diagnostics

MDPI

Article

PACE-MoE: A Plan-Adaptive Mixture-of-Experts Framework for Cost-Aware Multimodal Pubertal Development Assessment Under Modality and Resource Constraints

Hongbo Zhu [1,2], Yuhan Fu [2], Xintong Ma [2], Yuanguo Bi [1], Le Ha [3], Lin Li [4,*] and Xiaotong Wei [5,*]

[1] School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China; zhuhongbo@cse.neu.edu.cn (H.Z.); biyuanguo@mail.neu.edu.cn (Y.B.)
[2] School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China; fuyuhan122@sina.com (Y.F.); mxt121118@hotmail.com (X.M.)
[3] Department of Medical Engineering, General Hospital of Northern Theater Command of the Chinese People's Liberation Army, Shenyang 110016, China; halescut@126.com
[4] Department of Developmental Pediatrics, Shengjing Hospital of China Medical University, Shenyang 110004, China
[5] Institute of Health Professions Education Assessment and Reform, China Medical University, Shenyang 110134, China
* Correspondence: lilin@cmu.edu.cn (L.L.); xtwei@cmu.edu.cn (X.W.)

Abstract

Background/Objectives: Pubertal development assessment relies on multimodal evidence such as clinical text, laboratory tests, ultrasound, and X-ray. In clinical practice, inference costs vary across evidence sources, while real-world data often suffer from missing modalities and uneven sample distributions. To address these challenges, we develop a cost-efficient adaptive expert-selection framework that maintains diagnostic performance under resource constraints. **Methods and Materials:** We propose PACE-MoE (Plan-Adaptive Cost-Efficient Mixture-of-Experts), a two-stage dynamic MoE architecture. Stage 1 invokes only a symptom expert for low-cost screening. In Stage 2, candidate expert plans are generated via structural masking and stochastic combinatorial sampling, filtered under modality and resource constraints, and scored by a resource-aware deep reinforcement learning (DRL) router to select the highest-scoring feasible plan. Predictions from both stages are combined via logit-space weighting. We evaluated the model on 1249 pubertal development cases (970 females, 279 males), which include clinical text, laboratory tests, ultrasound, and X-ray with partial modality missingness, using five-fold cross-validation and additional ablation, resource configuration, and robustness experiments. **Results:** PACE-MoE reduced the normalized expert-invocation proxy cost from 3.7297 to 1.4836 (a 60.2% reduction) compared with Always-Upgrade (which invokes the complete expert set for every case), while maintaining a comparable F1-score (0.8255 vs. 0.8368). At 90% optional-modality missingness, with clinical text retained, the F1-score was 0.7079 and the Mean Cost was 0.6513. **Conclusions:** PACE-MoE provides a two-stage dynamic MoE framework for cost-aware pubertal development assessment, reducing the normalized expert-invocation proxy cost while maintaining comparable diagnostic performance. Due to the single-center retrospective data and proxy-based cost modeling, multi-center validation remains necessary.

Keywords: pubertal development assessment; multimodal clinical inference; Mixture-of-Experts; cost-aware inference; modality and resource constraints

Check for updates

Academic Editor: Jae-Ho Han

Received: 30 September 2026
Revised: 12 November 2026
Accepted: 13 November 2026
Published: 23 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Pubertal development assessment represents a critical clinical task in pediatric endocrinology, aiming to identify developmental abnormalities such as precocious puberty in a timely manner to guide treatment. In real-world healthcare systems, medical institutions can differ substantially in computational capacity and diagnostic resources: while tertiary hospitals may have access to high-performance computing and comprehensive imaging infrastructure, primary care clinics often operate with more limited computing resources and less direct access to high-cost examinations. This disparity makes conventional multimodal models—which assume complete inputs and rely on resource-intensive local fine-tuning—difficult to deploy across resource-constrained settings. Clinically, a definitive diagnosis requires integrating heterogeneous evidence, where bone age assessment and pelvic/gonadal ultrasound serve as common diagnostic aids [1]. Clinical studies indicate that precocious puberty diagnosis relies on comprehensive evaluations combining developmental signs, endocrine tests, and imaging findings [2]. Consequently, pubertal development assessment inherently possesses multimodal and resource-sensitive decision-making attributes.

Recently, deep learning has achieved significant progress in medical image analysis and clinical prediction tasks. Convolutional neural networks have demonstrated expert-level diagnostic performance in tasks including skin lesion identification, chest radiograph abnormality detection, and breast cancer screening [3]. Concurrently, the growth of large-scale clinical databases and medical imaging resources has further propelled medical artificial intelligence research, with clinical databases like MIMIC-III providing a foundational basis for modeling clinical text and structured variables [4]. These efforts demonstrate that data-driven models can learn effective representations from complex medical data to facilitate assisted screening and risk prediction.

Compared to single-modality modeling, multimodal learning integrates complementary information from diverse sources. Multimodal machine learning literature typically identifies heterogeneous representation learning, cross-modal alignment, information fusion, and missing-modality robustness as core challenges [5]. In clinical scenarios, textual records reflect presenting concerns, medical history, and physical-examination findings; laboratory tests provide objective physiological indicators; and medical imaging provides complementary anatomical and developmental information [6].

