MetaAlign: Prototype-Centric Multimodal Alignment for Robust Few-Shot Adaptation under Domain and Modality Shift

Md Abdullah Al Mahmud Pias
North South University
Dhaka, Bangladesh
abdullah.pias@northsouth.edu

Mahmodul Hasan Jobayer
North South University
Dhaka, Bangladesh
hasan.jobayer@northsouth.edu

Syed Muhammad Bilal Haider
Washington University of Science and Technology
Alexandria, Virginia, USA
bilalhaider@live.com

Fahim Istiak
University of Oulu
Oulu, Finland
istiakfahim2@gmail.com

Dewan Aminul Islam
Florida Atlantic University
Boca Raton, Florida, USA
dewanaminul12@gmail.com

Abdul Ahad*
North South University
Dhaka, Bangladesh
ahad.abdul01@northsouth.edu

***Abstract*—Medical imaging models often lose accuracy when the deployment site differs from the training site, when the imaging modality changes, or when only a handful of target labels exist. We propose MetaAlign, a prototype-centric multimodal framework that combines Hybrid Gated Fusion (HGF), Multimodal Prototype Alignment (MPA), Domain-Aware Prototype Augmentation (DPA), and domain-adversarial training. HGF fuses image, text, and metadata features and tolerates missing inputs. MPA keeps class geometry consistent between modality-specific and fused prototypes, and DPA perturbs prototypes using a descriptor of the target support set. We evaluate three transfer settings at 1, 3, and 5 shots: MIMIC-CXR → CheXpert, radiograph → CT (MosMedData), and ISIC 2019 → HAM10000. MetaAlign reaches 5-shot accuracies of 82.1%, 65.5%, and 83.7%, respectively, and obtains the highest accuracy among the compared methods at every shot level. Ablations indicate that each component contributes. A magnitude-matched Gaussian control shows a small, directionally consistent advantage for target-conditioned DPA, nominally significant only at 3-shot ($p = 0.045$, approximate unpaired test). Given the approximate statistics and the known limits of the cross-modality and cross-source benchmarks, we read the results as support for prototype-level adaptation, not as clinical validation.**

***Index Terms*—Cross-domain few-shot learning, Multimodal learning, Prototype-based learning, Domain adaptation, Medical image analysis, Modality robustness.**

I. Introduction

Deep learning models for medical imaging are usually developed and validated under favorable conditions: curated datasets, fixed acquisition protocols, and training and deployment data drawn from the same source [1], [2]. In practice, institutions differ in scanners, calibration, patient demographics, and documentation habits. The resulting distribution shift can make performance estimates from one site unreliable at another [3], and it is a major obstacle between benchmark results and clinical use.

*Corresponding author: Abdul Ahad (ahad.abdul01@northsouth.edu).

Data scarcity makes the problem harder. Annotating a new clinical setting takes expert time, regulatory coordination, and specialized labor, which are often scarce at deployment. Few-shot learning targets this constraint [4]–[6], but standard formulations assume that source features transfer cleanly to the target. Under a large domain shift that assumption often fails [7], [8]: class boundaries learned on the source can collapse or distort in the target, because discriminative signal is entangled with domain-specific appearance statistics. Domain adaptation encourages distributional invariance [9], [10], yet aligning global feature distributions can suppress the class-specific signal that classifiers depend on [11]. Balancing invariance against discriminability is difficult when labeled target data are limited [12].

Clinical data are also multimodal [2]. Radiology reports [13], acquisition metadata, and structured records carry complementary information, and recent image–report pretraining shows how much the text side can contribute [14]–[16]. Most multimodal methods, however, assume that every training modality is still available at inference [17]. That assumption is fragile across institutions: a receiving center may lack structured reports, follow different documentation standards, or use a different imaging modality altogether [18]. A model with hard modality dependencies may then fail without any visible warning. The same issue appears outside medicine: prior work on word-level Bangla sign language recognition combines pose-based video features with audio to improve accuracy [19].

Existing work tends to treat these issues separately. The setting we address combines all of them: few-shot adaptation under domain shift, modality discrepancy, and missing auxiliary inputs. A suitable framework should preserve the class-discriminative structure that few-shot classification needs, fuse heterogeneous modalities without requiring all of them [20], and calibrate class representations to the target distribution [21].

We propose *MetaAlign*, built on one principle: when labels are scarce, adaptation should act on class prototypes rather

than on raw feature distributions [22], [23]. Prototypes carry the class geometry that a few-shot classifier relies on, so intervening there may let the model change what needs to change (cross-domain class alignment) without indiscriminately removing domain variation [24], [25]. Fig. 1 gives an overview. MetaAlign has three main components. *Hybrid Gated Fusion* (HGF) combines cross-modal attention with a global concatenated representation through a learned gate, which adjusts each pathway's contribution and accommodates partial modality availability. *Multimodal Prototype Alignment* (MPA) keeps modality-specific and fused class prototypes geometrically consistent [26], [27], so that different representations of one class tend to lead to a consistent decision boundary. *Domain-Aware Prototype Augmentation* (DPA) generates target-conditioned prototype variations from the support set, giving the model a structured prior over the variation it may meet at inference without extra labeled data [28]. A domain-adversarial term complements these components.

We evaluate MetaAlign on three transfer scenarios: institutional transfer (MIMIC-CXR [29] $\rightarrow$ CheXpert [30]), cross-modality transfer (chest radiograph $\rightarrow$ CT via MosMedData), and cross-source dermatological transfer (ISIC 2019 $\rightarrow$ HAM10000). At 1, 3, and 5 shots, MetaAlign outperforms the compared few-shot and domain-adaptation baselines in all three settings in our experiments, reaching 71.5%, 77.9%, and 82.1% on MIMIC-CXR $\rightarrow$ CheXpert. Removing any single module lowers accuracy in our ablations. We also report calibration, interpretability, and computational cost, since deployability matters alongside accuracy.

The contributions of this work are:

- *MetaAlign*, a prototype-centric framework that addresses domain shift, modality discrepancy, and missing auxiliary inputs within one architecture.
- Hybrid Gated Fusion and Multimodal Prototype Alignment, which integrate heterogeneous representations while keeping class-level geometry consistent.
- Domain-Aware Prototype Augmentation, which uses target support information to generate adaptive prototype variations under limited supervision. Against a magnitude-matched Gaussian perturbation baseline, target-conditioned DPA is higher at all three shot levels; the gap is nominally significant at 3-shot ($p = 0.045$, approximate unpaired test, no multiple-comparison correction) and not significant at 1- and 5-shot.
- An analysis of accuracy, calibration, interpretability, and efficiency, to judge suitability for resource-constrained deployment.

The intuition behind the design is simple: with scarce supervision, a model should adapt what it knows about each class instead of trying to erase what it knows about each domain. Our results suggest that this prototype-centric view may be a useful basis for few-shot medical transfer, compared with methods that operate only on global feature distributions.

II. Related Work

A. Cross-Domain Few-Shot Learning

Cross-domain few-shot learning (CDFSL) relaxes the assumption of standard few-shot methods that source and target data share one distribution. Early approaches aligned features through adversarial training or domain-invariant representations [9], [10]. Metric-based methods, above all Prototypical Networks [4], showed that class prototypes estimated from a small support set are a natural basis for adaptation under scarce supervision. Meta-Baseline [6], FEAT [8], and systematic CDFSL benchmarks [3] suggested that prototype-based representations tend to generalize more reliably than optimization-heavy episodic alternatives. More recent work adds prototype-aware adaptation through hierarchical alignment, dual-level representation objectives, and supervised contrastive losses, with gains in class separability across heterogeneous source–target pairs [21], [24], [26]. A limitation remains: adaptation usually acts on global feature distributions rather than on class-specific geometry. When a shift involves a change of modality or acquisition physics and not only of imaging environment, distribution-level alignment may weaken the very class structure that few-shot classifiers need. MetaAlign is motivated by this gap and treats prototype geometry, rather than feature statistics, as the main object of adaptation.

