CAUSALLY GUIDED CHROMOSOMAL ABNORMALITY DETECTION FOR UNSEEN CONFIGURATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Chromosomal structural abnormality detection informs genetic diagnosis and counseling. However, scarce abnormal samples cannot fully cover naturally occurring configurations across rearrangement locations and intervals, while patient backgrounds and preparation and imaging conditions alter their appearance. Detection thus requires generalization to unseen configurations and robustness to environmental shifts. We observe progressive degradation under these shifts and near-chance performance on unseen configurations. Our causally guided two-stage framework combines homologous comparison with normal-statistics calibration, using abundant normal pairs as a baseline to recast normal-versus-abnormal classification as distinguishing normal from abnormal variation. This reduces reliance on chromosome-specific appearance and simplifies learning from scarce abnormal samples. Stage I combines skeleton-guided counterfactual generation, configuration generalization learning, and synthetic bias suppression to expand configuration coverage and learn transferable structural evidence. Stage II freezes the encoder and calibration statistics, refining the classification head through quality and reference constraints for environmental robustness. We introduce ChromoShift, a full-pipeline dataset comprising skeleton–centromere annotations, normal pretraining data, and real abnormality data under distribution shifts. Experiments show gains in both test domains, achieving 90.83% AUROC on unseen configurations. The framework supports clinical assessment of known abnormalities and screening for potentially novel configurations to guide further genetic research. Code: `https://anonymous.4open.science/r/CausalChrom-3D66/`.

1 INTRODUCTION

Chromosomal abnormality detection is an important component of the etiological diagnosis of genetic disorders and genetic counseling, with clinical value in the genetic evaluation of congenital anomalies, developmental disorders, and certain rare diseases (Silva et al., 2019). Chromosomal abnormalities include numerical and structural abnormalities. Structural rearrangements can cause disease by disrupting gene integrity or altering regulatory relationships, even without changing total genetic material (Redin et al., 2017). Studies of rare-disease families further highlight the diagnostic value of inversions and other structural variants (Pagnamenta et al., 2024). Karyotype analysis identifies abnormalities through chromosome number, morphology, and banding patterns (Silva et al., 2019). Compared with numerical abnormalities, structural abnormalities demand finer visual discrimination of internal band arrangements and structural relationships (Bechar et al., 2023; Li et al., 2024).

Inversions are challenging structural abnormalities in which the segment between two breakpoints is reinserted in reverse orientation, reversing the local band order while potentially preserving an overall length and appearance similar to those of the normal homolog. Short inversion intervals, similar bands, and limited imaging detail further complicate detection. Different breakpoints and intervals produce diverse configurations that scarce available abnormal samples cannot adequately cover (Uzolas et al., 2022; Zhang et al., 2026). Reliable detection must therefore generalize to unseen configurations while remaining stable across patient backgrounds and preparation and imaging conditions, supporting both known-abnormality recognition and screening for candidates with previously uncharacterized configurations (Pagnamenta et al., 2024; Bechar et al., 2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

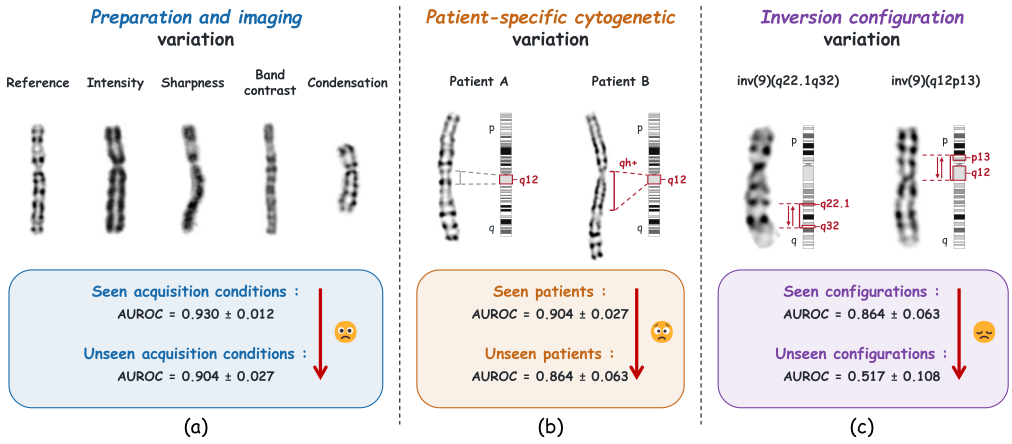


Figure 1: Generalization under shifts in (a) preparation and imaging conditions, (b) patient-specific cytogenetic background, and (c) inversion configuration. Baseline performance declines as generalization demands increase, approaching chance on unseen configurations. Results are from a motivating experiment using ResNet50.

Deep learning has advanced chromosomal structural abnormality detection, but generalization remains limited (Bechar et al., 2023; Li et al., 2024; Yang et al., 2026). As Figure 1 shows, preparation and imaging alter appearance and band visibility, patient background introduces morphological and heterochromatic variation (Tofanelli et al., 1993; Verma & Luke, 1991), and inversion configuration determines abnormality location and structural manifestation. In five-fold cross-validation with progressively stricter generalization requirements, baseline AUROC decreases from 0.930 ± 0.012 to 0.904 ± 0.027 under unseen acquisition conditions, to 0.864 ± 0.063 with patient-disjoint splits, and to 0.517 ± 0.108, near chance, on unseen configurations under patient isolation. Performance degradation under environmental shifts and failure on unseen configurations motivate learning structural evidence from limited abnormal data for cross-configuration transfer and environmental robustness.

We propose a causally guided two-stage framework grounded in structural formation and image acquisition. Patient background and inversion configuration shape the underlying structure, while acquisition conditions affect its observed appearance, motivating cross-configuration structural learning and cross-environment robustness. Homologous differences contain both normal variation and abnormal rearrangements. Normal-statistics calibration establishes a baseline from abundant normal pairs to distinguish these changes, reducing reliance on chromosome-specific appearance. Stage I learns transferable structural evidence through skeleton-guided counterfactual generation, configuration generalization learning, and synthetic bias suppression. Stage II freezes the encoder and calibration statistics and refines the classification head through quality and reference constraints, stabilizing the use of learned structural evidence under environmental shifts.

We further introduce ChromoShift, a full-pipeline dataset for generalization research comprising skeleton–centromere annotations, normal-case pretraining data, and real abnormality data under distribution shifts. The real abnormality component covers 84 patients and 25 inversion configurations, supporting research on patient and configuration shifts.

Our main contributions are as follows:

1. We reveal performance degradation under environmental shifts and recognition failure on unseen configurations through progressively stricter five-fold evaluation. Causal analysis of patient background, acquisition conditions, and inversion configuration guides our design.
2. We propose a causally guided two-stage framework built on homologous comparison and normal-statistics calibration. Counterfactual structural learning establishes transferable evidence; quality and reference constraints then stabilize cross-environment predictions.
3. We introduce ChromoShift, a full-pipeline dataset integrating detailed structural annotations, normal pretraining data, and real abnormalities across patients and configurations to support representation learning, abnormality generation, and generalization research.

2 Related Work

2.1 Chromosomal Abnormality Detection

Abnormality-specific modeling, homologous comparison, and synthetic augmentation improve chromosomal abnormality detection. Bechar et al. (2023) compare homologous chromosomes using a Siamese network; Li et al. (2024)'s HomNet adds homologous region alignment, cross-cell aggregation, and artificial abnormality pretraining. Chen et al. (2024)'s BR-ChromNet combines image and banding features with a conditional random field to localize abnormalities, while CHROMA improves transferability through large-scale self-supervised pretraining (Yang et al., 2026). Uzolas et al. (2022)'s Deep Anomaly Generation synthesizes chromosomes conditioned on shape and banding patterns, exploring unseen banding patterns to address abnormal sample scarcity.

These studies advance detection through representation learning and artificial abnormality augmentation, but data expansion alone does not ensure recognition of real unseen configurations. Limited training data cannot fully cover inversion configurations arising from different breakpoint and interval combinations. We focus on unseen-configuration generalization and robustness to variation in patient background and preparation and imaging conditions. We reveal recognition failure on unseen configurations and performance degradation under environmental shifts, providing a causal analysis of these factors alongside data resources and evaluation protocols.

2.2 Causal Learning and Counterfactual Generation

Causal learning helps explain bias and generalization failure in medical imaging. Castro et al. (2020) analyze transfer bias through the data-generating process; Wang et al. (2022) improve generalization through invariant transformations that preserve causally relevant features while changing other factors; Pawlowski et al. (2020) introduce deep structural causal models that support interventions on generative factors and counterfactual inference. To address observation shifts and limited configuration coverage, we combine stable prediction under patient and acquisition shifts with cross-configuration structural learning grounded in normal variation, promoting transferable abnormality evidence across environments and configurations.

Artificial abnormalities provide a direct way to learn abnormal variation. Chang et al. (2021) combine label-changing counterfactual generation with label-preserving invariant augmentation to reduce background dependence. Zhang et al. (2026)'s Perturb-and-Restore generates abnormalities through chromosome segment rearrangement and diffusion restoration, selecting training samples through energy-guided sampling. We combine normal-statistics calibration with structural counterfactual learning: abundant normal pairs establish a reference, while controlled rearrangements supplement abnormal variation. This reframes normal–abnormal classification as distinguishing normal from abnormal variation, reducing reliance on chromosome-specific appearance and seen abnormality patterns so that synthetic data support transferable structural learning.

3 ChromoShift Dataset

ChromoShift comprises skeleton–centromere annotations, normal-case pretraining data, and real abnormality data for distribution shift evaluation. Clinical images come from collaborating clinical centers, with annotations involving three clinical experts. Table 1 summarizes the components and inversion configurations; karyotype counts refer to source records and image counts to individual chromosomes.

Data overview. Structural annotations include ordered skeletons, endpoints, centromeres, and foreground masks. Normal-case data capture morphological, banding, and imaging variation across patients and cells, supporting pretraining tasks including normal structural learning and abnormality generation. An additional 133,484 images from five public sources supplement pretraining. Real abnormality data cover 25 inversion configurations across all 24 chromosome classes, with multiple cells per patient and configurations shared across patients supporting evaluation under patient, acquisition, and configuration shifts.

Table 1: Data components and inversion configurations of ChromoShift.

Component	Patients	Karyotypes	Images	Abnormal	Configs.
Skeleton–centromere	–	580	4,006	0	–
Normal pretraining	105	8,466	389,278	0	–
Abnormality shift data	84	1,251	57,546	1,251	25
Inversion type	Configurations				
Paracentric (12 configurations)	3:q13.2q26.2 11:q13.1q21 22:q12.2q13.1	4:q21.1q27 13:q14q21 X:p22.3p11.4	5:q13.1q15 15:q15q24	8:p22p21.1 19:p13.3p13.2	9:q22.1q32 21:q11.2q22.2
Pericentric (13 configurations)	1:p13q21 10:p21q23 18:p11.2q11.2	2:p11.2q13 12:p13.3q13.1 20:p13q11.2	6:p23q21 14:p13q21 Y:p11.2q11.23	7:p14q22 16:p11.2q12.1	9:q12p13 17:p11.2q21.3

4 METHODS

Figure 3 outlines our causally guided two-stage framework for unseen-configuration recognition and environmental robustness. Homologous comparison and normal-statistics calibration support structural learning through counterfactuals and configuration generalization, followed by classifier adjustment through quality perturbations and reference substitution with representations and statistics fixed.

4.1 PROBLEM FORMULATION AND CAUSAL ANALYSIS

Problem formulation. Given a target chromosome image x_T and a reference x_R of the same chromosome type, the model extracts a paired representation F and predicts the target abnormality probability:

$$F = \Phi_{\theta,\Omega}(x_T, x_R), \qquad \hat{y} = s(x_T \mid x_R) = g_\psi(F). \tag{1}$$

Here, θ and ψ parameterize the encoder and classification head, and Ω contains calibration and global standardization statistics. $A \in \{0, 1\}$ denotes abnormality status and a_T the target label. We use the actual homolog as reference when available, otherwise a standard ideogram. From limited patients and configurations, we seek structural evidence that transfers to unseen configurations and remains stable across patient backgrounds and preparation and imaging conditions.

Causal modeling. Figure 2 describes the relationships among abnormality formation, image acquisition, and model prediction. A indicates inversion status; B specifies chromosome type, breakpoints, and the inverted interval's relationship to the centromere; P denotes patient-specific biological and cytogenetic background; and M describes the underlying band arrangement, morphology, and relative centromere position. Q represents preparation, staining, imaging, and extraction conditions. X, F, and $\hat{Y}$ denote the observed image, representation, and prediction. The domain indicator E points to P, Q, and B, marking their distribution shifts across environments (Castro et al., 2020; Pearl & Bareinboim, 2022). Structural formation and image acquisition follow

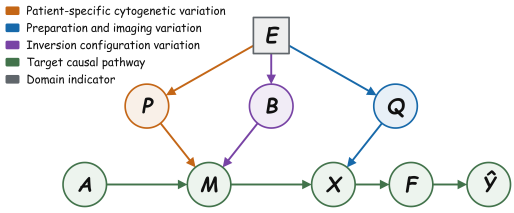


Figure 2: Causal selection diagram of patient, acquisition, and configuration shifts.

$$M = f_M(A, B, P, U_M), \qquad X = f_X(M, Q, U_X), \tag{2}$$

where U_M and U_X capture unmodeled structural and imaging factors, and B specifies the inversion configuration when $A = 1$. The intended evidence pathway $A \to M \to X \to F \to \hat{Y}$ bases predictions on abnormality-induced structural changes.

