ESTABLISH, BREAK, REBUILD, CONSOLIDATE: RELATIONAL SELF-SUPERVISED LEARNING FOR CROSS-DATASET SKIN-LESION REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Skin-lesion representations should preserve clinically meaningful relationships across acquisition conditions, yet dermoscopic and clinical or smartphone images can differ substantially in illumination, scale, viewpoint, color response, and visible structure. We study self-supervised representation learning from independently collected and unpaired datasets without using diagnosis labels to define cross-dataset correspondences. We introduce Cross-dataset INter-view relational Distillation with Evidence-masked Recovery and Stabilization (CINDERS), which learns representations by explicitly modeling relationships among samples. An exponential-moving-average teacher defines the relational organization of intact multi-view images and ranks local visual evidence. A shared student first establishes this intact relation; the highest-ranked patches are then physically removed before a second frozen backbone pass; the student rebuilds the intact relation from the surviving evidence; and a symmetric contrastive objective consolidates the intact and rebuilt representations. We evaluate the learned representation primarily through cross-acquisition retrieval at rank 1 (R@1). On DermaMNIST-C and PAD-UFES-20, CINDERS improves the two cross-acquisition R@1 directions from $0.2665/0.2799$ for frozen DINOv3 to $0.5330/0.4940$, increasing their harmonic mean from 0.2730 to 0.5128. Under the matched self-supervised learning protocol, CINDERS achieves a four-direction mean R@1 of 0.6014, compared with 0.5527 for iBOT, while its cross-acquisition harmonic mean of 0.5128 exceeds the 0.3716 obtained by Relational Knowledge Distillation. On the external Derm7pt–MILK10K evaluation, the four-direction mean R@1 increases from 0.2961 for frozen DINOv3 to 0.3665, and the cross-acquisition harmonic mean increases from 0.2638 to 0.3621. Together, the internal and external results show that recovering inter-sample relationships from incomplete visual evidence can improve cross-acquisition representation compatibility without diagnosis supervision or manually specified cross-dataset positive pairs.

1 INTRODUCTION

Skin-lesion recognition depends on fine-grained characteristics such as border morphology, pigment distribution, local texture, and color, but these characteristics are also affected by image acquisition. Dermoscopy provides controlled magnification and illumination and reveals structures that may be less apparent in ordinary photographs, whereas clinical and smartphone images vary more strongly in scale, viewpoint, camera response, surrounding skin, and lighting. Consequently, a representation can organize images partly according to acquisition characteristics rather than clinically meaningful visual relationships. Acquisition and anatomical-site shifts are known to affect the generalization of skin-lesion models (Chamarthi et al., 2024; Sultana et al., 2025).

Classification measures whether diagnostic information can be extracted from a representation, but a fitted decision boundary does not directly reveal whether independently collected acquisition domains occupy compatible neighborhoods. We therefore use cross-acquisition retrieval as

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.

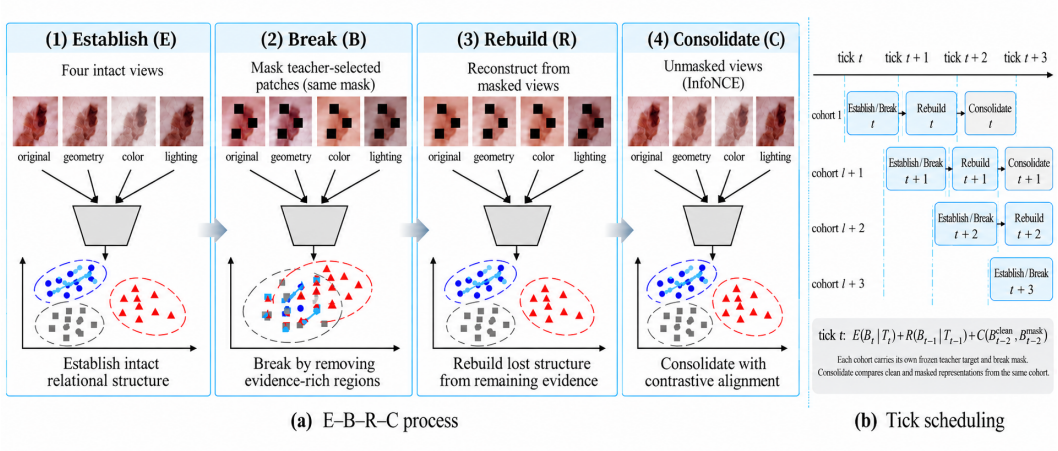


Figure 1: Overview of CINDERS. Establish learns the intact inter-sample relation defined by the exponential-moving-average teacher. Break removes teacher-selected visual evidence before a second frozen backbone pass. Rebuild recovers the intact relation from the surviving evidence, and Consolidate aligns intact and rebuilt representations of the same sample. Successive cohorts are temporally staggered so that each cohort preserves the Establish–Break–Rebuild–Consolidate order while different operations are optimized concurrently.

a representation-level evaluation: a query acquired in one domain must retrieve diagnosis-related images acquired in another without fitting a classifier to bridge the two domains.