B. Prototype-Based Representation Learning

Class prototypes are a lightweight and interpretable summary that can be estimated from little supervision, which made prototype-based learning a standard paradigm for few-shot recognition [4]. Later work added multi-prototype representations for intra-class variability, Gaussian uncertainty estimates, and semantic side information for prototype refinement [27], [28]. In medical imaging, prototype methods have been applied to chest pathology classification, lesion segmentation, and dermatological diagnosis [12], [22], [23], [25], [31], [32], often combined with attention, feature-mapping networks, or contrastive objectives. Most of this work implicitly assumes that the domain used to estimate prototypes resembles the domain where they are applied. Across institutions, modalities, or patient populations, source prototypes become poorly calibrated anchors for target classification. MetaAlign therefore treats prototype geometry as something to adapt, not only to estimate.

C. Multimodal Learning under Incomplete Modalities

Combining complementary clinical modalities, such as images, radiology reports, acquisition metadata, and structured records, generally improves representation quality. Cross-modal attention, graph-based relational reasoning, and contrastive multimodal objectives have all shown gains when heterogeneous streams can be used jointly [18], [20], and vision-language pretraining on chest radiographs and reports has pushed this further [14]–[16]. These methods typically assume that the full set of training modalities is present at inference. In deployment this often does not hold: reports may be missing, metadata may be inconsistently filled in, or a cross-institutional transfer may introduce a modality that training never saw. Fixed concatenation or static attention weights can degrade under such conditions, which is why robustness to missing modalities has become an active research topic [17]. MetaAlign's Hybrid Gated Fusion addresses it by learning to weight each pathway according to availability and informativeness, instead of assuming that all inputs are present.

A related example outside medicine is word-level Bangla sign language recognition. Hasan et al. [19] merged the public SignBD-Word collection with their own recordings to form a 200-word dataset with two modalities, video and audio. Body keypoints extracted from the video frames with OpenPose and MediaPipe were classified by a CNN-LSTM or a Vision Transformer, and a VGGish network handled the audio. The branches were trained independently, and their class probabilities were merged at the decision stage. The Vision Transformer alone reached 88.52% accuracy and the audio branch alone 81.91%, while the fused ViT–VGGish model reached 94.71% accuracy and a 94.52% F1 score. This shows how much a second modality can add when one stream is insufficient, which is the same motivation behind our use of text and metadata. The authors also report that background noise in the audio reduced performance, an example of the modality-reliability problem that a learned gate is meant to handle. Our setting differs in three ways. Because the branches meet only at the decision stage, the design offers no explicit way to align class structure in a shared feature space. The evaluation uses both modalities together on a single merged dataset and does not examine domain shift, few-shot adaptation, or missing modalities. The authors themselves note that the work is limited to word-level prediction and has not been tested on embedded devices. MetaAlign addresses the complementary setting: it fuses modalities earlier through a gated mixture of attention and concatenation, tolerates absent inputs by masking, and adapts class prototypes to a new domain from a few labeled samples.

D. Few-Shot Learning in Medical Imaging

The cost of clinical annotation makes few-shot learning especially relevant for medical AI. Recent methods based on prototype learning, contrastive self-supervision, foundation-model adaptation, Fourier-guided style transfer, and cross-attention have improved results under limited supervision in radiology, pathology, and dermatology [23], [25], [31], [32]. Foundation models have also been adapted with few labeled samples [2], [33]. Even so, generalization across hospitals, imaging protocols, and patient populations remains an open problem: models that do well within one dataset can degrade, sometimes noticeably, on external distributions with a nominally identical task. The literature also tends to handle few-shot adaptation, multimodal integration, and domain adaptation separately, and few studies examine settings where all three are active at once. This compound setting is the one MetaAlign targets.

E. Limitations of Existing Approaches

The survey above points to four limitations that motivate our framework. First, dominant domain adaptation strategies align feature distributions, an objective that does not necessarily preserve, and may suppress, class-specific semantic relationships. Second, most multimodal architectures assume complete modality availability, which cross-institutional and cross-modality deployment often violates. Third, conventional prototype methods treat prototypes as static point estimates and give no way to adapt their geometry to a new target domain under scarce supervision. Fourth, existing augmentation strategies work in input or intermediate feature spaces, which limits their ability to model the structured, class-level variation that domain shift induces at the prototype level. Together, these gaps leave few-shot adaptation under joint domain and modality shift, with incomplete auxiliary information, relatively unexplored.

F. Positioning of MetaAlign

MetaAlign treats the class prototype, not the global feature distribution, as the primary object of adaptation. HGF handles heterogeneous and partially missing modalities, MPA keeps modality-specific and fused class geometry consistent, and DPA moves prototypes toward target-domain structure using only the support set. Together they form a single prototype-centric pipeline for few-shot transfer under domain shift, modality asymmetry, and limited supervision.

III. Methodology

A. Problem Formulation

We cast cross-domain few-shot adaptation as episodic classification. A model trained on a well-annotated source domain $\mathcal{D}_S$ is adapted to an unseen target domain $\mathcal{D}_T$, where $\mathcal{D}_S \neq \mathcal{D}_T$ and the marginal feature distributions may differ substantially, $P_S(X) \neq P_T(X)$ [3]. Each episode samples a support set S and a disjoint query set Q from the target domain. For an N-way, K-shot task,

$$S = \bigcup_{k=1}^{N} \{(x_{k,i}, y_k)\}_{i=1}^{K}, \qquad Q = \{(x_j, y_j)\}_{j=1}^{|Q|}. \tag{1}$$

The goal is to predict query labels using only the supervision in S, which mirrors data-scarce clinical deployment [33]. MetaAlign extends this to heterogeneous inputs: each example $x = (x_I, x_T, x_M)$ consists of an image, a clinical text report, and structured metadata, and joint availability is not assumed. The framework works with any available subset $\mathcal{M}_{\text{avail}} \subseteq \{I, T, M\}$, which covers both institutional transfer and cross-modality scenarios in which auxiliary inputs may be absent entirely.

B. Overview of MetaAlign

MetaAlign rests on a single design principle: adaptation should act on class prototypes, which are the representation most directly tied to the few-shot decision. Fig. 1 summarizes the pipeline, which has three components. *Hybrid Gated Fusion* (HGF) integrates modality representations through an adaptive gate that balances cross-modal semantic interaction against global context, and it remains usable when modalities are missing. *Multimodal Prototype Alignment* (MPA) enforces geometric consistency between modality-specific and fused class prototypes. *Domain-Aware Prototype Augmentation* (DPA) generates target-conditioned prototype variations from the support set, which covers part of the variation expected at inference without additional labels [21], [28].

At run time, modality-specific encoders map the available inputs into a shared latent space, and HGF produces a fused embedding. Class prototypes are built from the fused support embeddings, kept consistent through MPA, and adapted through

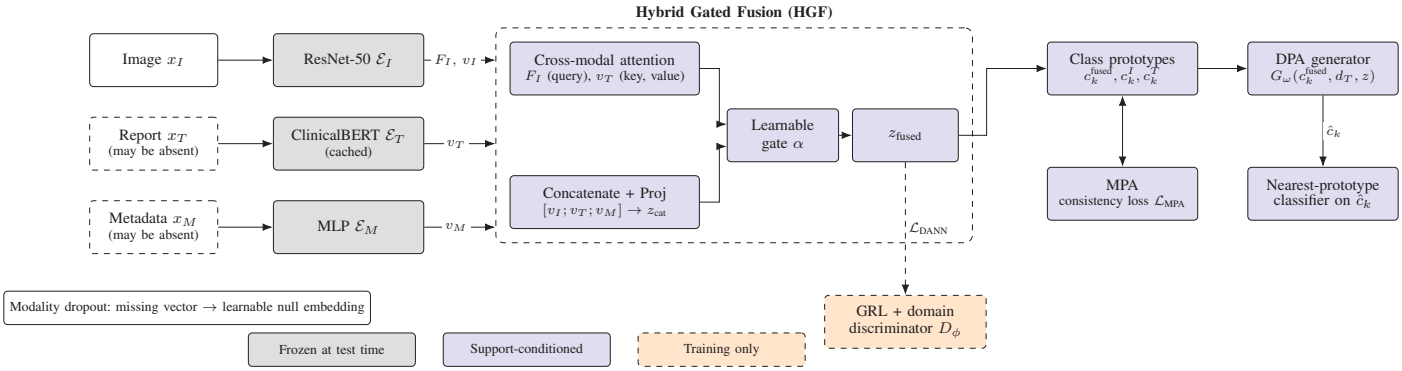


Fig. 1. Overview of MetaAlign. Image, text, and metadata encoders (gray) are frozen at test time; ClinicalBERT embeddings are precomputed. Hybrid Gated Fusion (HGF) contains the cross-modal attention branch, the concatenation branch, and the learnable gate α that mixes them into z_{fused}. Class prototypes built from z_{fused} and the modality-specific vectors are kept consistent by Multimodal Prototype Alignment (MPA) and shifted toward the target domain by Domain-Aware Prototype Augmentation (DPA). Query samples are classified by their nearest adapted prototype. Normalization statistics are also re-estimated from the support set at test time. HGF, MPA, and DPA (blue) are learned during meta-training and operate on the support set at test time; the domain-adversarial branch (orange, dashed) is used only during training.

