CDPM-Align: Multi-Scale Guidance-Aligned Diffusion Pretraining for Robust Few-Shot Anatomical Landmark Detection

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Anatomical landmark detection is a fundamental task in medical image analysis supporting a wide range of diagnostic and interventional workflows. Although recent methods have achieved sub-millimetric localisation, accuracy alone is not sufficient for clinical deployment, requiring reliability and robustness in prediction. Despite its clinical relevance, the impact of representation learning in this context is still underexplored. In this work, we introduce CDPM-align, a multi-scale guidance-aligned conditional diffusion pre-training for anatomical landmark detection. Our experimental setup focuses on a few images and a few annotation regimes. Specifically, we employ three popular heterogeneous small-scale benchmark datasets for representation learning via conditional generative pre-training. Furthermore, we consider low-annotation scenarios for the downstream task of landmark detection, with 10 and 25 annotated images, reflecting realistic trade-offs between clinical effort and resource constraints for annotations. Our results confirm that generative pre-training enables the model to learn a robust representation. This improves both accuracy and uncertainty on the downstream tasks, advancing towards safe and efficient clinical deployment.

Keywords: Anatomical landmark detection · Generative pretraining · Diffusion models · Few-shot learning · Uncertainty quantification.

1 Introduction

Anatomical landmark detection in medical X-ray images underpins cephalometric treatment planning [27], cardiothoracic ratio estimation [24], and skeletal maturity assessment [18]. Although modern deep learning methods can achieve sub-millimetre localisation accuracy [21,25,16,31,29], point-wise error alone does not guarantee clinical reliability [6,7,9]. A clinically trustworthy model must not only predict an accurate coordinate, but also concentrate probability mass tightly around the true landmark, reflecting calibrated spatial uncertainty. To quantify this property, the Expected Radial Error (ERE) was introduced for anatomical landmark detection [19]. ERE measures the expected Euclidean distance between the predicted landmark and a sample from the predicted heatmap, jointly capturing localisation accuracy and distributional sharpness, providing a

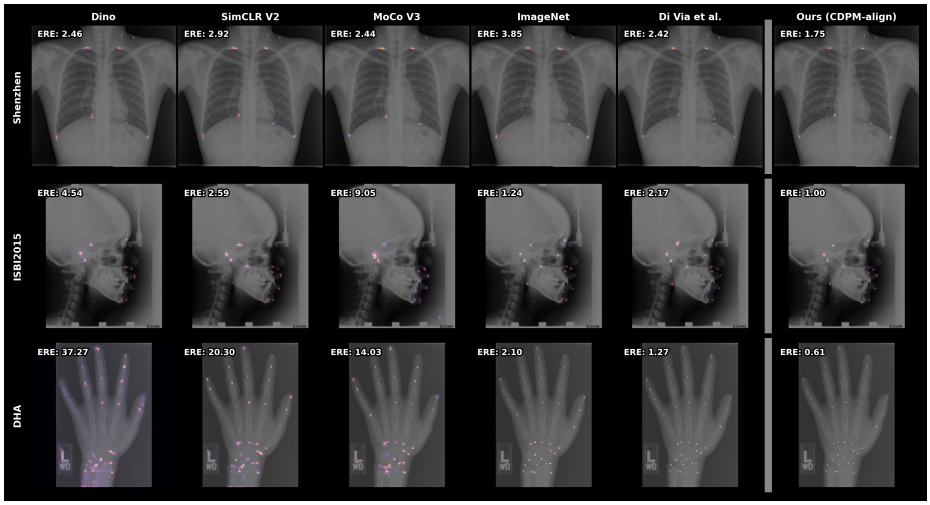

Fig. 1. Predicted heatmap overlays at 10-shot across datasets (rows) and selected methods (columns). Our CDPM-align (rightmost) yields consistently lower uncertainty.