We model environmental shifts through the joint distribution of P, Q, B, assuming stable conditional structural and imaging mechanisms. For chromosome type k, marginalizing over structure, patient

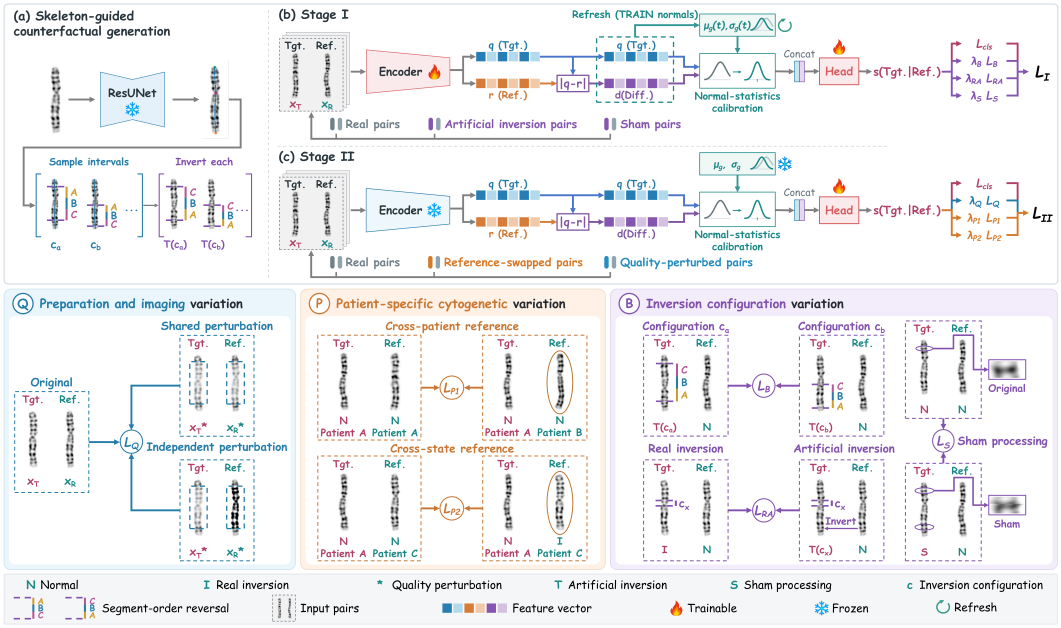


Figure 3: Overview of the proposed framework. Stage I learns cross-configuration structural representations; Stage II stabilizes predictions under environmental changes. Homologous comparison and normal-statistics calibration support both stages.

background, acquisition conditions, and configuration gives the class-conditional image distribution in environment e:

$$p_e(x \mid a, k) = \int p(x \mid m, q, k)\, p(m \mid a, b, p, k)\, p_e(p, q, b \mid a, k)\, \mathrm{d}(m, p, q, b). \tag{3}$$

Discrete variables are summed rather than integrated. Thus, patient, acquisition, and configuration shifts can change observations within the same abnormality class despite stable generative mechanisms. $P \to M$ introduces normal individual structural variation, and $Q \to X$ alters visibility and appearance, motivating environmental robustness. $B \to M$ changes rearrangement location, extent, and form, motivating cross-configuration structural learning under limited training coverage.

4.2 Homologous Comparison and Normal-Statistics Calibration

Conventional classifiers learn normal–abnormal boundaries from features shaped by patient background, acquisition conditions, and inversion configuration, intertwining discrimination with training-specific patterns. We extract structural differences through homologous comparison and calibrate the range of normal variation using abundant normal pairs. This provides a common reference for distinguishing normal from abnormal variation, reducing reliance on specific appearance and simplifying discrimination.

A shared encoder produces target and reference features and a directed pair representation:

$$q = e_\theta(x_T), \qquad r = e_\theta(x_R), \qquad d = |q - r|, \qquad h = [q; d]. \tag{4}$$

Absolute values are elementwise, and the semicolon denotes concatenation. d captures relative variation, while q preserves target structure to identify which member is abnormal. Swapping members provides a representation for each chromosome.

Homologous differences arise from both normal variation and abnormal rearrangements, requiring the range of normal variation as a baseline to distinguish between them. For groups g defined by chromosome and reference type, we estimate the mean μ_g and scale σ_g from real normal training pairs in the current fold with equal patient weighting:

$$z = \frac{h - \mu_g}{\sigma_g}, \qquad F = D(z), \qquad s(x_T \mid x_R) = \mathrm{sigmoid}(w^\top F + b). \tag{5}$$

Here, division is elementwise, D is shared global standardization estimated from real training members, and w, b are the linear head's weights and bias. Shrinkage toward pooled normal variance and a positive scale floor stabilize small-group estimates. During training, normal-statistics estimation excludes the target patient; validation and testing use statistics estimated from all normal training pairs in the current fold.

Calibration expresses target features and homologous differences relative to the normal center and scale. In the original feature space, the shared head has group-specific slopes and intercepts determined by normal statistics, reducing the need to learn these adjustments from scarce abnormalities. Normal pairs establish the comparison scale, while real and artificial abnormalities supply signals for distinguishing normal from abnormal variation.

All directed supervised samples (x_T, x_R, a_T) use weighted binary cross-entropy:

$$\mathcal{L}_{\text{cls}} = \mathbb{E}_\omega[\ell(s, a_T)], \qquad \ell(s, a) = -a \log s - (1 - a) \log(1 - s).$$

$\mathbb{E}_\omega$ averages with normalized sample weights. All subsequent supervised pairs share this objective, with labels assigned to the target member.

4.3 Cross-Configuration Representation Learning

4.3.1 Skeleton-Guided Counterfactual Generation

Complex chromosome curvature challenges endpoint coverage and centromere localization in conventional skeleton extraction methods. We train a lightweight ResUNet (Zhang et al., 2018) to jointly predict skeletons, endpoints, and centromeres, then freeze it for counterfactual construction. In Figure 3(a), normalized skeleton arc length defines $u \in [0, 1]$, and configuration c specifies chromosome type and interval $[u_1, u_2]$. T_c maps positions within the interval to $u_1 + u_2 - u$, reverses local structural order, and reconstructs the image. Intervals containing or excluding the centromere yield pericentric or paracentric inversions, respectively.

For generated samples, we check structural validity, then assess observable abnormality evidence through reversal-induced band differences and relative centromere displacement. Samples meeting the visibility criteria are designated high-visibility samples for subsequent supervision and constraints. Each valid artificial abnormality $T_c(x_N)$, generated from a normal chromosome x_N, is paired with a normal reference x_R to form $(T_c(x_N), x_R, 1)$ and $(x_R, T_c(x_N), 0)$ for $\mathcal{L}_{\text{cls}}$. Varying breakpoints and intervals supplies abnormal variation beyond the coverage of scarce real samples.

4.3.2 Configuration Generalization Learning

Average classification supervision can obscure performance differences across expanded configurations. We group configurations by centromere crossing, interval length, and midpoint, and use current group classification loss to emphasize underperforming configurations. Each step samples at most two high-visibility groups, balancing patients within each. With b_i denoting the mean classification loss over sample i's two supervised directions and $\mathcal{B}_j$ the samples in group j, we define

$$R_j = \frac{1}{|\mathcal{B}_j|} \sum_{i \in \mathcal{B}_j} b_i, \qquad \mathcal{L}_B = \max(R_1, R_2) - \frac{R_1 + R_2}{2}.$$

Both groups contribute to $\mathcal{L}_{\text{cls}}$ with prescribed weights, while $\mathcal{L}_B$ emphasizes harder configurations by penalizing the higher risk's excess over the group mean, without requiring identical prediction confidence. With only one available group, $\mathcal{L}_B = 0$ and classification supervision remains.

4.3.3 Synthetic Bias Suppression

Counterfactual rearrangement introduces processing artifacts that may reveal generation source. We combine real–artificial alignment and sham controls to concentrate learning signals on structural changes. Using the clinical breakpoints of a real abnormal chromosome, we generate a matched artificial abnormality from its normal homolog and compare real and artificial pairs in both directions, targeting the abnormal and normal members in turn. Their probabilities $p_{i,d}^{\text{real}}$ and $p_{i,d}^{\text{art}}$ define

$$\mathcal{L}_{\text{RA}} = \frac{1}{2n} \sum_{i=1}^{n} G_i \sum_{d=1}^{2} \left[\left(|p_{i,d}^{\text{art}} - p_{i,d}^{\text{real}}| - \varepsilon \right)_+ \right]^2. \tag{6}$$

Here, i indexes instances, d indexes prediction directions, G_i selects instances meeting visibility, breakpoint-information, and centromere-confidence requirements, n counts samples before filtering, and $(v)_+ = \max(v, 0)$. Both predictions receive gradients; $\varepsilon = 0.1$ tolerates response differences.

Sham processing preserves band order while reproducing local sampling, reconstruction, and boundary processing on the target, reference, or both members of real normal–normal pairs. For target probabilities $p_i, p_{S,i}$ before and after processing,

$$\mathcal{L}_S = \sum_{i \in \mathcal{I}_S} \omega_i \left[(p_{S,i} - \mathrm{sg}(p_i) - \varepsilon)_+ \right]^2, \qquad \mathcal{I}_S = \{i : p_i \le 0.5\}. \tag{7}$$

With stop-gradient sg and sample weights ω_i, this penalizes excessive probability increases relative to clean inputs correctly classified as normal, suppressing artifact-induced false positives.

4.4 Cross-Environment Robust Discrimination

Stage II fixes representations and statistics, learning a stable classification direction through controlled acquisition and reference changes. For normal and abnormal targets, Q perturbs either or both inputs through intensity, contrast, mild blur, and noise changes, using shared or independent parameters for joint perturbations. Memberwise logit consistency is enforced under label- and structural-evidence-preserving perturbations. Reference constraints P keep the target fixed: P_1 replaces its same-patient normal reference with a normal chromosome of the same type from another patient; P_2 fixes a normal target and switches between another patient's normal and abnormal homologs. These constraints reduce sensitivity to reference-patient variation and abnormality status, respectively.

For complete control-pair inputs V, V' and $\Delta F = F(V') - F(V)$, fixing the Stage I bias b_{I} gives

$$\begin{aligned} \eta(V) &= w^\top F(V) + b_{\mathrm{I}}, \\ \mathcal{L}_j &= \sum_i a_{i,j} \left[\eta(V_i') - \eta(V_i) \right]^2 = w^\top G_j w, \\ G_j &= \sum_i a_{i,j} \Delta F_{i,j} \Delta F_{i,j}^\top, \qquad j \in \{P_1, P_2, Q\}. \end{aligned} \tag{8}$$

Here, η is the abnormality logit (log-odds), and $a_{i,j} \ge 0$ weights control pairs. Q aggregates perturbation modes and member directions. Reference substitution changes only the difference branch, whereas quality perturbations can change both branches, controlling distinct sources of prediction variation.

For label-preserving controls, let $J = \mathbb{E}[(\eta' - \eta)^2] < \infty$, with finite average cross-entropy risk on at least one side. The Lipschitz properties of sigmoid and cross-entropy in the logit yield

$$\mathbb{E}|s' - s| \le \frac{1}{4}\sqrt{J}, \qquad |R' - R| \le \sqrt{J}, \tag{9}$$

where R, R' are average cross-entropy risks before and after the change. Thus, controlling logit variation stabilizes both probabilities and classification loss.

4.5 Optimization Objectives

Configuration learning must preserve subtle abnormal structural differences, whereas environmental constraints reduce sensitivity to patient and imaging variation. To reduce interference from environmental constraints during structural representation learning, we first learn cross-configuration representations, then fix them and adjust the classifier so that these constraints act on the use of learned structural evidence.

Stage I jointly updates the encoder and classification head, refreshing normal statistics as the encoder changes, with objective:

$$\mathcal{L}_{\mathrm{I}} = \mathcal{L}_{\mathrm{cls}} + \lambda_B \mathcal{L}_B + \lambda_{\mathrm{RA}} \mathcal{L}_{\mathrm{RA}} + \lambda_S \mathcal{L}_S.$$

Stage II fixes the encoder, normal statistics, and head bias, updating only classification weights, with objective:

$$\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\mathrm{cls}} + \lambda_Q \mathcal{L}_Q + \lambda_{P1} \mathcal{L}_{P1} + \lambda_{P2} \mathcal{L}_{P2}.$$

Here, λ weights each constraint, and Stage II classification supervision uses only real samples.

Table 2: Generalization results on ChromoShift (five-fold mean ± standard deviation, %). Sens. and Spec. denote sensitivity and specificity. Bold and underlined values indicate the best and second-best means, respectively.

	Cross-Environment Generalization					Unseen-Configuration Generalization				
Method	AUROC	AUPRC	Sens.	Spec.	F1	AUROC	AUPRC	Sens.	Spec.	F1
LR (Cox, 1958)	71.66 ± 8.22	7.49 ± 2.29	48.12 ± 11.99	79.62 ± 10.98	10.34 ± 4.09	52.47 ± 5.67	2.44 ± 0.42	21.74 ± 8.46	79.42 ± 9.48	4.45 ± 1.54
SVM (Cortes & Vapnik, 1995)	71.52 ± 8.53	6.72 ± 2.39	49.27 ± 16.04	79.33 ± 8.84	9.54 ± 3.46	52.34 ± 5.58	2.40 ± 0.35	22.47 ± 7.52	78.71 ± 7.37	4.17 ± 0.55
RF (Breiman, 2001)	75.37 ± 6.46	6.34 ± 1.85	36.96 ± 16.10	85.64 ± 8.14	9.70 ± 1.77	57.70 ± 6.88	2.84 ± 0.51	21.46 ± 13.35	85.94 ± 7.51	5.45 ± 1.16
ResNet50 (He et al., 2016)	90.90 ± 5.17	45.98 ± 17.30	22.74 ± 17.96	**99.83 ± 0.25**	31.60 ± 21.42	54.77 ± 9.01	3.39 ± 0.98	8.93 ± 4.75	95.79 ± 2.51	5.73 ± 1.80
DenseNet121 (Huang et al., 2017)	91.27 ± 4.78	48.64 ± 14.43	31.79 ± 16.04	99.76 ± 0.18	42.82 ± 16.71	58.76 ± 14.91	5.54 ± 2.14	12.71 ± 8.04	94.77 ± 4.58	7.31 ± 1.85
Swin-T (Liu et al., 2021)	80.73 ± 10.19	20.24 ± 20.42	33.17 ± 24.39	94.66 ± 2.12	16.54 ± 11.57	51.42 ± 9.86	2.81 ± 0.75	9.03 ± 7.12	94.50 ± 1.97	4.54 ± 2.78
ChroSiameseAD (Bechar et al., 2023)	62.88 ± 21.62	24.61 ± 24.96	17.32 ± 19.85	99.61 ± 0.28	24.73 ± 27.04	49.08 ± 8.88	4.69 ± 1.56	0.74 ± 0.90	99.29 ± 0.30	1.31 ± 1.58
BR-ChromNet (Chen et al., 2024)	69.45 ± 10.84	10.54 ± 7.44	13.34 ± 14.29	99.03 ± 0.45	15.05 ± 14.63	53.00 ± 12.88	2.61 ± 1.06	0.53 ± 0.78	99.05 ± 0.29	0.72 ± 1.06
HomNet (Li et al., 2024)	70.02 ± 6.20	12.49 ± 5.77	18.82 ± 25.18	93.99 ± 8.24	10.04 ± 9.92	55.79 ± 24.57	10.42 ± 7.05	24.14 ± 26.64	93.10 ± 6.57	15.22 ± 13.31
DCF (Hao et al., 2025)	49.50 ± 7.89	2.41 ± 0.46	42.41 ± 37.52	55.74 ± 43.25	4.03 ± 0.88	52.39 ± 8.87	2.45 ± 0.80	45.08 ± 45.33	55.44 ± 42.54	3.28 ± 1.91
CHROMA (Yang et al., 2026)	89.43 ± 6.32	36.57 ± 11.23	26.47 ± 14.92	99.65 ± 0.32	35.16 ± 13.63	75.04 ± 10.44	25.72 ± 21.36	21.77 ± 20.36	**99.58 ± 0.24**	27.46 ± 21.98
Ours	**93.41 ± 5.36**	**56.10 ± 21.33**	**52.98 ± 17.14**	99.03 ± 0.55	**53.90 ± 15.58**	**90.83 ± 5.74**	**46.77 ± 25.32**	**45.44 ± 22.36**	98.77 ± 0.70	**45.01 ± 19.40**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We use patient-disjoint five-fold cross-validation on ChromoShift: all images from each patient remain in one training, validation, or test set per fold, and each patient is tested once. We define two test domains to evaluate cross-environment and unseen-configuration generalization. The former tests seen configurations in unseen patients under their acquisition conditions; the latter additionally requires test configurations to be absent from real training and validation data. For each method, one model is trained per fold and tested on both domains; all methods use identical splits. Evaluation metrics include AUROC, AUPRC, sensitivity, specificity, and F1.

