Diagnostic Evidence-based Prototype Learning via Dirichlet Process for Multimodal Bone Tumor Subtyping

Feng Wu[1], Ruiling She[2], Ruiyang Zhang[1], Liang Peng[1], Qiang Ma[1], Yinghua Zhao[2] (✉), and Lequan Yu[1] (✉)

[1] The University of Hong Kong, Hong Kong SAR, China
{fengwu96,ruiyang_zhang}@connect.hku.hk, {larrypl,qiangma,lqyu}@hku.hk
[2] Southern Medical University, Guangzhou, China
ruilingshe@163.com, zhaoyh@smu.edu.cn

Abstract. Multimodal imaging is essential for accurate primary bone tumor subtyping, yet its clinical utility is limited by cross-modal diagnostic conflicts and frequent missing modalities. Existing deterministic multimodal fusion methods fail to model predictive uncertainty, often yielding over-confident predictions under incomplete inputs or inter-modality disagreements. In this paper, we propose DPDEP, a novel Dirichlet Process (DP)-driven diagnostic evidence-based prototype learning framework for multimodal bone tumor subtyping. Inspired by clinical differential diagnosis, DPDEP projects multimodal features onto modality-specific prototypes to capture heterogeneous diagnostic signatures. Instead of deterministic fusion, we introduce a DP-based evidential arbitration mechanism that dynamically quantifies prototype reliability and resolves inter-modality conflicts by prioritizing the most trustworthy evidence. To handle missing modalities, we introduce a DP-guided asymmetric evidential distillation strategy that transfers robust joint-modality reasoning from teacher to student by selectively dropping the more reliable modality without hallucinating missing-modal evidence. We collected a well-curated large-scale in-house bone tumor dataset comprising 4,443 clinical cases. Experiments demonstrate that DPDEP achieves the best performance among representative baselines under our PBT setting on both three-class malignancy grading and nine-class fine-grained histotype subtyping, with strong robustness and reliability in real-world incomplete multimodal settings. Our code is available at https://github.com/HKU-MedAI/PBT-DPDEP.

Keywords: Bone Tumor Classification · Incomplete Multimodal · Dirichlet Process · Prototype Learning

1 Introduction

Primary bone tumors (PBTs) are relatively uncommon, yet they are the third leading cause of cancer-related death among individuals under twenty [12, 15].

(✉) Corresponding authors.

Accurate malignancy grading and histological subtyping are critical for guiding individualized treatment [3, 4]. PBTs present complex and often overlapping radiographic features. As a result, a single imaging modality is frequently insufficient [18, 24]. A reliable radiographic diagnosis therefore depends on complementary information from multimodal imaging [22, 6]. In clinical practice, radiologists typically use Computed Tomography (CT) to assess matrix mineralization and periosteal reaction [1], while using Magnetic Resonance Imaging (MRI) to delineate soft-tissue masses and the extent of marrow involvement [7].

To leverage multimodal information, existing methods typically fuse features via concatenation [20] or cross-attention [26, 19], assuming all modalities are available. However, in clinical practice, missing modalities are common [25], for instance, MRI is frequently unavailable in primary bone tumor (PBT) diagnosis due to high cost, limited accessibility, or contraindications [2]. To handle incomplete inputs, recent works explore feature-level interventions, including late fusion [8], shared-space disentanglement [23], and self-distillation [13]. Moreover, attention and Transformer-based approaches dynamically aggregate available modalities by extracting modality-invariant features [27], utilizing region-aware attention [5], or employing masked mechanisms and indicators to explicitly bypass missing inputs [14, 18]. Despite their success, these approaches still rely on deterministic feature aggregation and provide limited interpretable diagnostic evidence. Prototype-based methods [28, 17, 9] improve interpretability by mapping features to typical diagnostic patterns, but they still rely on deterministic aggregation and do not explicitly assess evidence reliability or resolve cross-modal disagreements when modalities provide conflicting signals. Furthermore, their robustness and adaptability in real-world missing-modality scenarios remain underexplored.

To overcome these limitations, we propose `DPDEP`, which improves PBT subtyping reliability by learning prototype-level diagnostic evidence and arbitrating cross-modal fusion through a Dirichlet Process (DP)-driven evidential mechanism. The main contributions of this work are summarized as follows: 1) We introduce a semantically aligned diagnostic prototype learning module that projects multimodal features onto modality-specific prototypes to capture heterogeneous diagnostic signatures. 2) We formulate multimodal fusion as a dynamic evidential arbitration process. Using a DP stick-breaking prior, the proposed mechanism quantifies prototype reliability to prioritize trustworthy signals and suppress cross-modal conflicts. 3) We develop a DP-guided asymmetric evidential distillation strategy. By selectively dropping the more reliable modality during training, it transfers joint-modality reasoning from a teacher to a student network, enabling robust inference without hallucinating missing-modal evidence. 4) Extensive experiments on a well-curated large-scale in-house bone tumor dataset demonstrate consistent improvements on three-class malignancy grading and nine-class fine-grained histotype subtyping, validating robustness in real-world missing-modality settings.