more faithful assessment of reliability. For instance, a model with accurate mean predictions but diffuse heatmaps would incur a high ERE, indicating decreased reliability. Despite its clinical relevance, how pretraining influences ERE and robustness, particularly under limited annotation budgets, remains underexplored. Traditionally, medical landmark detection has relied on supervised pretraining, either on large natural image datasets, such as ImageNet, or on domain-specific X-rays [20,10]. However, these approaches often fail to capture nuances across different anatomical regions. Self-supervised learning (SSL) methods, including DINO [2], SimCLR v2 [3], MoCo v3 [4], and pixel-level embeddings such as SAM [30], learn transferable dense features without extensive labels [26], yet their performance plateaus under extreme data scarcity and structural heterogeneity. Universal detectors [32] and foundation models like MedSapiens [11] aggregate multiple datasets for broad generalisation but primarily optimise point-wise accuracy, neglecting local distributional sharpness and uncertainty. Generative pretraining with diffusion models offers a complementary approach to learn spatially organised semantics suitable for dense prediction [1]. Di Via et al. [8] showed unconditional diffusion can improve few-shot landmark detection, but without conditioning, anatomical modes are conflated and anatomical region-specific structural representations remain underconstrained. Across these methods, the effect of pretraining on predictive uncertainty and robustness has been unexplored, motivating approaches that enforce both global context and local structural consistency for reliable heatmap estimation. Motivated by these limitations, we propose **CDPM-align**, a conditional diffusion pretraining framework tailored to few-shot anatomical landmark detection. Starting from small-scale heterogeneous collection of data, we condition diffusion training exploiting the specific

dataset indexes as labels, thus enabling the model to learn dataset-specific representations. We exploit the classifier-free guidance signal as an explicit structural descriptor and introduce a multi-scale alignment objective that enforces directional features consistency across independently sampled diffusion timesteps and UNet hierarchy levels. This constraint encourages class-conditional structure that is stable across noise levels while preserving spatial fidelity. The pretrained backbone is subsequently fine-tuned end-to-end for the anatomical landmark detection task, using as few as 10–25 annotated images per dataset. Our main contributions can be summarised as follows:

- We introduce a conditional generative pretraining strategy for small-scale heterogeneous X-ray datasets, augmented with a multi-scale guidance alignment loss that improves localisation accuracy, uncertainty concentration, and robustness in low-shot regimes.
- We validate the proposed method across three clinical dimensions (accuracy, predictive uncertainty, and robustness) on three public benchmarks under 10- and 25-shot protocols, showing consistent improvements over supervised, SSL, and state-of-the-art baselines.

2 Method

CDPM-align consists of a pretraining stage followed by downstream adaptation (Sec. 2.3). The pretraining stage itself has two phases (Sec. 2.1–2.2): the first trains a conditional diffusion probabilistic model (CDPM) with the standard generative objective; the second fine-tunes the same model for a short additional phase with the guidance alignment loss activated. The pretrained backbone is then adapted for few-shot anatomical landmark detection. A schematic overview of CDPM-align is provided in Fig. 2.

2.1 Conditional Diffusion Probabilistic Model (CDPM)

We build on denoising diffusion probabilistic models (DDPMs) [14] with class-conditional generation via classifier-free guidance [15]. The forward process produces a noisy sample $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, linear schedule β_1=10^{-4} to β_T=0.028, T=500. A noise-prediction network $\epsilon_\theta(x_t, t, c)$ is trained with class label $c \in \{0, 1, 2, 3\}$ (0 = unconditional; 1–3 = dataset index), with 10% dropout to c=0 enabling CFG at inference, minimising $\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t, t, c)\|^2]$. The backbone is a UNet [23] with FiLM [22] embeddings for timestep and class conditioning.

2.2 CDPM-Align: Multi-scale Alignment on the Guidance Signal

For each image x_0 belonging to dataset y, we sample two timesteps $t_1, t_2 \sim p(t)$, produce the noisy images x_{t_1} and x_{t_2}, and perform four UNet forward passes—$f_\theta(x_{t_i}, t_i, c)$ for $i \in \{1, 2\}$ and $c \in \{y, \varnothing\}$—obtaining conditional h_{cond}