This setting creates an additional challenge for self-supervised learning (SSL). DermaMNIST-C and PAD-UFES-20 are independently collected, are not paired at the lesion-instance level, and only partially share diagnostic categories. Using diagnosis labels to construct cross-dataset positive pairs would introduce supervision into representation learning. At the same time, simply forcing invariance to acquisition-sensitive properties such as color or geometry can suppress clinically meaningful evidence (Van Assel et al., 2025). The objective is therefore not to remove acquisition information indiscriminately, but to learn relationships among samples that remain useful when acquisition conditions change.

We address this problem with Cross-dataset INter-view relational Distillation with Evidence-masked Recovery and Stabilization (CINDERS). CINDERS first learns the relational organization of intact samples, deliberately removes visual evidence identified as important by a slowly evolving teacher, and then requires the same student to reconstruct the intact relation from the remaining evidence. Consolidate subsequently couples the intact and rebuilt representations. Thus, the supervisory signal comes from relationships among unlabeled samples rather than from diagnosis-defined cross-dataset correspondences.

Our contributions are threefold. First, we formulate cross-acquisition SSL as relational recovery from incomplete evidence: an intact inter-sample relation is learned and subsequently recovered after informative visual evidence is removed. Second, we realize this formulation through one shared student following an ordered Establish–Break–Rebuild–Consolidate process, with evidence removal before backbone contextualization and cohort-wise temporal scheduling. Third, we evaluate the learned representation through internal and external cross-acquisition retrieval, together with controlled self-supervised baselines and component ablations, to determine whether the relational learning process improves compatibility across acquisition conditions.

2 Related Work

Skin-lesion learning across acquisition conditions. Skin-lesion models span clinical-image classifiers, smartphone datasets, and recent dermatology foundation models (Esteva et al., 2017; Pacheco et al., 2020; Yan et al., 2025a). Nevertheless, differences in imaging device, acquisition protocol, anatomical site, and clinical versus dermoscopic appearance remain important sources of dis-

tribution shift (Chamarthi et al., 2024; Sultana et al., 2025). Most prior studies evaluate these shifts through classification or domain adaptation. Our focus is representation-level compatibility: whether independently collected acquisition domains preserve useful inter-sample relationships without diagnosis-defined cross-dataset correspondences during representation learning.

Relational self-supervised learning. Relational Knowledge Distillation transfers teacher-derived distances and angles (Park et al., 2019), while Relational Self-Supervised Learning matches teacher and student similarity distributions (Zheng et al., 2021). Neighbor-based methods construct positive relationships from the evolving representation itself (Dwibedi et al., 2021). These approaches show that relations among samples can provide supervision beyond individual-instance matching. CINDERS follows this relational perspective but uses the relation obtained from intact samples as a target that must subsequently be recovered after selected visual evidence has been removed.

Masked, teacher–student, and temporally organized self-supervision. DINO uses a momentum teacher for self-distillation (Caron et al., 2021), while Masked Siamese Networks and iBOT combine teacher–student learning with masked inputs or tokens (Assran et al., 2022; Zhou et al., 2022); masked autoencoders instead reconstruct missing image content (He et al., 2022). CINDERS does not treat teacher–student distillation, masking, or reconstruction as novel in isolation. Its distinction is the ordered use of teacher-guided evidence removal as an intervention between establishing and recovering an intact inter-sample relation, followed by explicit intact–rebuilt consolidation. Guo et al. (Guo et al., 2024) showed that heterogeneous self-supervised objectives can benefit from explicit optimization order. CINDERS preserves this ordering at the cohort level so that each cohort progresses through Establish, Break, Rebuild, and Consolidate while different cohorts are processed concurrently.

3 Problem Formulation

Let $\mathcal{D}_D$ and $\mathcal{D}_C$ denote dermoscopic and clinical/smartphone image domains. Their datasets are unpaired at the lesion-instance level and only partially share diagnostic categories. During representation learning, image content and acquisition source are available, but diagnosis labels are withheld.

For image x_i, let $z_i = f_\theta(x_i) \in \mathbb{R}^d$ denote its learned representation. Rather than constructing diagnosis-defined cross-dataset positive pairs, we model how each sample is positioned relative to the other samples in the representation space. The desired representation should preserve useful relational organization when the available visual evidence changes. CINDERS therefore first establishes an intact inter-sample relation, deliberately disrupts the evidence from which that relation was obtained, and then learns to recover the intact relation from the surviving evidence.

Retrieval evaluates the resulting representation rather than defining its training target. For query acquisition domain A and gallery acquisition domain B, $R_{AB}@k$ denotes retrieval at rank k. Thus, R_{DD} and R_{CC} measure within-acquisition retrieval, whereas R_{DC} and R_{CD} measure the two cross-acquisition directions. Diagnosis labels are used only after representation learning to determine retrieval relevance and for downstream evaluation.

4 Method

CINDERS learns representations through four ordered operations: *Establish*, *Break*, *Rebuild*, and *Consolidate*. The same trainable network is used throughout these operations. The operations differ in the visual evidence presented to the network and in the objective applied to the resulting representation.