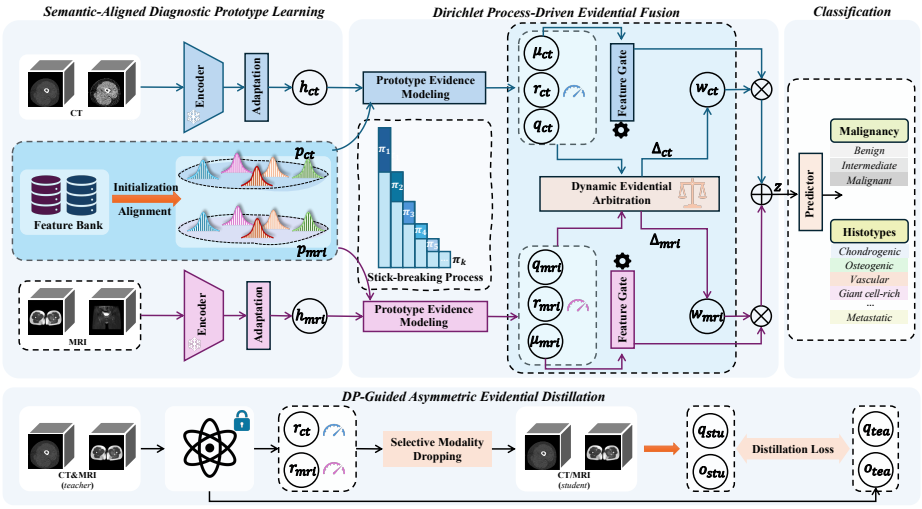


Fig. 1. Overview of DPDEP.

2 Methodology

Fig. 1 overviews DPDEP. It first projects multimodal features onto semantically aligned prototypes to encode heterogeneous diagnostic signatures. A Dirichlet Process (DP)-driven evidential fusion then estimates prototype reliability and arbitrates cross-modal conflicts. Finally, a DP-guided asymmetric distillation strategy transfers joint-modality reasoning from teacher to student for robust inference under missing modalities.

2.1 Problem Formulation

Given an incomplete multimodal dataset $\mathcal{D} = \{(\mathcal{X}_{ct}^{(i)}, \mathcal{X}_{mri}^{(i)}, \mathbf{m}^{(i)}, y^{(i)})\}_{i=1}^{N}$, our goal is to predict the tumor label $y \in \{1, \ldots, C\}$ under realistic modality missingness. Here, $\mathcal{X}_{ct}$ contains dual-window CT volumes (bone/soft-tissue), and $\mathcal{X}_{mri}$ denotes paired T1- and T2-weighted MRI sequences. The binary mask is $\mathbf{m} = (m_{ct}, m_{mri}) \in \{0,1\}^2$ with $\|\mathbf{m}\|_1 \geq 1$, where $m_s = 0$ indicates that modality s is unavailable. The model learns $p_\theta(y \mid \mathcal{X}_{ct}, \mathcal{X}_{mri}, \mathbf{m})$ without explicit missing-modality imputation.

2.2 Semantic-Aligned Diagnostic Prototype Learning

Prototype Initialization. For each modality $s \in \{ct, mri\}$ with two subsequences (e.g., $\mathcal{X}^{ct} = \{\mathcal{X}_{bone}^{ct}, \mathcal{X}_{soft}^{ct}\}$), we extract frozen 3D encoder features $\mathcal{E}_s$ (e.g., Hulu-Med [10]) and concatenate them as $\mathbf{x}^s = [\mathcal{E}_s(\mathcal{X}_1^s) \| \mathcal{E}_s(\mathcal{X}_2^s)] \in \mathbb{R}^{2d_f}$. After ℓ_2-normalization, we perform K-means independently for each modality on

the training set only to obtain initial centers $\{\tilde{\boldsymbol{\mu}}_k^{ct}\}_{k=1}^K$ and $\{\tilde{\boldsymbol{\mu}}_k^{mri}\}_{k=1}^K$. Cross-modal correspondence is established by Hungarian matching:

$$\pi^* = \arg\min_{\pi \in \Pi_K} \sum_{i=1}^{K} \left(1 - \cos(\tilde{\boldsymbol{\mu}}_i^{ct}, \tilde{\boldsymbol{\mu}}_{\pi(i)}^{mri})\right). \tag{1}$$

We reorder MRI prototypes as $\hat{\boldsymbol{\mu}}_i^{mri} = \tilde{\boldsymbol{\mu}}_{\pi^*(i)}^{mri}$ and keep $\hat{\boldsymbol{\mu}}_i^{ct} = \tilde{\boldsymbol{\mu}}_i^{ct}$. Both modalities are then projected into a shared d-dimensional space by PCA:

$$\mathbf{p}_k^s = \ell_2\big((\hat{\boldsymbol{\mu}}_k^s - \bar{\mathbf{x}})\mathbf{W}\big) \in \mathbb{R}^d, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{2d_f \times d}$ and $\bar{\mathbf{x}}$ are estimated from concatenated training-set prototype centers. These aligned prototypes initialize the learnable centers $\{\boldsymbol{\mu}_k^s\}_{k=1}^K$.