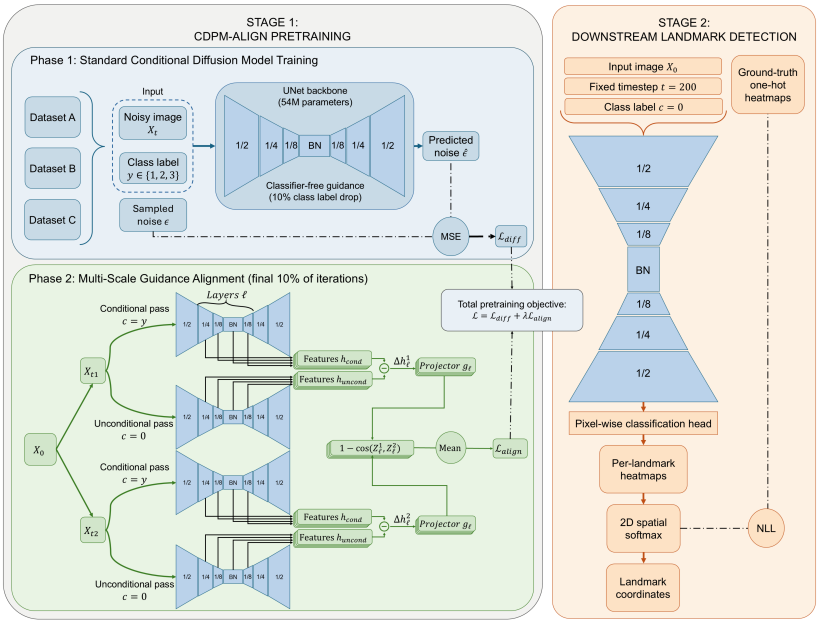

Fig. 2. Overview of CDPM-align. *Left*: Two-phase multi-dataset pretraining. Standard conditional diffusion training is followed by an alignment phase (final 10% of iterations) enforcing directional consistency of the guidance signal across timesteps at four UNet scales. *Right*: End-to-end few-shot fine-tuning via pixel-wise classification.

and unconditional h_{uncond} features at each timestep. The guidance difference $\Delta h = h_{\text{cond}} - h_{\text{uncond}}$ encodes dataset-specific anatomy; its magnitude varies with timestep while our intuition is that enforcing directional consistency across independently sampled timesteps encourages the UNet encoder to learn representations invariant to noise level. We bias sampling toward mid-range timesteps $[T/4, 3T/4]$, since very early timesteps yield near-clean images where the guidance difference Δh carries little class information, while very late timesteps are dominated by noise; the mid-range provides the richest dataset-discriminative signal. Features are extracted at four hierarchy levels:

$$\mathcal{S} = \{\text{encoder}_{1/4}, \text{encoder}_{1/8}, \text{bottleneck}, \text{decoder}_{1/8}\}.$$

The per-layer guidance difference at level $\ell \in \mathcal{S}$ is $\Delta h_\ell^{(i)} = f_\theta(x_{t_i}, t_i, y)[\ell] - f_\theta(x_{t_i}, t_i, \varnothing)[\ell]$ for $i \in \{1, 2\}$. At each level independently, $\Delta h_\ell^{(i)}$ is processed by a level-specific projection head g_ℓ—comprising global average pooling (GAP), a two-layer MLP_ℓ ($C_\ell \to 512 \to 256$, GELU), and ℓ_2-normalisation—to yield $z_\ell^{(i)} = g_\ell(\Delta h_\ell^{(i)})$. The alignment loss averages the per-level cosine dissimilarity between the two views:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{S}|} \sum_{\ell \in \mathcal{S}} \big(1 - \cos(z_\ell^{(1)}, z_\ell^{(2)})\big).$$

During the alignment phase, $\mathcal{L}_{\text{diff}}$ is computed over all four noise predictions (for timesteps t_1, t_2 and for condition $c = 0, c = y$), since the unconditional passes required for Δh are already computed and supervising them preserves both generation pathways needed by CFG. The total objective is:

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}.$$

The projection heads g_ℓ are discarded after pretraining; full UNet weights are transferred downstream. Intuitively, our feature alignment is mostly effective when the diffusion model has learned to estimate the noise added to input images. Furthermore, it brings an additional computational overhead, as it requires to perform four forward passes. As such, we employ our alignment loss only when we reach the final 10% of training iterations, when we fine-tune the CDPM model pre-trained in the first phase with $\mathcal{L} = \mathcal{L}_{\text{diff}}$. In this way, we manage to reduce the computational overhead from 4× to 1.3× with respect to naive training.

2.3 Downstream Landmark Detection

Following [1], we use a fixed forward timestep (t=200) before feature extraction, yielding semantically organised intermediate representations. For anatomical landmark detection, the full pretrained UNet is fine-tuned end-to-end with a pixel-wise classification head, as in [19]. During fine-tuning and inference, the unconditional class label (c=0) is used, decoupling the downstream task from pretraining labels, theoretically enabling future application to unseen datasets, while preventing over-reliance on the class token in few-shot settings.
Metrics. We report MRE (mean radial error), ERE (expected radial error, quantifying spatial uncertainty as the expected distance between a heatmap-sampled point and the predicted landmark), SDR@2 (successful detection rate within 2 mm or 2 px), and P95 (95th-percentile error, reflecting worst-case tail behaviour). Metrics are reported in millimetres, where possible (ISBI2015 and DHA, where the DICOM pixel spacing is available for pixel-millimetres conversion), and in pixels where such information is not available (Shenzhen).