4.1 Backbone, Multi-View Representation, and Teacher–Student Networks

For each image x_i, we construct original, geometry-, color-, and lighting-transformed views $\{x_i^o, x_i^g, x_i^c, x_i^\ell\}$. Each view is encoded by a frozen DINOv3 Vision Transformer (ViT-L/16) backbone (Siméoni et al., 2025), denoted by f_B,

$$H_i^v = f_B(x_i^v) = (g_i^v, L_i^v), \qquad L_i^v = \{\ell_{i,v,p}\}_{p=1}^{196},$$

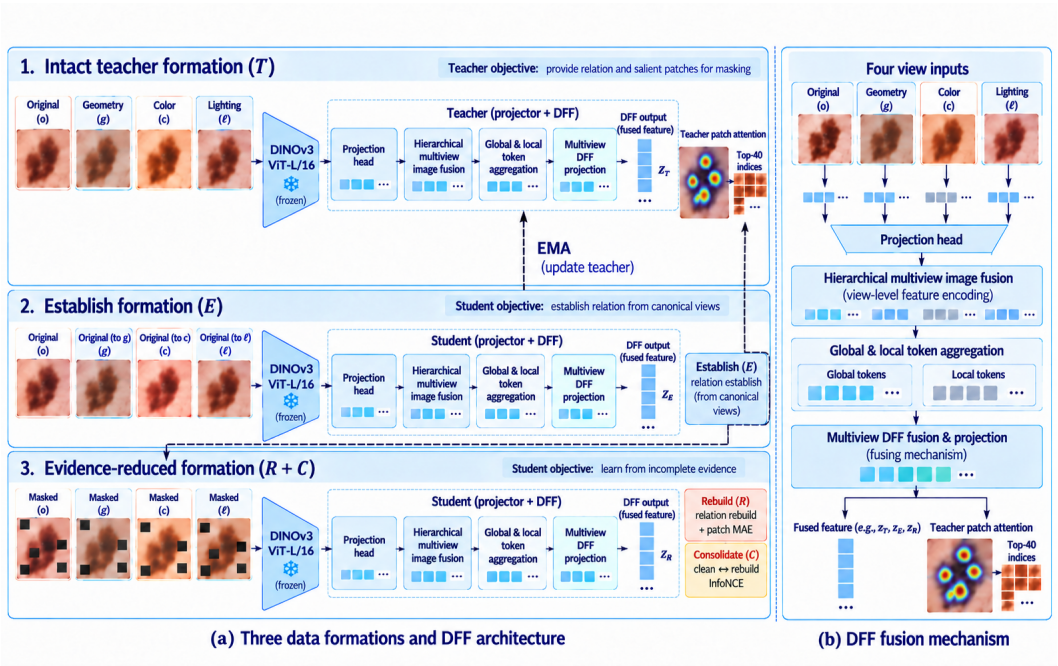


Figure 2: **DFF architecture and CINDERS data formations. (a)** The intact EMA teacher receives the four true views and provides a detached teacher feature together with patch attention for Top-40 evidence selection. Establish feeds the original representation to all four slots of the same student DFF–projector; Rebuild uses the four evidence-reduced views after a second frozen backbone pass. The projector output is a normalized feature, from which batch relations are subsequently computed. **(b)** DFF hierarchically encodes the four views, aggregates global and local tokens, and fuses the resulting view features. The teacher's local attention $\beta_{i,v,p}$ supplies the Break ranking.

where g_i^v and L_i^v denote the global and local patch tokens, respectively. Geometry transformations use nearest-valid boundary extension to avoid artificial blank borders.

A trainable Dynamic Feature Fusion (DFF) module combines global and local information across the four views, and a projection head produces the normalized representation. We denote this trainable student by F_θ and maintain an exponential moving average (EMA) teacher $F_{\bar{\theta}}$ with the same trainable architecture. The frozen backbone is shared by both paths and is never updated. In addition to the fused representation, the teacher DFF produces local attention scores $\beta_{i,v,p}$, which are later used to rank patch evidence for Break. The detailed DFF attention equations are provided in the supplementary material.

The teacher, Establish, and Rebuild use three input formations:

$$\bar{z}_i^T = F_{\bar{\theta}}(H_i^o, H_i^g, H_i^c, H_i^\ell), \qquad z_i^E = F_\theta(H_i^o, H_i^o, H_i^o, H_i^o), \tag{1}$$

and

$$z_i^R = F_\theta(H_i^{o,-}, H_i^{g,-}, H_i^{c,-}, H_i^{\ell,-}), \tag{2}$$

where $H_i^{v,-}$ denotes the backbone representation obtained after selected patch locations are removed from view v before a second frozen backbone pass. We refer to representations computed before evidence removal as *intact* and those computed afterward as *evidence-reduced*.

Thus, the teacher, Establish, and Rebuild use the same backbone and DFF representation machinery but differ in their input formation: the teacher observes four intact views, Establish uses the intact original-view formation, and Rebuild uses four evidence-reduced views.