Feature Adaptation. Given the input CT and MRI images, we first extract their concatenated features $\mathbf{x}^s$ via the frozen encoders. To adapt these representations for the downstream task and bridge the domain gap, $\mathbf{x}^s$ is mapped into the shared prototype space using a trainable adaptation head $\mathcal{A}_s$ (implemented as a 1-layer MLP). This step outputs the adapted features $\mathbf{h}^s = \mathcal{A}_s(\mathbf{x}^s) \in \mathbb{R}^d$, effectively serving as the entry point for direct interaction with the learnable prototypes $\{\boldsymbol{\mu}_k^s\}$ in the subsequent DP-driven prototype evidence modeling.

2.3 Dirichlet Process-Driven Evidential Fusion

Given $\mathbf{h}^s$ and learnable prototype parameters $\{(\boldsymbol{\mu}_k^s, \boldsymbol{\sigma}_k^s, \lambda_k^s)\}_{k=1}^K$, we perform evidential fusion via DP-based evidence modeling and conflict-aware calibration. Each $\boldsymbol{\sigma}_k^s$ is constrained to be positive via softplus.

Prototype Evidence Modeling. We model prototype importance with a truncated stick-breaking process. For modality s, $\pi_k^s = v_k^s \prod_{j<k}(1 - v_j^s)$ for $k < K$, where $v_k^s = \sigma(\lambda_k^s)$, and $\pi_K^s = \prod_{j<K}(1 - v_j^s)$. The prototype evidence score is

$$\phi_k^s = -\frac{1}{2} \left\|(\mathbf{h}^s - \boldsymbol{\mu}_k^s)/\boldsymbol{\sigma}_k^s\right\|_2^2 - \mathbf{1}^\top \log \boldsymbol{\sigma}_k^s + \log \pi_k^s. \tag{3}$$

This yields the modality reliability $r_s = \log \sum_{j=1}^K \exp(\phi_j^s)$, and the prototype posterior $q_k^s = \exp(\phi_k^s - r_s)$. This mechanism dynamically favors diagnostically informative prototypes over uniform smoothing.

Dynamic Evidential Arbitration for Cross-Modal Conflicts. We summarize modality evidence as $\mathbf{e}_s = \sum_{k=1}^K q_k^s \boldsymbol{\mu}_k^s$. If both modalities are present, the conflict score is $c = 1 - \cos(\mathbf{e}_{ct}, \mathbf{e}_{mri})$, otherwise, we set $c = 0$. The modality interaction weight is

$$\alpha_s = \frac{m_s \exp(r_s)}{m_{ct} \exp(r_{ct}) + m_{mri} \exp(r_{mri})}, \quad s \in \{ct, mri\}. \tag{4}$$

We then define the calibrated reliability

$$\Delta_s = \begin{cases} \beta \log(\alpha_s)(1 + \gamma c), & m_s = 1, \\ 0, & m_s = 0, \end{cases} \tag{5}$$

where β and γ are the calibration strength and conflict amplification factor. Here, reliability calibration denotes evidence-level reliability adjustment for arbitration rather than calibration of final predictive probabilities. We set $\beta = \gamma = 1.0$ in all experiments.

Evidence-Weighted Diagnostic Representation Fusion. We modulate each prototype with a learnable gate $\mathbf{g}_k^s \in \mathbb{R}^d$: $\tilde{\boldsymbol{\mu}}_k^s = \boldsymbol{\mu}_k^s \odot \sigma(\mathbf{g}_k^s)$. This produces prototype tokens $\mathbf{t}_k^s = q_k^s \tilde{\boldsymbol{\mu}}_k^s$ and log-scores $\ell_k^s = \log q_k^s + b_k^s + \Delta_s$, where b_k^s is a scalar bias. The fused representation $\mathbf{z}$ is

$$\mathbf{z} = \sum_{s \in \{ct, mri\}} \sum_{k=1}^{K} w_k^s \, \mathbf{t}_k^s, \qquad w_k^s = \frac{m_s \exp(\ell_k^s)}{\sum_{s',k'} m_{s'} \exp(\ell_{k'}^{s'})}. \tag{6}$$

A 2-layer MLP classifier f_{cls} then maps $\mathbf{z}$ to logits $\mathbf{o} = f_{\text{cls}}(\mathbf{z}) \in \mathbb{R}^C$.