However, real-world clinical data rarely satisfy the ideal assumptions of modality completeness, consistent quality, and synchronized examinations. On one hand, individual patient examinations vary, leading to inherently missing modalities across laboratory tests, ultrasound, and X-ray images; on the other hand, the acquisition, preprocessing, and inference costs differ across modalities. To tackle the problem of incomplete modalities, various robust multimodal learning methods have been proposed. ModDrop enhances model adaptability to missing inputs by randomly dropping modalities during training [7]; HeMIS targets heterogeneous-modality medical image segmentation, leveraging shared statistical representations to alleviate input inconsistencies arising from diverse modality combinations [8]; and SMIL further addresses multimodal representation learning in scenarios with severe missing modalities [9]. Although these methods effectively enhance model robustness under missing modalities, their primary objective remains achieving fusion given available modalities, with insufficient consideration given to the invocation cost and budget control of different modalities during the inference phase.

From a deployment perspective, fixed invocation of all available modalities is not always reasonable. For screening-oriented medical tasks, systems typically need to perform rapid initial screening across large-scale populations. Invoking the complete set of high-cost experts for all samples not only escalates the average inference overhead but also potentially

reduces model deployability in resource-constrained scenarios. Cost-sensitive feature acquisition offers relevant insights into this problem. Joint Active Feature Acquisition (JointAFA) jointly optimizes active feature acquisition and classification objectives, enabling models to decide whether to acquire new features based on observed characteristics [10]; methods like Classification with Costly Features (CWCF) further model the feature selection process as a reinforcement learning problem to balance predictive benefit against feature cost [11]. These approaches demonstrate that not all information must be indiscriminately invoked during prediction; however, their decision targets are typically raw features or feature groups, making them difficult to directly adapt to medical expert systems characterized by multiple modalities, multiple experts, and complex fusion paths.

Similar to cost-sensitive feature acquisition, dynamic inference research focuses on allocating computational resources on a per-sample basis. Adaptive Computation Time provides an early concept of dynamic computation steps for sequential models [12]; BranchyNet enables simple samples to exit early by embedding multiple exit branches within the network [13]; MSDNet further leverages multi-scale dense connections to support resource-efficient classification [14]. Additionally, SkipNet learns to skip specific network modules per sample [15], while BlockDrop learns dynamic execution paths within residual networks [16]. Although these methods reduce redundant computation inside a single model, they primarily focus on computational pruning at the level of network depth, modules, or residual blocks. Conversely, clinical multimodal scenarios require decision-making and budget control regarding whether to invoke additional modalities and which high-cost experts to select.

More recent multimodal methods have begun to explore sample-adaptive routing and expert weighting under incomplete inputs. DynMM introduces data-dependent multimodal fusion paths to reduce redundant computation during inference, while SimMLM proposes a dynamic mixture of modality experts together with a more-versus-fewer ranking constraint to improve robustness under missing modalities [17,18]. Nevertheless, these methods still do not explicitly address clinically constrained expert invocation budgets, feasibility filtering under modality availability, or structured review-path selection in a two-stage diagnostic workflow.

The Mixture-of-Experts (MoE) model provides a natural framework for sample-level conditional computation. Classical MoE assigns distinct expert weights to different inputs via a gating network, allowing the model to selectively leverage expert knowledge conditioned on sample features [19]. Subsequently, Sparsely-Gated MoE introduces sparse gating mechanisms into large-scale neural networks, enhancing model capacity while controlling computational overhead by activating only a subset of experts [20]. GShard and Switch Transformer further advance the application of sparse expert models in large-scale pre-training, demonstrating the scaling benefits of conditional computation [21,22]. In vision tasks, V-MoE also validates the effectiveness of sparse expert routing in image modeling [23]. However, existing MoE research primarily focuses on model capacity expansion, training stability, and large-scale computational efficiency, lacking explicit modeling of cost differences across medical multimodal branches, modality availability constraints, and budget-controllable inference behavior.

For medical assisted diagnosis systems, classification accuracy alone is insufficient. Model output probabilities must also possess reliability to support threshold selection, risk stratification, and the review of boundary cases. Existing studies indicate that although deep neural networks possess strong discriminative capabilities, their predictive probabilities suffer from overconfidence or miscalibration [24]. Temperature scaling is a simple and effective calibration method that improves the alignment between predictive probabilities and empirical accuracy without altering the ranking capability of the model [24]. Mean-

while, MC Dropout serves as a Bayesian approximation to estimate model uncertainty [25], and Deep Ensembles further enhance the stability of uncertainty estimation [26]. Selective prediction methods also explore how to identify hard samples and allow the model to reject or defer decisions from a risk-control perspective [27]. These studies lay the foundation for review mechanisms in medical screening but are generally not unified with multimodal expert invocation, cost budgets, and expert combination selection into a cohesive dynamic inference framework.