4.2 Relational Target and Losses

For a balanced batch of B normalized representations, the pairwise cosine relation is

$$K_{ij} = z_i^\top z_j.$$

For each anchor i, the similarities to the remaining samples define a relational distribution

$$q_{ij} = \frac{\exp(K_{ij}/\tau_R)}{\sum_{k \neq i} \exp(K_{ik}/\tau_R)}, \qquad j \neq i,$$

which describes the anchor's relative organization with respect to the rest of the batch without using diagnosis labels. The intact teacher distribution supervises the student through

$$\mathcal{L}_{\mathrm{QKL}} = \frac{1}{B} \sum_i D_{\mathrm{KL}} \left(q_i^T \| q_i^S \right).$$

Because matching the normalized distribution alone can preserve relative neighborhood ordering while allowing the underlying cosine geometry to drift, we additionally use

$$\mathcal{L}_K = \frac{1}{B(B-1)} \sum_{i \neq j} \mathrm{SmoothL1} \left(\widehat{K}_{ij}^S, \widehat{K}_{ij}^T \right),$$

giving

$$\mathcal{L}_{\mathrm{rel}} = \mathcal{L}_{\mathrm{QKL}} + 0.5 \mathcal{L}_K.$$

The hierarchy-consistency term $\mathcal{L}_{\mathrm{hier}}$ denotes the consistency penalty between corresponding intermediate DFF hierarchy outputs of the student and detached teacher, whereas $\mathcal{L}_{\mathrm{feat}}$ matches their final projected representations. Their loss weights and roles in the selected implementation are summarized in the supplementary implementation details. The auxiliary masked reconstruction term $\mathcal{L}_{\mathrm{MAE}}$ is evaluated only at removed patch locations.

4.3 Establish, Break, Rebuild, and Consolidate

Establish. For each newly admitted cohort, the intact teacher relation is detached and retained as the target for that cohort. The student learns this relation from the intact original-view formation using

$$\mathcal{L}_E = 2\mathcal{L}_{\mathrm{rel}}^E + 0.25\mathcal{L}_{\mathrm{hier}}^E + 0.10\mathcal{L}_{\mathrm{feat}}^E.$$

Break. For every view, the intact teacher ranks the 196 patch locations using $\beta_{i,v,p}$ and selects the Top-40. These locations are physically removed before a second frozen backbone pass; no zero-valued patches, noise, or mask tokens are inserted. The remaining 156 patches retain their original positional identities. Break therefore changes the evidence available to the representation but introduces no optimization objective of its own.

Rebuild. The same student processes the four evidence-reduced views and is trained toward the intact teacher target,

$$\mathcal{L}_R = 2\mathcal{L}_{\mathrm{rel}}^R + 0.25\mathcal{L}_{\mathrm{hier}}^R + 0.05\mathcal{L}_{\mathrm{feat}}^R + 0.10\mathcal{L}_{\mathrm{MAE}}^R.$$

Establish and Rebuild share the intact teacher-derived relational target, but their objectives and evidence differ. Establish learns this structure from complete evidence, whereas Rebuild operates on four evidence-reduced views and jointly optimizes relational recovery, teacher–student feature consistency, and masked reconstruction of the removed content.

Consolidate. Let Z^c and Z^r denote the current student's intact and rebuilt representations. Same-image intact–rebuilt pairs are positives and other samples are negatives. Consolidate uses the symmetric contrastive objective

$$\mathcal{L}_{\mathrm{NCE}} = \frac{\mathrm{CE}(Z^c Z^{r\top}/\tau_C, I) + \mathrm{CE}(Z^r Z^{c\top}/\tau_C, I)}{2 \log B}, \qquad \mathcal{L}_C = \mathcal{L}_{\mathrm{NCE}} + 0.10\mathcal{L}_{\mathrm{MAE}}^C.$$

Thus, Rebuild restores the teacher-defined inter-sample relation, whereas Consolidate directly couples the intact and rebuilt state of each individual sample. The masked-reconstruction term is auxiliary to both stages: it regularizes the evidence-reduced path during Rebuild and is evaluated again when that path is re-computed at the later Consolidate state.

4.4 Temporally Staggered Optimization

The Establish–Break–Rebuild–Consolidate order is preserved for each cohort without switching the entire model between long training phases. A cohort reaches Rebuild only after an update containing its Establish objective and reaches Consolidate only after a subsequent update containing its Rebuild objective.

After pipeline fill, optimization step t combines

$$\mathcal{L}^{(t)} = \mathcal{L}_E^{(t)} + \mathcal{L}_R^{(t-1)} + 0.5\mathcal{L}_C^{(t-2)}.$$

All active objectives share the same student parameters and persistent stochastic gradient descent (SGD) optimizer state. Their gradients are accumulated before one student update, after which the EMA teacher is updated. The cohort queue persists across epoch boundaries and is drained after the final batch so that every cohort completes the full sequence.

Relations are computed over a balanced logical batch of $B = 128$ images, containing 64 dermoscopic and 64 clinical samples. Backbone computation may be microbatched for memory efficiency, but the features are concatenated before relation construction.

