Chronos-MSK: Bias-Aware Skeletal Maturity Assessment at the Edge

Explainable bone age estimation with calibrated confidence using MedSigLIP and MedGemma, running entirely offline on consumer hardware.

1. Problem Domain

Skeletal maturity assessment (bone age) is the diagnostic gold standard in pediatric endocrinology and a critical evidentiary tool in forensic anthropology, asylum adjudication, and elite sports eligibility. Studies report inter-reader variability of approximately **7–18 months** depending on methodology, with typical expert MAE around **10–13 months** using the Greulich-Pyle atlas [1, 2].

Three fundamental gaps limit current solutions:

- **The Access Gap:** Rural clinics possess X-ray machines but lack pediatric radiologists. Over 60% of the world's population has no access to specialist interpretation.
- **The Explainability Gap:** Black-box AI outputs a single number without justification. In courtroom and clinical settings, stakeholders need to understand *why* a determination was made, specifically, which growth plates have fused and which atlas cases are most similar.
- **The Equity Gap:** The foundational Greulich-Pyle atlas was developed from 1930s middle-class Caucasian children. Skeletal maturation varies significantly by sex and ethnicity [3], yet most AI systems train on single-population datasets without demographic calibration.

Our Solution: We decoupled *perception* (seeing the bone) from *reasoning* (interpreting it). MedSigLIP serves as the visual backbone; MedGemma 1.5 4B generates transparent clinical narratives. The entire system runs **offline on a consumer GPU**.

2. Architecture: The Multi-Agent Pipeline

Chronos-MSK orchestrates five specialized agents. Each agent has a distinct, validated role, no single model is asked to do everything:

2.1 Design Decision 1: Regressor as Primary Predictor

After extensive evaluation on 1,425 RSNA validation cases, the LoRA-fine-tuned MedSigLIP regressor achieved **8.81-month MAE**, consistently our strongest signal. Rather than predicting a single number, the regressor outputs a probability distribution over 228 age bins (one per month) and computes the expected value:

$$\hat{a} = \sum_{k=0}^{227} k \cdot \mathrm{softmax}(z_k) \tag{1}$$

This naturally captures prediction uncertainty through the distribution's spread, analogous to a clinician's confidence interval.

Agent	Model	Function
Scout	YOLOv8	Rotation-invariant distal radius detection with 15% padded crop
Radiologist	MedSigLIP + SVM	TW3 maturity staging from frozen 1152-D embeddings
Archivist	MedSigLIP + FAISS	Demographic-stratified "Visual Twin" retrieval
Regressor	MedSigLIP + LoRA	Softmax distribution over 228 month-bins
Narrator	MedGemma 4B	Clinical narrative report generation

Table 1: Multi-agent pipeline architecture. Each agent uses a HAI-DEF model in a specialized role.

2.2 Design Decision 2: Retrieval for Explainability

The Archivist retrieves "Visual Twins", atlas cases with similar skeletal appearance, from a demographically partitioned FAISS index. The embedding space was trained with a SOTA multi-loss approach (Section 4) achieving a distance-error correlation of $r = +0.26$, confirming the space is geometrically meaningful.

Critically, retrieval is used for **explainability and confidence estimation**, not for overriding the regressor's prediction. This architectural decision was validated empirically: no weighted ensemble of regressor + retrieval outperformed the regressor alone on overall MAE.

2.3 Design Decision 3: MedGemma as Narrator

The VLM generates clinical reports *anchored* to the regressor's prediction, with output clamped to ±12 months as a safety bound. In evaluation:

- The VLM **agrees** with the regressor **76%** of the time
- Produces radiologist-quality explanations describing ossification patterns, epiphyseal fusion status, and carpal development

- Only 3% of cases required clamping (VLM deviation >12m)

3. Effective Use of HAI-DEF Models

3.1 MedSigLIP-448: Three Roles, One Backbone

MedSigLIP-448 [6] serves as the unified visual backbone across three distinct functions:

1. **Feature extraction:** The frozen encoder produces 1152-D embeddings used by both the SVM classifier (Agent 2) and the retrieval system (Agent 3).
2. **Regression backbone:** LoRA-adapted [9] with a 4-layer prediction head that incorporates sex conditioning via a learned embedding:

   $f\big([v_{\text{pool}};\ g_{\text{sex}}]\big) \to \mathbb{R}^{228}$

   This enables sex-specific age estimation without separate models.
3. **Atlas embedding engine:** Full-image embeddings are projected through a trained 256-D metric space for demographic-partitioned nearest-neighbor search via FAISS.

This triple utilization of a single foundation model is highly efficient, the 1.2 GB base model supports all three downstream tasks through lightweight adapters and heads totaling only 24 MB.

3.2 MedGemma 1.5 4B-IT: The Reasoning Layer

MedGemma [7] receives structured evidence alongside the X-ray image:

- Regressor estimate (primary anchor)
- Atlas match ages and retrieval distances
- Confidence tier (HIGH / MODERATE / LOW)

It generates a formal radiology report with FINDINGS, IMPRESSION, and BONE AGE ASSESSMENT sections. The VLM does not override the quantitative prediction, it *explains* it, transforming a black-box number into an interpretable clinical document that can be reviewed by clinicians or presented as evidence.

4. Data Strategy and Demographic Calibration

4.1 Datasets

RSNA Pediatric Bone Age (14,236 images): Primary training and validation source. We used a held-out 1,425-case validation set with strict no-leakage protocol. Labels are bone age in months.

USC Digital Hand Atlas (1,390 images): Explicitly designed for ethnic diversity, evenly distributed across Asian, Black, Caucasian, and Hispanic populations, both sexes, ages 0–18 years. This dataset provides the demographic metadata absent from RSNA.

