis Types	RA or suspected RA

Table C32: MURA hand-subset image statistics for the training and validation partitions.

Split	Images	Normal	Abnormal	Normal ratio
Training	5,543	4,059	1,484	0.732
Validation	460	271	189	0.589
Total	**6,003**	**4,330**	**1,673**	**0.721**

partition, whereas evaluation annotations are drawn from the disjoint validation partition. The corresponding split definitions and overlap controls are documented in Table C21.

E.3 Training configuration

Optimization. Pretraining uses AdamW with separate learning rates for the student encoder and auxiliary modules, including the latent regressor and expert projection layers. The student is frozen during the initial two epochs and subsequently optimized jointly with the auxiliary modules. The optimizer settings, learning rates, batch size, and training schedule are specified in Table A4. Checkpoints are saved periodically throughout pretraining.

Masking and losses. The training configuration specifies the active expert encoders, masking strategy, masking ratio, and coefficients of the alignment, reconstruction, CLS, and EMA consistency terms. The main configuration employs edge-aware masking with a ratio of 0.5 and an edge-mixing coefficient of 0.5. Ablation variants modify the corresponding components while retaining the remaining training settings, as described in Section A.5.

E.4 Code entry points and artifacts

Code entry points. The pretraining pipeline is configured through runtime arguments specifying dataset paths, pretrained checkpoints, expert selection, masking settings, loss coefficients, optimizer parameters, and output directories. Downstream experiments use task-specific training and evaluation procedures with the corresponding pretrained encoder weights. The configurations for the main model and pretraining variants follow the settings reported in Section A.

Saved artifacts. Training checkpoints contain the student and EMA encoder states, masked latent regressor, expert projection modules, optimizer state, and training progress information. The EMA encoder from the final pretraining checkpoint is used as the default HandXFM backbone for

Table C33: Model-label mapping for the 17 RSNA-derived joint/landmark detection categories.

Joint label	Anatomical definition
DIP_5	Distal interphalangeal joint of the little finger
DIP_4	Distal interphalangeal joint of the ring finger
DIP_3	Distal interphalangeal joint of the middle finger
DIP_2	Distal interphalangeal joint of the index finger
PIP_5	Proximal interphalangeal joint of the little finger
PIP_4	Proximal interphalangeal joint of the ring finger
PIP_3	Proximal interphalangeal joint of the middle finger
PIP_2	Proximal interphalangeal joint of the index finger
PIP_1	Interphalangeal joint of the thumb (dataset label `PIP_1`)
MCP_5	Metacarpophalangeal joint of the little finger
MCP_4	Metacarpophalangeal joint of the ring finger
MCP_3	Metacarpophalangeal joint of the middle finger
MCP_2	Metacarpophalangeal joint of the index finger
MCP_1	Metacarpophalangeal joint of the thumb
Radius	Distal tip of the radius bone
Ulna	Distal tip of the ulna bone
Wrist	Center point of the carpal (wrist) region

downstream adaptation. Training logs record the optimization process, including the individual loss components, while downstream evaluation produces the task-specific metrics reported in the main paper and supplementary experiments.

Code and checkpoint availability. The HandXFM pretraining implementation, dependencies, and setup instructions are publicly available at `https://github.com/YUESAI/HandXFM`. The pretrained visual encoder, model configuration, and usage documentation are available at `https://huggingface.co/Yuesai/HandXFM`.

F Broader Impact

HandXFM provides a reusable visual representation for hand radiographs, with potential applications in skeletal maturity estimation, rheumatoid arthritis assessment, abnormality screening, and anatomical localization. By transferring knowledge from existing biomedical and radiography-domain models, the framework may reduce the need to develop separate visual representations for individual hand X-ray tasks. This could facilitate research on musculoskeletal imaging applications where task-specific annotated data are limited.

However, HandXFM is a research model and has not been established as a standalone clinical diagnostic system. Its performance may be affected by differences in patient demographics, disease prevalence, imaging equipment, and acquisition protocols. In particular, underrepresented anatomical abnormalities or patient populations may exhibit different error characteristics from those observed in the evaluated benchmarks. The present results therefore should not be interpreted as evidence of clinical readiness or uniformly reliable performance across patient subgroups.

Responsible clinical translation would require independent external validation, systematic evaluation of demographic and acquisition-related performance differences, and prospective assessment under clinical workflows. Automated predictions should be interpreted with appropriate clinical oversight, particularly when errors could affect diagnostic or treatment decisions.