We use ViT-L/16 (Dosovitskiy et al., 2020) as a shared encoder for the target and reference, with a linear head for abnormality classification. Stage I uses AdamW (Loshchilov & Hutter, 2017) with encoder/head learning rates of $10^{-6}/10^{-5}$, batch size 24, and at most 10 epochs. Counterfactual construction, reference substitution, and normal-statistics estimation use only the current fold's training data.

5.2 COMPARISON WITH EXISTING METHODS

Table 2 compares traditional classifiers, general-purpose deep networks, and chromosome analysis methods. We achieve the highest mean AUROC, AUPRC, sensitivity, and F1 in both domains. Cross-environment AUROC, AUPRC, and F1 reach 93.41%, 56.10%, and 53.90%, exceeding the strongest comparator, DenseNet121, by 2.14, 7.46, and 11.08 percentage points, respectively.

Unseen configurations widen the gap. ResNet50 and DenseNet121 exceed 90% AUROC across environments but fall to 54.77% and 58.76% on unseen configurations, indicating limited transfer to unseen configurations. Our AUROC, AUPRC, and F1 reach 90.83%, 46.77%, and 45.01%, surpassing CHROMA, the strongest comparator, by 15.79, 21.05, and 17.55 percentage points.

The severe imbalance between normal and abnormal samples can favor conservative decision rules biased toward the normal class, yielding high specificity but missing many abnormalities (He & Garcia, 2009). By learning from diverse abnormal configurations, our method detects abnormalities more actively but also produces more false positives. Its specificity is therefore slightly lower than that of some comparators, but remains above 98%, with higher sensitivity and F1.

5.3 ABLATION STUDY

Table 3 assesses pretraining and individual components. Normal-data pretraining raises base-model AUROC from 54.67% to 76.11% across environments and from 53.78% to 64.73% on unseen configurations; the full framework reaches 93.41% and 90.83%. These results show that normal-data pretraining provides an effective foundation, while the full framework further improves generalization in both domains.

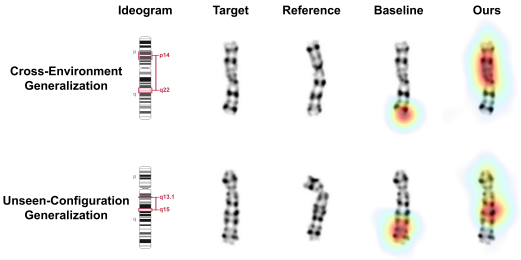


Figure 4: Positive spatial responses from RISE-style random masking.

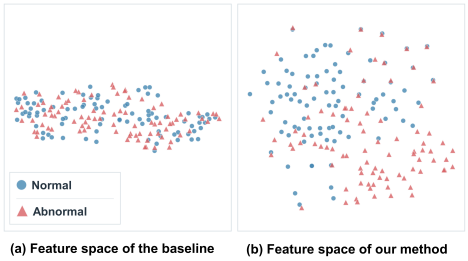


Figure 5: t-SNE visualization of classifier-weighted contributions.

Normal-statistics calibration (NSC) mainly benefits unseen configurations: removing it lowers AUROC/AUPRC by 8.61/9.16 percentage points, while cross-environment AUROC slightly rises and AUPRC falls. This supports normal variation as a reference for structural transfer. Removing cross-configuration representation learning (CCRL) lowers unseen-configuration AUROC/AUPRC by 3.14/4.47 percentage points, supporting its role in learning transferable structural evidence.

Table 3: Ablation results on ChromoShift (five-fold mean ± standard deviation, %). Bold and underlined values indicate the best and second-best means, respectively.

	Cross-Environment Generalization		Unseen-Configuration Generalization	
Variant	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
Base model	54.67 ± 11.16	2.76 ± 0.77	53.78 ± 9.11	2.57 ± 0.57
Base + pretraining	76.11 ± 11.87	21.15 ± 13.03	64.73 ± 13.39	6.81 ± 4.87
w/o NSC	$\mathbf{93.84 \pm 4.56}$	55.47 ± 24.29	82.22 ± 14.85	37.61 ± 21.94
w/o CCRL	92.78 ± 3.18	53.85 ± 18.04	87.69 ± 5.02	42.30 ± 24.70
w/o P	91.35 ± 5.47	54.96 ± 21.27	$\underline{90.69 \pm 5.84}$	$\underline{46.58 \pm 25.45}$
w/o Q	92.70 ± 5.30	$\underline{55.67 \pm 21.09}$	89.82 ± 5.67	45.64 ± 25.26
Full model	$\underline{93.41 \pm 5.36}$	$\mathbf{56.10 \pm 21.33}$	$\mathbf{90.83 \pm 5.74}$	$\mathbf{46.77 \pm 25.32}$

Reference constraints P and quality constraint Q complement structural learning with environmental robustness. Removing P's cross-patient and cross-state reference constraints lowers cross-environment AUROC by 2.06 points, with smaller changes on unseen configurations. Removing Q lowers AUROC by 0.71 and 1.01 points in the two domains and reduces AUPRC in both. These results highlight the complementary roles of cross-configuration representation learning and environmental robustness constraints.

5.4 Interpretability and Qualitative Analysis

We examine spatial responses and discriminative representations. Figure 4 uses RISE-style random masking to visualize responses associated with target abnormality scores (Petsiuk et al., 2018). In examples from both domains, the base model responds mainly near chromosome ends, whereas ours responds more to internal bands that better match the annotated intervals' relative locations, indicating different spatial emphases in discrimination.

Figure 5 shows t-SNE (Van der Maaten & Hinton, 2008) projections of normal–abnormal homologous pairs from the unseen-configuration domain. Contribution vectors are the elementwise product of classification head inputs and weights; both models use identical samples and projection settings. The base model's classes are strongly intermixed, while ours show clearer separation in the projection, consistent with improved unseen-configuration recognition.

6 Conclusion

We propose a causally guided two-stage framework addressing environmental sensitivity and unseen-configuration failure in chromosomal structural abnormality detection. Homologous comparison and normal-statistics calibration recast normal–abnormal classification as distinguishing normal from abnormal variation; counterfactual configuration learning and environmental constraints improve structural evidence transfer. We introduce the full-pipeline ChromoShift dataset and show gains in both test domains, particularly on unseen configurations. Together, the framework and dataset support assessing known abnormalities and screening potentially novel configurations.

AI USE STATEMENT

LLMs were used solely to assist with writing and polishing the manuscript, including correcting grammatical errors and improving clarity and readability. They were not used for research ideation, methodology development, experimental design or execution, or analysis and interpretation of results. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of this paper.

ETHICS STATEMENT

This study uses chromosome images provided by collaborating clinical centers, which have authorized data sharing for research purposes. Information that directly identifies patients has been removed from the research data. Images and associated annotations are available upon request from the corresponding author, subject to institutional data-sharing policies.

REPRODUCIBILITY STATEMENT

Section 4 describes the method, and Section 5 specifies the evaluation protocol. Appendix A provides the theoretical assumptions and derivations; Appendix B details data curation and patient-disjoint splits; Appendices C and D describe skeleton extraction, pretraining, and implementation settings; and Appendix E reports additional ablations. Code is available at `https://anonymous.4open.science/r/CausalChrom-3D66/`.

REFERENCES

Akila Subasinghe Arachchige, Jagath Samarabandu, Joan Knoll, Wahab Khan, and Peter Rogan. An image processing algorithm for accurate extraction of the centerline from human metaphase chromosomes. In *2010 IEEE International Conference on Image Processing*, pp. 3613–3616. IEEE, 2010.

Xiang Bai, Longin Jan Latecki, and Wen-Yu Liu. Skeleton pruning by contour partitioning with discrete curve evolution. *IEEE transactions on pattern analysis and machine intelligence*, 29(3): 449–462, 2007.

Mohammed El Amine Bechar, Jean-Marie Guyader, Marwa El Bouz, Nathalie Douet-Guilbert, Ayman Al Falou, and Marie-Berengere Troadec. Highly performing automatic detection of structural chromosomal abnormalities using siamese architecture. *Journal of Molecular Biology*, 435 (8):168045, 2023.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Chun-Hao Chang, George Alexandru Adam, and Anna Goldenberg. Towards robust classification model by counterfactual and invariant data generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15207–15216. IEEE, 2021.

Siyuan Chen, Tingting Hu, Na Li, Xin Gao, and Yongguo Yu. Br-chromnet: Banding resolution localization of chromosome structural abnormality with conditional random field. *Journal of molecular biology*, 436(20):168733, 2024.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232, 1958.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Enrico Grisan, Enea Poletti, and Alfredo Ruggeri. Automatic segmentation and disentangling of chromosomes in q-band prometaphase images. *IEEE Transactions on information technology in Biomedicine*, 13(4):575–581, 2009.

Frans CA Groen, Ton K ten Kate, Arnold WM Smeulders, and Ian T Young. Human chromosome classification based on local band descriptors. *Pattern Recognition Letters*, 9(3):211–222, 1989.

Zicheng Guo and Richard W Hall. Parallel thinning with two-subiteration algorithms. *Communications of the ACM*, 32(3):359–373, 1989.

Yue Hao, Xin Wang, Ge Song, Zhiyuan Li, Lei Wang, Lingwei Li, Yongqi Nie, and Peng Wang. A general framework for chromosomal anomaly detection based on dual constraints of nearest-neighbor and regionality. *Engineering Applications of Artificial Intelligence*, 156:111165, 2025.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

Wim H Hesselink and Jos BTM Roerdink. Euclidean skeletons of digital image and volume data in linear time by the integer medial axis transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2204–2217, 2008.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2261–2269. Ieee, 2017.

Dakai Jin, Krishna S Iyer, Cheng Chen, Eric A Hoffman, and Punam K Saha. A robust and efficient curve skeletonization algorithm for tree-like objects using minimum cost paths. *Pattern recognition letters*, 76:32–40, 2016.

Jau-hong Kao, Jen-hui Chuang, and Tsaipei Wang. Chromosome classification based on the band profile similarity along approximate medial axis. *Pattern Recognition*, 41(1):77–89, 2008.

W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6): 996–1006, 2002.

Boaz Lerner, Hugo Guterman, I Dinstein, and Yitzhak Romem. Medial axis transform-based features and a neural network for human chromosome classification. *Pattern Recognition*, 28(11): 1673–1683, 1995.

Juren Li, Fanzhe Fu, Ran Wei, Yifei Sun, Zeyu Lai, Ning Song, Xin Chen, and Yang Yang. Chromosomal structural abnormality diagnosis by homologous similarity. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 5317–5328, 2024.

Chengchuang Lin, Gansen Zhao, Zhirong Yang, Aihua Yin, Xinming Wang, Li Guo, Hanbiao Chen, Zhaohui Ma, Lei Zhao, Haoyu Luo, et al. Cir-net: Automatic classification of human chromosome based on inception-resnet architecture. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(3):1285–1293, 2020.

Chengchuang Lin, Gansen Zhao, Aihua Yin, Zhirong Yang, Li Guo, Hanbiao Chen, Lei Zhao, Shuangyin Li, Haoyu Luo, and Zhaohui Ma. A novel chromosome cluster types identification method using resnext wsl model. *Medical Image Analysis*, 69:101943, 2021.

Chengchuang Lin, Hanbiao Chen, Jiesheng Huang, Jing Peng, Li Guo, Zhirong Yang, Jiahua Du, Shuangyin Li, Aihua Yin, and Gansen Zhao. Chromosomenet: A massive dataset enabling benchmarking and building basedlines of clinical chromosome classification. *Computational Biology and Chemistry*, 100:107731, 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

C Metz, M Schaap, T Van Walsum, and W Niessen. Two point minimum cost path approach for cta coronary centerline extraction. *The Insight Journal*, 12, 2008.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Robert L Ogniewicz and Markus Ilg. Voronoi skeletons: theory and applications. In *CVPR*, volume 92, pp. 63–69, 1992.

Alistair T Pagnamenta, Jing Yu, Susan Walker, Alexandra J Noble, Jenny Lord, Prasun Dutta, Mona Hashim, Carme Camps, Hannah Green, Smrithi Devaiah, et al. The impact of inversions across 33,924 families with rare disease from a national genome sequencing project. *The American Journal of Human Genetics*, 111(6):1140–1164, 2024.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. In *Probabilistic and causal inference: The works of Judea Pearl*, pp. 451–482. 2022.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Claire Redin, Harrison Brand, Ryan L Collins, Tammy Kammin, Elyse Mitchell, Jennelle C Hodge, Carrie Hanscom, Vamsee Pillalamarri, Catarina M Seabra, Mary-Alice Abbott, et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature genetics*, 49(1):36–45, 2017.

Wei Shen, Xiang Bai, Rong Hu, Hongyuan Wang, and Longin Jan Latecki. Skeleton growing and pruning with bending potential ratio. *Pattern Recognition*, 44(2):196–209, 2011.

Marisa Silva, Nicole de Leeuw, Kathy Mann, Heleen Schuring-Blom, Sian Morgan, Daniela Giardino, Katrina Rack, and Ros Hastings. European guidelines for constitutional cytogenomic analysis. *European Journal of Human Genetics*, 27(1):1–16, 2019.

Sergio Tofanelli, Roscoe Stanyon, Maura Agostini, Marcello Giovanni Franceschi, and Giorgio Paoli. Variability of da/dapi and c heterochromatic regions: a population study. *Human biology*, pp. 635–646, 1993.

Lukas Uzolas, Javier Rico, Pierrick Coupé, Juan C SanMiguel, and György Cserey. Deep anomaly generation: an image translation approach of synthesizing abnormal banded chromosome images. *IEEE Access*, 10:59090–59098, 2022.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ram S Verma and Sunny Luke. Heteromorphisms of pericentromeric heterochromatin of chromosome 19. *Genetic Analysis: Biomolecular Engineering*, 8(6):179–180, 1991.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Ruoyu Wang, Mingyang Yi, Zhitang Chen, and Shengyu Zhu. Out-of-distribution generalization with causal invariant transformations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 375–385. IEEE, 2022.

Xingwei Wang, Bin Zheng, Shibo Li, John J Mulvihill, and Hong Liu. Development and assessment of an integrated computer-aided detection scheme for digital microscopic images of metaphase chromosomes. *Journal of Electronic Imaging*, 17(4):043008–043008, 2008.

Xingwei Wang, Bin Zheng, Shibo Li, John J Mulvihill, Marc C Wood, and Hong Liu. Automated classification of metaphase chromosomes: optimization of an adaptive computerized scheme. *Journal of biomedical informatics*, 42(1):22–31, 2009.