5 Experiments

5.1 Datasets, Protocol, and Metrics

The internal self-supervised learning and evaluation setting uses the leakage-corrected DermaMNIST-C split (Abhishek et al., 2025), derived from DermaMNIST/MedMNIST v2 (Yang et al., 2023), as the dermoscopic domain and PAD-UFES-20 (Pacheco et al., 2020) as the clinical/smartphone domain. Their train/validation/test counts are $8{,}215/573/1{,}227$ and $1{,}608/230/460$, respectively. Diagnosis labels are excluded from representation learning and are used only after training for retrieval scoring and linear-probe evaluation. Derm7pt and MILK10K are reserved for external evaluation.

For internal retrieval, $R_{XY}@1$ denotes retrieval at rank 1 with acquisition domain X as query and Y as gallery. Thus, R_{DD} and R_{CC} are the two within-acquisition directions, whereas R_{DC} and R_{CD} are the two cross-acquisition directions. We report the four-direction mean $\mathrm{Mean4} = (R_{DD} + R_{CC} + R_{DC} + R_{CD})/4$ and the cross-acquisition harmonic mean $H_{DC,CD} = 2R_{DC}R_{CD}/(R_{DC} + R_{CD})$. For downstream diagnosis evaluation, we report the arithmetic mean of macro one-vs-rest area under the receiver operating characteristic curve (AUC) from linear probes on DermaMNIST-C and PAD-UFES-20. Training hyperparameters, Bayesian parameter selection, transformation ranges, auxiliary reconstruction settings, and validation-only checkpoint selection are provided in the supplementary material.

5.2 Matched Self-Supervised Objective Controls

To isolate the effect of the learning objective, Table 1 compares CINDERS with representative contrastive, non-contrastive, relational-distillation, and masked teacher–student objectives while holding the frozen backbone, DFF, data splits, training horizon, checkpoint-selection rule, and evaluation protocol fixed. These experiments are matched objective controls rather than reproductions of each method's complete native training stack.

CINDERS achieves the highest four-direction mean R@1 at 0.6014, compared with 0.5527 for iBOT and 0.5484 for RKD. Several baselines strongly favor one cross-acquisition direction: SimSiam and BYOL reach $R_{DC}@1 = 0.8050$ and 0.7810 but fall to $R_{CD}@1 = 0.0670$ and 0.1190, respectively, while iBOT reaches $0.7704/0.1365$. RKD provides the strongest competing bidirectional result at $0.3967/0.3494$ and $H_{DC,CD} = 0.3716$. CINDERS instead reaches $0.5330/0.4940$ and $H_{DC,CD} = 0.5128$. The principal gain is therefore a more compatible bidirectional cross-acquisition geometry rather than uniformly higher retrieval in every direction. The ImageBind (Girdhar et al., 2023) and ResNet-50 (He et al., 2016) backbone swaps further show that the framework can be instantiated with different encoder families, although the magnitude of the gain depends on the geometry supplied by the frozen backbone.

Table 1: **Matched self-supervised objective controls on DermaMNIST-C (D) and PAD-UFES-20 (C).** The primary experiments use a frozen DINOv3 backbone (Siméoni et al., 2025); CINDERS backbone swaps additionally use ResNet-50 (He et al., 2016) and ImageBind (Girdhar et al., 2023). The within-acquisition column is $(R_{DD} + R_{CC})/2$, Mean_4 averages all four R@1 directions, and $H_{DC,CD}$ is the harmonic mean of the two cross-acquisition directions.

		Within-acquisition	Cross-acquisition		Aggregate		Linear probe
Method	Backbone	Mean R@1	R_{DC}@1	R_{CD}@1	Mean_4	$H_{DC,CD}$	Mean AUC
Standard self-supervised objective controls							
Random DFF initialization	DINOv3	.6931	.3462	.2836	.5040	.3118	.9150
SimCLR (Chen et al., 2020)	DINOv3	.6702	.1045	.2164	.4153	.1409	.8600
BYOL (Grill et al., 2020)	DINOv3	.6340	.7810	.1190	.5420	.2065	.8280
Barlow Twins (Zbontar et al., 2021)	DINOv3	.6990	.1050	.2090	.4280	.1398	.8900
SimSiam (Chen and He, 2021)	DINOv3	.6080	**.8050**	.0670	.5220	.1237	.8060
DINO distillation (Caron et al., 2021)	DINOv3	.6891	.4070	.2687	.5135	.3237	.9070
Relational and masked teacher–student objective controls							
ReSSL (Zheng et al., 2021)	DINOv3	**.7379**	.2117	.2129	.4751	.2123	.9208
RKD (Park et al., 2019)	DINOv3	.7237	.3967	.3494	.5484	.3716	.9260
MSN (Assran et al., 2022)	DINOv3	.6629	.2348	.3293	.4724	.2741	.8925
iBOT (Zhou et al., 2022)	DINOv3	.6520	.7704	.1365	.5527	.2319	.9281
CINDERS and backbone swaps							
$\mathrm{CINDERS}_{\mathrm{ResNet50}}$	ResNet-50	.6324	.2211	.2008	.4217	.2105	.8725
$\mathrm{CINDERS}_{\mathrm{ImageBind}}$	ImageBind	.5895	.4773	.4418	.5245	.4589	.8900
CINDERS (ours)	DINOv3	.6893	.5330	**.4940**	**.6014**	**.5128**	**.9599**

Table 2: Component analysis under the matched cycle-50 protocol. Mean_4 is the arithmetic mean of the four directional R@1 values, and $H_{DC,CD}$ is the harmonic mean of the two cross-acquisition directions.