2.4 Dirichlet Process-Guided Asymmetric Evidential Distillation

To preserve evidential fusion under missing modalities, we apply in-network distillation only to *naturally paired* CT-MRI samples. A stop-gradient *teacher* pass uses the complete mask $\mathbf{m} = (1, 1)$ to produce reliabilities r_s^{tea}, prototype posteriors $q_k^{s,tea}$, fused representation $\mathbf{z}^{tea}$, and logits $\mathbf{o}^{tea}$. A *student* pass is then optimized on synthetically degraded masks $\mathbf{m}^{stu}$.

Selective Modality Dropping. During training, when both modalities are available, we sample a Bernoulli variable with probability p_{drop}. If activated, we drop the *more* reliable modality: $s_i^* = \arg\max_s r_s^{tea,(i)}, m_{s_i^*}^{stu,(i)} = 0$. This forces the student to infer from weaker evidence.

Uncertainty-Weighted DP Distillation. We define a reliability-weighted posterior

$$\hat{q}_{s,k} = \frac{m_s \, \sigma(r_s) \, q_k^s}{\sum_{s',k'} m_{s'} \, \sigma(r_{s'}) \, q_{k'}^{s'}}. \tag{7}$$

We then minimize an uncertainty-weighted KL divergence between teacher and student:

$$\mathcal{L}_{dp\text{-}dist} = \sum_{s,k} \tilde{u}_{s,k} \, \hat{q}_{s,k}^{tea} \left(\log \hat{q}_{s,k}^{tea} - \log \hat{q}_{s,k}^{stu} \right), \tag{8}$$

where $\tilde{u}_{s,k} \propto 1 - \sigma(r_s^{tea})$ and is normalized to mean 1. We further apply temperature scaled logit distillation:

$$\mathcal{L}_{kd} = \tau^2 \mathrm{KL}\big(\mathrm{softmax}(\mathbf{o}^{tea}/\tau) \,\|\, \mathrm{softmax}(\mathbf{o}^{stu}/\tau)\big). \tag{9}$$

2.5 Optimization Objective and Technical Details

The total loss is

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{kd}\mathcal{L}_{kd} + \lambda_{dp\text{-}dist}\mathcal{L}_{dp\text{-}dist} + \lambda_{div}\mathcal{L}_{div} + \lambda_{dp}\mathcal{L}_{dp}. \tag{10}$$

We use class-weighted cross-entropy for classification. The diversity loss $\mathcal{L}_{div}$ penalizes high cosine similarity among intra-modality gates. To regularize prototype allocation, we impose a Beta$(1, \eta)$ prior on the stick-breaking variables:

$$\mathcal{L}_{dp} = -(\eta - 1) \sum_{s,k} \log(1 - v_k^s), \tag{11}$$

encouraging a sparse prototype distribution. During training, $\mathcal{L}_{kd}$ and $\mathcal{L}_{dp\text{-}dist}$ are applied only to naturally paired samples, while $\mathcal{L}_{ce}$, $\mathcal{L}_{div}$, and $\mathcal{L}_{dp}$ are applied to all samples.

3 Experiments

3.1 Experimental Setup

Datasets. This study utilizes anonymized medical imaging data from the Third Affiliated Hospital of Southern Medical University, collected between January 2010 and March 2025. Inclusion criteria required a definitive clinical diagnosis of bone tumor and the availability of pre-treatment CT or MRI scans. We excluded cases involving post-treatment recurrence, orthopedic implants (e.g., bone cement or metallic hardware), or severe motion/respiratory artifacts. Following the 2020 WHO Classification [21], tumors were annotated by malignancy grade (benign, intermediate, or malignant) and histotypes. While the WHO defines ten histogenetic categories, we omitted the fibrogenic group due to insufficient samples for robust evaluation. The final dataset comprises 4,443 cases—3,092 CT-only and 1,351 paired CT-MRI—partitioned into training, validation, and testing sets with a 7:1:2 ratio. The detail dataset distribution is summarized in Table 1.

Evaluation Metrics. Model performance is quantified via AUROC and ACC for overall assessment. We additionally employ BACC and Macro-F1 to ensure a robust evaluation across all tumor types. Results are reported with the corresponding 95% confidence intervals based on 1,000 bootstrap iterations.

Table 1. Summary of the Datasets.