Changchun Yang, Weiqian Dai, Yilan Zhang, Siyuan Chen, Jingdong Hu, Junkai Su, Yuxuan Chen, Ao Xu, Na Li, Xin Gao, et al. A comprehensive foundation model for generalizable cytogenetics in precision oncology with chroma. *NPJ Precision Oncology*, 2026.

Tongjie Y Zhang and Ching Y. Suen. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 27(3):236–239, 1984.

Yilan Zhang, Hanbiao Chen, Changchun Yang, Yuetan Chu, Siyuan Chen, Jing Wu, Jingdong Hu, Na Li, Junkai Su, Yuxuan Chen, et al. Perturb-and-restore: Simulation-driven structural augmentation framework for imbalance chromosomal anomaly detection. *arXiv preprint arXiv:2604.00854*, 2026.

Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual u-net. *IEEE geoscience and remote sensing letters*, 15(5):749–753, 2018.

A Theoretical Analysis and Proofs

This appendix examines how normal-statistics calibration, structural construction, and prediction constraints affect different parts of the learning process. Section A.2 derives configuration-group supervision, the P/Q constraints, and their transfer conditions. Section A.3 analyzes the coordinate transformation induced by normal-statistics calibration and its estimation error.

A.1 Notation and Modeling Assumptions

Inputs and predictor. An ordered supervised sample is denoted by (x_T, x_R, a_T), where x_T is the target image, x_R is a reference of the same chromosome type, and $a_T \in \{0, 1\}$ is the target inversion label. The shared encoder produces $q = e_\theta(x_T)$ and $r = e_\theta(x_R)$. The pair representation is $h = [q; |q - r|] \in \mathbb{R}^{2d_e}$, where d_e is the feature dimension and the absolute value is applied elementwise. Swapping the members leaves the difference branch unchanged and replaces q with r in the target branch, yielding a separate prediction for each chromosome.

The statistical group is defined as $g = (k, t_R)$, where k denotes chromosome type and t_R distinguishes a real chromosome reference from a standard ideogram. For male samples, X and Y chromosomes lack a homologous reference of the same type, so standard ideograms are used. Cross-patient reference substitution preserves chromosome type and reference type and therefore retains the statistical group. Let $S_g = \mathrm{diag}(\sigma_g)$ denote the scale matrix used for normal-statistics calibration,

and let $D(z) = \Lambda^{-1}(z - m_D)$ denote shared global standardization, where $\Lambda = \mathrm{diag}(s_D)$. The predictor in the main text can then be written as

$$\begin{aligned} z &= S_g^{-1}(h - \mu_g), & F &= \Lambda^{-1}(z - m_D), \\ \eta &= w^\top F + b, & s(x_T \mid x_R) &= \mathrm{sigmoid}(\eta). \end{aligned} \tag{A.1}$$

Here, η is the logit, s is the abnormality probability, and $\Omega = \{\mu_g, \sigma_g, m_D, s_D\}_g$ contains all statistics. The group statistics μ_g, σ_g are estimated from real normal training targets and their valid normal references in the current fold, including real normal homologs and standard ideograms in the corresponding statistical groups. Inputs pairing normal targets with abnormal references are excluded from the normal baseline. Within each group, normal moments are aggregated with equal weight per patient and equal weight per valid normal input within each patient. The pooled moments used for shrinkage assign equal weight to each group–patient entry. Let v_g and v_{pool} denote the elementwise group and pooled variances computed from the corresponding first and second moments. The elementwise scale is

$$\sigma_g = \max\left\{\sqrt{(1-\alpha)\max(v_g, 0) + \alpha \max(v_{\mathrm{pool}}, 0)}, \epsilon_\sigma\right\}.$$

We use $\alpha = 0.25$ and $\epsilon_\sigma = 10^{-5}$. The global statistics m_D, s_D are estimated from real training members, including normal and abnormal targets, with equal weights across classes, across patients within each class, and across members within each patient–class combination. The global scale floor is 10^{-6}.

Training inputs use normal moments computed after excluding the target patient, denoted by $\mu_{g,-i}, \sigma_{g,-i}$. Validation and testing use statistics from all normal training pairs. The exclusion applies to both the conditional group and the shrinkage pool. Reference substitution retains the exclusion index i_T of the target patient, while the substituted input determines the statistical group:

$$F(V') = D\big(C_{g(V'),-i_T}(h(V'))\big).$$

The exclusion index does not change with the reference patient. Below, we fix the statistical snapshot and exclusion rule and write μ_g, σ_g for brevity. Stage I updates the encoder and classification head and refreshes the statistical snapshot at the end of each epoch. Stage II fixes θ, Ω, and the Stage I bias b_I, and optimizes only the head weights w. We denote the Stage I weights by w_I.

Generative factors and environments. Following the main text, A denotes inversion status, B inversion configuration, P patient background, Q preparation and imaging conditions, M the underlying structure, X the observed image, and E the data domain. The analysis conditions on the known chromosome type k. When $A = 0$, inversion breakpoints and interval relations take null values, while k is retained. The structural equations are

$$M = f_M(A, B, P, U_M), \qquad X = f_X(M, Q, U_X). \tag{A.2}$$

We formalize the stable conditional mechanisms assumed in the main text by requiring conditional kernels K_M, K_X that do not depend on the environment e:

$$\begin{aligned} p_e(m \mid a, b, p, q, k) &= K_M(m \mid a, b, p, k), \\ p_e(x \mid m, a, b, p, q, k) &= K_X(x \mid m, q, k). \end{aligned} \tag{A.3}$$

Environmental variation is represented by the joint distribution $p_e(p, q, b \mid a, k)$, allowing observational associations among patient background, acquisition conditions, and configuration. This assumption also constrains the conditional distributions induced by unmodeled residual factors; identical structural-function forms alone do not imply Eq. (A.3). The domain node in the diagram marks the sources of variation, and the derivations below use this conditional factorization directly.

Constructed comparisons and scope of analysis. Reference substitution fixes the target image, label, and chromosome type. Cross-patient reference substitution (P1) compares a normal reference from the same patient with a normal reference from another patient. Cross-state reference substitution (P2) fixes a normal target and compares the normal and abnormal homologs of another patient as references. Normal supervision and calibration support learning patient-related normal variation, while P1/P2 constrain dependence on reference selection. Cross-patient substitution may

change both reference biology and image quality; it provides a patient-motivated reference comparison while leaving the target background unchanged. Q constructs quality-perturbed views of training pairs containing normal or abnormal targets by perturbing the target, reference, or both members. Joint perturbations use shared or independent parameters; noise realizations remain independently sampled at the pixel level in the shared-parameter mode. Motivated by the imaging mechanism $X = f_X(M, Q, U_X)$, these transformations provide proxy comparisons for acquisition variation while preserving structural state, subject to the label- and evidence-preservation condition below. S applies sham processing to the target, reference, or both members of real normal–normal pairs. These operations require preservation of the target label and the structural evidence needed for recognition. The structural operator T_c transforms a normal source into an artificial inversion with configuration c, followed by quality filtering based on inversion-evidence visibility. Instances with breakpoints matched to real abnormalities that satisfy the alignment criteria are additionally used for real–artificial alignment.

For a comparison $j \in \{P1, P2, Q, S, \mathrm{RA}\}$ with a shared target label, write $(V, V', a) \sim \Gamma_j$. Here, V and V' are complete model inputs, a is the target label shared by V and V', rather than a label shared by the target and reference members, and Γ_j is the analysis distribution normalized by the corresponding comparison weights. A complete input includes the target and reference images, the statistical group, and the training-time target-patient index required for normal-statistics calibration. The comparison construction and constraint are specified separately for each module. Binary cross-entropy is

$$\ell(s, a) = -a \log s - (1 - a) \log(1 - s). \tag{A.4}$$

The average classification losses before and after an operation are

$$R = \mathbb{E}_{\Gamma_j}[\ell(s(V), a)], \qquad R' = \mathbb{E}_{\Gamma_j}[\ell(s(V'), a)].$$

These quantities characterize changes in classification loss induced by constructed operations. The weights and coefficients in the actual training objectives follow the definitions of the respective modules.

These risks concern constructed comparisons. To analyze cross-domain generalization, we also consider the average classification loss over an entire environment. Since the predictor uses both a target and a reference, this analysis must account for the distribution of complete paired inputs, including target images, reference selection, pairing, and the statistical information used for normal-statistics calibration.

For a fixed chromosome type k, let $\mathsf{P}_e(dV \mid A_T = a, k)$ be the distribution of complete inputs with target label a in environment e. Define

$$R_e = \sum_{a=0}^{1} \pi_a \int \ell(s(V), a)\, \mathsf{P}_e(dV \mid A_T = a, k),$$

$$\pi_a \geq 0, \qquad \sum_{a=0}^{1} \pi_a = 1.$$

The weights π_0 and π_1 correspond to the normal and abnormal classes and remain fixed across environments. Thus, R_e is a weighted sum of the two class-conditional average losses. Results across chromosome types can be aggregated using specified chromosome-type weights.

Comparison constraints reduce prediction fluctuations under specified changes. Their effect in unseen environments depends on whether the changes constructed during training cover or approximate the relevant changes in those environments. For example, transferring cross-patient reference constraints to new patients requires relating the feature changes induced by those patients to the changes represented in training comparisons. The results below connect stability on constructed comparisons to risk changes in unseen environments through explicit coverage conditions.

All scales used for normal-statistics calibration and global standardization are positive. Conditions on finite moments, probability ranges, location–scale models, and generative coverage are stated in the corresponding results.

A.2 Derivations and Proofs of the Main Results

Using the conditional-kernel factorization in Eq. (A.3), we fix chromosome type k and marginalize the structural and background variables to obtain

$$p_e(x \mid a, k) = \int K_X(x \mid m, q, k) K_M(m \mid a, b, p, k) \cdot p_e(p, q, b \mid a, k)\, \mathrm{d}(m, p, q, b). \tag{A.5}$$

Integrals over discrete variables are replaced by sums. Environments change the mixture weights of patient background, acquisition conditions, and configurations, so class-conditional image distributions can vary even when the conditional mechanisms remain stable. Our method modifies learning through normal-reference coordinates, artificial configurations, and controlled pairing. The following results characterize their respective effects.

A.2.1 Structural Construction and Configuration Transfer

Configuration-group supervision. To expose weaknesses that an overall average loss can obscure, we partition artificial inversions into at most eight geometric groups according to whether the interval spans the centromere, interval length, and interval-center position, and compute a classification loss for each group. At each step, we sample at most two available groups containing high-visibility instances and balance patients within each group. High visibility means that the band-reversal difference or expected centromere displacement meets predefined visibility criteria, selecting artificial inversions with clear structural evidence. Group difficulty is measured by the current average classification loss, with greater weight assigned to the group with the larger loss rather than a predefined difficulty ordering.

For an artificial inversion T_i and a normal reference N_i, define the bidirectional supervised risk

$$b_i = \frac{1}{2}\left\{\ell(s(T_i \mid N_i), 1) + \ell(s(N_i \mid T_i), 0)\right\}.$$

This supervision requires recognition of both the artificial inversion and the normal member when the artificial inversion serves as its reference. When two groups are available, let R_1, R_2 be their sampled mean losses. The configuration loss is

$$\begin{aligned} \mathcal{L}_{B,\mathrm{vis}} &= 0.1\left[0.8\frac{R_1 + R_2}{2} + 0.2\max(R_1, R_2)\right] \\ &= 0.1 \max_{u \in [0.4, 0.6]} \{uR_1 + (1-u)R_2\}. \end{aligned} \tag{A.6}$$

The mean term maintains supervision for both groups, while the maximum term emphasizes the currently harder group. The outer coefficient 0.1 controls the overall strength.

Equation (A.6) decomposes into an average classification term $0.1(R_1 + R_2)/2$ and an additional configuration term $0.02\mathcal{L}_B$, where $\mathcal{L}_B = \max(R_1, R_2) - (R_1 + R_2)/2$. The former is included in $\mathcal{L}_{\mathrm{cls}}$ in the main text, while the latter emphasizes the currently harder group with weight $\lambda_B = 0.02$.

Proof. The right-hand side of Eq. (A.6) is linear in u. If $R_1 \geq R_2$, its maximum occurs at $u = 0.6$; otherwise, it occurs at $u = 0.4$. Thus, the loss assigns 60% of the weight to the harder group and 40% to the other group. It encourages correct recognition across configurations without requiring identical prediction probabilities.

When only one group is available, the term becomes $0.1R_1$. This robust weighting applies to the two geometric groups sampled at the current step.

Real–artificial alignment. To reduce dependence on differences between real and artificial image sources, we compare real and artificial pairs with matched breakpoints and the same normal reference. The two prediction directions use the abnormal and normal members as targets, respectively, with abnormality probabilities $p_{i,d}^{\mathrm{real}}, p_{i,d}^{\mathrm{art}}$, where $d \in \{1, 2\}$. Let $G_i \in \{0, 1\}$ indicate that an instance satisfies the requirements for high visibility, complete breakpoint metadata, and centromere confidence. Define

$$\mathcal{L}_{\mathrm{RA}} = \frac{1}{n}\sum_{i=1}^{n} G_i \frac{1}{2}\sum_{d=1}^{2}\left[(|p_{i,d}^{\mathrm{art}} - p_{i,d}^{\mathrm{real}}| - 0.1)_+\right]^2. \tag{A.7}$$

Here, $(u)_+ = \max(u, 0)$, and n is the number of sampled instances before filtering. Ineligible instances are excluded through $G_i = 0$, while the denominator remains n. The term penalizes only probability differences beyond a tolerance of 0.1. Gradients propagate through both the real and artificial branches, while classification supervision supplies the signal for correct predictions. This term enters the Stage I objective with weight $\lambda_{\mathrm{RA}} = 0.01$.

Under the normalized analysis distribution of eligible comparisons, let

$$J_\varepsilon = \mathbb{E}\left[(|p' - p| - \varepsilon)_+^2\right].$$

If $p, p' \in [\delta, 1-\delta]$ for $0 < \delta < 1/2$, then

$$\mathbb{E}|p' - p| \le \varepsilon + \sqrt{J_\varepsilon}, \qquad |R' - R| \le \frac{\varepsilon + \sqrt{J_\varepsilon}}{\delta}. \tag{A.8}$$

For a fixed distribution of eligible comparisons, smaller J_ε gives tighter bounds on prediction and average classification-loss differences. The quantity J_ε excludes the retention rate and the outer training coefficient. The first inequality follows from $|p' - p| \le \varepsilon + (|p' - p| - \varepsilon)_+$ and the Cauchy–Schwarz inequality. The second uses the $1/\delta$-Lipschitz continuity of cross-entropy on this interval.

Suppressing sham-induced false positives. Sham processing preserves band order while introducing processing artifacts, discouraging their use as inversion evidence. For a real normal–normal pair, let p_i and $p_{S,i}$ be the target abnormality probabilities before and after processing. The constraint is applied only to inputs satisfying $p_i \le 0.5$, whose index set is denoted by $\mathcal{I}_S$:

$$\mathcal{L}_S = \sum_{i \in \mathcal{I}_S} \omega_i \left[(p_{S,i} - \mathrm{sg}(p_i) - 0.1)_+\right]^2. \tag{A.9}$$

Here, ω_i denotes a predetermined sample weight, and sg denotes stop-gradient. This term enters the Stage I objective with weight $\lambda_S = 0.01$. The clean prediction is produced by the current model and acts as a fixed comparison value in this loss. The constraint applies to clean inputs correctly classified as normal. It allows small probability fluctuations after processing and penalizes only increases beyond the tolerance, thereby reducing artifact-induced increases in abnormality scores.