In summary, while existing studies provide important foundations from the perspectives of missing-modality learning, cost-sensitive feature acquisition, dynamic inference, and Mixture-of-Experts (MoE) models, gaps remain in real-world clinical multimodal diagnostic scenarios. First, many multimodal methods assume that modality invocation costs during inference are negligible, lacking explicit modeling of heterogeneous expert costs. Second, dynamic inference methods mostly concentrate on computational pruning within a single network rather than sample-adaptive invocation across different medical modality experts. Finally, although cost-sensitive feature acquisition considers budget constraints, it typically does not explicitly address candidate plan generation, modality availability constraints, and feasible expert-plan selection within a multimodal expert pool. Therefore, an important remaining problem is how to perform budget-aware and sample-adaptive expert-plan selection under missing modalities while maintaining deployability across settings with different local resource capacities.

To tackle these problems, we propose PACE-MoE (Plan-Adaptive Cost-Efficient Mixture-of-Experts), a two-stage dynamic mixture-of-experts framework for pubertal development assessment under resource constraints and missing modalities. The framework first invokes a low-cost symptom text expert for Stage-1 screening, after which all samples proceed to Stage 2 for dynamic evidence review. In Stage 2, candidate expert plans are generated through structural masking and stochastic combinatorial sampling, filtered by Operations Research (OR)-based constraints under modality availability and resource limitations, and evaluated by a resource-aware router to select the highest-scoring feasible plan for each sample. Finally, the Stage-1 and Stage-2 predictions are fused in the logit space. With frozen backbone encoders, PACE-MoE performs sample-adaptive expert-plan selection using lightweight expert heads and a resource-aware routing module.

The main contributions of this paper are as follows:

- We propose PACE-MoE, a two-stage dynamic mixture-of-experts framework that separates low-cost symptom screening from adaptive evidence review and enables sample-level selection of compact or enriched expert plans according to the available evidence and resource constraints.
- We design a combinatorial candidate plan generation mechanism together with OR-based constraint filtering. The former generates diverse expert plans through structural masking and stochastic combinatorial sampling without enumerating the full combinatorial space, while the latter explicitly incorporates modality availability, mandatory expert requirements, expert-count limits, and budget constraints to construct a feasible candidate set for subsequent plan scoring.
- We develop a resource-aware feasible-plan scoring and selection mechanism and combine it with two-stage logit-space fusion. The scoring mechanism balances diagnostic utility against the normalized expert-invocation proxy cost to support sample-adaptive selection from the feasibility-filtered candidate set, while logit-space fusion integrates the initial screening result with the selected review evidence.

2. Methods

2.1. Problem Definition

Let the i-th sample be composed of multiple modalities, including clinical notes, laboratory tests, ultrasound images, and X-ray scans. Due to the presence of missing modalities in real-world clinical data, the modality availability of sample i is denoted as a_i^m, where $a_i^m = 1$ indicates that the m-th modality is available, and $a_i^m = 0$ indicates that the modality is missing. The system comprises N expert models, denoted as $E = \{E_1, E_2, \ldots, E_N\}$. Each expert corresponds to a single modality or a cross-modal combination input and outputs a binary classification prediction. To characterize the relative resource consumption of different experts, we assign a fixed execution cost c_j to each expert E_j. This cost approximates the preprocessing, encoding, and inference overhead required for expert invocation, serving as a unified proxy for resource comparison across different routing strategies. For sample x_i, the set of experts actually invoked by the model is denoted as $S_i \subseteq E$, with the model finally outputting the predictive probability p. The objective of this study is to balance diagnostic performance and the average normalized expert-invocation proxy cost by selecting additional experts according to sample-specific evidence availability and feasibility constraints.

2.2. Overall Framework

The proposed PACE-MoE framework adopts a two-stage inference pipeline consisting of low-cost symptom screening and dynamic evidence review, as illustrated in Figure 1. In Stage 1, only the symptom expert is invoked to obtain a preliminary prediction at a low inference cost. All samples subsequently proceed to Stage 2, where candidate expert plans are generated, filtered according to feasibility constraints, and evaluated by a resource-aware router to determine the expert combination used for evidence review.