Task	Label	Total
Malignancy	Benign	1462
	Intermediate	399
	Malignant	2582
Histotypes	Chondrogenic	644
	Osteogenic	593
	Vascular	229
	Giant cell-rich	405
	Notochordal	67
	Other mesenchymal	308
	Hematopoietic	166
	Small round cell sarcomas	61
	Metastatic	1970
Total		4443

Implementation Details. We use a batch size of 16 and a learning rate of 5×10^{-4} with the Adam optimizer [11]. For the proposed objective, we set $\lambda_{kd} = 0.3$, $\lambda_{dp\text{-}dist} = 1.0$, $\lambda_{div} = 0.2$, $\lambda_{dp} = 0.5$, with distillation temperature $\tau = 1.5$ and selective drop probability $p_{\text{drop}} = 0.5$. We apply early stopping if validation AUROC does not improve for 15 consecutive epochs. All experiments are conducted on a single NVIDIA L40 GPU.

Table 2. Comparison between our method with other baselines on malignancy grade and histotypes classification. The **Best** and 2nd best methods under each setup are bold and underlined. The results are reported with 95% confidence intervals.

Model	Malignancy				Histotypes			
	AUROC(↑)	ACC(↑)	Macro-F1(↑)	BACC(↑)	AUROC(↑)	ACC(↑)	Macro-F1(↑)	BACC(↑)
Late Fusion [16]	83.80$_{(80.43,87.09)}$	75.12$_{(71.57,78.68)}$	62.74$_{(57.86,67.38)}$	64.54$_{(59.18,70.28)}$	81.55$_{(78.94,84.10)}$	49.76$_{(45.40,53.80)}$	26.36$_{(23.49,29.09)}$	34.81$_{(30.46,38.58)}$
Interleaved [26]	84.25$_{(80.98,87.44)}$	71.24$_{(67.53,74.80)}$	59.43$_{(55.09,63.81)}$	62.66$_{(57.34,68.08)}$	80.03$_{(76.96,82.99)}$	49.27$_{(44.75,53.31)}$	<u>29.62</u>$_{(25.55,33.16)}$	<u>37.57</u>$_{(32.27,42.73)}$
RFNet [5]	84.58$_{(81.44,87.53)}$	74.96$_{(71.40,78.51)}$	<u>62.80</u>$_{(58.26,67.14)}$	64.45$_{(59.47,69.69)}$	81.00$_{(78.19,83.92)}$	52.02$_{(47.82,55.74)}$	28.24$_{(24.72,31.39)}$	31.58$_{(26.27,37.88)}$
DrFuse [23]	85.59$_{(82.55,88.43)}$	75.61$_{(71.73,79.16)}$	61.43$_{(56.65,66.22)}$	61.83$_{(56.82,67.00)}$	82.03$_{(79.42,84.64)}$	53.47$_{(49.27,57.35)}$	28.26$_{(24.45,31.59)}$	31.08$_{(26.17,36.64)}$
Pasion [13]	<u>85.96</u>$_{(83.02,88.66)}$	73.99$_{(70.27,77.54)}$	<u>62.80</u>$_{(57.96,67.19)}$	<u>65.74</u>$_{(59.94,71.12)}$	81.61$_{(78.53,84.63)}$	<u>54.28</u>$_{(49.92,58.00)}$	29.13$_{(25.73,32.38)}$	33.21$_{(28.09,38.65)}$
mmFormer [27]	83.76$_{(80.32,87.02)}$	73.99$_{(70.11,77.54)}$	60.42$_{(56.03,64.90)}$	61.73$_{(56.76,66.87)}$	80.35$_{(77.22,83.25)}$	48.30$_{(43.78,52.02)}$	26.17$_{(22.41,29.37)}$	35.92$_{(30.75,40.56)}$
M2FTrans [14]	84.94$_{(81.81,88.08)}$	73.83$_{(69.95,77.38)}$	61.73$_{(57.10,66.19)}$	64.56$_{(59.43,69.74)}$	81.45$_{(78.82,83.94)}$	45.23$_{(41.03,49.27)}$	27.11$_{(23.69,30.78)}$	34.44$_{(28.39,41.03)}$
PBTC-TransNet [18]	84.29$_{(81.14,87.40)}$	<u>76.41</u>$_{(72.70,79.81)}$	59.88$_{(55.31,64.41)}$	60.14$_{(55.82,64.77)}$	<u>82.77</u>$_{(79.75,85.60)}$	52.18$_{(47.98,56.06)}$	27.67$_{(24.13,31.05)}$	34.38$_{(28.27,41.66)}$
DPDEP(Ours)	**87.39**$_{(84.82,89.83)}$	**77.71**$_{(74.31,80.94)}$	**64.80**$_{(59.74,69.89)}$	**66.45**$_{(61.09,71.72)}$	**83.06**$_{(80.54,85.61)}$	**55.25**$_{(51.21,59.13)}$	**35.41**$_{(31.46,39.20)}$	**40.25**$_{(33.78,47.21)}$
P-value (vs. Best)	< 1E-6	< 1E-6	< 1E-6	5.85E-03	2.42E-02	< 1E-6	< 1E-6	< 1E-6

Compared Methods. We compare our proposed DPDEPwith eight baselines: (1) Late Fusion [16], (2) Interleaved [26], (3) RFNet [5], (4) DrFuse [23], (5) Pasion [13], (6) mmFormer [27], (7) M2FTrans [14], and (8) PBTC-TransNet [18].