Under the distribution of these eligible comparisons, define $J_S = \mathbb{E}[(p_S - p - \varepsilon)_+^2]$. For $\gamma > 0$ and a threshold τ,

$$\Pr(p \le \tau,\ p_S > \tau) \le \Pr(\tau - \varepsilon - \gamma < p \le \tau) + \frac{J_S}{\gamma^2}. \tag{A.10}$$

Proof. If the clean score lies at least $\varepsilon + \gamma$ below the threshold, a new false positive requires $(p_S - p - \varepsilon)_+ > \gamma$. Applying Markov's inequality to its square gives the result. The bound separates the allowed tolerance, samples near the decision boundary, and substantial processing-induced false positives.

Structural coverage and generation error. The preceding supervision and comparison constraints act on constructed training samples. Their benefit for unseen configurations additionally depends on recognition performance over generated configurations, configuration coverage, and the discrepancy between actual generation and ideal structural interventions.

The actual operator $T_c(x_N; \xi)$ constructs an artificial inversion from skeleton interval c, where ξ represents reconstruction randomness. As a reference, let v fix patient background, acquisition conditions, and the remaining exogenous factors. The ideal structural-intervention image for configuration c is

$$X_c^{\mathrm{int}}(v) = f_X\left(f_M(1, c, P(v), U_M(v)), Q(v), U_X(v)\right). \tag{A.11}$$

This expression sets inversion status to 1 and configuration to c while holding the remaining conditions fixed. We consider configurations that are valid for the source and retain recognizable structural evidence.

Fix the chromosome type and complete predictor. Let P_*^a, $\mathsf{P}_{\mathrm{int}}^a$, and $\mathsf{P}_{\mathrm{gen}}^a$ denote the class-conditional distributions of the target evaluation, ideal structural design, and actual generation design in the same feature space. If these distributions have finite first moments and risks use the same class weights

π_a, then

$$R_* \le R_{\text{gen}} + \|w\|_2 \sum_{a=0}^{1} \pi_a \big[W_1(\mathsf{P}^a_*, \mathsf{P}^a_{\text{int}}) + W_1(\mathsf{P}^a_{\text{int}}, \mathsf{P}^a_{\text{gen}}) \big]. \tag{A.12}$$

Here, W_1 is the first Wasserstein distance. The first distance measures the coverage gap between the ideal generation design and the target evaluation distribution, including differences in configuration, source patients, and reference design. The second measures the approximation error of actual generation relative to the ideal intervention.

Proof. For a fixed label, cross-entropy is Lipschitz in the features with constant at most $\|w\|_2$. The risk difference between two distributions is therefore bounded by their Wasserstein distance multiplied by $\|w\|_2$ (Villani et al., 2009). Comparing the target, ideal, and generated distributions in sequence and weighting over classes yields Eq. (A.12). RA constrains output differences for matched instances, whereas the generation error here concerns entire feature distributions.

To isolate the role of configuration coverage, fix the source distribution, reference protocol, and statistical context. Let $\mathsf{P}^{\text{int}}_c$ denote the ideal abnormal-feature distribution for configuration c. Let R^{int}_c and $R^{\text{gen}}_{c_j}$ denote the average cross-entropy risks for label 1 under the corresponding ideal and generated distributions. Assume that, within a valid subset of configurations,

$$W_1(\mathsf{P}^{\text{int}}_c, \mathsf{P}^{\text{int}}_{c'}) \le L_B d_{\mathcal{C}}(c, c'),$$

where $d_{\mathcal{C}}$ measures distance between configurations. If every target configuration lies within distance r_B of some generated configuration c_j, and

$$\max_j W_1(\mathsf{P}^{\text{int}}_{c_j}, \mathsf{P}^{\text{gen}}_{c_j}) \le \epsilon_{\text{gen}},$$

then the cross-entropy risk for the corresponding abnormal targets satisfies

$$\sup_c R^{\text{int}}_c \le \max_j R^{\text{gen}}_{c_j} + \|w\|_2 (L_B r_B + \epsilon_{\text{gen}}). \tag{A.13}$$

Proof. For each target configuration c, choose a generated configuration c_j within distance r_B. First compare their ideal distributions, then compare the ideal and generated distributions for c_j. Apply the preceding risk-difference bound and take the supremum.

Equation (A.13) bounds risk on ideal unseen configurations by the maximum risk on generated configurations, configuration coverage error, and generation error. Configuration expansion therefore works together with effective supervision and generation quality control. The geometric-group weighting in Eq. (A.6) emphasizes harder sampled groups and balances supervision over covered configurations. Group-average risk, however, does not directly replace the maximum risk over individual configurations; transfer still depends on the stated coverage and generation-error conditions.

A.2.2 Pairwise Constraints and Environmental Transfer

Stage II fixes the encoder, the statistics used for normal-statistics calibration and global standardization, and the classification-head bias, and updates only the classification-head weights. The logit is therefore $\eta(V) = w^\top F(V) + b_I$. The P/Q constraints adjust w to reduce sensitivity to specified input changes.

Reference-substitution constraints. P1 fixes the target and replaces a normal reference from the same patient with a normal reference from another patient. P2 fixes a normal target and switches between the normal and abnormal members of the same external patient as references. Let $V^{(0)}_{i,j}$ and $V^{(1)}_{i,j}$ denote the inputs before and after substitution. Define

$$\begin{aligned} \Delta F_{i,Pj} &= F(V^{(1)}_{i,j}) - F(V^{(0)}_{i,j}), \\ G_{Pj} &= \sum_{i \in \mathcal{I}_{Pj}} a_i \Delta F_{i,Pj} \Delta F^\top_{i,Pj}, \qquad j \in \{1, 2\}. \end{aligned} \tag{A.14}$$

Here, $a_i \geq 0$ denotes a sample weight, and $\mathcal{I}_{Pj}$ is the set of valid comparisons. It follows that

$$w^\top G_{Pj} w = \sum_{i \in \mathcal{I}_{Pj}} a_i (\eta_i^{(1)} - \eta_i^{(0)})^2.$$

Thus, the matrix form is exactly equivalent to a squared logit-difference penalty, constraining prediction changes for the same target under reference substitution.

Quality-perturbation constraints. Let $\mathcal{I}_Q$ denote the training pairs used for quality perturbations, including normal–normal pairs and pairs containing abnormal members. Using each member as the target gives two ordered inputs V_{i1} and V_{i2}. Let $m \in \mathcal{M}_Q$ index target-only, reference-only, and joint perturbations, with shared or independent parameters for joint perturbations. Let V_{id}^m be the perturbed input, where $d \in \{1, 2\}$ denotes the member direction. Define

$$\begin{aligned} \Delta F_{id,Qm} &= F(V_{id}^m) - F(V_{id}), \\ G_{Qm} &= \sum_{i \in \mathcal{I}_Q} \sum_{d=1}^{2} a_{id} \Delta F_{id,Qm} \Delta F_{id,Qm}^\top. \end{aligned} \tag{A.15}$$

Here, $a_{id} \geq 0$ is the corresponding member weight. Therefore,

$$w^\top G_{Qm} w = \sum_{i \in \mathcal{I}_Q} \sum_{d=1}^{2} a_{id} \left[\eta(V_{id}^m) - \eta(V_{id})\right]^2.$$

Q constrains each member's logit before and after perturbation and aggregates the penalties with weights. In normal–abnormal pairs, each member retains its own target label: the constraint stabilizes each prediction without requiring the two members to receive equal scores.

Branch roles and proof. The pair representation consists of the target branch q and the difference branch $|q - r|$. Reference substitution preserves the target and its calibration context and therefore changes only the difference branch. Quality perturbations can change both target and reference encodings and consequently affect both branches. Fixed elementwise calibration does not mix the branches, so these relations also hold after calibration.

For any constructed comparison, the fixed bias cancels when logits are subtracted:

$$\eta(V') - \eta(V) = w^\top [F(V') - F(V)]. \tag{A.16}$$

Thus, P directly constrains logit changes induced by reference variation through the difference branch, while Q constrains changes induced by quality variation through both branches. Both have the form $w^\top G w$, penalizing the projection of specified feature changes onto the classification direction. Classification supervision simultaneously supplies the signal for distinguishing normal and abnormal targets.

For a same-label, single-target comparison $(V, V', a) \sim \Gamma_j$ constructed by P/Q, denote the logits by η, η' and probabilities by $p = \sigma(\eta), p' = \sigma(\eta')$. Let

$$J_\eta = \mathbb{E}_{\Gamma_j} (\eta' - \eta)^2 < \infty,$$

and assume that at least one of the two cross-entropy risks is finite. Then

$$\mathbb{E}_{\Gamma_j} |p' - p| \leq \frac{1}{4} \sqrt{J_\eta}, \qquad |R' - R| \leq \sqrt{J_\eta}. \tag{A.17}$$

Here, R, R' are average cross-entropy risks before and after the change under the same comparison distribution. The sigmoid and the logit-based cross-entropy $\ell(\sigma(\eta), a)$ have Lipschitz constants $1/4$ and 1, respectively. The result follows from the Cauchy–Schwarz inequality and requires no additional assumption that probabilities remain bounded away from 0 and 1. Controlling logit changes therefore controls changes in probabilities and average classification losses for the constructed reference substitutions and quality perturbations. This applies to both reference-substitution effects under P and memberwise perturbation effects under Q.

Nonnegative sample weights ensure that G is positive semidefinite. The population form additionally requires finite second moments of the corresponding feature changes. Its spectral expansion is

$$w^\top G w = \sum_l \lambda_l (w^\top u_l)^2,$$

where u_l and λ_l are eigenvectors and eigenvalues. The eigenvalue λ_l measures the weighted squared variation of constructed comparisons along u_l: larger variation incurs a stronger penalty on the classification head's sensitivity in that direction. This is consistent with conditional variance regularization, which promotes robustness by constraining prediction variation for the same object under different conditions (Heinze-Deml & Meinshausen, 2021).

Transfer to new variation directions. P/Q constructs a finite set of reference substitutions and quality perturbations during training. Transfer to unseen changes depends on whether those changes lie in feature directions covered by training and have controlled magnitude.

For the positive semidefinite matrix G_j of a module, let $G_j^\dagger$ denote its pseudoinverse and define

$$\mathcal{E}_j(\kappa) = \{\delta \in \mathrm{range}(G_j) : \delta^\top G_j^\dagger \delta \le \kappa^2\}, \qquad \kappa \ge 0.$$

Here, δ is the feature difference induced by an input change, $\mathrm{range}(G_j)$ contains variation directions covered by training comparisons, and κ controls the allowed magnitude relative to training variation. This set is an ellipsoid determined by the training comparisons. For any fixed classification head,

$$\sup_{\delta \in \mathcal{E}_j(\kappa)} |w^\top \delta| = \kappa \sqrt{w^\top G_j w}. \tag{A.18}$$

A smaller pairwise penalty $w^\top G_j w$ therefore bounds the maximum logit change within this covered set.

Proof. For $\delta \in \mathrm{range}(G_j)$,

$$w^\top \delta = (G_j^{1/2} w)^\top \big((G_j^\dagger)^{1/2} \delta\big).$$

The Cauchy–Schwarz inequality gives the upper bound. When $w^\top G_j w > 0$, the choice $\delta = \kappa G_j w / \sqrt{w^\top G_j w}$ attains it. When $w^\top G_j w = 0$, both sides are zero.

New changes may also contain directions absent from training. Let $\Pi_j = G_j G_j^\dagger$ be the orthogonal projection onto the covered directions. If

$$\delta^\top G_j^\dagger \delta \le \kappa^2, \qquad \|(I - \Pi_j)\delta\|_2 \le \epsilon_\perp,$$

then

$$|w^\top \delta| \le \kappa \sqrt{w^\top G_j w} + \epsilon_\perp \|(I - \Pi_j) w\|_2. \tag{A.19}$$

The proof decomposes δ into the covered component $\Pi_j \delta$ and the uncovered component $(I - \Pi_j)\delta$, then applies Eq. (A.18) and the Cauchy–Schwarz inequality, respectively. The first term is controlled by the training pairwise constraint, while the second captures the additional response to uncovered variation.

For a new same-label comparison, let $\delta = F(V') - F(V)$. If these coverage conditions hold almost surely under its distribution and at least one cross-entropy risk is finite, the right-hand side of Eq. (A.19) also bounds the difference in average classification risk. One quarter of this quantity bounds the average probability change. This result applies to both reference substitution under P and memberwise quality perturbations under Q.

Thus, reference substitutions involving unseen patients and unseen quality changes can be controlled when their induced feature changes satisfy the directional and magnitude coverage conditions. For P, this concerns the response of a fixed target to reference substitution; for Q, it concerns each member's response to quality perturbations. The causal motivation guides comparison construction, while the coverage conditions specify the scope of transfer.

A.3 WHY NORMAL-STATISTICS CALIBRATION HELPS

Normal-statistics calibration uses normal data to establish comparison coordinates for target features and homologous differences. The normal mean defines the reference origin, and the normal standard deviation defines the unit of deviation. Figure 6 illustrates these two operations, and Figure 7 shows how the resulting coordinates support a shared classification rule. Throughout this section, $g = (k, t_R)$ denotes a calibration group determined by chromosome number and reference type.

Variation represented by the normal baseline. Let H denote the pair features of a normal input. Within group g, first sample a patient I uniformly, then uniformly sample a valid normal input from that patient. The elementwise law of total variance gives

$$v_g = \mathbb{E}_I[\mathrm{Var}(H \mid I, g)] + \mathrm{Var}_I\big(\mathbb{E}[H \mid I, g]\big). \tag{A.20}$$

Proof. Average the identity

$$\mathbb{E}[H^2 \mid I, g] = \mathrm{Var}(H \mid I, g) + \big(\mathbb{E}[H \mid I, g]\big)^2$$

over patients and subtract the squared overall mean, with all operations applied elementwise. The normal scale therefore captures both within-patient fluctuations and normal differences between patients. Patients with more images do not dominate the baseline solely through sample count. The shrinkage pool likewise averages first and second moments equally across group–patient entries before computing the variance, retaining differences between entry means.

Under the local additive approximation $q = u_T + b_i + \epsilon_T$ and $r = u_R + b_i + \epsilon_R$, we have $q - r = u_T - u_R + \epsilon_T - \epsilon_R$. The shared term b_i cancels only in the difference branch. Member-specific variation remains, and normal statistics provide a scale for the variation observed in normal training pairs.

Figure 6 illustrates one coordinate of the pair feature h. In panel (a), the groups differ in both feature distributions and normal reference statistics. Subtracting μ_g in panel (b) places each group's normal mean at zero: sample coordinates now express deviations from that group's normal baseline. Dividing by σ_g in panel (c) expresses each deviation in units of the group's normal standard deviation. Thus, an inversion example at h^* moves from h^* to $h^* - \mu_g$ and then to $(h^* - \mu_g)/\sigma_g$. This establishes a comparable scale while allowing the full sample distributions to retain different shapes. Abnormality supervision learns which deviations are discriminative on this scale.