Variant	R_{DD}@1	R_{CC}@1	R_{DC}@1	R_{CD}@1	Mean_4	$H_{DC,CD}$	Mean_{AUC}
CINDERS	.7979	.5806	.5330	.4940	.6014	.5128	.9599
−MAE	.7954	.5695	.5390	.4859	.5975	.5111	.9173
MAE-only Rebuild	.7767	.5541	.7027	.3936	.6068	.5046	.9212
Random-40 Break	.7938	.5585	.5236	.4378	.5784	.4769	.9129
Same-state update	.8011	.5894	.5587	.4217	.5927	.4806	.9247
−Consolidate	.8085	.5762	.6033	.2892	.5693	.3910	.9262

5.3 Component Analysis

Table 2 separates the effects of evidence selection, relational recovery, temporal ordering, Consolidate, and masked reconstruction. Random-40 decreases $\mathrm{Mean4}$ from 0.6014 to 0.5784 and $H_{DC,CD}$ from 0.5128 to 0.4769. Replacing temporally staggered optimization with a same-state update lowers $H_{DC,CD}$ to 0.4806. Removing Consolidate produces the largest directional collapse, reducing R_{CD}@1 to 0.2892 and $H_{DC,CD}$ to 0.3910. Removing MAE has little effect, whereas MAE-only Rebuild is strongly asymmetric at $0.7027/0.3936$ for R_{DC}/R_{CD}. These results support teacher-guided evidence selection, relational Rebuild, temporal ordering, and Consolidate as complementary components, with reconstruction acting primarily as auxiliary restorative supervision.

5.4 Frozen Encoder Context

Table 3 places the learned CINDERS representation in the context of general-purpose and domain-specific pretrained encoders, including DINOv3 (Siméoni et al., 2025), ImageBind (Girdhar et al., 2023), ResNet-50 (He et al., 2016), MedImageInsight (Codella et al., 2024), PanDerm (Yan et al., 2025a), MONET (Kim et al., 2024), DermLIP (Yan et al., 2025b), and BiomedCLIP (Zhang et al., 2023), evaluated on the same internal directional retrieval task.

Frozen DINOv3-L/16 gives $R_{DC}/R_{CD} = 0.2665/0.2799$, $\mathrm{Mean4} = 0.4866$, and $H_{DC,CD} = 0.2730$, compared with $0.5330/0.4940$, 0.6014, and 0.5128 after CINDERS training. Cross-acquisition compatibility nevertheless varies substantially across pretrained encoders: DermLIP is strongly directional at $0.7686/0.2164$, BiomedCLIP is more balanced with $H_{DC,CD} = 0.5306$, and

Table 3: **Frozen-encoder context and learned CINDERS representation on the internal benchmark.** Pretrained encoder rows are evaluated without CINDERS optimization; the final row reports the representation learned from the selected DINOv3 starting backbone. Mean_4 averages the four directional R@1 values and $H_{DC,CD}$ summarizes the two cross-acquisition directions.

Encoder / representation	R_{DD}@1	R_{CC}@1	R_{DC}@1	R_{CD}@1	Mean_4	$H_{DC,CD}$	Mean_{AUC}
DINOv3-L/16 (Siméoni et al., 2025)	.8305	.5696	.2665	.2799	.4866	.2730	.9351
ImageBind-H (Girdhar et al., 2023)	.8264	.5891	.1868	.2313	.4584	.2067	.9339
ResNet-50 (He et al., 2016)	.8060	.4935	.1448	.2052	.4124	.1698	.8530
MedImageInsight (Codella et al., 2024)	**.9747**	**.9130**	**.9700**	**.9179**	**.9439**	**.9432**	**.9963**
PanDerm (Yan et al., 2025a)	.8297	.6087	.1851	.2948	.4796	.2274	.9464
MONET (Kim et al., 2024)	.8688	.5870	.6530	.2463	.5887	.3576	.9265
DermLIP (Yan et al., 2025b)	.8908	.5978	.7686	.2164	.6184	.3377	.9347
BiomedCLIP (Zhang et al., 2023)	.8256	.6022	.6855	.4328	.6365	.5306	.9084
CINDERS (ours)	.7979	.5806	.5330	.4940	.6014	.5128	.9599

Table 4: **External Derm7pt↔MILK10K cross-dataset stress test.** R^{XD}_{XY}@1 is the bidirectional cross-dataset average defined in the text; $\mathrm{Mean}^{\mathrm{XD}}_4$ averages all four acquisition directions and $H^{\mathrm{XD}}_{DC,CD}$ their two cross-acquisition directions. Reference models are contextual rather than matched-training baselines. Best/second/third values are **bold**/underlined/*italicized*.