Table C34: Output indices for the 19 anatomically distinct bone-segmentation classes.

Label ID	**Anatomical definition**
0	Metacarpal bone of the index finger
1	Metacarpal bone of the middle finger
2	Metacarpal bone of the thumb
3	Metacarpal bone of the ring finger
4	Metacarpal bone of the little finger
5	Proximal phalanx of the middle finger
6	Proximal phalanx of the index finger
7	Proximal phalanx of the ring finger
8	Proximal phalanx of the thumb
9	Proximal phalanx of the little finger
10	Middle phalanx of the middle finger
11	Middle phalanx of the ring finger
12	Middle phalanx of the index finger
13	Distal phalanx of the thumb
14	Middle phalanx of the little finger
15	Distal phalanx of the middle finger
16	Distal phalanx of the ring finger
17	Distal phalanx of the index finger
18	Distal phalanx of the little finger

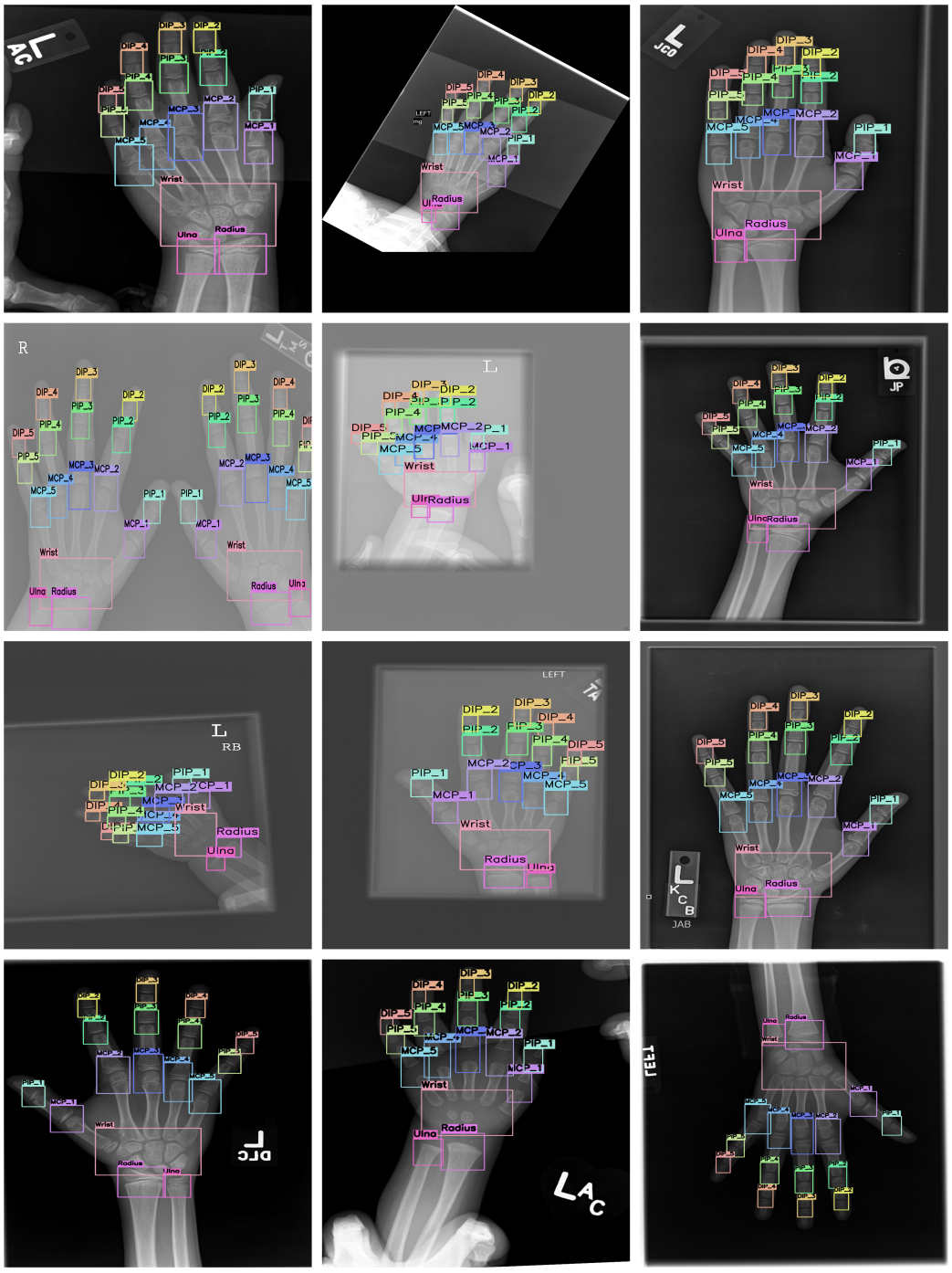

Figure B3: Representative predictions on the held-out RSNA-based joint-detection evaluation subset. The panels illustrate localization under changes in hand orientation, finger configuration, and radiographic appearance.