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate on three public benchmarks: *Shenzhen* [17] (279 anteroposterior chest radiographs, 6 landmarks), *ISBI2015* [27] (400 lateral cephalograms, 19 landmarks), and the *Digital Hand Atlas* (DHA) [12] (910 hand radiographs, 37 landmarks). All images are resized to 256×256 pixels preserving the original aspect ratio. The pooled unlabeled pretraining corpus, excluding the test set, comprises 988 images across all three datasets. *Baselines and few-shot protocol, and implementation.* We evaluate under 10- and 25-shot annotation budgets, reporting mean ± std over 5 independent runs. External baselines include: (i) supervised UNet with ResNet-101 encoder (ImageNet pretrained); (ii)

Table 1. Landmark detection on **Shenzhen**. Best results per column are bolded.

Method	10-shot				25-shot			
	MRE↓	ERE↓	SDR@2↑	P95↓	MRE↓	ERE↓	SDR@2↑	P95↓
Zhu et al. [32]	67.82 ± 18.55	132.46 ± 30.43	9.43 ± 3.26	370.73 ± 57.18	37.45 ± 11.27	39.61 ± 10.27	19.59 ± 4.98	266.69 ± 92.00
Di Via et al. [8]	3.44 ± 0.57	3.01 ± 0.67	52.00 ± 4.12	7.92 ± 0.80	3.10 ± 0.59	2.71 ± 0.41	52.93 ± 3.35	8.12 ± 0.79
ResNet-101 (ImageNet)	5.18 ± 1.24	4.07 ± 0.43	43.87 ± 3.76	10.23 ± 1.56	3.09 ± 0.27	3.16 ± 0.34	45.67 ± 2.21	6.91 ± 0.17
ResNet-101 (DINO) [2]	4.92 ± 0.51	5.13 ± 0.69	43.87 ± 2.44	15.15 ± 4.16	4.21 ± 0.78	4.23 ± 1.22	44.07 ± 2.96	12.25 ± 4.37
ResNet-101 (MoCo v3) [4]	5.19 ± 1.75	5.17 ± 1.65	42.73 ± 2.06	20.98 ± 18.40	4.22 ± 0.39	4.11 ± 0.63	44.00 ± 2.24	9.85 ± 1.59
ResNet-101 (SimCLR v2) [3]	5.50 ± 0.34	5.59 ± 0.67	38.20 ± 2.78	20.63 ± 0.86	3.58 ± 0.36	3.74 ± 0.40	44.00 ± 2.48	8.86 ± 0.89
CDPM Scratch	6.39 ± 1.35	6.40 ± 1.81	50.40 ± 2.62	26.25 ± 10.27	5.02 ± 2.98	4.75 ± 2.22	53.20 ± 2.81	18.06 ± 20.69
CDPM (NIH, λ=0) Sec. 3.3	$\mathbf{2.89 \pm 0.36}$	$\mathbf{2.49 \pm 0.49}$	$\mathbf{53.40 \pm 3.18}$	$\mathbf{6.41 \pm 0.08}$	2.65 ± 0.21	$\mathbf{2.50 \pm 0.34}$	53.53 ± 3.26	6.77 ± 0.34
CDPM-align (Ours, λ=5)	3.01 ± 0.35	2.96 ± 0.61	53.07 ± 2.23	7.17 ± 0.54	$\mathbf{2.58 \pm 0.28}$	2.58 ± 0.36	$\mathbf{56.00 \pm 2.49}$	$\mathbf{6.51 \pm 0.46}$

three self-supervised UNet baselines [13] (DINO, MoCo v3, SimCLR v2) pretrained on the same pooled radiograph corpus with a ResNet-101 backbone; (iii) unconditional diffusion pretraining [8]; and (iv) a universal multi-anatomy detector [32]. ResNet-101 is used to match CDPM's ∼52M parameters and skip-connection capacity, as diffusion backbones cannot directly reuse discriminatively pretrained weights. Internal ablations include CDPM Scratch (random initialisation). CDPM-align is pretrained for 50k iterations (45k standard diffusion, 5k alignment fine-tuning) with a batch size of 16 and the AdamW optimiser. Downstream fine-tuning runs for 200 epochs with early stopping, the NLL loss, and AdamW with an lr of 10^{-4} and a batch size of 8. Code will be released upon acceptance to support reproducibility.

3.2 Main Results

Tables 1–3 compare all methods across datasets and annotation budgets. Both Friedman and Kruskal–Wallis statistical tests confirms global performance differences ($p \leq 0.001$), and CDPM-align achieves consistent superiority (5/5 runs) over all supervised, self-supervised, and state-of-the-art baselines [32,8].