DPA to reflect the target domain. Queries are then assigned to the nearest adapted prototype.

Notation. $\mathcal{E}_I$, $\mathcal{E}_T$, and $\mathcal{E}_M$ denote the image, text, and metadata encoders. F_I is the spatial feature map of an image, v_I, v_T, and v_M are the corresponding image, text, and metadata vectors, and z_{attn}, z_{cat}, and z_{fused} are the attention, concatenation, and fused embeddings. Prototypes are written c_k^{fused}, c_k^I, and c_k^T for class k, and $\hat{c}_k$ is the DPA-adapted prototype.

Training objective. Training combines the episodic cross-entropy loss $\mathcal{L}_{\text{CE}}$ of the prototype classifier (Eq. 8), the alignment loss $\mathcal{L}_{\text{MPA}}$, and the adversarial loss $\mathcal{L}_{\text{DANN}}$:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{MPA}} + \lambda_2 \mathcal{L}_{\text{DANN}}, \tag{2}$$

where λ_1 and λ_2 set the relative weight of prototype alignment and adversarial adaptation. The following subsections define each term in turn.

C. Visual Feature Extraction

A ResNet-50 encoder $\mathcal{E}_I(\cdot)$ [6], [34] maps an input image x_I to a spatial feature map F_I, and pooling gives a global vector v_I:

$$F_I = \mathcal{E}_I(x_I) \in \mathbb{R}^{H \times W \times d_v}, \qquad v_I = \text{Pool}(F_I). \tag{3}$$

We keep the spatial map on purpose. Abnormalities in medical images are often localized, and compressing the image to one global vector before cross-modal interaction would discard fine-grained spatial information that may be needed to ground text in specific regions. The map F_I feeds the cross-modal attention, while v_I supplies global context. Images are resized to 224×224 before encoding.

D. Clinical Text Representation

When a clinical report is available, a ClinicalBERT encoder $\mathcal{E}_T(\cdot)$ [13] produces $v_T = \mathcal{E}_T(x_T)$. We chose ClinicalBERT over general-domain language models because it was pretrained on clinical text and therefore captures the vocabulary and reporting conventions of radiology, including descriptions of finding type, location, and severity that an image alone may leave ambiguous. Text embeddings are precomputed and cached, which avoids repeated transformer forward passes during episodic training. When no text is available, the text pathway is masked and the remaining modalities provide the fused representation.

E. Metadata Representation

Structured metadata are encoded by a lightweight MLP $\mathcal{E}_M(\cdot)$, giving $v_M = \mathcal{E}_M(x_M)$. Categorical variables pass through learned embeddings, and continuous variables are normalized to zero mean and unit variance. The output is projected into the shared latent space so that fusion is consistent. If metadata are unavailable, the pathway is masked instead of imputed, because imputed values may mislead more than a learned treatment of absence.

F. Hybrid Gated Fusion

Cross-modal attention and global aggregation capture different kinds of multimodal information. Attention grounds semantic concepts in specific anatomical regions, whereas aggregation summarizes context across all available representations. Neither is better in general; which one helps depends on the sample, the available modalities, and the clinical task. HGF uses both. Given the spatial map F_I and the text vector v_T, cross-modal attention is

$$z_{\text{attn}} = \text{Pool}\big(\text{Attn}(F_I, v_T)\big), \tag{4}$$

where the spatial visual features act as queries, the text vector provides the keys and values, and the attended map is pooled into one vector [18]. The global multimodal representation is

$$z_{\text{cat}} = \text{Proj}([v_I; v_T; v_M]). \tag{5}$$

A learnable gating scalar α sets the contribution of each branch:

$$\alpha = \sigma(W_g[z_{\text{attn}}; z_{\text{cat}}] + b_g), \qquad z_{\text{fused}} = \alpha z_{\text{attn}} + (1-\alpha) z_{\text{cat}}. \tag{6}$$

The model can therefore lean on fine-grained cross-modal interaction when the text is informative and on global aggregation when it is not, without committing to a fixed fusion rule [20]. Missing modalities are masked before fusion by replacing the corresponding vector with a learnable null embedding (the modality-dropout step in Fig. 1), so HGF operates on whichever inputs exist.

G. Few-Shot Prototype Construction

Class prototypes are the means of the support embeddings of each class:

$$c_k^{\text{fused}} = \frac{1}{K}\sum_{i=1}^{K} z_{k,i}^{\text{fused}}, \quad c_k^I = \frac{1}{K}\sum_{i=1}^{K} v_{k,i}^I, \quad c_k^T = \frac{1}{K}\sum_{i=1}^{K} v_{k,i}^T. \tag{7}$$

A query embedding z_q is classified by its distance $d(\cdot,\cdot)$ to each prototype [4]:

$$p(y=k \mid z_q) = \frac{\exp(-d(z_q, c_k))}{\sum_{j=1}^{N}\exp(-d(z_q, c_j))}. \tag{8}$$

Because the decision structure is built entirely from the support set, the classifier tends to reflect the class distribution of the target domain rather than source-domain bias stored in fixed parameters [7].

H. Multimodal Prototype Alignment

Fusing modalities does not by itself keep modality-specific class structure coherent with the fused prototype space. Under domain shift each modality responds differently, and inconsistencies that reach the fused prototype can destabilize the decision boundary [26], [27]. MPA adds a soft geometric constraint at the class level:

$$\mathcal{L}_{\text{MPA}} = \sum_{k=1}^{N}\left[\left\|c_k^I - c_k^{\text{fused}}\right\|_2^2 + \beta\left\|c_k^T - c_k^{\text{fused}}\right\|_2^2\right], \tag{9}$$

where β weights the textual term. MPA does not force the modality-specific representations to become identical, since that could remove the complementary information that multimodal learning is meant to exploit. It only asks their class geometries to stay mutually consistent, so each modality keeps its own content. This matters under domain shift, where full feature-level invariance risks discarding clinically meaningful signal.

I. Domain-Aware Prototype Augmentation

With only K support examples per class, the empirical prototype is a noisy estimate of the true target-domain class center, especially when K is small and the shift is large. DPA addresses this by moving each prototype with a learned, target-conditioned displacement. It works in three steps for each class k.

Step 1: summarize the target. A target-domain descriptor d_T is computed as the mean of the fused prototypes of all classes in the episode, $d_T = \frac{1}{N}\sum_{k=1}^{N} c_k^{\text{fused}}$. This gives the generator a coarse summary of where the target support set lies in the shared latent space.

Step 2: generate a displacement. A learnable generator G_ω receives the class prototype, the descriptor d_T, and a noise vector $z \sim \mathcal{N}(0, I)$, and outputs a displacement Δc_k.

Step 3: apply it. The adapted prototype is the original one plus a scaled displacement:

$$\Delta c_k = G_\omega(c_k^{\text{fused}}, d_T, z), \qquad \hat{c}_k = c_k^{\text{fused}} + \gamma\Delta c_k, \tag{10}$$

where γ controls the perturbation magnitude.

G_ω is a two-hidden-layer MLP with ReLU activations and hidden width 256. Its input is the concatenation $[c_k^{\text{fused}}; d_T; z] \in \mathbb{R}^{2d+128}$, where $d = 512$ is the shared latent dimension and z is 128-dimensional, and its output is $\Delta c_k \in \mathbb{R}^d$. Sampling z lets DPA represent a distribution of plausible class centers instead of a single point estimate, which can regularize the prototype against the sampling variability of small support sets [23], [32]. Because d_T is an input, the displacement can depend on the target distribution as seen through the support set instead of being arbitrary noise. Section IV-F tests this against an unconditional alternative.