3.2 Comparison with Representative Baselines

Table 2 presents a detailed comparison between DPDEP and representative baselines, demonstrating our model's superior performance across all evaluated metrics for both malignancy grading and histotype subtyping tasks. All main results are evaluated on the full held-out test set under natural clinical missingness, using each case's true modality mask without synthetic dropping or imputation at inference. In malignancy grading, it achieves a leading 87.39% AUROC and 64.80% Macro-F1, significantly outperforming top baselines like Pasion. This improvement is consistent with the conflict-aware evidential arbitration mechanism. The advantage is even more pronounced in the challenging nine-class histotyping, where DPDEP surpasses the second-best results by substantial margins (Macro-F1: 35.41% vs. 29.62%; BACC: 40.25% vs. 37.57%). Unlike deterministic fusion-based methods that degrade under missing MRI, our DP-guided distillation successfully preserves joint-modality reasoning, ensuring robust inference with incomplete inputs.

3.3 Ablation Study

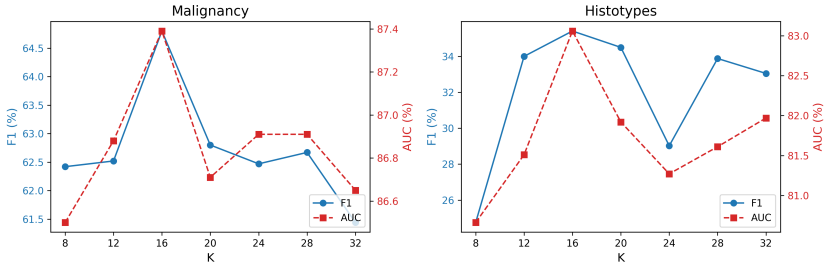
Analysis of Proposed Modules. Table 3 examines the impact of the proposed components within our framework. The integration of Diagnostic Prototype Learning (DPL) enhances feature representation, yielding consistent gains even in the unimodal CT-only setting. Incorporating DP-Driven Evidential Fusion (DPEF) further boosts performance by explicitly resolving cross-modal conflicts via dynamic evidential arbitration. Significantly, the full framework with DP-Guided Asymmetric Evidential Distillation (DPAED) outperforms the unimodal baseline by a large margin, confirming that our approach effectively leverages joint-modality information even when MRI is absent at inference.

Analysis of Hyperparameters. The influence of the prototype count K is assessed in Fig. 2. Optimal performance is observed at $K = 16$ for both tasks.

Table 3. Ablation study of the proposed modules. The **Best** and 2nd best methods under each setup are bold and underlined. The results are reported with 95% confidence intervals. "DPL" denotes the Diagnostic Prototype Learning. "DPEF" denotes the DP-driven Evidential Fusion. "DPAED" denotes the DP-guided Asymmetric Evidential Distillation. Here the baseline (with all ×) denotes 2 layer MLP in CT-only setting, and Late Fusion [16] in multimodal setting.

Model				Malignancy				Histotypes			
DPL	DPEF	DPAED	CT only	AUROC(↑)	ACC(↑)	Macro-F1(↑)	BACC(↑)	AUROC(↑)	ACC(↑)	Macro-F1(↑)	BACC(↑)
×	×	×	✓	$84.11_{(80.94,87.18)}$	$74.61_{(71.21,77.84)}$	$60.44_{(56.65,65.48)}$	$61.06_{(56.83,65.68)}$	$80.73_{(78.16,83.10)}$	$50.53_{(46.50,54.41)}$	$27.59_{(23.95,31.07)}$	$33.93_{(29.40,37.83)}$
✓	×	×	✓	$84.62_{(81.56,87.52)}$	$74.93_{(71.54,78.16)}$	$61.15_{(57.23,66.30)}$	$61.50_{(57.13,66.00)}$	$80.90_{(78.23,83.31)}$	$50.70_{(46.50,54.74)}$	$28.89_{(25.20,32.30)}$	$35.47_{(30.77,39.78)}$
×	×	×	×	$83.80_{(80.43,87.09)}$	$75.12_{(71.57,78.68)}$	$62.74_{(57.86,67.38)}$	$64.54_{(59.18,70.28)}$	$81.55_{(78.94,84.10)}$	$49.76_{(45.40,53.80)}$	$26.36_{(23.49,29.09)}$	$34.81_{(30.46,38.58)}$
✓	×	×	×	$85.41_{(82.63,88.06)}$	$76.58_{(73.02,80.13)}$	$63.03_{(58.62,67.25)}$	$65.85_{(61.34,70.31)}$	$82.08_{(79.59,84.75)}$	$52.83_{(48.47,56.70)}$	$29.99_{(26.40,33.57)}$	$35.17_{(29.63,41.58)}$
✓	✓	×	×	$\underline{85.63}_{(82.84,88.28)}$	$\underline{76.89}_{(73.01,80.44)}$	$\underline{63.67}_{(59.07,68.10)}$	$\underline{66.03}_{(60.77,71.07)}$	$\underline{82.24}_{(79.60,84.94)}$	$\underline{53.80}_{(49.76,57.67)}$	$\underline{33.68}_{(30.40,37.26)}$	$\underline{37.23}_{(33.23,42.05)}$
✓	✓	✓	×	$\mathbf{87.39}_{(84.82,89.83)}$	$\mathbf{77.71}_{(74.31,80.94)}$	$\mathbf{64.80}_{(59.74,69.89)}$	$\mathbf{66.45}_{(61.09,71.72)}$	$\mathbf{83.06}_{(80.54,85.61)}$	$\mathbf{55.25}_{(51.21,59.13)}$	$\mathbf{35.41}_{(31.46,39.20)}$	$\mathbf{40.25}_{(33.78,47.21)}$