Proposition A.3.1 (Location–scale calibration). Assume that normal features in comparable groups satisfy $H_g^0 = a_g + T_g U_0$, where T_g is a positive diagonal scale matrix and U_0 has a shared distribution across groups, zero mean, and unit elementwise variance. For structurally compatible member states a and configurations c, further assume

$$H_g^{a,c} = a_g + T_g U_{a,c} + R_g^{a,c}. \tag{A.21}$$

Here, $U_{a,c}$ is a standardized structural response shared across groups, and $R_g^{a,c}$ is an unexplained residual. Ideal normal calibration gives

$$T_g^{-1}(H_g^{a,c} - a_g) = U_{a,c} + T_g^{-1} R_g^{a,c}. \tag{A.22}$$

Let $S_g = \mathrm{diag}(\sigma_g)$ be the actual shrinkage scale and define $K_g = S_g^{-1} T_g$. With an accurate mean estimate, actual calibration yields

$$Z_g = K_g U_{a,c} + S_g^{-1} R_g^{a,c}. \tag{A.23}$$

Proof. Substitute the structural decomposition into the calibration operator. With ideal location and scale matching, shared structural responses are expressed in common coordinates. Actual alignment depends on $K_g - I$ and the standardized residual. The common-response conclusion requires the stated distributional assumption; matching means and elementwise variances alone does not make complete distributions identical.

Proposition A.3.2 (Conditional classification rules determined by normal statistics). Fix the encoder, statistical snapshot, and target-patient exclusion context. The calibrated linear head is

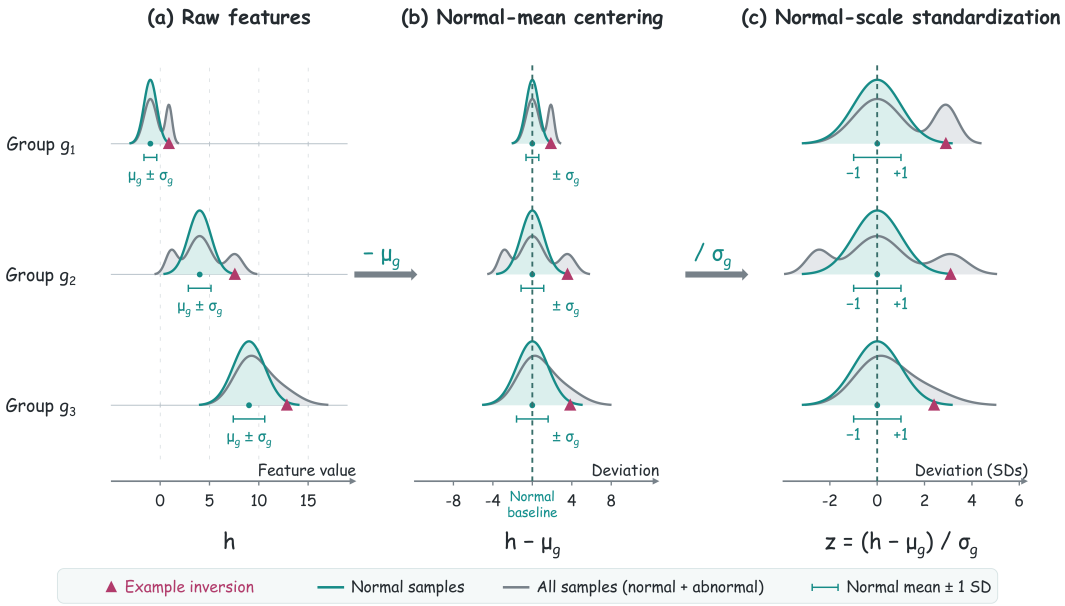


Figure 6: Normal means and scales define comparison coordinates. The schematic shows raw features (a), normal-mean centering (b), and normal-scale standardization (c) for three calibration groups. Teal curves represent normal samples, gray curves represent all samples, and triangles track example inversions. The dashed lines in (b,c) mark the normal reference origin. Brackets show the normal mean $\pm$ one standard deviation under each transformation; they indicate a scale, not a classification boundary. Only training normal inputs determine the calibration statistics. Curves illustrate horizontal coordinate changes rather than estimated probability densities.

equivalent to

$$\begin{aligned} \eta(h, g) &= \beta_g^\top h + b_g, \\ \beta_g &= S_g^{-\top} \Lambda^{-\top} w, \\ b_g &= b - w^\top \Lambda^{-1} m_D - \beta_g^\top \mu_g. \end{aligned} \tag{A.24}$$

Proof. Expand $w^\top \Lambda^{-1}[S_g^{-1}(h - \mu_g) - m_D] + b$ and collect the feature-dependent and constant terms.

The same head therefore shares a decision rule in normal-calibrated coordinates, while inducing group-specific slopes and intercepts in the original feature space. These adjustments are determined by normal statistics rather than fitted separately from scarce abnormal samples in each group.

Figure 7 gives a one-dimensional example. Let the training normal statistics be $(\mu_{g_1}, \sigma_{g_1}) = (0, 1)$ and $(\mu_{g_2}, \sigma_{g_2}) = (10, 3)$. The illustrated normal and inversion samples have feature values $(1, 3)$ in the first group and $(13, 19)$ in the second. A single threshold cannot separate all four samples in the raw coordinates. Centering alone gives $(1, 3)$ and $(3, 9)$, leaving a normal and an inversion sample at the same value. Dividing by the respective normal standard deviations gives $(1, 3)$ in both groups, which a shared threshold at $z = 2$ separates. The normal sample markers are individual observations; the means and scales come from the training normal inputs.

Given the complete statistical context and positive scales, the transformation is invertible and adds no new label information. Its role is to provide useful comparison coordinates for a shared classifier with limited capacity. Normal data determine the group-specific coordinate adjustments, while labeled normal and abnormal inputs train the shared decision rule.

Proposition A.3.3 (Statistical error and refresh). Let $C(h) = (h - \mu) \oslash \sigma$, and let $\widehat{C}$ use estimated statistics $\widehat{\mu}, \widehat{\sigma}$. If both scales are bounded below by $s_{\min} > 0$ and

$$\|\widehat{\mu} - \mu\|_\infty \le e_\mu, \qquad \|\widehat{\sigma} - \sigma\|_\infty \le e_\sigma,$$

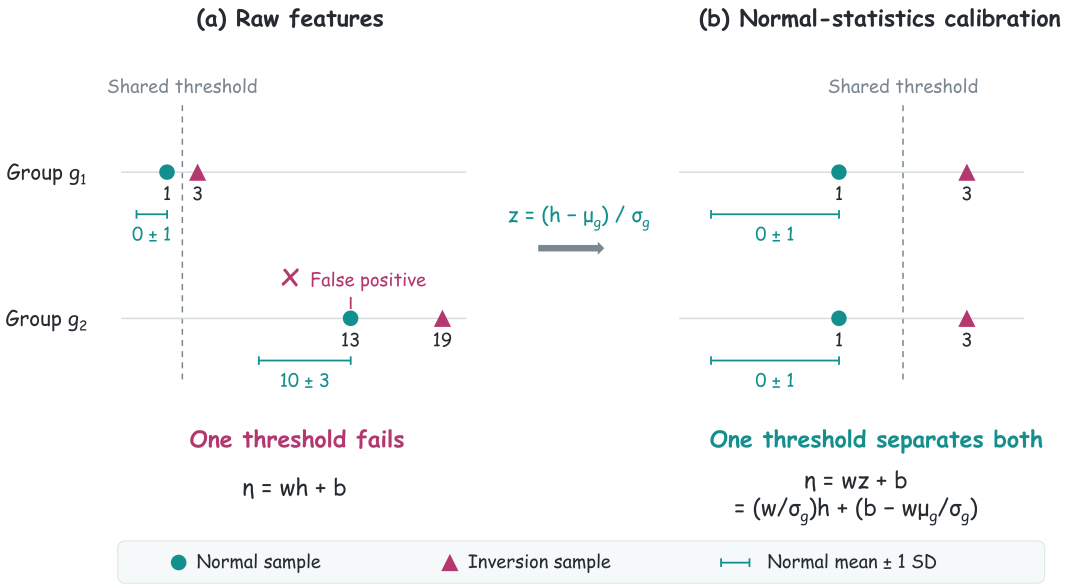


Figure 7: Normal-statistics calibration supports a shared decision rule. In this illustrative example, a threshold suitable for group g_1 misclassifies the normal sample in group g_2 in the raw coordinates (a). Group-specific centering and scaling place the shown normal and inversion samples at $z = 1$ and $z = 3$, respectively, allowing a shared threshold (b). The expanded scalar head makes the induced slope and intercept explicit. For this illustration, fixed global standardization is absorbed into the head coefficients w, b; Eq. (A.24) gives the full form.

then

$$\|\widehat{C}(h) - C(h)\|_\infty \le \frac{e_\mu}{s_{\min}} + \frac{e_\sigma}{s_{\min}^2}\|h - \mu\|_\infty. \tag{A.25}$$

Proof. Apply the elementwise identity

$$\frac{h - \widehat{\mu}}{\widehat{\sigma}} - \frac{h - \mu}{\sigma} = \frac{\mu - \widehat{\mu}}{\widehat{\sigma}} + (h - \mu)\left(\frac{1}{\widehat{\sigma}} - \frac{1}{\sigma}\right)$$

and bound the two terms separately. With the same global standardization and classification head, Hölder's inequality and the $1/4$-Lipschitz continuity of the sigmoid further yield

$$|\widehat{s} - s| \le \frac{1}{4}\|\Lambda^{-\top} w\|_1 \left[\frac{e_\mu}{s_{\min}} + \frac{e_\sigma}{s_{\min}^2}\|h - \mu\|_\infty\right]. \tag{A.26}$$

This bound shows how errors in normal statistics propagate to predictions, while a positive scale floor limits error amplification in small-scale coordinates. Dynamic refresh updates both normal and global statistics. The two transformations can be combined as

$$\begin{aligned} m_{\text{eff}} &= \mu_g + \sigma_g \odot m_D, \\ t_{\text{eff}} &= \sigma_g \odot s_D, \\ F &= (h - m_{\text{eff}}) \oslash t_{\text{eff}}. \end{aligned} \tag{A.27}$$

When the effective scales have a positive lower bound, applying the same error bound to the combined statistics from old and refreshed snapshots characterizes the coordinate difference for the same current feature. Refreshing after each epoch realigns statistics with the current encoder. Once the encoder and statistics are fixed in Stage II, encoder updates no longer cause statistical staleness.

Connection to unseen configurations. Normal samples establish the reference origin and scale illustrated in Figure 6, while real abnormalities and artificial inversions teach abnormality-discriminative directions in these coordinates. Normal-statistics calibration and configuration expansion address background scale differences and limited configuration coverage, respectively, and are complementary. Since the encoder also receives abnormality supervision, the learned representation remains dependent on training configurations.

Table B.1: Test-domain definitions. Checkmarks indicate satisfied conditions. Acquisition variation refers to natural differences in new patients.

Condition	Cross-Environment Generalization	Unseen-Configuration Generalization
Patients absent from real training and validation (P)	✓	✓
Natural acquisition variation retained (Q)	✓	✓
Test configurations present in real training (B)	✓	–
Test configurations absent from real training and validation (B)	–	✓
Evaluation focus	Environmental robustness	Configuration generalization

B Data Splits and Evaluation Protocol

B.1 Data Composition, Annotations, and Curation

Skeleton and centromere annotations. This annotated subset of normal-case data contains 4,006 grayscale images of 300×300 pixels covering all 24 chromosome classes. Skeletons are stored as 256×2 NumPy arrays of ordered (x, y) coordinates, with endpoints defined by the first and last points. Centromere JSON annotations record normalized skeleton arc length, skeleton index, and two-dimensional coordinates. Each image has an aligned pixel-level foreground mask; both are stored as PNG files.

Normal-case pretraining data. Images from 105 normal cases capture morphological, banding, and imaging variation across patients and cells, supporting pretraining tasks including normal structural distribution learning and abnormality generation. We additionally collect 133,484 single-chromosome images from five public sources: ChromosomeNet (Lin et al., 2022), CIR-Net (Lin et al., 2020), BioImLab (Grisan et al., 2009), 3CNet (Lin et al., 2021), and CHROMA (Yang et al., 2026). This supplementary pool consists predominantly of samples with normal morphology and retains source information and available chromosome labels. Differences in band appearance, staining, and imaging quality broaden the variation observed during pretraining and support representation learning across sources.

Public-data curation includes standardizing grayscale formats, file names, and chromosome labels, and compiling source and label manifests. We locally remove nonbiological residues, such as extraction lines clearly separated from the chromosome body, while retaining original images and processing records for traceability. Public data are used for pretraining; patient-disjoint generalization evaluation uses clinical data with patient-level identifiers and abnormality annotations.

Real abnormality shift data. Images from 84 inversion cases cover 25 configurations across all 24 chromosome classes, including 12 paracentric and 13 pericentric inversions. Multiple cells per patient and configurations shared across patients support evaluation under patient, acquisition, and configuration shifts. Table 1 lists configurations using chromosome:breakpoint notation, e.g., 3:q13.2q26.2 denotes inv(3)(q13.2q26.2).

B.2 Patient Splits and Test Domains

Section 3 of the main text describes the dataset composition and annotations. Here, we detail the splits, input construction, and evaluation protocol.

We use five-fold patient-disjoint evaluation. Within each fold, all karyotypes and single-chromosome images from the same patient belong to the same subset. Each patient appears in the test set exactly once across the five folds. Both test domains evaluate new patients; configuration overlap distinguishes cross-patient generalization from the additional challenge of configuration generalization.

Cross-environment test domain. Test patients are excluded from training, but their inversion configurations are represented by other patients in the training set of that fold. This setting retains

naturally occurring preparation and imaging differences in new patients and evaluates combined robustness to patient background and accompanying acquisition variation, corresponding to the P/Q generalization problem.

Unseen-configuration test domain. In addition to patient separation, the inversion configurations of test patients are absent from the real training and validation sets of the corresponding fold. A configuration is defined by chromosome type and the inversion breakpoint combination. The training set may contain normal chromosomes of the same type or other inversion configurations. This setting adds the generalization challenge associated with B and evaluates recognition of unseen real inversion configurations.

Each fold uses the same trained model for both test domains. We select the checkpoint with the highest validation AUROC and use a fixed classification threshold of 0.5 across all folds and both test domains. The domains contain different configuration sets; their results characterize performance under different generalization settings rather than estimate independent effects of P, Q, and B.

B.3 Input Construction and Use of Statistics

Base pairs are constructed by chromosome type within the same patient and karyotype. Each member serves as the target in one direction and as the reference in the other, with the label always indicating whether the target has an inversion. Swapping the direction of a normal–abnormal homologous pair changes the target label accordingly. Male X and Y chromosomes use their respective standard ideograms as references and are not paired with each other as homologs. Standard reference ideograms and those in Figures 1 and 4 were drawn using UCSC hg38 cytoband data (Kent et al., 2002).