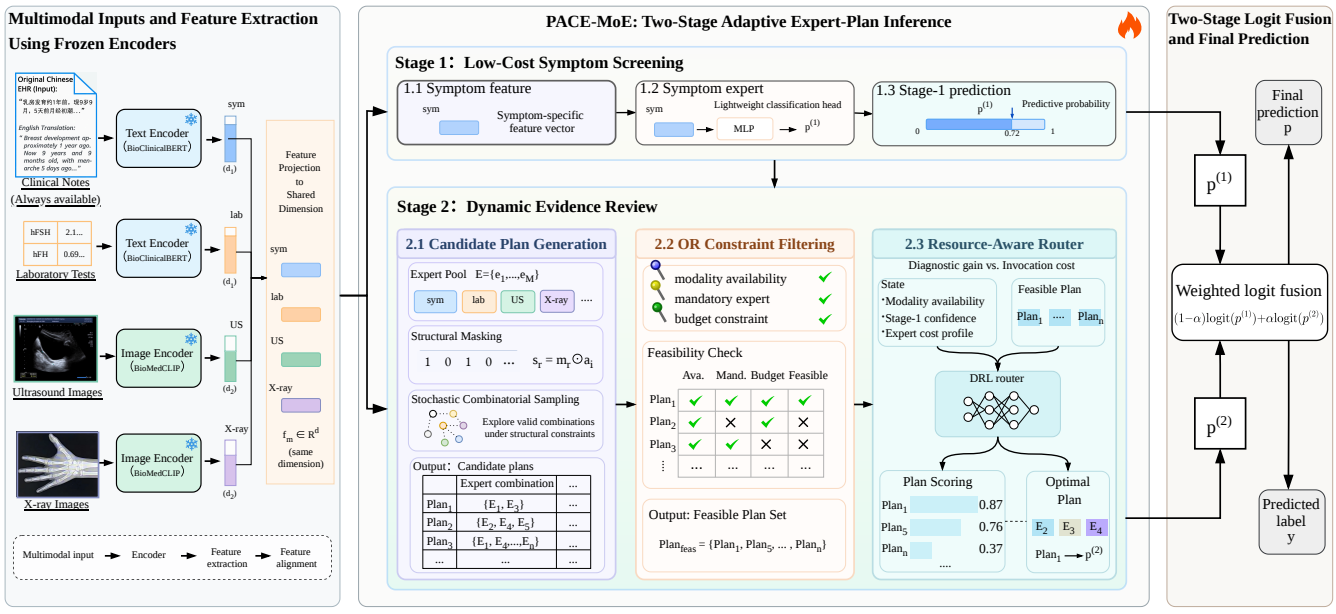


Figure 1. Overview of the proposed PACE-MoE framework. The framework comprises a two-stage pipeline: Stage 1 performs low-cost symptom screening, and Stage 2 generates candidate expert plans, filters them under feasibility constraints, and selects the final expert combination with a resource-aware router. The predictions from the two stages are fused in the logit space. All feature encoders are frozen during routing. Arrows indicate the direction of information flow, and ellipses (...) represent additional candidate plans or combinations not shown individually.

In Stage 1, the clinical text is encoded to obtain a symptom-specific feature representation, which is then fed into a lightweight classification head to produce the Stage-1 prediction $p^{(1)}$. Since only the symptom expert is used, this stage provides a low-cost preliminary assessment for the subsequent evidence review.

In Stage 2, PACE-MoE performs dynamic evidence review through three consecutive steps: candidate plan generation, OR-based constraint filtering, and resource-aware plan selection. Candidate expert combinations are first generated through structural masking and stochastic combinatorial sampling. The generated plans are then filtered according to modality availability, mandatory expert requirements, and resource constraints. Finally, the resource-aware router evaluates the feasible plans and selects the plan with the highest routing score for the current sample, producing the Stage-2 prediction $p^{(2)}$.

Finally, the predictions from the two stages are fused in the logit space to obtain the final prediction.

2.3. Expert Models and Modality Representations

A decoupled implementation is adopted to separate pre-trained feature encoders from lightweight expert classification heads. The multimodal feature encoders remain frozen during the training of the expert heads and routing strategy. Heterogeneous inputs are first mapped into vector representations by the frozen encoders and are subsequently processed by the corresponding lightweight expert heads.

For imaging modalities, BioMedCLIP is utilized to extract medical image features. BioMedCLIP inherits the vision-language contrastive learning paradigm of CLIP and is further pre-trained on large-scale biomedical image-text pairs, making it highly effective for semantic representation learning in medical imaging [28,29]. Ultrasound and X-ray images are processed by the vision encoder to derive respective representations, which are then fed into the corresponding imaging experts. For clinical text, BioClinicalBERT is employed to extract semantic representations. This architecture built upon Transformer and BERT [30,31] undergoes domain adaptation via clinical and biomedical corpora, capturing contextual semantics in clinical narratives [32–34]. For laboratory test data, structured laboratory panels are aggregated into textual summaries and encoded by BioClinicalBERT, transforming laboratory measurements into semantic features aligned with clinical text.

The expert models utilize lightweight classification heads. Single-modality experts yield predictions conditioned on individual modal features, whereas fusion experts output cross-modal predictions based on multi-modality feature combinations. As the focus lies on dynamic inference and resource-aware routing, all baseline methods share identical feature encoders and the expert pool; discrepancies reside primarily in candidate plan generation, constraint filtering, routing selection, and output fusion mechanisms.

2.4. Candidate Review Plan Generation

In the second stage, the number of possible expert combinations grows exponentially with the size of the expert pool, making exhaustive enumeration increasingly inefficient. Therefore, we formulate stage-two review as a candidate plan selection problem over a finite and structured candidate space.