Fig. 2. The effects of the number of prototypes K.



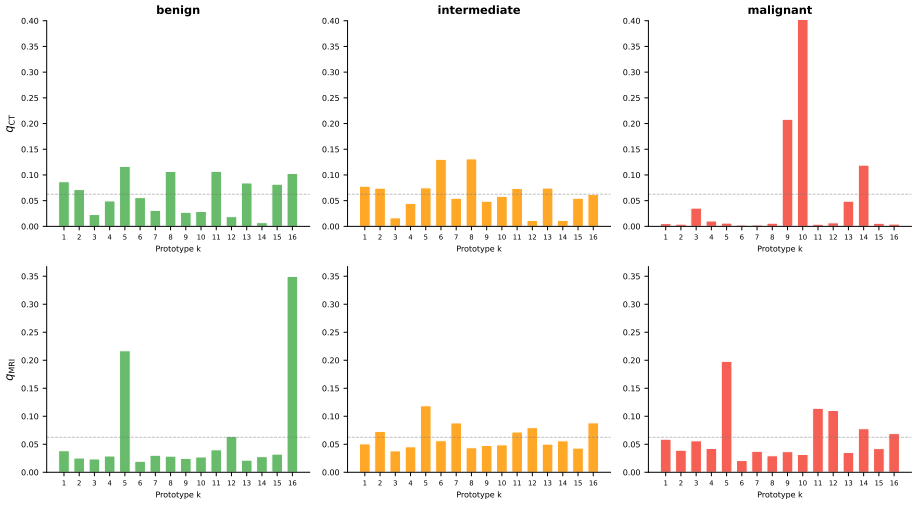
Performance degrades with fewer prototypes due to insufficient capacity to capture diverse tumor patterns, while an excessive number introduces redundancy that dilutes diagnostic focus. This suggests that $K = 16$ provides the most effective granularity, aligning with the clinical logic of differential diagnosis by offering distinct reference points for robust evidence arbitration.

Analysis of DP-Driven Evidence Allocation. Fig. 3 visualizes the posterior prototype assignments across modalities, revealing distinct sparsity and complementarity. Driven by the truncated stick-breaking prior of the Dirichlet Process, the model concentrates confidence on specific informative prototypes rather than spreading it uniformly. Notably, CT and MRI activate different prototype subsets for the same tumor class, indicating that DPDEP successfully arbitrates modality-specific diagnostic signatures rather than naively homogenizing them.

4 Conclusion

In this work, we propose DPDEP, a novel diagnostic evidence prototype learning framework for robust primary bone tumor subtyping. By integrating a Dirichlet Process-driven evidence arbitration mechanism with an asymmetric evidential distillation strategy, our approach maintains reliable diagnostic reasoning even with incomplete inputs. Extensive evaluations demonstrate its superior performance in both malignancy grading and histotype subtyping tasks, highlighting its promise for reliable decision support in incomplete multimodal settings. A

Fig. 3. DP-inferred posterior prototype assignments across modalities, demonstrating sparse and complementary diagnostic evidence allocation.



limitation of this study is that the cohort is from a single clinical center; external multi-center validation remains necessary future work.