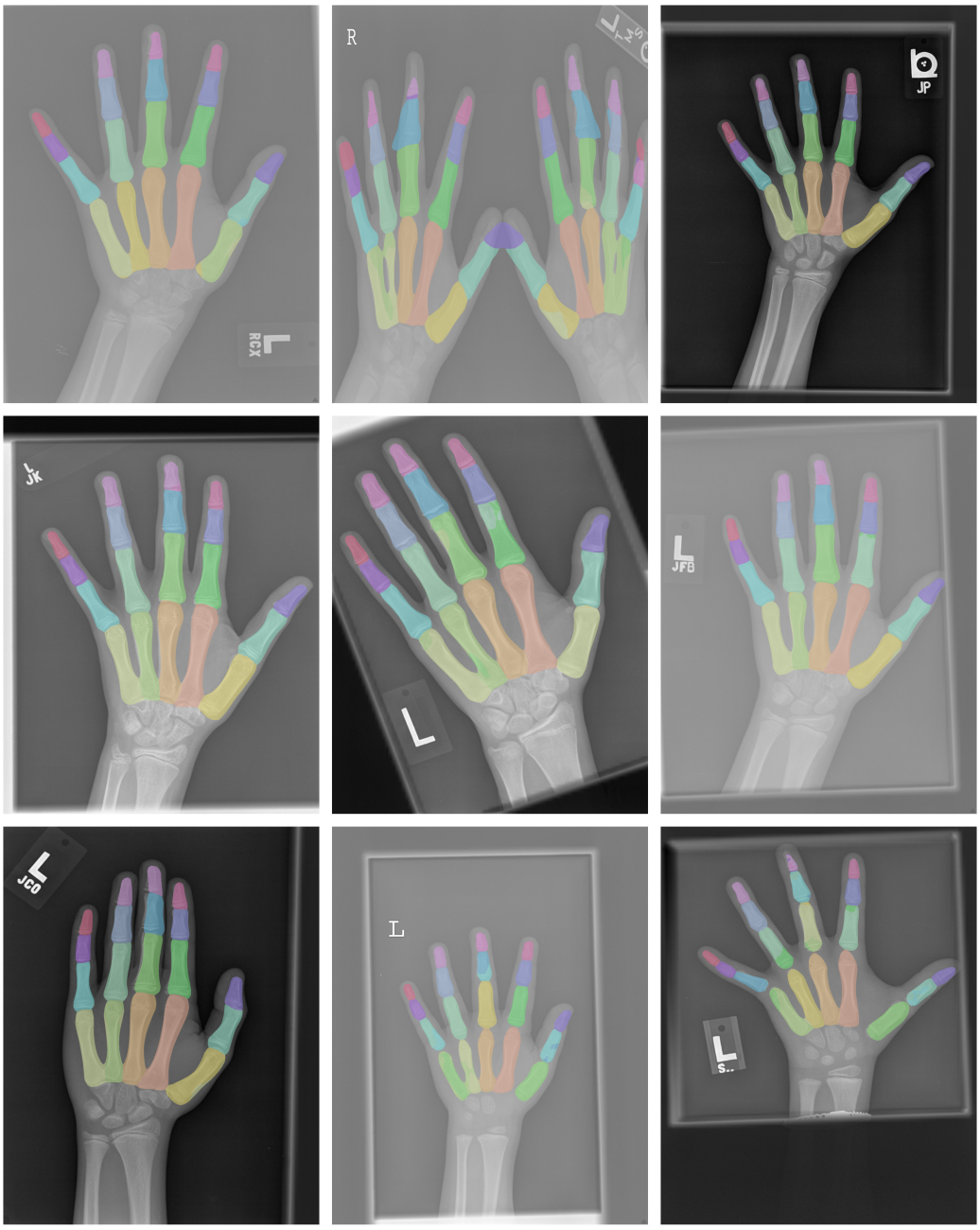

Figure B4: Qualitative 19-class bone-segmentation results on held-out radiographs. The color-coded masks distinguish individual annotated bone categories and illustrate predictions across varying image appearances and hand configurations.