J. Domain-Adversarial Adaptation

To reduce domain-specific bias that remains after prototype-level adaptation, MetaAlign adds a domain-adversarial objective. A domain discriminator $D_\phi(\cdot)$ learns to separate source from target representations, while a gradient reversal layer (GRL) between the encoder and the discriminator trains the encoder to make them indistinguishable. The GRL acts as an identity map in the forward pass and negates scaled gradients in the backward pass [9], [10]:

$$\mathcal{L}_{\text{DANN}} = -\mathbb{E}_{z_s}[\log D_\phi(z_s)] - \mathbb{E}_{z_t}[\log(1 - D_\phi(z_t))]. \tag{11}$$

D_ϕ is a single-hidden-layer MLP (Linear(512→256), ReLU, Linear(256→1)) that operates on the shared $d = 512$-dimensional latent representation and outputs one domain logit. Together with $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{MPA}}$ it forms the training objective in Eq. (2).

K. Test-Time Adaptation

At deployment, MetaAlign receives the labeled target-domain support set for each few-shot episode and performs lightweight, support-conditioned adaptation without retraining the primary feature encoders. The visual and auxiliary modality encoders remain frozen throughout target-domain adaptation. No gradient-based fine-tuning of these encoders is performed using the target support set.

For each episode, the support samples are first encoded using the frozen representations. Class prototypes are then constructed from the resulting support embeddings and refined using Multimodal Prototype Alignment (MPA) and target-conditioned Domain-Aware Prototype Augmentation (DPA). Target normalization statistics are updated using the available target support data to account for low-level distributional shift. Only these support-conditioned prototype-level components and normalization statistics are adapted; the primary feature encoder weights are not updated.

The resulting adapted prototypes are used to classify the query samples according to

$$\hat{y}_q = \arg\min_k d(z_q, \hat{c}_k), \tag{12}$$

where z_q denotes the query representation, $\hat{c}_k$ is the refined prototype for class k, and $d(\cdot, \cdot)$ denotes the prototype distance function. Query samples are withheld entirely from prototype construction and adaptation and are used only for evaluation.

No separate gradient-based optimization loop is performed at test time; consequently, there is no test-time optimizer, test-time learning rate, or encoder update step to tune. The Adam optimizer with a learning rate of 1×10^{-4} reported in Table I refers to the meta-training stage rather than to test-time adaptation. Thus, MetaAlign adapts to the target domain through support-conditioned prototype construction and refinement while keeping the primary feature encoders frozen. This design limits the number of degrees of freedom available to the adaptation procedure and reduces the risk of overfitting the very small K-shot support set.

L. Episodic Training and Evaluation Protocol

MetaAlign is trained and evaluated with episodic sampling that simulates low-data target adaptation. Each episode is an N-way, K-shot task with $K \in \{1, 3, 5\}$, and N depends on the benchmark: MIMIC-CXR $\rightarrow$ CheXpert [29], [30] is 5-way, MIMIC-CXR $\rightarrow$ MosMedData is 2-way, and ISIC 2019 $\rightarrow$ HAM10000 [35], [36] is 3-way. Each class contributes K labeled support samples and 15 query samples, so an episode has $N \times K$ support and $N \times 15$ query samples, with no overlap between them. Query samples are never used for prototype construction. For each benchmark and shot setting we sample 600 evaluation episodes and report mean accuracy with a 95% confidence interval.

M. Computational Considerations

MetaAlign is designed for practical deployment. All experiments run on a consumer-grade NVIDIA RTX 3050 with 4 GB of VRAM, which suggests that the method is accessible to groups without large GPU infrastructure. Four choices keep memory use low: images are processed at 224×224, the episodic batch contains one episode, text embeddings are precomputed and cached, and test-time adaptation touches only lightweight prototype-level components while the primary encoders stay frozen. That MetaAlign improves over competitive baselines under these constraints suggests that prototype-centric adaptation remains viable in resource-limited settings.

IV. Experiments and Results

A. Experimental Setup

We evaluate MetaAlign on three few-shot transfer benchmarks that probe different kinds of shift: institutional shift within one imaging modality, cross-modality transfer between different imaging physics, and cross-source variation within one clinical specialty [3], [33]. Evaluating only within-modality transfer could overstate how general the approach is. The cross-modality and cross-source settings test whether prototype-centric adaptation still helps when the source–target gap is large.

Datasets. MIMIC-CXR (about 5,000–10,000 curated frontal radiographs with paired reports and patient-level clinical metadata) [29] is the main source domain for Benchmarks 1 and 2. Radiographs are intensity-normalized and resized to 224×224, reports are tokenized and encoded offline with ClinicalBERT [13], and structured metadata are normalized and projected into the shared latent space. Source samples are used to build meta-training episodes, not to train a conventional supervised classifier, so the learned representation matches the episodic structure of target evaluation. CheXpert (Benchmark 1) [30] provides institutional shift within chest radiography: the task is the same, but patient demographics, acquisition hardware, labeling conventions, and disease prevalence differ. MosMedData (Benchmark 2; about 500–1,000 CT slices) [37] introduces a categorical modality shift from 2D radiography to axial CT. The selected CT slices are treated as 2D target inputs and passed through the same visual encoder interface as the source images after resizing and intensity normalization. MosMedData has no paired reports and lacks the source-domain tabular metadata, so the text and metadata pathways are masked during target evaluation. ISIC 2019 $\rightarrow$ HAM10000 (Benchmark 3) [35], [36] examines cross-source dermatological transfer. The ISIC 2019 collection itself incorporates HAM10000 together with BCN20000 and MSK images [38], so the two datasets are related by construction. Results on this benchmark therefore reflect cross-source adaptation, not independent external validation.

Class definitions and source-to-target mapping. To ensure consistency across the cross-domain benchmarks, the class definitions and source-to-target correspondences are fixed before episodic evaluation. The exact mappings used in the three benchmarks are illustrated in Fig. 2. For the MIMIC-CXR $\rightarrow$ CheXpert institutional-shift benchmark, five shared diagnostic categories are evaluated: Cardiomegaly, Edema, Pleural Effusion, Pneumonia, and Atelectasis. These categories use a one-to-one correspondence between the source and target domains under the shared label definitions.

For the MIMIC-CXR $\rightarrow$ MosMedData cross-modality benchmark, the task is formulated as a two-way classification problem. In the source domain, Normal corresponds to No Finding and Abnormal corresponds to Pneumonia-positive cases. In the target domain, Normal corresponds to CT-0, while Abnormal corresponds to CT-1 through CT-4. This constitutes a coarse binary correspondence rather than an identical diagnostic label mapping. Moreover, this benchmark represents a radiograph-to-CT modality shift; therefore, its results are interpreted as a cross-modality domain-adaptation evaluation and do not imply direct diagnostic equivalence between radiographs and CT images.

For the ISIC 2019 $\rightarrow$ HAM10000 cross-source benchmark, three corresponding diagnostic categories are used: MEL, NV, and BKL in ISIC 2019, and mel, nv, and bkl in HAM10000, respectively. These represent corresponding diagnoses across the two datasets.

The class mapping is fixed before episodic evaluation and is applied consistently to both support and query samples. No class definitions are modified according to individual episodes or query samples.