To construct this candidate space, we use structural binary masking together with stochastic combinatorial sampling. Rather than enumerating all possible expert combinations, the sampler combines a set of predefined candidate plans with randomly sampled binary masks, while respecting sample-specific modality availability and structural constraints. Let $\mathbf{E} = [E_1, \ldots, E_N]^T$ denote the expert vector and $\mathbf{a}_i \in \{0,1\}^N$ denote the modality-availability vector of sample i. Guided by $\mathbf{a}_i$ and the structural constraints, candi-

date plans are represented by a binary selection matrix $\mathbf{M} \in \{0,1\}^{R\times N}$. The r-th candidate plan P_r is defined as:

$$P_r = (\mathbf{s}_r, \mathbf{m}_r^{\text{mand}}), \tag{1}$$

where $\mathbf{s}_r = \mathbf{m}_r \odot \mathbf{a}_i \in \{0,1\}^N$ is the executable selection vector, with $\mathbf{m}_r$ denoting the r-th row of $\mathbf{M}$ and $\odot$ denoting the Hadamard product. The vector $\mathbf{m}_r^{\text{mand}}$ denotes the mandatory expert mask. This formulation restricts candidate plans to the modalities available for the current sample and provides a structured candidate space for subsequent feasibility filtering and plan selection.

2.5. Feasible Plan Filtering

After candidate plans are constructed under sample-specific modality availability, some plans may still violate the remaining execution constraints. Different experts incur varying execution costs, and each review plan must satisfy mandatory expert requirements, expert-count limits, and budget constraints. Therefore, we perform constraint-aware feasibility filtering and pruning to obtain an executable candidate set for each sample.

For a candidate plan P_r, the executable selection vector $\mathbf{s}_r$ is already restricted by the modality availability vector $\mathbf{a}_i$. The plan must additionally satisfy the mandatory expert requirements; specifically, $\mathbf{s}_r$ must satisfy:

$$\mathbf{m}_r^{\text{mand}} \leq \mathbf{s}_r \leq \mathbf{m}_r, \tag{2}$$

where $\leq$ denotes element-wise inequality. The lower bound requires all mandatory experts to be included, while the upper bound prevents experts outside the candidate mask from being invoked.

On this basis, the plan is also subject to budget constraints. Let $S_{i,r} = \{E_j \mid s_{r,j} = 1\}$ denote the set of experts actually invoked under plan P_r for sample i. The actual expert invocation cost of the plan is denoted as:

$$\mathcal{C}(S_{i,r}) = \sum_{E_j \in S_{i,r}} c_j. \tag{3}$$

Given a per-sample review budget B, a feasible plan must satisfy $\mathcal{C}(S_{i,r}) \leq B$. The number of invoked experts is also bounded by a predefined upper limit $K_{\max}$, i.e., $|S_{i,r}| \leq K_{\max}$.

These constraints ensure that the resulting plans are executable for the current sample and remain within the specified resource limits. Through constraint-aware feasibility checking and pruning, the resulting executable plans form the feasible candidate set for subsequent plan scoring and selection.

2.6. Review Plan Scoring

After candidate plan generation and feasibility filtering, the model obtains a set of executable review plans. The resource-aware deep reinforcement learning (DRL) router implements plan selection as a single-step contextual policy optimization problem. For each patient, the router observes the current context once, selects one discrete plan from the feasible candidate set, and receives an immediate utility signal combining predictive loss and normalized expert-invocation proxy cost. This formulation is equivalent to a contextual bandit or a one-step Markov decision process; it does not contain cross-patient state transitions, delayed rewards, or discounted long-term returns.

Let u_i denote the context of sample i, including modality availability, the initial screening output, and the expert cost profile. Let $P_r \in \mathcal{P}_i^{\text{feas}}$ denote a feasible plan and $S_{i,r}$ its corresponding expert set. The immediate utility for selecting plan P_r is defined as:

$$\mathcal{R}_i(P_r \mid u_i) = -\mathcal{L}(y_i, p_{i,r}) - \lambda \mathcal{C}(S_{i,r}), \tag{4}$$

where u_i denotes the patient context, P_r is a feasible candidate plan, y_i is the ground-truth label, $p_{i,r}$ is the fused predictive probability obtained after selecting P_r, and $S_{i,r}$ is the corresponding expert set. The first term measures the immediate predictive loss, whereas $\mathcal{C}(S_{i,r}) = \sum_{E_j \in S_{i,r}} c_j$ denotes the normalized expert-invocation proxy cost. The coefficient λ controls the trade-off between predictive utility and expert invocation cost. The policy was optimized using the REINFORCE policy-gradient algorithm [35], with a moving baseline to reduce gradient variance and entropy regularization to encourage exploration [36]. Policy-gradient optimization was used because the action is a discrete expert plan and the candidate space is modified by non-differentiable feasibility filtering and explicit constraints.

During inference, the learned DRL contextual policy is used as a masked scorer for the feasible candidate plans. For each feasible plan $P_r \in \mathcal{P}_i^{\text{feas}}$, the router evaluates the plan under its corresponding selection mask and produces a routing score $s_\theta(P_r \mid u_i)$. The final review plan is selected as:

$$P_i^* = \arg \max_{P_r \in \mathcal{P}_i^{\text{feas}}} s_\theta(P_r \mid u_i). \tag{5}$$

The selected plan P_i^* determines the expert set S_i executed in the second-stage review.