Method	R^{XD}_{DD}@1	R^{XD}_{CC}@1	R^{XD}_{DC}@1	R^{XD}_{CD}@1	$\mathrm{Mean}^{\mathrm{XD}}_4$	$H^{\mathrm{XD}}_{DC,CD}$
DINOv3-L/16 (Siméoni et al., 2025)	.3590	.2905	.2363	.2987	.2961	.2638
ImageBind-H (Girdhar et al., 2023)	.3111	.2862	.2188	.2729	.2722	.2429
ResNet-50 (He et al., 2016)	.2687	.2018	.2241	.1777	.2181	.1982
MedImageInsight (Codella et al., 2024)	**.7279**	**.4466**	**.6195**	**.4626**	**.5642**	**.5297**
PanDerm (Yan et al., 2025a)	.3400	.3041	.3137	.3355	.3233	.3242
MONET (Kim et al., 2024)	.3362	.2993	.2644	.2360	.2840	.2494
DermLIP (Yan et al., 2025b)	.5399	.4121	*.3845*	.4130	.4374	.3983
DermLIP+PanDerm (Yan et al., 2025b;a)	*.4522*	.3818	.3943	.3288	*.3893*	.3586
BiomedCLIP (Zhang et al., 2023)	.3044	.2953	.3584	.2644	.3056	.3043
CINDERS (ours)	.3387	*.4028*	.3672	*.3572*	.3665	*.3621*

MedImageInsight remains substantially stronger in absolute performance with $\mathrm{Mean4} = 0.9439$. The comparison therefore contextualizes the improvement obtained from the selected DINOv3 starting representation rather than implying that CINDERS is the strongest available dermatology representation.

External cross-dataset stress test. Derm7pt and MILK10K each contain dermoscopic (D) and clinical/smartphone (C) images. We define $R^{\mathrm{XD}}_{XY}@1 = \frac{1}{2}[R(X_{\mathrm{D7}} \to Y_{\mathrm{M}}) + R(X_{\mathrm{M}} \to Y_{\mathrm{D7}})]$, where D7 and M denote Derm7pt and MILK10K, respectively. Thus, DD and CC retain acquisition type while changing dataset, whereas DC and CD change both dataset and acquisition. Queries are scored only when a harmonized diagnosis exists in the destination gallery; all gallery samples remain retrieval candidates. The pretrained reference encoders provide representation context rather than matched-training baselines.

Relative to frozen DINOv3, CINDERS raises $\mathrm{Mean4}^{\mathrm{XD}}$ from 0.2961 to 0.3665 and $H^{\mathrm{XD}}_{DC,CD}$ from 0.2638 to 0.3621. The two cross-acquisition directions increase from $0.2363/0.2987$ to $0.3672/0.3572$, while their absolute gap contracts from 0.0624 to 0.0100. The external experiment therefore reproduces the central internal behavior—more compatible bidirectional cross-acquisition neighborhoods—on two independently collected datasets. MedImageInsight and DermLIP remain stronger in absolute external retrieval, so the result supports improved transfer from the DINOv3 starting representation rather than universal superiority over specialized pretrained encoders.

Post-hoc sensitivity and qualitative analysis. The Top-40 evidence-removal setting is fixed for all primary experiments above. A post-hoc sweep over evidence-removal ratios, teacher-attention visualizations, and representative Top-5 retrieval examples for all four directional evaluations are provided in the supplementary material. The sweep characterizes sensitivity to intervention strength and is not used to reselect the primary model.

6 Discussion and Limitations

The principal effect of CINDERS is observed in cross-acquisition neighborhood organization. The method does not use diagnosis labels to construct positive pairs or optimize a diagnostic decision boundary; instead, it models relationships among samples and trains the student to recover those relationships after informative evidence is removed. The matched-baseline results show that high retrieval in one cross-acquisition direction can coexist with poor reverse transfer, whereas CINDERS substantially improves their bidirectional balance. The component analysis further separates this behavior from simple masking or reconstruction: teacher-guided Break improves over random removal, relational Rebuild is more balanced than reconstruction-only recovery, and Consolidate prevents a large directional collapse.

The frozen-encoder and external evaluations provide complementary context. Relative to frozen DINOv3, CINDERS substantially improves internal and external cross-acquisition compatibility, and the Derm7pt–MILK10K experiment shows that the same qualitative behavior extends beyond the two internal datasets. Specialized medical and dermatology encoders nevertheless remain stronger in absolute performance, indicating that the contribution is the relational learning process and its effect on the selected starting representation rather than a universally superior pretrained encoder.