Accuracy. CDPM-align achieves the best MRE in 5 of 6 dataset–budget settings; the only exception is Shenzhen at 10-shot, where CDPM (NIH) benefits from larger-scale pretraining. Gains are substantial: +22% on ISBI2015 10-shot (2.11 vs. 2.70 mm) and +42% on DHA 10-shot (2.51 vs. 4.34 mm). At 25-shot, CDPM-align reaches 1.54 mm on ISBI2015 (77.52% SDR), approaching the 2 mm clinical threshold with only 25 images. On Shenzhen, CDPM-align achieves 3.01 px (10-shot), outperforming the full-label method of [5] (3.82 px). Although the foundation model MedSapiens [11] reports 1.24 mm (ISBI2015) and 3.72 px (Shenzhen) with full supervision, our method is competitive using only 10–25 annotations.

Reliability. At 25-shot, CDPM-align achieves ERE below 1 mm on both ISBI2015 (0.95 mm, a 25% improvement over [8] reaching 1.26 mm) and DHA (0.97 mm, a 34% improvement over [8] reaching 1.48 mm). Consistently, ERE $\leq$ MRE across conditions, indicating tight probability mass concentration. Figure 1 provides visual corroboration: CDPM-align systematically produces tight, unimodal heatmaps in contrast to the diffuse distributions of other methods. P95 improvements track this pattern across all benchmarks.

Table 2. Landmark detection on **ISBI2015**. Best results per column are bolded.

Method	10-shot				25-shot			
	MRE↓	ERE↓	SDR@2↑	P95↓	MRE↓	ERE↓	SDR@2↑	P95↓
Zhu et al. [32]	17.20 ± 4.29	48.33 ± 9.31	38.25 ± 3.29	91.94 ± 11.10	2.63 ± 0.20	3.23 ± 0.26	66.96 ± 2.14	5.29 ± 0.28
Di Via et al. [8]	2.70 ± 0.34	1.97 ± 0.19	66.57 ± 1.39	5.41 ± 0.47	1.80 ± 0.10	1.26 ± 0.10	74.06 ± 1.74	4.34 ± 0.20
ResNet-101 (ImageNet)	2.83 ± 0.30	2.38 ± 0.37	64.20 ± 1.19	5.42 ± 0.68	2.86 ± 0.43	2.37 ± 0.38	68.83 ± 1.28	4.87 ± 0.36
ResNet-101 (DINO) [2]	4.71 ± 0.13	3.96 ± 0.23	54.29 ± 1.29	18.88 ± 1.90	2.36 ± 0.24	1.94 ± 0.25	66.98 ± 1.28	5.09 ± 0.40
ResNet-101 (MoCo v3) [4]	5.23 ± 0.55	4.39 ± 0.60	52.56 ± 1.84	24.62 ± 6.16	2.44 ± 0.20	2.12 ± 0.23	66.15 ± 1.76	5.18 ± 0.51
ResNet-101 (SimCLR v2) [3]	4.74 ± 0.50	3.85 ± 0.56	55.04 ± 1.74	19.73 ± 4.90	2.41 ± 0.17	1.89 ± 0.14	66.53 ± 1.39	5.30 ± 0.37
CDPM Scratch	4.87 ± 0.89	3.80 ± 0.89	64.23 ± 2.32	16.64 ± 6.00	1.89 ± 0.14	1.46 ± 0.15	74.91 ± 1.35	4.51 ± 0.20
CDPM (NIH, λ=0) Sec. 3.3	2.32 ± 0.15	1.78 ± 0.27	**68.12 ± 1.40**	5.20 ± 0.20	1.70 ± 0.06	1.19 ± 0.04	75.49 ± 1.43	4.04 ± 0.15
CDPM-align (Ours, λ=5)	**2.11 ± 0.29**	**1.38 ± 0.14**	67.48 ± 3.20	**4.75 ± 0.37**	**1.54 ± 0.02**	**0.95 ± 0.07**	**77.52 ± 0.74**	**3.90 ± 0.07**

Table 3. Landmark detection on **DHA**. Best results per column are bolded.