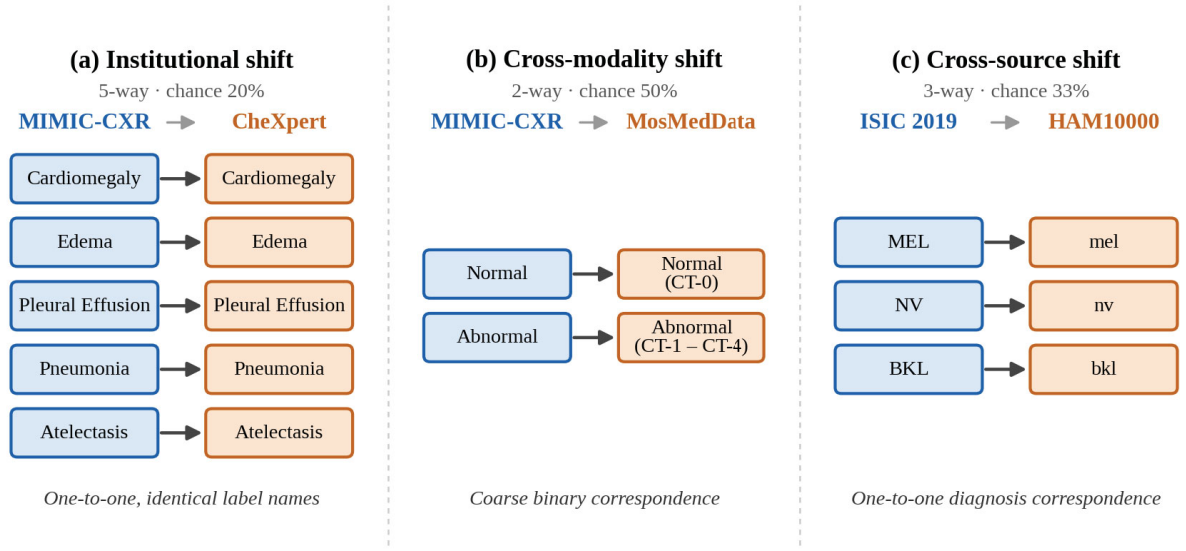


Fig. 2. Class definitions and source-to-target correspondences used in the three benchmarks: (a) institutional shift with five shared chest-radiograph categories, (b) cross-modality shift with a coarse binary correspondence, and (c) cross-source shift with three corresponding skin-lesion diagnoses. The mappings are fixed before episodic evaluation and are not changed between support and query sets.

Episodic evaluation. All evaluations follow the N-way, K-shot protocol described in Section III-L [4], [39]. The MosMedData pool of roughly 500–1,000 slices is small relative to 600 sampled evaluation episodes, so episodes for this benchmark are drawn with replacement across the evaluation run. Support and query sets are disjoint within each episode, but the same slice may appear in several episodes, as support in one and query in another. The 600 episodes are therefore not fully independent, and the confidence intervals for Benchmark 2 should be read as a lower bound on the true sampling uncertainty, not as an exact independent-sample estimate. We expect this concern to be smaller for Benchmarks 1 and 3, whose pools are larger.

Data-split considerations. Episodes are sampled at the level of individual images or slices. Patient-level separation between support and query samples, and between meta-training and evaluation pools, has not been verified in this study. Likewise, no dedicated image-level de-duplication between ISIC 2019 and HAM10000 was applied. Where a patient contributes several images, or an image appears in both collections, accuracies may be somewhat optimistic. We return to this point in Section V-D.

Statistical evaluation. For each benchmark and shot setting, 600 evaluation episodes were sampled, and we report mean accuracy with 95% confidence intervals computed over episodes [4]. To go beyond overlapping intervals, we also ran approximate two-sample z-tests using the reported means and the standard errors implied by the 95% intervals ($n = 600$). On this analysis MetaAlign's advantage over CrossDomain is large relative to the estimated sampling noise at every shot level on all three benchmarks ($p < 0.001$ in all nine comparisons), and the same holds against Multimodal ProtoNet on Benchmark 1. These p-values should be treated with caution. Per-episode paired predictions were not retained, so the tests cannot exploit pairing and treat the two methods as independent samples. We did not use paired bootstrap or permutation tests, and we did not resample at the patient or study level. Episodes drawn from a finite pool are also not fully independent. Each of these factors can make the reported p-values too small. We therefore regard them as approximate supporting evidence and base our conclusions mainly on the consistency of the gains across shot levels and benchmarks.

Baselines. MetaAlign is compared with ProtoNet [4], MAML [5], Meta-Baseline [6], FEAT [8], CrossDomain, and Multimodal ProtoNet (a prototypical baseline that receives the same image, text, and metadata inputs). These cover nearest-prototype, optimization-based, attention-based, domain-adaptation, and multimodal prototype approaches, and all are run under the same episodic protocol so that differences reflect the learning strategy and not the evaluation conditions. CrossDomain is our implementation of a domain-adversarial prototypical baseline: a ProtoNet backbone [4] with a domain discriminator and gradient reversal layer [9], [10], trained to align source and target features before prototype construction, following the feature-alignment strategy described in cross-domain few-shot benchmarking [3]. It receives the same multimodal inputs as MetaAlign (image, clinical text, and metadata, with unavailable modalities masked identically; Sections III-D and III-E) but combines them by simple concatenation instead of HGF. Unlike MetaAlign, it has no prototype-level alignment, augmentation, or learned gating, and prototypes are plain per-class means of the aligned, concatenated features. This design isolates prototype-centric adaptation and gated fusion as the sources of MetaAlign's gains.

Gaussian prototype augmentation baseline. To isolate the contribution of DPA's learned target conditioning, we build an

TABLE I. HYPERPARAMETER AND TRAINING CONFIGURATION

Parameter	Value	Description
β	0.4	MPA text alignment (Eq. 9)
γ	0.15	DPA perturbation (Eq. 10)
z noise	$\mathcal{N}(0, I)$, 128-dim	DPA generator input
λ_1	0.25	MPA loss weight (Eq. 2)
λ_2	0.05	DANN loss weight (Eq. 2)
Optimizer	Adam, lr $= 10^{-4}$	
Training eps	8,990	Meta-training
Eval. eps	600	Per benchmark
Batch size	1 episode	VRAM-constrained
Seed	42	

unconditional Gaussian perturbation baseline. For each prototype we draw isotropic noise $\boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \mathbf{I})$, normalize it to unit length, and scale it by the Euclidean norm of the corresponding DPA displacement:

$$\boldsymbol{\delta}_k^{\text{Gaussian}} = \|\boldsymbol{\delta}_k^{\text{DPA}}\|_2 \cdot \frac{\boldsymbol{\epsilon}_k}{\|\boldsymbol{\epsilon}_k\|_2}, \qquad \hat{c}_k^{\text{Gaussian}} = c_k^{\text{fused}} + \boldsymbol{\delta}_k^{\text{Gaussian}}.$$

The perturbation magnitude then matches DPA's per-prototype displacement exactly, and the baseline adds no trainable parameters. It tests whether the direction learned by DPA gives a benefit beyond generic stochastic perturbation.

Implementation. The visual encoder is a ResNet-50 [34] pretrained on ImageNet and fine-tuned during meta-training. ClinicalBERT [13] embeddings are precomputed offline and cached as fixed vectors, which removes repeated transformer forward passes and lowers per-episode memory use. Structured metadata pass through a two-layer MLP with ReLU activations. At test time the primary encoders are frozen and no parameters are updated; only the prototypes and normalization statistics are recomputed from the support set. All experiments run on a single NVIDIA RTX 3050 (4 GB VRAM). The hyperparameters are $\beta = 0.4$ (Eq. 9), $\gamma = 0.15$ (Eq. 10), 128-dimensional DPA noise from $\mathcal{N}(0, I)$, and loss weights $\lambda_1 = 0.25$ and $\lambda_2 = 0.05$ (Eq. 2). Table I lists them with the optimizer settings, episode counts, and random seed. They were chosen through preliminary validation-set tuning on MIMIC-CXR → CheXpert and held fixed for all evaluations. We did not run an exhaustive grid search, and a systematic sensitivity analysis is left for future work.

Model capacity. Excluding the shared ResNet-50 and ClinicalBERT backbones (frozen at test time), the adaptation layers contain no additional parameters for ProtoNet, MAML, and Meta-Baseline, a set-to-set Transformer for FEAT, 787K for Multimodal ProtoNet, 919K for CrossDomain, and 2.73M for MetaAlign. The last figure is under 11% of the ResNet-50 backbone and is concentrated in HGF, DPA, and the domain discriminator. These counts are derived from the architectures and are not measured from checkpoints.