Several limitations remain. First, the relational target is constructed within a finite balanced batch and is therefore batch-relative. Second, the primary Top-40 intervention strength is fixed rather than adapted to individual images; the post-hoc mask-ratio sweep confirms that retrieval is sensitive to the amount of removed evidence and motivates future adaptive intervention policies. Third, the exponential moving average teacher and student co-evolve, so cohort-wise Establish–Break–Rebuild–Consolidate ordering does not isolate a cohort from updates contributed by neighboring cohorts. Finally, broader validation across backbones, acquisition devices, and external datasets is needed to determine how generally the relational recovery process transfers beyond the present implementation.

7 Conclusion

We introduced CINDERS, a self-supervised representation-learning framework that Establishes an intact inter-sample relation, Breaks teacher-selected evidence before backbone contextualization, Rebuilds the relation with auxiliary restorative supervision, and Consolidates intact and rebuilt states. On DermaMNIST-C/PAD-UFES-20, CINDERS reaches $\mathrm{Mean4} = 0.6014$ versus 0.5527 for iBOT and $H_{DC,CD} = 0.5128$ versus 0.3716 for RKD; on Derm7pt–MILK10K, it improves frozen-DINOv3 $\mathrm{Mean4}^{\mathrm{XD}}/H^{\mathrm{XD}}_{DC,CD}$ from $0.2961/0.2638$ to $0.3665/0.3621$. Together with the component and sensitivity analyses, these results support relational recovery from incomplete evidence as a useful mechanism for improving cross-acquisition representation compatibility.

AI Use Statement

In this work, we used generative AI tools for brainstorm and refine scientific ideas and conceptual framing. We have not used generative AI tools to draft sections of the paper, to generate synthetic datasets and for proving mathematical claims; the remaining required-disclosure tasks were not applicable to this work. Additionally, we used generative AI tools for literature search and discovery and to edit manuscript text for clarity and readability. We reviewed all AI-assisted work. References and related-work suggestions obtained through AI-assisted search were checked against the original literature, and the authors independently verified all reported results and made the final scientific and methodological decisions. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Mahmoud Assran et al. Masked Siamese Networks for label-efficient learning. In *ECCV*, 2022.

Mahmoud Assran et al. The hidden uniform cluster prior in self-supervised learning. In *ICLR*, 2023.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021.

Sireesha Chamarthi, Katharina Fogelberg, Titus J. Brinker, and Julia Niebling. Mitigating the influence of domain shift in skin lesion classification: A benchmark study of unsupervised domain adaptation methods. *Informatics in Medicine Unlocked*, 44:101430, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *CVPR*, pp. 15750–15758, 2021.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *ICCV*, pp. 9588–9597, 2021.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

Jean-Bastien Grill et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, volume 33, 2020.

Zuwei Guo, Nahid Ul Islam, Michael B. Gotway, and Jianming Liang. Stepwise incremental pretraining for integrating discriminative, restorative, and adversarial learning. *Medical Image Analysis*, 95:103159, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Tim Lebailly, Thomas Stegmüller, Behzad Bozorgtabar, Jean-Philippe Thiran, and Tinne Tuytelaars. Adaptive similarity bootstrapping for self-distillation based representation learning. In *ICCV*, pp. 16505–16514, 2023.

Lucas Maes, Ian Hajra, Arnesh Batra, Hugues Van Assel, Damien Scieur, and Randall Balestriero. Relational representation learning. In *UniReps Workshop at NeurIPS*, 2025.

Andre G. C. Pacheco et al. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, 2020.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pp. 3967–3976, 2019.

Oriane Siméoni et al. DINOv3. *arXiv:2508.10104*, 2025.

Vlad Sobal, Mark Ibrahim, Randall Balestriero, Vivien Cabannes, Diane Bouchacourt, Pietro Astolfi, Kyunghyun Cho, and Yann LeCun. $\mathbb{X}$-Sample Contrastive Loss: Improving contrastive learning with sample similarity graphs. In *ICLR*, 2025.

Nurjahan Sultana, Wenqi Lu, Xinqi Fan, and Moi Hoon Yap. Domain adaptation for skin lesion: Evaluating real-world generalisation. In *CVPR Workshops*, pp. 3433–3443, 2025.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, volume 30, 2017.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, pp. 1365–1374, 2019.

Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint-embedding vs reconstruction: Provable benefits of latent space prediction for self-supervised learning. In *NeurIPS*, 2025.

Jiancheng Yang et al. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. ReSSL: Relational self-supervised learning with weak augmentation. In *NeurIPS*, 2021.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *ICLR*, 2022.

Kumar Abhishek, Aditi Jain, and Ghassan Hamarneh. Investigating the quality of DermaMNIST and Fitzpatrick17k dermatological image datasets. *Scientific Data*, 12:196, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *CVPR*, pp. 15180–15190, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Chanwoo Kim, Soham U. Gadgil, Alex J. DeGrave, Jesutofunmi A. Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nature Medicine*, 30:1154–1165, 2024.

Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. Derm1M: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In *ICCV*, pp. 12681–12690, 2025.

Sheng Zhang et al. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image–text pairs. *arXiv:2303.00915*, 2023.

Siyuan Yan et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31:2691–2702, 2025.

Noel C. F. Codella et al. MedImageInsight: An open-source embedding model for general domain medical imaging. *arXiv:2410.06542*, 2024.