2.7. Plan Degeneration and On-Demand Review

In PACE-MoE, Stage 2 does not invoke a fixed set of extended experts for every sample. Instead, the model selects an executable expert plan from the feasible candidate set for each sample, allowing the number and composition of invoked experts to adapt to the sample state, modality availability, and routing score.

Let E_{sym} denote the symptom expert and E_{ext} denote the set of extended experts, including the laboratory, ultrasound, X-ray, and cross-modal fusion experts. When a compact plan receives the highest routing score among the feasible candidates, the selected Stage-2 plan can degenerate to a low-cost expert combination. For example, a symptom-only or other small expert combination can be selected when it satisfies the feasibility constraints and achieves the highest routing score.

Conversely, when a plan incorporating additional evidence satisfies the feasibility constraints and receives a higher routing score, the model can select one or more extended experts for the review.

Therefore, the on-demand review mechanism in PACE-MoE does not partition samples into reviewed and unreviewed groups using a hard trigger. Instead, all samples proceed to Stage 2, where the routing policy selects a feasible expert plan according to the sample state and resource constraints. Simple or modality-limited samples may be assigned compact low-cost plans, whereas samples that benefit from richer evidence can be assigned plans involving additional experts.

2.8. Output Fusion

When the selected Stage-2 plan is the symptom-only plan, the second-stage output is identical to the first-stage symptom assessment. For samples assigned plans involving additional experts, the Stage-2 output provides supplementary evidence. To prevent the Stage-2 output from directly overriding the initial assessment, we fuse the outputs of the two stages in the logit space. This design provides a more stable way to integrate the addi-

tional evidence, since high-cost modalities may provide useful information but may also introduce noise, quality variations, or inconsistencies with the low-cost symptom evidence.

Let $p_i^{(1)}$ and $p_i^{(2)}$ denote the predictive probabilities of the first and second stages, respectively. They are first mapped into the logit space and subsequently combined via weighted fusion:

$$l_i = (1 - \alpha)\,\mathrm{logit}(p_i^{(1)}) + \alpha\,\mathrm{logit}(p_i^{(2)}), \tag{6}$$

$$p_i = \sigma(l_i). \tag{7}$$

The fusion coefficient $\alpha \in [0, 1]$ controls the relative contributions of the two stages, and $\sigma(\cdot)$ denotes the sigmoid function. A smaller α places greater weight on the first-stage screening result, whereas a larger α places greater weight on the second-stage review evidence. The primary experiments used $\alpha = 0.8$, which was fixed during test-set evaluation.

Compared with direct weighting in the probability space, logit-space fusion reduces the effects of probability saturation and preserves a more consistent contribution from the two-stage predictions. By combining the outputs on the log-odds scale, the final prediction can incorporate supplementary review evidence without allowing the second-stage output to dominate solely because of probability-scale effects.

3. Experimental Settings

3.1. Dataset and Task Configuration

This study focuses on the assessment of pubertal development status. The experimental dataset was collected from the developmental pediatrics outpatient clinic of a partner hospital and comprises 1249 puberty-related clinical cases. We formulate the clinical diagnostic task as a binary classification problem to determine whether a subject has precocious puberty. Each sample corresponds to an individual subject, and the model outputs a predictive probability based on the available multimodal evidence.

Precocious puberty was defined clinically as the development of secondary sexual characteristics before 8 years of age in girls or 9 years in boys [1,2]. In routine practice, the diagnosis was established by the treating pediatric clinician using age at onset and pubertal progression together with physical examination, growth pattern, endocrine findings, bone age, and pelvic or gonadal ultrasonography when available. The clinical reference standard for this retrospective study was the final diagnosis recorded in the hospital information system. Records with a final diagnosis indicating precocious puberty were assigned to the positive class; records without such a diagnosis or containing explicit negation or exclusion were assigned to the negative class. The diagnosis field was used only for label construction and was excluded from model inputs. The cohort contained 495 positive and 754 negative cases. Because the diagnostic work-up followed routine care, not every patient underwent identical confirmatory tests.

The observed sex distribution arose from the retrospective source cohort and was not deliberately imposed during sampling. The cohort comprised 970 female patients, including 477 positive and 493 negative cases, and 279 male patients, including 18 positive and 261 negative cases. No sex-based resampling was applied. Because only 18 male patients were positive, separate sex-specific model training and tuning would have produced unstable estimates; we therefore retained a common model and assessed sex-related performance differences through post hoc subgroup analysis.

Age was available for all 1249 patients, with a mean of 10.08 years (SD 2.18), a median of 10.00 years (interquartile range 9.00–12.00), and a range of 1.42–18.00 years. The numbers of patients aged under 8 years, 8 to under 9 years, and 9 years or older were 143, 117, and

989, respectively. Clinical text, laboratory tests, X-ray images, and ultrasound images were available for 1249 (100.0%), 1088 (87.1%), 941 (75.3%), and 505 (40.4%) patients, respectively.