Real source images and reference candidates for cross-patient substitution, artificial inversion construction, and quality perturbation are restricted to the current training fold. Validation and testing retain the predefined real data and base pairings, without these training-time constructions. Unseen configurations are defined with respect to real patient data. Counterfactual configurations are sampled using training data without guidance from test patients or test breakpoints; overlap arising from this sampling is not excluded.

The statistics used for normal-statistics calibration are estimated only from valid normal inputs in the current training fold. Training excludes the target patient's contribution to normal statistics, while validation and testing use statistics from all normal training inputs in that fold. Subsequent global standardization parameters are also estimated solely from training data. Statistics are refreshed as the encoder changes in Stage I and fixed during Stage II and evaluation; validation and test images are not used for re-estimation. At inference, the model receives the target and reference images and selects the corresponding calibration statistics using the known chromosome type and reference type, without requiring abnormality labels or inversion breakpoints.

B.4 Evaluation Unit and Metrics

Inversion is the positive class. Evaluation is performed at the level of individual real chromosomes using the abnormality score obtained when each chromosome serves as the target. Standard ideograms provide references only and are not treated as real evaluation members. All methods are compared using the same evaluation unit and test-sample scope.

We evaluate detection performance using AUROC, AUPRC, sensitivity, specificity, and F1, with AUPRC computed as average precision. Metrics are calculated separately for each test domain and fold. We report the mean and sample standard deviation across five folds to characterize performance and variability across patient splits.

C Skeleton and Centromere Estimation

C.1 Model Architecture and Training

We use a basic ResUNet as a multi-task prediction network to jointly estimate the skeleton, two endpoints, and centromere from a single grayscale chromosome image. Inputs have a resolution of

300×300 and undergo grayscale normalization. The encoder extracts multiscale features, and the decoder restores spatial resolution through upsampling and skip connections. Separate prediction heads produce a skeleton heatmap, two endpoint heatmaps, and a centromere heatmap at the input resolution.

During training, an auxiliary foreground-mask head uses the dataset masks to provide shape supervision for shared features. This head is used only for training: masks are neither model inputs nor part of centerline decoding at inference, so foreground extraction is not required before deployment.

Skeleton supervision is obtained by rasterizing and smoothing the manually annotated ordered curves. Endpoint and centromere supervision use two-dimensional Gaussian heatmaps centered at the annotated locations. Each prediction task combines focal binary cross-entropy (Lin et al., 2017) and soft Dice losses (Milletari et al., 2016). The total objective is

$$\mathcal{L} = \mathcal{L}_s + \lambda_e(\mathcal{L}_{e_0} + \mathcal{L}_{e_1}) + \lambda_c\mathcal{L}_c + \lambda_m\mathcal{L}_m, \tag{C.1}$$

where $\mathcal{L}_s$, $\mathcal{L}_{e_0}$, $\mathcal{L}_{e_1}$, $\mathcal{L}_c$, and $\mathcal{L}_m$ supervise the skeleton, two endpoints, centromere, and mask, respectively, and $\lambda_e, \lambda_c, \lambda_m$ are their weights.

The model is trained using 4,006 manually reviewed annotations with AdamW, an initial learning rate of 2×10^{-4}, a batch size of 16, and 200 epochs. Exponential moving average parameters are used for inference. Training augmentations include rotation, scaling, translation, horizontal flipping, and grayscale perturbations. Geometric transformations are applied consistently to images, masks, and annotation coordinates. Detailed parameters and implementation are provided in the public code.

C.2 Centerline Decoding and Landmark Localization

We convert the predicted heatmaps into an ordered centerline and its landmarks. Peaks are first extracted from the two endpoint heatmaps and refined using local skeleton responses. Skeleton probabilities are then converted into negative log costs, and a minimum-cost path is found between the endpoints, favoring regions with high predicted skeleton probability. Decoding operates directly on the skeleton heatmap and requires no additional foreground segmentation or morphological thinning.

The extracted path is smoothed and resampled at equal arc-length intervals into 256 ordered coordinates. Its orientation follows the order of the two endpoint prediction heads. The peak of the centromere heatmap is projected onto the nearest centerline point to obtain the centromere coordinate and its normalized longitudinal position. The final output contains the centerline, two endpoints, and centromere, providing a common spatial reference for subsequent band sampling and local structural operations.

C.3 Qualitative Comparison of Centerline Extraction

We select eight conventional baselines spanning seven mechanisms: principal-axis estimation (Pearson, 1901), cross-sectional centers (Groen et al., 1989; Kao et al., 2008), morphological thinning (Zhang & Suen, 1984; Guo & Hall, 1989; Wang et al., 2008; 2009), distance-transform medial axes (Hesselink & Roerdink, 2008; Lerner et al., 1995), minimum-cost paths (Jin et al., 2016; Metz et al., 2008), computational geometry (Ogniewicz & Ilg, 1992), and skeleton pruning (Bai et al., 2007; Arachchige et al., 2010; Shen et al., 2011). They comprise a PCA principal-axis baseline, the four-directional cross-sectional center method of Kao et al. (2008), the improved thinning procedure of Wang et al. (2009), a distance-transform medial axis, a distance-weighted minimum-cost path, a contour-based Voronoi skeleton, and DCE and BPR pruning. The distance-transform baseline uses the medial-axis implementation in scikit-image. The cross-sectional and Wang-style thinning methods are adapted from their paper descriptions, while DCE and BPR use the authors' released core code.

All conventional baselines use the same foreground masks obtained by thresholding, selecting the largest connected component, and filling internal holes. Their outputs are converted into a single ordered centerline. Results are overlaid on the same original images with consistent display scales and line styles to compare global orientation, local trajectories, and endpoint coverage.

Figure 8 shows three selected examples: an elongated chromosome, a small compact chromosome, and a markedly curved chromosome. For the elongated example, most methods recover the overall longitudinal direction, with differences mainly in endpoint coverage and local curvature. For the small compact example, the cross-sectional method provides incomplete coverage, and several conventional methods produce local detours; our model yields a more complete longitudinal path with fewer detours. For the curved example, the PCA line cannot follow the chromosome shape. Most nonlinear methods recover the main curvature, although some exhibit local oscillations or endpoint deviations. Our model maintains a coherent centerline in these examples and jointly localizes the endpoints and centromere.

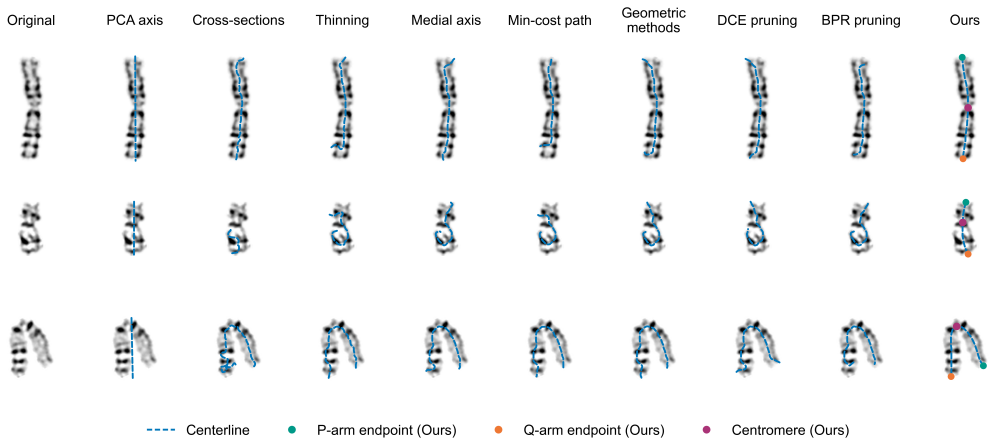


Figure 8: Qualitative comparison of chromosome centerline extraction. The three rows show an elongated, a small compact, and a markedly curved chromosome, respectively. Blue dashed lines indicate centerlines. For our method, green and orange markers indicate the two predicted endpoints, and magenta markers indicate the predicted centromere.

D Implementation Details

D.1 Network and Inputs

We use ViT-L/16 as the shared feature encoder. The encoder first learns basic representations of chromosome morphology and band structure from normal chromosome pretraining data, and the pretrained weights initialize the downstream inversion detection model. Grayscale images are converted to three channels, resized to 224×224 using bicubic interpolation, and normalized using the ImageNet mean and standard deviation (Deng et al., 2009). The normalized CLS token provides a 1,024-dimensional image feature. The target feature and the absolute difference feature form a 2,048-dimensional pair representation, which undergoes normal-statistics calibration and global standardization before being passed to a linear classification head. The skeleton–centromere model remains frozen during downstream training.

D.2 Compared Methods and Pretraining

Implementation of compared methods. We compare traditional classifiers (LR, SVM, and RF), general-purpose deep networks (ResNet50, DenseNet121, and Swin-T), and chromosome analysis methods (ChroSiameseAD, BR-ChromNet, HomNet, DCF, and CHROMA). General-purpose networks classify individual chromosome images. Compared methods retain the encoders and core architectures used in their original papers' main experiments. All deep methods, including general-purpose networks and chromosome analysis methods, undergo pretraining on the same normal chromosome data within their respective architectures. Methods using the same encoder architecture share the same pretrained checkpoint. Each method is then adapted to abnormality detection using the corresponding fold's real training data. Procedures requiring skeleton information use the same

skeleton extraction model. The main comparison evaluates complete methods under a common evaluation protocol, whereas component ablations within our framework fix the encoder architecture and pretrained initialization.

Normal chromosome pretraining comparison. To isolate the effect of pretraining, Base model and Base + pretraining both use a ViT-L/16 encoder with single-image input and a linear binary classification head, following the same downstream training procedure. Base model starts from random initialization, whereas Base + pretraining initializes its encoder using the complete pretraining dataset, comprising images from 105 normal patients and the supplementary public data described in Appendix B.1, and uses a newly initialized classification head. Pretraining reconstructs original grayscale pixels in masked regions using a weighted mean squared error over those regions, with greater weight assigned to chromosome foreground pixels. Both models fine-tune the encoder and classification head without homologous references, normal-statistics calibration, artificial abnormalities, or additional constraints. We find that normal-image pretraining alone partially alleviates performance degradation on unseen configurations, but recognition remains limited and substantially below the full framework.

D.3 Two-Stage Optimization

Stage I jointly optimizes the encoder and classification head using AdamW for up to 10 epochs. Normal-statistics calibration and global standardization statistics are refreshed at the end of each epoch, and the model is selected based on validation performance. Stage II fixes the encoder, statistical parameters, and classification-head bias, and optimizes only the head weights using L-BFGS (Liu & Nocedal, 1989) for up to 100 iterations. The quality perturbation constraint uses memberwise consistency.

D.4 Hyperparameter Settings

All folds use the same hyperparameter settings, and model selection uses only the validation set. Table D.1 summarizes the main settings.

Table D.1: Main hyperparameter settings for the downstream detection model.

Parameter	Setting
Encoder	ViT-L/16
Input resolution	224×224
Image / pair feature dimensions	1,024 / 2,048
Stage I optimizer	AdamW
Encoder / head learning rates	10^{-6} / 10^{-5}
Weight decay	0
Main training batch size	24
Maximum training epochs	10
Gradient norm clipping threshold	1.0
Stage II optimizer	L-BFGS
Maximum Stage II iterations	100
λ_B	0.02
λ_{RA}, λ_S	0.01 each
$\lambda_{P1}, \lambda_{P2}, \lambda_Q$	0.01 each
Sham-processing probability tolerance	0.1

D.5 Structural Visibility and Geometric Grouping

Structural visibility. We extract a 128-bin band-intensity profile along the skeleton and apply Gaussian smoothing with a standard deviation of one bin. Within the inversion interval, we compute the correlation change $C = 1 - \mathrm{corr}(s, s^{\mathrm{rev}})$, normalized mean absolute difference D, and number of changed bins K. Absolute differences are normalized by $r = \max(P_{95} - P_5, 0.05)$, where the percentiles are computed over the full smoothed profile. A bin is counted as changed when its absolute

difference is at least $\max(0.08r, 0.01)$. For normalized breakpoints $u_1 < u_2$ and centromere position c, the centromere displacement is $\Delta_c = |u_1 + u_2 - 2c|$ when $u_1 < c < u_2$, and zero otherwise. A sample has high visibility if $(C \geq 0.20,\ D \geq 0.08,\ K \geq 4)$ or $\Delta_c \geq 0.03$.

Among the remaining samples, those satisfying $\Delta_c < 0.01$ and either $K < 2$ or $(C < 0.08,\ D < 0.03)$ are classified as having insufficient visible evidence; all others have low visibility. Low-visibility samples receive reduced classification weights, while samples with insufficient evidence are excluded from synthetic classification supervision. Configuration weighting uses high-visibility samples. Real–artificial alignment additionally requires available breakpoint-band information and a centromere localization confidence of at least 0.05.

Geometric grouping. Generated configurations are divided into eight groups according to whether the inversion spans the centromere, whether its normalized length $u_2 - u_1$ is at least 0.4, and whether its midpoint $(u_1+u_2)/2$ is at least 0.5. Each step samples up to two groups containing high-visibility samples, with patient-balanced sampling within each group.

E Additional Ablation Studies

We examine backbone choice, normal-image pretraining, normal-statistics calibration, pair representations, training strategies, and artificial abnormality generation through ablation studies, and further analyze cross-configuration generalization across chromosome types. For the ablations, we follow the evaluation protocol in Appendix B and report AUROC and AUPRC as mean $\pm$ sample standard deviation over five folds in the cross-environment and unseen-configuration test domains. All metric values are percentages.

E.1 Backbone Models and Normal-Chromosome Pretraining

We compare the basic classification performance of ResNet-152, DenseNet-201, and ViT-L/16 to examine normal-image pretraining and encoder choice. All models take individual chromosome images as input and use the same patient-disjoint five-fold splits and two test domains, without homologous references, normal-statistics calibration, artificial abnormalities, or additional constraints. Methods in the main comparison retain the encoders specified in their original papers: methods using the same encoder architecture and configuration share the corresponding pretrained weights, while specialized encoders are pretrained on the same normal chromosome data.

Table E.1: Basic classification performance across backbones with and without normal-image pretraining. Results are five-fold mean $\pm$ standard deviation, in percentages.