Method	10-shot				25-shot			
	MRE↓	ERE↓	SDR@2↑	P95↓	MRE↓	ERE↓	SDR@2↑	P95↓
Zhu et al. [32]	33.85 ± 13.68	90.66 ± 9.75	44.48 ± 6.51	161.83 ± 33.82	2.87 ± 0.43	4.95 ± 0.89	83.13 ± 0.98	4.04 ± 0.40
Di Via et al. [8]	4.34 ± 1.04	3.01 ± 0.81	79.94 ± 3.25	19.11 ± 11.34	2.22 ± 0.15	1.48 ± 0.09	88.28 ± 0.41	3.67 ± 0.10
ResNet-101 (ImageNet)	6.07 ± 0.64	5.76 ± 0.30	78.44 ± 0.36	26.39 ± 3.29	3.95 ± 1.12	4.39 ± 1.26	84.07 ± 1.48	7.65 ± 4.80
ResNet-101 (DINO) [2]	23.63 ± 6.96	22.38 ± 5.80	34.95 ± 8.35	99.49 ± 25.50	10.16 ± 6.97	10.24 ± 6.34	63.83 ± 15.43	46.98 ± 23.90
ResNet-101 (MoCo v3) [4]	24.60 ± 8.45	23.48 ± 7.37	34.86 ± 11.53	102.67 ± 24.80	9.28 ± 4.21	9.64 ± 4.32	64.50 ± 10.54	45.29 ± 17.23
ResNet-101 (SimCLR v2) [3]	20.24 ± 3.22	19.91 ± 3.19	39.36 ± 4.77	86.83 ± 8.75	9.47 ± 4.37	9.82 ± 4.56	64.97 ± 10.17	46.99 ± 18.75
CDPM Scratch	8.60 ± 2.57	5.83 ± 2.06	71.19 ± 4.57	44.34 ± 13.61	4.22 ± 0.85	2.96 ± 0.69	84.50 ± 1.56	14.43 ± 9.88
CDPM (NIH, λ=0) Sec. 3.3	4.08 ± 0.75	3.24 ± 0.57	80.50 ± 2.20	16.41 ± 4.22	1.62 ± 0.11	1.19 ± 0.14	89.80 ± 0.42	3.18 ± 0.16
CDPM-align (Ours, λ=5)	**2.51 ± 0.30**	**2.16 ± 0.31**	**86.47 ± 1.08**	**4.46 ± 0.90**	**1.53 ± 0.10**	**0.97 ± 0.08**	**91.13 ± 0.69**	2.77 ± 0.20

Robustness. SSL methods using discriminative objectives exhibit severe performance degradation on DHA at 10-shot: DINO, MoCo v3, and SimCLR v2 reach MRE of 20–25 mm with P95 exceeding 86 mm, compared to 4.46 mm for CDPM-align. This failure is consistent across all 5 independent runs and persists at 25-shot (9–10 mm MRE). When pooled multi-anatomical regions X-rays span heterogeneous structural modes, discriminative SSL objectives collapse onto global image identity, suppressing fine-grained intra-region geometry. Both CDPM-align and Di Via et al. maintain stable performance, with CDPM-align providing additional gains from the alignment objective.

3.3 Ablation Studies

Alignment weight λ. We analyse the effect of the alignment weight λ on Shenzhen (Table 4), the smallest and structurally simplest benchmark, to isolate dataset-independent trends. Without alignment (λ=0), MRE is 3.95 px at 10-shot. Increasing λ improves MRE and SDR up to λ=5 (best MRE: 3.01 px; SDR: 53.07%; P95 at 25-shot: 6.51 px). Beyond this, performance drops (λ=10: 3.42 px MRE), indicating that over-constraining the guidance reduces pixel-level fidelity. This trade-off reflects the differing roles of features along the diffusion trajectory: larger λ forces low-noise features to rigidly follow coarse global semantics, over-concentrating probability mass and minimising ERE, but limiting the flexibility required for precise sub-millimetre localisation. ERE is lowest at λ=10, highlighting that minimal uncertainty does not necessarily translate into optimal spatial accuracy. Based on these observations, we adopt λ=5 as the operating point that best balances accuracy, uncertainty, and robustness.

Table 4. Ablation of alignment weight λ on **Shenzhen**.

λ	10-shot				25-shot			
	MRE↓	**ERE↓**	**SDR@2↑**	**P95↓**	**MRE↓**	**ERE↓**	**SDR@2↑**	**P95↓**
0	3.95 ± 0.84	3.35 ± 1.12	51.60 ± 3.52	9.30 ± 1.56	2.75 ± 0.30	2.69 ± 0.32	53.20 ± 3.68	7.20 ± 0.44
1	3.82 ± 1.07	3.28 ± 1.06	51.53 ± 2.69	9.44 ± 3.98	2.61 ± 0.17	2.59 ± 0.28	53.47 ± 3.20	7.00 ± 0.69
3	3.25 ± 0.86	2.88 ± 0.56	52.20 ± 0.80	7.03 ± 0.39	2.89 ± 0.54	2.57 ± 0.39	53.47 ± 1.56	7.26 ± 0.46
5	3.01 ± 0.35	2.96 ± 0.61	53.07 ± 2.23	7.17 ± 0.54	2.58 ± 0.28	2.58 ± 0.36	56.00 ± 2.49	6.51 ± 0.46
10	3.42 ± 0.46	2.78 ± 0.22	52.67 ± 1.75	7.23 ± 0.44	2.74 ± 0.32	2.37 ± 0.10	53.93 ± 0.80	7.23 ± 0.56