Each case contains incomplete multimodal information, including clinical text, laboratory tests, ultrasound images, and X-ray radiographs. Clinical text is available for all cases, while laboratory tests are available for the vast majority of cases. In contrast, ultrasound and X-ray images have relatively limited coverage, primarily corresponding to gonadal and bone-age examinations. The resulting uneven modality coverage and partial missingness provide a realistic setting for evaluating the sample-adaptive inference capability of PACE-MoE.

The reference label was derived from the physician's final diagnosis field, which was excluded from all model inputs. The Stage-1 clinical text was constructed from non-diagnostic fields, including the reason for visit, medical history, and non-conclusive physical-examination information. Columns containing diagnostic conclusions, management recommendations, follow-up, re-examination, or treatment plans were excluded. Laboratory and imaging data were processed separately by their designated experts.

To ensure reliable evaluation, we adopt a fixed five-fold cross-validation protocol with patient-level splitting. In each fold, the data are divided into training, validation, and test subsets using predefined split files, with each patient assigned to only one subset within the fold to prevent patient-level data leakage. The label distribution and modality missingness patterns are kept approximately consistent across folds. All baseline methods are evaluated under the same split protocol and input feature space. The validation subset in each fold is used for threshold selection, probability calibration, and hyperparameter tuning, whereas the corresponding test subset is reserved exclusively for final evaluation.

3.2. Baseline Methods

To evaluate the performance of the proposed PACE-MoE framework under different inference strategies, several categories of baseline methods are established. All comparative methods are evaluated under the same five-fold cross-validation split, with threshold selection performed on the validation set. Among them, methods belonging to learning-to-defer, cost-sensitive feature acquisition, active feature acquisition, and recent dynamic multimodal learning are included for comparison. Classical baselines are constructed following selective prediction, budget-constrained classification, and active feature acquisition paradigms [10,11,27,37], while recent public-code multimodal methods are adapted to our fixed expert-feature protocol for matched-cost evaluation [17,18].

Stage-1 only uses only the first-stage low-cost symptom expert for prediction without invoking any subsequent high-cost experts. This baseline represents the foundational screening performance under minimum cost constraints.

Always-Upgrade method invokes the full expert set for all samples. This corresponds to a static fusion strategy without dynamic routing, utilizing the maximum possible multimodal evidence. It serves as a high-cost reference point without sample-adaptive routing.

L2D (Learn-to-Defer) employs a learning-to-defer mechanism from the low-cost symptom expert to the complete expert pool. When the first-stage screening output meets specific upgrade criteria, the complete expert combination is invoked for review. This baseline incorporates principles of selective prediction and deferred decision-making, highlighting the distinction between conventional binary upgrade mechanisms and our dynamic expert combination selection [27].

Joint-AFA-like and CWCF-like adapt active feature acquisition and cost-sensitive classification to sequential modality acquisition [10,11,37]. Both methods begin with symptom evidence and make stop-or-acquire decisions using only the evidence acquired at the current step and modality-availability information; outputs from unacquired experts are

excluded from the routing state. Joint-AFA-like uses a learned sequential acquisition policy, whereas CWCF-like uses a cost-aware uncertainty-based stopping rule.

DynMM-adapted follows the public DynMM design and uses a sample-dependent gate to select between a low-cost branch and an enriched multimodal branch [17]. The gate receives the low-cost branch output and modality-availability information, without observing the enriched branch output before the branch decision. The branch definitions are adapted to the available expert pool.

SimMLM-adapted follows the public SimMLM framework and models each available modality as a lightweight expert under missing-modality conditions [18]. Its selector uses symptom-side context and modality-availability information, while unselected expert outputs are excluded from the selection state. A cost-aware regularization term and a more-versus-fewer ranking constraint are used under the same data splits and validation protocol.

PACE-MoE (Ours). The proposed two-stage dynamic mixture-of-experts framework. The first stage leverages low-cost clinical text for a baseline assessment. In the second stage, it generates candidate plans via a combinatorial sampler based on structural masking, performs OR constraint filtering, and applies DRL resource-aware routing to select the highest-scoring feasible expert plan under budget constraints, followed by two-stage output fusion in the logit space.

3.3. Implementation Details

A unified expert pool was used across all experiments, comprising a symptom text expert, a laboratory test expert, an ultrasound imaging expert, an X-ray imaging expert, and a cross-modal fusion expert. The normalized costs of these experts were set to 0.5, 0.5, 1.0, 1.0, and 2.0, respectively. These values define an ordinal resource scale: the symptom and laboratory branches form the low-cost tier, the two imaging branches use the unit reference cost, and the fusion branch receives the largest value because it combines multiple evidence sources. For each patient, Mean Cost was calculated as the sum of the proxy costs of the experts actually executed, with the Stage-1 symptom expert counted only once. The proxy values were fixed before method comparison and applied identically to all routing strategies; they were not fitted to the evaluation results, measured latency, clinical examination fees, or monetary expenditure. Actual inference latency and GPU memory were reported separately.