B. Benchmark 1: MIMIC-CXR → CheXpert

Table II shows that MetaAlign has the highest accuracy at every shot level, with narrow confidence intervals over the 600 episodes. At 1-shot it exceeds CrossDomain (67.2%) by 4.3 points and ProtoNet (62.4%) by 9.1 points. This is the hardest regime, since one support example per class leaves almost no target-domain signal. The 9.1-point gap over ProtoNet splits into two parts. CrossDomain's 4.8-point gain over ProtoNet reflects feature-level domain adaptation [3], and MetaAlign adds a further 4.3 points, which is consistent with a benefit from adapting class geometry at the prototype level [7]. The two parts are similar in size, although this decomposition is descriptive and does not show that the two effects are independent. The comparison with Multimodal ProtoNet is informative: MetaAlign is 4.7 points higher at 1-shot with access to the same modalities, which suggests that the gain does not come simply from using more modalities. Accuracy rises from 71.5% at 1-shot to 82.1% at 5-shot. MetaAlign's margin over CrossDomain is 4.3, 4.3, and 3.2 points at 1, 3, and 5 shots, so the advantage is largest when support is scarcest and narrows somewhat as more support examples are added. Section IV-A describes the approximate significance analysis, whose limits should be kept in mind when reading these differences.

TABLE II. FEW-SHOT ACCURACY ON MIMIC-CXR → CHEXPERT (MEAN ± 95% CI, 600 EPISODES)

Method	1-shot	3-shot	5-shot
ProtoNet	62.4 ± 0.8	69.8 ± 0.7	74.1 ± 0.6
MAML	60.2 ± 1.1	68.5 ± 0.9	72.9 ± 0.7
Meta-Baseline	64.1 ± 0.7	71.3 ± 0.6	76.5 ± 0.5
FEAT	65.8 ± 0.8	72.1 ± 0.6	77.8 ± 0.5
CrossDomain	67.2 ± 0.7	73.6 ± 0.6	78.9 ± 0.4
Multimodal ProtoNet	66.8 ± 0.9	73.1 ± 0.7	77.2 ± 0.6
MetaAlign	$\mathbf{71.5 \pm 0.6}$	$\mathbf{77.9 \pm 0.5}$	$\mathbf{82.1 \pm 0.4}$

C. Benchmark 2: MIMIC-CXR → MosMedData

TABLE III. FEW-SHOT ACCURACY ON MIMIC-CXR → MOSMEDDATA (MEAN ± 95%CI; 2-WAY, CHANCE = 50%).

Method	1-shot	3-shot	5-shot
ProtoNet	45.0 ± 1.2	52.0 ± 1.0	56.0 ± 0.9
MAML	42.0 ± 1.5	50.0 ± 1.2	53.0 ± 1.0
Meta-Baseline	47.0 ± 1.0	54.0 ± 0.9	58.0 ± 0.8
FEAT	49.0 ± 1.1	55.0 ± 0.9	59.5 ± 0.7
CrossDomain	50.5 ± 1.0	56.5 ± 0.8	61.0 ± 0.7
Multimodal ProtoNet	48.5 ± 1.1	54.5 ± 0.9	58.5 ± 0.8
MetaAlign	$\mathbf{55.0 \pm 0.9}$	$\mathbf{61.0 \pm 0.7}$	$\mathbf{65.5 \pm 0.6}$

This is the most challenging benchmark in our study. Table III reports the results. Projection radiographs and volumetric CT slices differ in more than statistics: the imaging physics differ, the representations encode different physical quantities, and global feature alignment is unlikely to bridge this gap fully [3]. MetaAlign still ranks first, improving on CrossDomain by 4.5 points at both 1-shot and 5-shot and increasing monotonically with the number of shots while the text and metadata pathways are fully masked. This agrees with the main architectural argument: operating on class prototypes instead of feature distributions [27] targets a representation layer that may be better placed to retain transferable semantic structure when low-level appearance is incompatible. The absolute numbers call for caution, however. With two classes, chance level is 50%. MetaAlign's 1-shot accuracy of 55.0% is only modestly

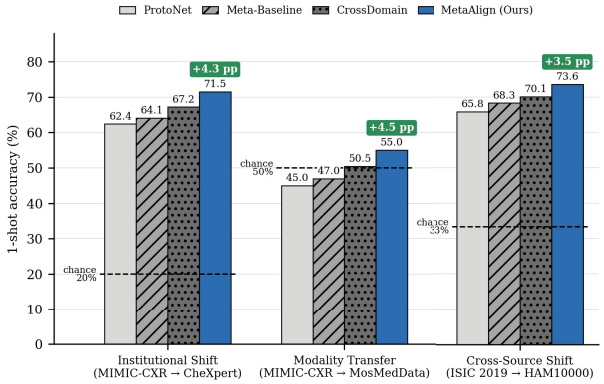


Fig. 3. 1-shot accuracy of ProtoNet, Meta-Baseline, CrossDomain, and MetaAlign on the three transfer benchmarks (institutional, cross-modality, and cross-source). Green labels give MetaAlign's gain in percentage points over CrossDomain, and dashed lines mark chance level for each task (20%, 50%, and 33% for 5-, 2-, and 3-way classification).

above chance, and the 1-shot accuracies of ProtoNet, MAML, Meta-Baseline, FEAT, and Multimodal ProtoNet are at or below it. At 1-shot this transfer therefore sits close to what a shared 2D encoder can support, and the relative gains should be read in that light. These results do not imply diagnostic equivalence between radiographs and CT, nor transfer without loss. They indicate that prototype-centric adaptation can extract whatever transferable class structure exists. Because the MosMedData pool is small, episodes are also not fully independent (Section IV-A), and the reported intervals should be read accordingly.

D. Benchmark 3: ISIC 2019 → HAM10000

TABLE IV. FEW-SHOT ACCURACY ON ISIC 2019 → HAM10000 (MEAN ± 95% CI).

Method	1-shot	3-shot	5-shot
ProtoNet	65.8 ± 1.0	72.4 ± 0.8	76.9 ± 0.6
Meta-Baseline	68.3 ± 0.9	74.7 ± 0.7	79.1 ± 0.5
CrossDomain	70.1 ± 0.8	76.2 ± 0.6	80.5 ± 0.4
MetaAlign	$\mathbf{73.6 \pm 0.7}$	$\mathbf{79.8 \pm 0.5}$	$\mathbf{83.7 \pm 0.3}$

As Table IV shows, MetaAlign leads CrossDomain by 3.5, 3.6, and 3.2 points at 1, 3, and 5 shots. These margins are similar to those on Benchmark 1, even though lesion morphology, color-based diagnostic cues, and imaging hardware are quite different. This consistency suggests that the prototype-centric mechanism is not tuned narrowly to radiological conventions. The confidence intervals tighten from ± 0.7 at 1-shot to ± 0.3 at 5-shot, as expected when more support samples stabilize the estimates. As noted in Section IV-A, the relationship between ISIC 2019 and HAM10000 [35], [36], [38] limits these results to cross-source adaptation, and we do not extend the conclusions beyond that. Fig. 3 compares 1-shot accuracy on the three benchmarks.

E. Ablation Study

Table V gives the component ablation. Removing HGF causes the largest drop, 3.7 points at 1-shot and 3.5 at 5-shot.

TABLE V. COMPONENT ABLATION ON MIMIC-CXR → CHEXPERT

Configuration	1-shot	3-shot	5-shot
Full MetaAlign	71.5 ± 0.6	77.9 ± 0.5	82.1 ± 0.4
w/o HGF	67.8 ± 0.9	73.2 ± 0.7	78.6 ± 0.6
w/o MPA	69.1 ± 0.8	75.4 ± 0.6	80.3 ± 0.5
w/o DPA	69.8 ± 0.7	76.1 ± 0.6	80.9 ± 0.5
w/o DANN	70.3 ± 0.7	76.8 ± 0.5	81.4 ± 0.4
HGF only	66.9 ± 1.0	72.8 ± 0.8	77.9 ± 0.7

TABLE VI. DPA MECHANISM ABLATION ON MIMIC-CXR → CHEXPERT; GAUSSIAN PERTURBATIONS ARE MAGNITUDE-MATCHED.