Example	Radiograph	Question, ground truth, and top-5 predictions
A. Correct prediction		**Question:** Frontal, lateral and oblique radiographs of the left hand demonstrate chondrocalcinosis within the triangular fibrocartilage complex as well as degenerative changes at the scaphotrapezial joint evidenced by joint space narrowing and subchondral sclerosis. The remainder of the osseous and soft tissue structures are unremarkable. **GT:** Chondrocalcinosis **Top-5 predictions:** 1. Chondrocalcinosis (78.3%) ✓ 2. Transscaphoid–transtriquetrial perilunate fracture dislocation (9.2%) 3. Dislocation at the interphalangeal joint of the thumb with avulsion fracture (3.8%) 4. Perilunate dislocation (3.7%) 5. Rheumatoid arthritis (0.8%)
B. Correct prediction		**Question:** Same caption-derived question as Example A. **GT:** Chondrocalcinosis **Top-5 predictions:** 1. Chondrocalcinosis (84.0%) ✓ 2. Transscaphoid–transtriquetrial perilunate fracture dislocation (7.8%) 3. Bilateral Salter Harris Fracture Type 2A (3.3%) 4. Dislocation at the interphalangeal joint of the thumb with avulsion fracture (0.9%) 5. Rheumatoid Arthritis (0.7%)
C. Correct prediction		**Question:** Same caption-derived question as Example A. **GT:** Chondrocalcinosis **Top-5 predictions:** 1. Chondrocalcinosis (84.7%) ✓ 2. Transscaphoid–transtriquetrial perilunate fracture dislocation (9.2%) 3. Dislocation at the interphalangeal joint of the thumb with avulsion fracture (1.4%) 4. Bilateral Salter Harris Fracture Type 2A (1.3%) 5. Perilunate dislocation (1.2%)
D. Lower-confidence correct prediction		**Question:** Three views of the left hand clearly demonstrate the dislocation. Note the protuberance of the head of the proximal phalynx through the volar soft tissues of the digit. **GT:** Traumatic Phalanx Dislocation **Top-5 predictions:** 1. Traumatic Phalanx Dislocation (23.5%) ✓ 2. Bizarre Parosteal Osteochondromatous Proliferation (BPOP) (16.8%) 3. Osteopoikilosis (12.6%) 4. Dislocation at the interphalangeal joint of the thumb with avulsion fracture (7.6%) 5. Pathologic Fracture in Giant Cell Tumor (6.6%)

Figure B5: Qualitative predictions on four radiographs from the MedPix-derived closed-set VQA benchmark. Examples A–C share the same caption-derived question and diagnosis but correspond to different radiographs. All four examples have correct Top-1 predictions, with varying confidence. The percentages indicate the classifier's predicted probabilities.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the proposed HandXFM framework, the multi-expert pretraining strategy, and the experimental findings across downstream tasks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations and future directions in the supplemental material, including dataset scope, computational cost, and generalization considerations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical theorems or formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper and supplemental material describe datasets, preprocessing, architectures, training settings, optimization details, and evaluation protocols required to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The HandXFM pretraining implementation and setup instructions are publicly available at `https://github.com/YUESAI/HandXFM`, and the pretrained visual encoder checkpoint is released at `https://huggingface.co/Yuesai/HandXFM`. The supplemental material provides detailed experimental protocols.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper and supplemental material specify datasets, preprocessing, architectures, hyperparameters, optimization settings, and evaluation protocols.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The supplementary material reports five-seed evaluations for AIRINI-IReumHands erosion and JSN scoring, including mean and standard deviation for RMSE, MAE, R^2, and PCC. These experiments characterize variation across random initializations under a fixed data split.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The supplemental material describes the hardware and software environment, including GPU type and training setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics and uses publicly available medical imaging datasets for research purposes.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential positive clinical applications as well as limitations and risks related to dataset bias and clinical deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce generative models or high-risk assets that would require special safeguards for responsible release.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Public datasets, pretrained models, and prior methods used in this work are properly cited in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The pretrained HandXFM visual encoder is publicly released with model configuration and usage documentation at `https://huggingface.co/Yuesai/HandXFM`. The pretraining implementation, dependencies, and setup instructions are available at `https://github.com/YUESAI/HandXFM`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or human subject studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The work uses publicly available retrospective datasets and does not involve direct interaction with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the proposed methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.