Pretraining data scale. To further assess the benefit of the proposed feature alignment, we design an ablation study comparing CDPM-align against regular CDPM pre-trained with a larger dataset. As such, we train the same CDPM architecture, with λ=0, on NIH ChestX-ray14 [28] ($\sim$112k chest X-rays), referred to as CDPM (NIH). Results in Tables 1–3 show that CDPM-align (988 images) is competitive with CDPM (NIH) on Shenzhen 25-shot (2.58 vs. 2.65 px MRE), where CPDM (NIH) benefits from in-domain images, while providing slight improvements on the out-of-domain DHA 10-shot (2.51 vs. 4.08 mm) and ISBI2015 10-shot (2.11 vs. 2.32 mm), where NIH domain shift penalises the larger corpus. The alignment objective compensates for reduced data scale by producing more discriminative representations from limited heterogeneous data.

4 Conclusion

We presented CDPM-align, a conditional diffusion pretraining framework with multi-scale guidance alignment for anatomical landmark detection under realistic low-annotation clinical budgets. By enforcing directional consistency of the class-conditional guidance signal Δh across diffusion timesteps and UNet levels, the model learns stable, dataset-specific representations from pooled heterogeneous X-rays. Across Shenzhen, ISBI2015, and DHA in 10- and 25-shot regimes, CDPM-align consistently improves accuracy (MRE), detection rate (SDR), tail reliability (P95), and uncertainty concentration (ERE) over supervised, self-supervised, unconditional diffusion, and large-scale domain-specific pretraining. In contrast to conventional SSL, which degrades on heterogeneous datasets, CDPM-align remains robust across domains. Notably, pretrained on only 988 images, it matches or surpasses models trained on 112k NIH radiographs, with the largest gains on DHA where domain shift penalises out-of-domain corpora. Sub-millimetre ERE at 25-shot on ISBI2015 and DHA further demonstrates reliable localisation with minimal supervision. Overall, targeted in-domain conditional diffusion pretraining emerges as a practical and deployment-oriented alternative to large out-of-domain training for clinical imaging. Current limitations include evaluation restricted to 2D X-rays and a moderated increase in computational cost during alignment (1.3$\times$ with respect to naive CDPM training), pointing to future work on scaling to volumetric modalities, while further improving efficiency.

References

1. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: Proc. ICLR (2022)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294 (2021)
3. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. In: Advances in Neural Information Processing Systems. vol. 33, pp. 21296–21309 (2020)
4. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proc. ICCV. pp. 9640–9649 (2021)
5. Choi, S.B., Ham, G.S., Oh, K.: Learning structural relations for robust chest x-ray landmark detection. Electronics **15**(3), 589 (2025). https://doi.org/10.3390/electronics15030589
6. Clement, A., Willoughby, J., Voiculescu, I.: Confidence in Angle Predictions for Clinical Decision Support. In: Proc. MICCAI. pp. 116–124 (2025)
7. Di Via, R., Ciranni, M., Marinelli, D., Clement, A., Patel, N., Wyatt, J., Odone, F., Santacesaria, M., Voiculescu, I., Pastore, V.P.: Are x-ray landmark detection models fair? a preliminary assessment and mitigation strategy. In: Proc. ICCVW. pp. 272–278 (2025)
8. Di Via, R., Odone, F., Pastore, V.P.: Self-supervised pre-training with diffusion model for few-shot landmark detection in x-ray images. In: Proc. WACV. pp. 3886–3894 (2025)
9. Di Via, R., Pastore, V.P., Odone, F., Glyn-Jones, S., Voiculescu, I.: Automated landmark detection for assessing hip conditions: A cross-modality validation of MRI versus x-ray. arXiv preprint arXiv:2601.18555 (2026)
10. Di Via, R., Santacesaria, M., Odone, F., Pastore, V.P.: Is in-domain data beneficial in transfer learning for landmarks detection in x-ray images? In: Proc. ISBI. pp. 1–5 (2024). https://doi.org/10.1109/ISBI56570.2024.10635861
11. Elbatel, M., Wang, A., Liu, K., Mouheb, K., Almar-Munoz, E., et al.: MedSapiens: Taking a pose to rethink medical imaging landmark detection. arXiv preprint arXiv:2511.04255 (2025). https://doi.org/10.48550/arXiv.2511.04255
12. Gertych, A., Zhang, A., Sayre, J., Pospiech-Kurkowska, S., Huang, H.K.: Bone age assessment of children using a digital hand atlas. Comput. Med. Imaging Graph. **31**(4–5), 322–331 (2007). https://doi.org/10.1016/j.compmedimag.2007.02.012
13. Giakoumoglou, N., Giakoumoglou, P.: PySSL: A PyTorch implementation of self-supervised learning methods. https://github.com/giakou4/pyssl (2023)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
16. Huang, Z., Tang, T., Xu, R., Wei, Y., Yang, W., et al.: H3DE-Net: Efficient and accurate 3D landmark detection in medical imaging. arXiv preprint arXiv:2502.14221 (2025)
17. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quant. Imaging Med. Surg. **4**(6), 475–477 (2014). https://doi.org/10.3978/j.issn.2223-4292.2014.11.20