Variant	1-shot	3-shot	5-shot
No DPA	69.8 ± 0.7	76.1 ± 0.6	80.9 ± 0.5
Gaussian Prototype Augmentation	70.8 ± 0.7	77.1 ± 0.6	81.6 ± 0.5
Target-conditioned DPA (Ours)	$\mathbf{71.5 \pm 0.6}$	$\mathbf{77.9 \pm 0.5}$	$\mathbf{82.1 \pm 0.4}$

This suggests that fusion quality limits what MPA and DPA can refine: without effective multimodal fusion [18], [20], the prototype-level components work on an already degraded class representation. The HGF-only configuration (66.9% at 1-shot) is slightly below the w/o-HGF variant (67.8%), so MPA, DPA, and DANN together give at least as much as HGF alone, and no single mechanism accounts for the full gain. Removing MPA costs 2.4, 2.5, and 1.8 points at 1, 3, and 5 shots, removing DPA costs 1.7, 1.8, and 1.2 points, and removing DANN [9], [10] costs 1.2, 1.1, and 0.7 points. All three components therefore have their largest effect at the lower shot counts and a smaller one at 5-shot. One plausible reading is that class-geometric inconsistency and noisy prototype estimates matter most when few examples define each class, but our experiments do not test this explanation directly. The DANN effect is the smallest, yet it is consistent, which suggests that adversarial alignment [9], [10] works at a different representational level instead of duplicating the prototype-centric components. The full model has the highest mean accuracy in every row, although the confidence intervals of some reduced variants (for example w/o DANN) overlap with it, so differences of about one point should be regarded as suggestive.

Perturbation magnitudes. The mean Euclidean norm of the prototype displacement is approximately $\gamma = 0.15$ for both DPA and the Gaussian baseline (the value listed in Table I). Perturbation strength is therefore matched in the comparison below, and only the direction differs.

F. DPA Mechanism Validation

Table VI separates the effect of perturbation magnitude from that of learned target conditioning. Magnitude-matched Gaussian augmentation improves on no DPA by 1.0, 1.0, and 0.7 points at 1, 3, and 5 shots, and target-conditioned DPA adds a further 0.7, 0.8, and 0.5 points. With the same approximate z-test, the DPA–Gaussian gap is nominally significant at 3-shot ($z = 2.01$, $p = 0.045$) and not at 1-shot ($p = 0.137$) or 5-shot ($p = 0.126$). Three comparisons were made without correction, and the 3-shot value would not stay below 0.05 after a Bonferroni adjustment (threshold $0.05/3 \approx 0.017$). The unpaired, episode-level nature of the test adds further

uncertainty (Section IV-A). The consistent direction of the gap hints that learned target-aware displacement contributes beyond generic stochastic regularization, but the effect is small and the evidence is not conclusive.

G. Modality Availability Analysis

TABLE VII. MODALITY AVAILABILITY ANALYSIS (5-SHOT, MIMIC-CXR → CHEXPERT).

Configuration	Accuracy
Image Only	74.6 ± 0.8
Image + Tabular	79.3 ± 0.5
Image + Text	80.8 ± 0.4
All Modalities	82.1 ± 0.4
Text Only	63.2 ± 1.1
Tabular Only	58.7 ± 1.3

Table VII supports three observations. First, each auxiliary modality helps and accuracy rises as modalities are added. Tabular metadata adds 4.7 points over the image alone, clinical text adds 6.2 points, and using both adds 7.5 points. The combined gain is smaller than the sum of the individual gains (10.9 points), so the contributions are complementary but not additive, as one would expect when the two modalities partly carry overlapping information. The steady improvement is consistent with HGF integrating the extra information without letting one modality suppress the others [18]. Second, text-only (63.2%) and tabular-only (58.7%) configurations are well above the 20% chance level of the 5-way task but far below any image-based configuration, so neither auxiliary modality can replace the image. This is consistent with our choice of masking over imputation, since imputing from a weak signal could introduce misleading information. Third, the image-only configuration reaches 74.6%, which is higher than the image-only baselines ProtoNet (74.1%) and MAML (72.9%) in Table II, although it remains below Meta-Baseline, FEAT, and the two baselines that use all modalities. MetaAlign therefore tends to degrade gradually when auxiliary modalities are missing instead of failing outright. For sites without radiology reports or structured metadata, this may lower deployment risk, although some accuracy is lost. The contribution of DPA's target conditioning beyond stochastic perturbation is examined separately in Section IV-F (Table VI).

H. Calibration Analysis

Accuracy tells us whether a model is right, and calibration tells us whether its confidence can be trusted. This matters clinically wherever a confidence score decides whether a prediction is acted on or sent for human review [40]. A model that is systematically overconfident on its errors gives only a weak signal for deferral, whatever its overall accuracy. Fig. 4 compares ProtoNet, Meta-Baseline, and MetaAlign. In the reliability diagram all three methods fall below the diagonal, so each is somewhat overconfident, but MetaAlign is the closest to perfect calibration. Its ECE is 0.032, against 0.065 for Meta-Baseline and 0.087 for ProtoNet, which is about 51% lower than Meta-Baseline [40]. The figure covers these three methods only, so we make no claim about the other baselines. Domain-shifted models often become overconfident because their internal confidence signals no longer match the target distribution. A plausible explanation for MetaAlign's behavior is that building class boundaries directly from target support examples ties confidence to target-domain statistics instead of source-domain priors, although we did not test this explanation separately.

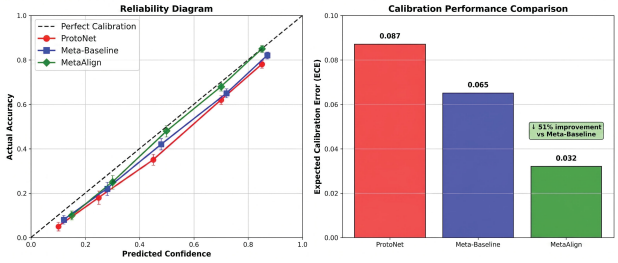


Fig. 4. Calibration analysis: reliability diagram and Expected Calibration Error (ECE).

I. Interpretability Analysis

We first examine prototype-based retrieval [4]. MetaAlign represents target-domain queries through their nearest class prototypes in the adapted embedding space, and matching query features to source-domain prototypes identifies the most similar source examples for each query. Fig. 5 shows retrieval results for two CheXpert queries (pleural effusion and cardiomegaly), each with its three nearest MIMIC-CXR cases and similarity scores.

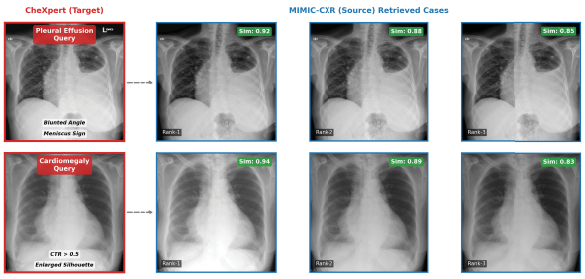


Fig. 5. Prototype-based retrieval of semantically similar source-domain cases.

The retrieved cases are not clinically equivalent to the queries, but the matches are consistent at the representation level. This suggests that the prototype space captures class-discriminative patterns that transfer across domains. We present these examples qualitatively and did not quantify retrieval quality.

Next, we use Grad-CAM [41] to visualize the image regions that contribute most to MetaAlign's predictions. Fig. 6 shows maps for four target-domain radiographs predicted as pleural effusion, cardiomegaly, pneumonia, and pulmonary edema.

In these examples the heatmaps concentrate on anatomically plausible regions such as the lung fields and cardiac silhouette, and not on the background or acquisition artifacts. This is not clinical validation, but it gives behavioral evidence that predictions are driven by relevant image content [41]. The maps are smooth and coarse, so they indicate the general region of

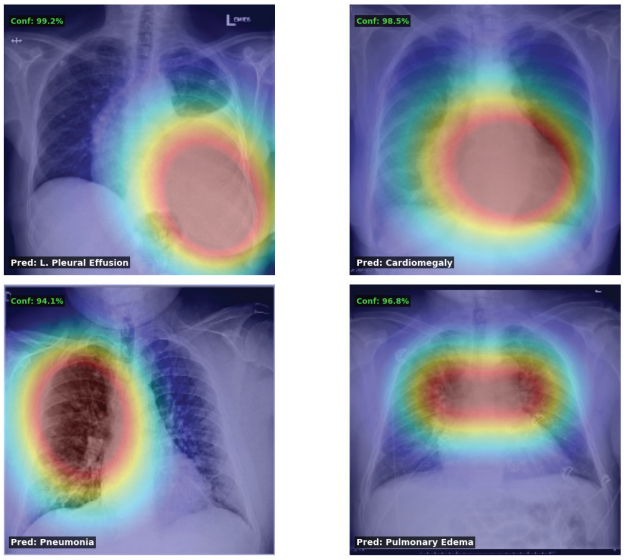


Fig. 6. Grad-CAM visualizations on representative target-domain radiographs.

interest and not lesion boundaries. Plausible attention does not guarantee diagnostic correctness, and a few representative maps cannot rule out spurious patterns elsewhere in the data.