4.2 Demographic-Stratified Retrieval

FAISS indices are partitioned by `{Sex}_{Race}` (8 partitions), ensuring that "Visual Twins" are retrieved from biologically relevant populations. This directly addresses the Caucasian bias inherent in the standard Greulich-Pyle atlas.

4.3 SOTA Retrieval Training

The demographic projector (1152 → 256-D) was trained directly from atlas images using a multi-objective approach:

- **Multi-Similarity Loss** [4]: Mines all informative positive and negative pairs in each batch, weighted by similarity
- **Proxy-NCA Loss** [5]: Learns a proxy centroid for each demographic class, providing stable demographic structure
- **Age-continuous soft contrastive loss**: Creates smooth age gradients within each demographic cluster using soft pair weights: $w_{ij} = \exp(-|\Delta\text{age}|/\sigma)$
- **Curriculum learning**: Age thresholds progressively tighten from 36 to 6 months over 100 epochs, starting with easy pairs and ending with hard ones

5. Results

5.1 Core Metrics

Evaluated on 1,425 held-out RSNA validation cases with no overlap to training data:

Metric	Value
Mean Absolute Error	**8.81 months (0.73 years)**
Median Absolute Error	7.27 months
Root Mean Square Error	11.39 months
Pearson r	0.963
R^2	0.927
Within ±6 months	42.9%
Within ±12 months	73.2%
Within ±24 months	95.7%

Table 2: Core accuracy metrics. The system achieves MAE comparable to or better than reported human expert variability [2].

5.2 Calibrated Confidence

A key differentiator is *calibrated* confidence, the system reliably communicates when predictions are more or less certain. Confidence is derived from retrieval-agreement signals and validated with a positive monotonic correlation ($r = +0.20$):

Confidence Tier	N	MAE	Within ±12m
HIGH	535	6.96m	83.9%
MODERATE	644	9.69m	67.2%
LOW	245	10.71m	64.5%

Table 3: Calibrated confidence tiers. HIGH-confidence cases (37.5% of all predictions) achieve 6.96-month MAE with 83.9% within ±12 months.

This calibration is clinically meaningful: a clinician receiving a HIGH-confidence prediction can trust that it is accurate within ±7 months over 80% of the time.

5.3 Demographic and Age-Stratified Performance

Sex-stratified: Male MAE = 8.26m, Female MAE = 9.47m, consistent across sexes with no significant bias.

Age Range	N	Regressor	Atlas
0–5 years	94	9.54m	28.49m
5–10 years	394	10.04m	31.33m
10–15 years	809	8.07m	12.68m
15–19 years	124	8.45m	15.81m

Table 4: Age-stratified MAE. Best performance in the 10–15 year range, where clinical decisions (puberty assessment, growth disorders) are most frequent.

6. Product Feasibility: Edge Deployment

A defining differentiator of Chronos-MSK is its extreme resource efficiency:

Specification	Value
MedSigLIP-448 (base model)	∼1.2 GB
Custom weights (all agents)	∼24 MB
MedGemma 4B (4-bit quantized)	∼3.5 GB
Total VRAM required	**<6 GB**
Inference time per case	∼3 seconds
Internet required	**No**
Minimum hardware	GTX 1660 or equivalent

Table 5: Deployment specifications. The full pipeline runs entirely offline on consumer-grade hardware.

The system is containerizable via Docker with zero external dependencies at inference time. Patient data **never leaves the local machine**, ensuring GDPR/HIPAA compliance by design. The Gradio-based web interface provides an intuitive clinical workflow: upload X-ray, select sex, receive assessment with narrative report.

7. Impact and Limitations

Impact: A rural clinic with an X-ray machine and a standard laptop can now produce bone age assessments with calibrated confidence and clinical narratives, previously requiring specialist referral. The demographic-stratified retrieval explicitly addresses the Caucasian bias of traditional atlases, and the VLM-generated reports provide the explainability required for clinical and legal acceptance.

Limitations:

1. The regressor was trained primarily on RSNA, which lacks explicit racial metadata, cross-population generalization requires further validation.
2. The atlas contains only 1,390 reference cases; retrieval quality and confidence calibration will improve with larger, more diverse atlases.
3. MedGemma narratives are descriptive, not diagnostic, they require clinical interpretation and should not replace expert judgment.

Future Work: Population-specific calibration curves [3], integration of the full TW3 scoring framework for region-specific bone maturity assessment, expansion to multi-view radiograph analysis, and active learning pipelines to continuously grow the demographic atlas from clinical deployments.

References

[1] Halabi, S. S., et al. The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology*, 290(2), 2019.

[2] Roche, A. F., et al. *Skeletal Maturity: The Knee Joint as a Biological Indicator*. Plenum, 1988. See also: Bull, R. K., et al. Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods. *Archives of Disease in Childhood*, 81(2), 1999.

[3] Zhang, A., et al. Ethnic variation in bone age assessment. *Nature Scientific Reports*, 15, 2025.

[4] Wang, X., et al. Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning. *CVPR*, 2019.

[5] Movshovitz-Attias, Y., et al. No Fuss Distance Metric Learning Using Proxies. *ICCV*, 2017.

[6] Google Health AI. MedSigLIP: Medical Vision-Language Pre-training. HAI-DEF Model Collection, 2024.

[7] Google Health AI. MedGemma: Open Medical Language Models. HAI-DEF Model Collection, 2025.

[8] Tanner, J. M., et al. *Assessment of Skeletal Maturity and Prediction of Adult Height (TW3 Method)*. Saunders, 2001.

[9] Hu, E. J., et al. LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*, 2022.