18. Li, Z., Chen, W., Ju, Y., Chen, Y., Hou, Z., Li, X., Jiang, Y.: Bone age assessment based on deep neural networks with annotation-free cascaded critical bone region extraction. Frontiers in Artificial Intelligence **6**, 1142895 (2023). https://doi.org/10.3389/frai.2023.1142895
19. McCouat, J., Voiculescu, I.: Contour-hugging heatmaps for landmark detection. In: Proc. CVPR. pp. 20597–20605 (2022)
20. Patel, N., Clement, A., Wyatt, J., Di Via, R., Marinelli, D., Ciranni, M., Pastore, V.P., Voiculescu, I.: A handful of data: Evaluating few-shot incremental landmark detection. In: Proc. ICIAP. pp. 127–139 (2025). https://doi.org/10.1007/978-3-032-10192-1_11
21. Payer, C., Štern, D., Bischof, H., Urschler, M.: Integrating spatial configuration into heatmap regression based cnns for landmark localization. Medical Image Analysis **54**, 207–219 (2019). https://doi.org/10.1016/j.media.2019.03.012
22. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: Proc. AAAI. vol. 32, pp. 8101–8108 (2018). https://doi.org/10.1609/aaai.v32i1.11671
23. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proc. MICCAI. pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28
24. Sanjas, A.M., Ida, A.M., Santhiya, S.G., Mercy, P.A.S., Nithya, S.A.: Deep learning–based automatic estimation of cardiometric coefficients from chest x-ray images. Journal of Advance and Future Research **3**(12), 692–702 (2025), https://rjwave.org/jaafr/papers/JAAFR2512414.pdf
25. Stansfield, E., Mitterer, J.A., Altahhan, A.: Landmark detection for medical images using a general-purpose segmentation model. arXiv preprint arXiv:2507.11551 (2025)
26. Truong, T., Mohammadi, S., Lenga, M.: How transferable are self-supervised features in medical image classification tasks? In: Proc. ML4H. vol. 158, pp. 54–74 (2021)
27. Wang, C.W., Huang, C.T., Lee, J.H., Li, C.H., Chang, S.W., et al.: A benchmark for comparison of dental radiography analysis algorithms. Medical Image Analysis **31**, 63–76 (2016). https://doi.org/10.1016/j.media.2016.02.004
28. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-Ray14: Hospital-scale chest x-ray database and benchmarks. In: Proc. CVPR. pp. 2097–2106 (2017)
29. Wyatt, J., Voiculescu, I.: Optimising for the unknown: Domain alignment for cephalometric landmark detection. arXiv preprint arXiv:2410.04445 (2024)
30. Yan, K., Cai, J., Jin, D., Miao, S., Guo, D., Harrison, A.P., Tang, Y., Xiao, J., Lu, J., Lu, L.: SAM: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. IEEE Trans. Med. Imaging **41**(10), 2658–2669 (2022). https://doi.org/10.1109/TMI.2022.3169003
31. Zhou, X., Huang, Z., Zhu, H., Yao, Q., Zhou, S.K.: Hybrid attention network: An efficient approach for anatomy-free landmark detection. arXiv preprint arXiv:2412.06499 (2024). https://doi.org/10.48550/arXiv.2412.06499
32. Zhu, H., Yao, Q., Mao, E., Zhou, S.K.: You only learn once: Universal anatomical landmark detection. In: Proc. MICCAI. pp. 84–94 (2021). https://doi.org/10.1007/978-3-030-87240-3_9