V. Discussion

A. Main Findings

Across the three transfer settings, MetaAlign improves on feature-alignment and conventional multimodal prototype baselines, which is consistent with the hypothesis that few-shot adaptation benefits from preserving and adapting class geometry explicitly. HGF produces the largest single ablation effect, so reliable multimodal fusion appears to be an important basis on which MPA and DPA build. MPA, DPA, and DANN each help most at low shot counts. DANN adds a smaller but consistent gain by regularizing residual domain bias at the feature level [9], [10].

B. Missing-Modality and Cross-Modality Behavior

The modality study shows gradual degradation instead of collapse. Image-only MetaAlign reaches 74.6% at 5-shot, adding tabular or text information raises this to 79.3% and 80.8%, and all modalities give 82.1%. This favors masking unavailable modalities over imputing uncertain values [17]. The MIMIC-CXR $\rightarrow$ MosMedData experiment needs a narrower reading. Radiographs and CT encode different physical information, so the result does not establish diagnostic equivalence, and 1-shot accuracy is only modestly above chance. It shows that prototype-level adaptation can preserve some transferable class-semantic structure when low-level appearance differs substantially [27], [33].

C. Practical Implications

MetaAlign needs only a few labeled examples per class at the target site and avoids full-network retraining during deployment. Cached backbone representations and prototype-level adaptation keep the procedure computationally light, and explicit support-set prototypes give an interpretable decision structure. These properties help institutions with limited annotation capacity, incomplete multimodal infrastructure, or constrained hardware, but they do not imply clinical readiness.

D. Limitations and Future Work

This study has several limitations. The evaluation cohorts are small compared with real clinical diversity, and MIMIC-CXR is the only benchmark with the full multimodal configuration. ISIC 2019 $\rightarrow$ HAM10000 is cross-source transfer within a related ecosystem, and because ISIC 2019 includes HAM10000, image-level overlap between source and target cannot be excluded without explicit de-duplication. MosMedData episodes are not fully independent, since a small slice pool is reused across 600 episodes, and the cross-modality setting passes CT slices through a 2D encoder trained on radiographs. Patient-level separation of support and query samples was not verified. The significance tests are approximate: they are unpaired, lack episode-level paired predictions, and do not resample at the patient or study level, so the reported p-values may be too optimistic. Hyperparameters were tuned only preliminarily on MIMIC-CXR $\rightarrow$ CheXpert, and the parameter counts are architecture-derived estimates, not checkpoint measurements. We also did not run parameter-matched baselines. Future work should therefore use larger multi-institutional cohorts with strict external validation, patient-level splits and de-duplication, paired statistical testing (paired bootstrap or permutation tests with patient-level resampling), parameter-matched controls, patient-level dynamic missing modalities, federated prototype adaptation, and integration with medical foundation models.

VI. Conclusion

MetaAlign offers a unified prototype-centric approach to few-shot medical imaging transfer under domain shift, modality mismatch, and incomplete auxiliary inputs. HGF fuses the available modalities adaptively, MPA preserves cross-modal class geometry, DPA generates target-conditioned prototype variations, and domain-adversarial training suppresses residual domain bias. On institutional, cross-modality, and cross-source benchmarks, MetaAlign outperforms the evaluated baselines at every shot level in our experiments, and the component and Gaussian-control ablations are consistent with each proposed mechanism adding value.

The results suggest that adapting the class geometry used directly for few-shot decisions can work better than relying only on global feature-distribution alignment. Validation on larger external and prospective cohorts across diverse institutions and patient populations is still needed. Within the limits described above, the study presents prototype-centric adaptation as a computationally practical direction for robust transfer under scarce supervision.

References

[1] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, vol. 79, 2022.

[2] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar, "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, pp. 259–265, 2023.

[3] Y. Guo, N. C. Codella, L. Karlinsky, J. V. Codella, J. R. Smith, K. Saenko, T. Rosing, and R. Feris, "A broader study of cross-domain few-shot learning," in *European Conf. Computer Vision*, 2020, pp. 124–141.

[4] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4077–4087.

[5] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Machine Learning*, 2017, pp. 1126–1135.

[6] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 9062–9071.

[7] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *Int. Conf. Learning Representations*, 2019.

[8] H. Ye, H. Hu, D. Zhan, and F. Sha, "Few-shot learning via embedding adaptation with set-to-set functions," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 8808–8817.

[9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.

[10] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 7167–7176.

[11] M. N. Rizve, S. Khan, F. S. Khan, and M. Shah, "Exploring complementary strengths of invariant and equivariant representations for few-shot learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 10836–10846.

[12] R. Wang and G. Zheng, "PFMNet: Prototype-based feature mapping network for few-shot domain adaptation in medical image segmentation," *Computerized Medical Imaging and Graphics*, 2024.

[13] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proc. 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 72–78.

[14] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1399–1406, 2022.

[15] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[16] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "MedKLIP: Medical knowledge enhanced language-image pre-training for X-ray diagnosis," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023.

[17] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, "Deep multimodal learning with missing modality: A survey," *Transactions on Machine Learning Research*, 2026.

[18] F. Liu *et al.*, "Multimodal feature interactive learning for few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, 2025.

[19] A. Hasan, M. H. Jobayer, M. A. A. M. Pias, T. Alam, and R. Khan, "Bangla sign language recognition with multimodal deep learning fusion," *Engineering Reports*, vol. 7, no. 4, p. e70139, 2025, doi: 10.1002/eng2.70139.

[20] X. Zhao *et al.*, "DPCA-Net: Dual-prototype consistency alignment network for robust multimodal few-shot action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, 2025.

[21] Y. Zhao *et al.*, "Dual adaptive representation alignment for cross-domain few-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[22] Y. Zhang *et al.*, "Prototype correlation matching and class-relation reasoning for few-shot medical image segmentation," *IEEE Trans. Med. Imaging*, 2024.

[23] K. R. Hasan *et al.*, "Prototypical few-shot learning for histopathology classification," *IEEE Access*, 2025.

[24] Z. Li, C. Zhang, Y. Wang, W. Li, Q. Du, Z. Fang, and Y. Chen, "Cross-domain few-shot hyperspectral image classification with cross-modal alignment and supervised contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, 2024, doi: 10.1109/TGRS.2024.3407201.

[25] Z. Liu *et al.*, "Contrastive coarse-to-fine medical segmentation with prototype guidance," *Neurocomputing*, 2026.

[26] H. Liu, J. He, Y. Li, and Y. Bi, "Multilevel prototype alignment for cross-domain few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, 2025.

[27] K. Shi *et al.*, "Multi-Gaussian prototype metric with dual optimization for cross-domain few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, 2025.

[28] Y. Li *et al.*, "Semantic guided prototype learning for cross-domain few-shot hyperspectral image classification," *Expert Systems with Applications*, 2025.

[29] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, 2019.

[30] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 33, 2019, pp. 590–597.

[31] Z. Cheng *et al.*, "Dual interspersion and flexible deployment for few-shot medical image segmentation," *IEEE Trans. Med. Imaging*, 2025.

[32] L. Li *et al.*, "Robust few-shot segmentation using Fourier-guided style fusion and prototype alignment," *Biomedical Signal Processing and Control*, 2026.

[33] F. Shakeri, Y. Huang, J. Silva-Rodríguez, H. Bahig, A. Tang, J. Dolz, and I. Ben Ayed, "Few-shot adaptation of medical vision-language models," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024, arXiv:2409.03868.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[35] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomedical Imaging (ISBI)*, 2018, pp. 168–172.

[36] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, p. 180161, 2018.

[37] S. P. Morozov *et al.*, "MosMedData: Chest CT scans with COVID-19 related findings dataset," arXiv preprint arXiv:2005.07046, 2020.

[38] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, "BCN20000: Dermoscopic lesions in the wild," arXiv preprint arXiv:1908.02288, 2019.

[39] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, and K. Kavukcuoglu, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3630–3638.

[40] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, 2017, pp. 1321–1330.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.