Acknowledgments. This work was supported in part by the Research Grants Council of Hong Kong (27206123, 17200125, C5055-24G, and T45-401/22-N), the Hong Kong Innovation and Technology Fund (GHP/318/22GD, ITS/544/24FP), Guangdong Natural Science Fund (No. 2024A1515011875), the Natural Science Foundation of Guangdong Province (No. 2026A1515011880), and the National Natural Science Foundation of China (No. 81871510 and No. 82172014).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allen, H., Barnthouse, N.C., Chan, B.Y.: Periosteal pathologic conditions: imaging findings and pathophysiology. Radiographics **43**(2), e220120 (2022)
2. Bădilă, A.E., Rădulescu, D.M., Niculescu, A.G., Grumezescu, A.M., Rădulescu, M., Rădulescu, A.R.: Recent advances in the treatment of bone metastases and primary bone tumors: an up-to-date review. Cancers **13**(16), 4229 (2021)
3. Ciftdemir, M., Tuncel, S.A., Usta, U.: Atypical osteoid osteomas. European Journal of Orthopaedic Surgery & Traumatology **25**(1), 17–27 (2015)
4. Costelloe, C.M., Madewell, J.E.: Radiography in the initial diagnosis of primary bone tumors. American Journal of Roentgenology **200**(1), 3–7 (2013)
5. Ding, Y., Yu, X., Yang, Y.: Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3975–3984 (2021)

6. Do, B.H., Langlotz, C., Beaulieu, C.F.: Bone tumor diagnosis using a naïve bayesian model of demographic and radiographic features. Journal of digital imaging **30**(5), 640–647 (2017)
7. Gaume, M., Chevret, S., Campagna, R., Larousserie, F., Biau, D.: The appropriate and sequential value of standard radiograph, computed tomography and magnetic resonance imaging to characterize a bone tumor. Scientific Reports **12**(1), 6196 (2022)
8. Hayat, N., Geras, K.J., Shamout, F.E.: Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In: Machine Learning for Healthcare Conference. pp. 479–503. PMLR (2022)
9. Hussaindeen, A., Iqbal, S., Ambegoda, T.D.: Multi-label prototype based interpretable machine learning for melanoma detection. International journal of advances in signal and image sciences **8**(1), 40–53 (2022)
10. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Picci, P., Manfrini, M., Donati, D.M., Gambarotti, M., Righi, A., Vanel, D., Dei Tos, A.P.: Diagnosis of musculoskeletal tumors and tumor-like conditions: clinical, radiological and histological correlations-the Rizzoli case archive. Springer Nature (2019)
13. Shi, J., Shang, C., Sun, Z., Yu, L., Yang, X., Yan, Z.: Passion: Towards effective incomplete multi-modal medical image segmentation with imbalanced missing rates. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 456–465 (2024)
14. Shi, J., Yu, L., Cheng, Q., Yang, X., Cheng, K.T., Yan, Z.: Mftrans: Modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation. IEEE journal of biomedical and health informatics **28**(1), 379–390 (2023)
15. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA: a cancer journal for clinicians **73**(1), 17–48 (2023)
16. Snoek, C.G., Worring, M., Smeulders, A.W.: Early versus late fusion in semantic video analysis. In: Proceedings of the 13th annual ACM international conference on Multimedia. pp. 399–402 (2005)
17. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11566–11578 (2024)
18. Song, L., Li, C., Tan, L., Wang, M., Chen, X., Ye, Q., Li, S., Zhang, R., Zeng, Q., Xie, Z., et al.: A deep learning model to enhance the classification of primary bone tumors based on incomplete multimodal images in x-ray, ct, and mri. Cancer Imaging **24**(1), 135 (2024)
19. Theodorou, B., Glass, L., Xiao, C., Sun, J.: Framm: Fair ranking with missing modalities for clinical trial site selection. Patterns **5**(3) (2024)
20. Trong, V.H., Gwang-hyun, Y., Vu, D.T., Jin-young, K.: Late fusion of multimodal deep neural networks for weeds classification. Computers and Electronics in Agriculture **175**, 105506 (2020)
21. WHO Classification of Tumours Editorial Board: Soft tissue and bone tumours, vol. 3. World Health Organization (2020)
22. Winn, R.J., McClure, J.: The nccn clinical practice guidelines in oncology: a primer for users. Journal of the National Comprehensive Cancer Network **1**(1), 5–13 (2003)

23. Yao, W., Yin, K., Cheung, W.K., Liu, J., Qin, J.: Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 16416–16424 (2024)
24. Zeng, Q., Yang, Y.: Incomplete multimodal bone tumor image classification based on attention fusion and feature sharing. In: Proceedings of the 2024 5th International Conference on Intelligent Medicine and Health. pp. 32–39 (2024)
25. Zhang, C., Chu, X., Ma, L., Zhu, Y., Wang, Y., Wang, J., Zhao, J.: M3care: Learning with missing modalities in multimodal healthcare data. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. pp. 2418–2428 (2022)
26. Zhang, X., Li, S., Chen, Z., Yan, X., Petzold, L.R.: Improving medical predictions by irregular multimodal electronic health records modeling. In: International conference on machine learning. pp. 41300–41313. PMLR (2023)
27. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 107–117. Springer (2022)
28. Zhang, Y., Xu, Y., Chen, J., Xie, F., Chen, H.: Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In: The Twelfth International Conference on Learning Representations (2024)