		Cross-Environment Generalization		Unseen-Configuration Generalization	
Backbone	Pretraining	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
ResNet-152	No	70.87 ± 18.01	25.12 ± 20.59	46.36 ± 10.95	2.20 ± 0.51
ResNet-152	Yes	82.57 ± 9.08	41.46 ± 14.70	61.22 ± 8.54	5.17 ± 3.52
DenseNet-201	No	76.95 ± 17.79	26.73 ± 20.45	53.12 ± 7.27	4.75 ± 3.76
DenseNet-201	Yes	81.25 ± 16.41	40.17 ± 31.81	60.07 ± 10.89	3.60 ± 1.35
ViT-L/16	No	54.67 ± 11.16	2.76 ± 0.77	53.78 ± 9.11	2.57 ± 0.57
ViT-L/16	Yes	76.11 ± 11.87	21.15 ± 13.03	64.73 ± 13.39	6.81 ± 4.87

As shown in Table E.1, the pretrained variants of all three backbones achieve higher mean AUROC in both test domains and higher mean AUPRC in the cross-environment domain. On unseen configurations, AUROC increases by 14.86, 6.95, and 10.95 percentage points for ResNet-152, DenseNet-201, and ViT-L/16, respectively, while AUPRC increases by 2.97 and 4.24 points for ResNet-152 and ViT-L/16 but decreases by 1.15 points for DenseNet-201. These results show a consistent improvement in mean AUROC across backbones, with effects varying by architecture and metric. All three pretrained classifiers remain well below the full framework's 90.83% AUROC and 46.77% AUPRC, indicating that normal-image pretraining provides a useful representation foundation but remains insufficient for unseen-configuration generalization.

The pretrained backbones perform differently across domains: ResNet-152 achieves the highest mean AUROC and AUPRC under cross-environment evaluation, whereas ViT-L/16 achieves the highest means on unseen configurations. We adopt ViT-L/16 for its global self-attention, which models relationships between band regions and provides a representation basis for homologous comparison and normal-statistics calibration. Its batch-independent LayerNorm also suits feature comparison and normal-statistics estimation across training branches.

E.2 Normal-Statistics Calibration

To distinguish the roles of grouping, centering, and scale correction, we compare four settings: no normal-statistics calibration; groupwise centering by subtracting only the normal mean; global centering and scale correction using shared normal statistics; and the full groupwise centering and scale correction. Groups are defined by chromosome type and reference source type, consistent with the normal-statistics calibration used in our method.

Table E.2: Comparison of normal-statistics calibration schemes. All metrics are percentages; higher is better.

	Cross-Environment Generalization		Unseen-Configuration Generalization	
Calibration	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
None	$\mathbf{93.84 \pm 4.56}$	55.47 ± 24.29	82.22 ± 14.85	37.61 ± 21.94
Groupwise centering	93.03 ± 2.85	55.83 ± 15.63	90.33 ± 6.31	44.98 ± 22.45
Global centering and scaling	93.12 ± 4.84	53.35 ± 25.45	82.70 ± 14.86	38.02 ± 21.62
Groupwise centering and scaling	93.41 ± 5.36	$\mathbf{56.10 \pm 21.33}$	$\mathbf{90.83 \pm 5.74}$	$\mathbf{46.77 \pm 25.32}$

The benefits of normal-statistics calibration are concentrated in unseen configurations. Compared with no calibration, full calibration improves AUROC and AUPRC by 8.61 and 9.16 percentage points, respectively. Groupwise centering alone accounts for most of this improvement; adding scale correction further increases the two metrics by 0.50 and 1.79 percentage points. In contrast, global calibration performs similarly to no calibration, supporting normal baselines tailored to chromosome type and reference type.

For seen configurations, AUROC is relatively similar across settings. Full calibration does not achieve the highest AUROC but yields the highest AUPRC. Overall, these results support its contribution to unseen-configuration recognition rather than uniform improvements across all evaluation conditions.

E.3 Pair Representations

We compare target features alone, q; concatenated target and reference features, $[q, r]$; absolute differences alone, $|q - r|$; and our representation, $[q, |q - r|]$, to examine the roles of target-specific information and homologous comparison. Square brackets denote feature concatenation.

Table E.3: Comparison of pair representations. All metrics are percentages; higher is better.

	Cross-Environment Generalization		Unseen-Configuration Generalization	
Representation	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
q	91.17 ± 3.80	46.67 ± 24.25	88.03 ± 4.99	43.70 ± 22.32
$[q, r]$	92.80 ± 3.42	39.24 ± 22.34	$\mathbf{90.94 \pm 5.13}$	45.93 ± 24.26
$\lvert q - r \rvert$	71.37 ± 9.14	12.04 ± 9.08	72.56 ± 11.91	10.06 ± 4.82
$[q, \lvert q - r \rvert]$	$\mathbf{93.41 \pm 5.36}$	$\mathbf{56.10 \pm 21.33}$	90.83 ± 5.74	$\mathbf{46.77 \pm 25.32}$

Compared with target features alone, adding the absolute difference branch improves AUROC and AUPRC by 2.24 and 9.43 percentage points for seen configurations and by 2.80 and 3.07 percentage points for unseen configurations, respectively. This indicates that homologous comparison complements the target representation.

Using only absolute differences substantially reduces performance in both domains. Under the same fixed statistical context, swapping members preserves the difference branch, so the linear head satisfies

$$\eta(x_1 \mid x_2) - \eta(x_2 \mid x_1) = \beta_{q,g}^{\top}\big(e_\theta(x_1) - e_\theta(x_2)\big),$$

where $\beta_{q,g}$ is the effective target-branch weight after calibration and global standardization. The difference branch provides a shared, pair-dependent logit adjustment, while the target branch distinguishes the members. Their combination uses comparative and member-specific information for target abnormality prediction. Direct concatenation, $[q, r]$, achieves a slightly higher unseen-configuration AUROC than our representation, whereas our representation performs better on the other three metrics, particularly seen-configuration AUPRC. These findings support the complementary roles of target information and explicit difference information.

E.4 Training Strategy

We compare Stage-I-only training, single-stage joint training, and three two-stage settings in which the encoder is always frozen, always trainable, or frozen only in Stage II. These comparisons examine representation adaptation, staged optimization, and encoder freezing in Stage II.

Table E.4: Comparison of training strategies. All metrics are percentages; higher is better.

	Cross-Environment Generalization		Unseen-Configuration Generalization	
Training strategy	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
Stage-I-only training	90.29 ± 5.47	53.78 ± 21.34	90.69 ± 5.74	46.46 ± 25.30
Single-stage joint training	89.65 ± 3.83	37.24 ± 16.72	88.65 ± 5.27	36.72 ± 16.84
Two-stage, encoder always frozen	87.79 ± 5.02	32.99 ± 18.92	88.63 ± 5.84	35.13 ± 16.26
Two-stage, encoder always trainable	92.75 ± 5.04	55.92 ± 21.20	90.30 ± 5.70	46.66 ± 25.39
Two-stage, freeze encoder in Stage II	$\mathbf{93.41 \pm 5.36}$	$\mathbf{56.10 \pm 21.33}$	$\mathbf{90.83 \pm 5.74}$	$\mathbf{46.77 \pm 25.32}$

The full two-stage strategy achieves the highest mean on all four metrics. Relative to Stage-I-only training, Stage II improves seen-configuration AUROC and AUPRC by 3.12 and 2.32 percentage points, whereas the corresponding gains for unseen configurations are only 0.14 and 0.31 percentage points. This indicates that Stage I already provides most of the unseen-configuration recognition capability, while Stage II contributes more to performance on new patients and accompanying preparation and imaging variation in the seen-configuration domain.

Single-stage joint training and keeping the encoder frozen throughout both stages yield weaker overall results, supporting representation adaptation for inversion detection followed by robustness optimization. Continuing to update the encoder in Stage II performs similarly to freezing it, indicating that freezing is not the sole source of improvement. In this experiment, freezing preserves the representation while yielding slightly higher mean metrics.

E.5 Effect of Artificial Abnormality Generation

Table E.5 compares geometric inversion generation with conditional GAN, diffusion, and flow matching, using w/o CCRL as the reference without artificial samples or their associated training objectives. All three conditional generators further improve performance in both test domains. Conditional FM achieves the highest cross-environment AUROC/AUPRC of 95.44%/61.50%, while conditional GAN achieves the highest means on unseen configurations at 91.64%/49.35%. These results show that more complex generation strategies can further improve detection performance, at the cost of additional generator parameters and training and sampling computation. We use simple geometric inversion as the default generation strategy throughout the main experiments.

Table E.5: Comparison of artificial abnormality generation strategies. The w/o CCRL variant removes artificial samples and their associated training objectives. Results are five-fold mean ± standard deviation (%). FM denotes flow matching. The highest and second-highest means in each column are bold and underlined, respectively.

	Cross-Environment Generalization		Unseen-Configuration Generalization	
Generation strategy	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
w/o CCRL	92.78 ± 3.18	53.85 ± 18.04	87.69 ± 5.02	42.30 ± 24.70
Geometric inversion (default)	93.41 ± 5.36	56.10 ± 21.33	90.83 ± 5.74	46.77 ± 25.32
Conditional GAN	94.73 ± 3.68	58.05 ± 18.33	$\mathbf{91.64 \pm 4.84}$	$\mathbf{49.35 \pm 25.42}$
Conditional diffusion	$\underline{95.00 \pm 3.02}$	$\underline{59.25 \pm 16.04}$	$\underline{91.58 \pm 4.99}$	$\underline{49.33 \pm 24.83}$
Conditional FM	$\mathbf{95.44 \pm 2.56}$	$\mathbf{61.50 \pm 13.90}$	91.09 ± 5.49	49.28 ± 24.86

Table E.6: ResNet50 transfer to unseen $\mathrm{inv}(9)(q12p13)$ with and without another chr9 inversion configuration in training. Results are mean ± sample standard deviation over three random seeds, in percentages. Sens. and Spec. denote sensitivity and specificity. Thresholds are selected on validation data.

chr9 abnormalities	AUROC	AUPRC	Sens.	Spec.	F1
None	62.90 ± 3.13	61.87 ± 2.99	20.36 ± 17.21	90.30 ± 10.82	28.65 ± 18.86
Other configuration	51.07 ± 13.53	52.56 ± 11.00	38.19 ± 1.20	60.23 ± 17.84	43.17 ± 3.80

E.6 Chromosome-Type Dependence in Cross-Configuration Generalization

An inversion configuration is jointly defined by chromosome type and breakpoint combination. Cross-configuration generalization therefore requires recognizing new breakpoint structures and involves how models use prior experience with abnormalities of the same chromosome type. We examine whether same-type abnormal experience supports configuration transfer, whether the full framework generalizes with and without such experience, and whether improvements extend across chromosome types.

Same-type abnormal experience and configuration transfer. To examine whether exposure to another abnormal configuration of the same chromosome type helps recognize a new configuration, we compare two ResNet50 training conditions on a fixed chr9 test set. The first excludes chr9 abnormalities; the second replaces an equal number of other-type abnormalities with $\mathrm{inv}(9)(q22.1q32)$ images. Both use identical normal images and retain the same total and abnormal training sample counts. The test set contains equal numbers of normal and abnormal chr9 images from 46 unseen patients, with abnormal configuration $\mathrm{inv}(9)(q12p13)$ absent from training. Both conditions use identical validation data and optimization settings, no artificial abnormalities, and three random seeds each. Introducing another chr9 configuration into training reduces mean AUROC from 62.90% to 51.07% and mean AUPRC from 61.87% to 52.56% (Table E.6). In this comparison, same-type abnormal experience does not improve new-configuration recognition and instead accompanies lower ranking performance. This observation indicates that learning the appearance of one inversion does not necessarily provide abnormality evidence transferable to other breakpoint combinations. Same-type abnormal experience alone is therefore insufficient to ensure configuration transfer; models still need to learn discriminative information extending beyond observed configurations.

Transfer with and without same-type abnormal support. Motivated by this observation, we further evaluate the full framework under the main evaluation protocol in two settings, one without real abnormal training examples of the same chromosome type and the other with another abnormal configuration of that type. We therefore partition the unseen-configuration test results of Full, ResNet50, and DenseNet121 according to whether the corresponding training fold contains same-type real abnormalities. Same-type normal images may be present in training in both settings.

Real inversion samples are scarce, and collecting multiple distinct configurations of the same chromosome type is particularly difficult. In the current data, only one unseen configuration is available

Table E.7: Unseen-configuration recognition with and without same-type real abnormal training support (%). Supported results come from Fold 1; unsupported results are five-fold mean ± sample standard deviation. Because the normal-control scope and positive prevalence differ, AUPRC values in this table are not directly compared with those in the main table.

	With support		Without support	
Method	AUROC	AUPRC	AUROC	AUPRC
ResNet50	45.45	8.32	63.27 ± 7.05	29.79 ± 7.46
DenseNet121	49.25	10.55	67.12 ± 12.06	36.36 ± 10.08
Full	**96.02**	**63.10**	$\mathbf{91.84 \pm 5.92}$	$\mathbf{78.88 \pm 14.13}$

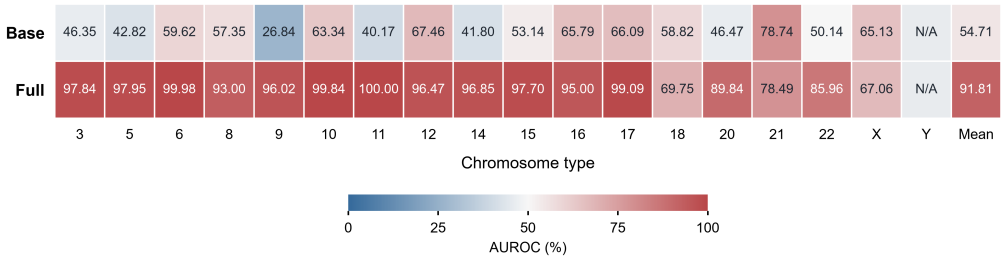


Figure 9: Within-type AUROC (%) on unseen configurations. Each type uses normal and abnormal members from its corresponding test fold. Mean is the unweighted average over 17 evaluable types; Y lacks normal members and is marked N/A.

for analysis with same-type abnormal training support. The unsupported subset contains 17 configurations and accounts for 96.74% of abnormal test members.

Within each fold, we evaluate each subset's abnormal members against normal members of the corresponding chromosome types from that fold's unseen-configuration test domain. All methods use the same jointly scorable members, without retraining or changing model selection. Without same-type real abnormal training support, Full achieves an AUROC of $(91.84 \pm 5.92)\%$ and an AUPRC of $(78.88 \pm 14.13)\%$, demonstrating transfer to chromosome types not covered by real abnormal training examples. On the supported chr9 configuration, Full achieves an AUROC of 96.02% and an AUPRC of 63.10%, showing that it also recognizes a new breakpoint combination beyond the observed abnormal configuration of the same type. Together, these results support the framework's cross-configuration transfer capability both without same-type abnormal examples and beyond existing same-type abnormal experience.

Discrimination across chromosome types. The preceding stratified results summarize performance under two training-support conditions. We further analyze individual chromosome types to examine whether improvements are broadly reflected in within-type normal–abnormal discrimination. We compute AUROC for Full and Base using each chromosome type's normal and abnormal members from its corresponding unseen-configuration test fold, and average equally across evaluable types. Chromosome Y lacks normal members and is therefore excluded from this analysis.

As shown in Figure 9, Full outperforms Base on 16 of the 17 evaluable chromosome types, increasing the unweighted mean of per-type AUROCs from 54.71% to 91.81%. These gains reflect improved within-type discrimination across most evaluable chromosome types. Together with the training-support analysis, these findings support configuration expansion and transferable structural learning for recognizing unseen real inversions, both with and without same-type real abnormal training examples.