All experiments were conducted using an AMD Ryzen 9 7950X3D processor (Advanced Micro Devices, Inc., Santa Clara, CA, USA), 128 GB RAM, and one NVIDIA GeForce RTX 4090 GPU with 24 GB memory (NVIDIA Corporation, Santa Clara, CA, USA). The software environment comprised Python 3.10.21 (Python Software Foundation, Beaverton, OR, USA), PyTorch 2.5.1 (PyTorch Foundation, San Francisco, CA, USA), CUDA 12.1, and cuDNN 9.1 (NVIDIA Corporation, Santa Clara, CA, USA).

BioMedCLIP and BioClinicalBERT were used as frozen feature encoders. Each expert-specific classification head contained one 256-unit hidden layer with ReLU activation and a dropout rate of 0.1, followed by a two-class output layer. The expert heads were optimized using AdamW with a learning rate of 1×10^{-4}, weight decay of 1×10^{-5}, a batch size of 32, and 50 training epochs.

The routing network received a 20-dimensional patient context and used one 64-unit ReLU hidden layer. It was optimized using AdamW with a learning rate of 3×10^{-4}, a batch size of 64, and 20 training epochs, with checkpoint selection based on validation F1-score. The policy objective used a cost coefficient of 0.05, an entropy coefficient of 0.01, and a moving-baseline momentum of 0.95. The primary configuration generated 32 stochastic candidate masks, used a router Top-k limit of four experts, and applied an

OR budget of 3.0 with at most three experts per executable plan and an additional fusion feasibility cost of 0.5. The second-stage fusion coefficient was fixed at $\alpha = 0.8$.

PACE-MoE adopts a two-stage inference pipeline. In Stage 1, only the symptom expert is invoked to obtain a low-cost screening prediction. All samples subsequently proceed to Stage 2, where candidate expert plans are generated, filtered according to feasibility constraints, and evaluated by a resource-aware router to select an executable plan for evidence review. The final prediction is obtained by fusing the outputs of the two stages in the logit space. All final decision thresholds are selected on the validation set by maximizing the F1-score, and temperature scaling is applied to mitigate probability miscalibration in deep models [24]. Beyond the primary configuration, we also establish various resource tiers to analyze the effects of different candidate plan sizes and expert selection scopes on diagnostic performance, average execution cost, and inference latency.

3.4. Evaluation Metrics

The proposed framework is evaluated across three dimensions: diagnostic performance, normalized expert-invocation proxy cost, and system resource overhead. We report Accuracy (Acc), Sensitivity (Recall, Rec), Specificity, Precision, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). The F1-score balances Precision and Sensitivity, while Sensitivity reflects positive-case detection and Specificity reflects correct exclusion of negative cases. Precision represents the proportion of predicted positive cases that are truly positive, and AUC summarizes discrimination across decision thresholds.

Calibration was assessed using ECE with 10 equal-width bins of top-label confidence and the Brier score. In each fold, temperature scaling was fitted on the validation subset and evaluated on the held-out test subset, following Guo et al. [24]. Patient-level bootstrap resampling with 5000 repetitions was used to estimate 95% confidence intervals.

Inference cost is quantified via the Mean Cost. Let S_i be the set of experts actually invoked for the i-th sample, c_j be the normalized invocation cost of expert E_j, and K be the total number of samples in the test set. The Mean Cost is defined as:

$$\text{Mean Cost} = \frac{1}{K} \sum_{i=1}^{K} \sum_{E_j \in S_i} c_j. \tag{8}$$

This metric characterizes the average normalized resource consumption incurred by expert invocations during the inference phase. A lower Mean Cost indicates higher resource efficiency in fulfilling the diagnostic objectives.

Regarding system resource overhead, we further record the average inference time per sample. Let T represent the total execution time of the inference process on the test set, and K represent the number of test samples. The average inference time per sample is defined as:

$$\text{Time}_{\text{avg}} = \frac{T}{K}. \tag{9}$$

It is important to note that the Mean Cost is not equivalent to the empirically measured inference time. The former reflects the normalized resource consumption associated with the selected expert plans, whereas the latter captures practical execution efficiency under the specific hardware environment, parallel implementation, and data loading workflows.

3.5. Robustness Experiment Settings

To evaluate stability under missing modalities, synthetic missingness was applied during testing only to the laboratory, ultrasound, and X-ray inputs that were originally available for each patient; clinical text was retained for all cases. Each eligible source

modality was independently masked at the specified patient-level missingness probability using deterministic masks. If a masked source modality had contributed to the derived fusion input, the fusion input was also treated as unavailable. The missingness ratios were 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9, and 0.99.

This experiment characterizes changes in diagnostic performance and Mean Cost as optional evidence becomes progressively unavailable. It also examines how the executable expert-plan space changes under increasing optional-modality missingness.

4. Results

This section reports the experimental results of PACE-MoE. We first compare it with the primary baseline strategies in terms of diagnostic performance and Mean Cost, followed by comparisons with cost-sensitive and dynamic multimodal methods under matched or near-matched cost conditions. We then examine the effects of candidate plan generation, OR-based feasibility filtering, and the routing strategy through ablation and router-substitution experiments. Finally, robustness under increasing optional-modality missingness is evaluated.

4.1. Comparison with Baseline Methods

The experimental results of PACE-MoE are presented in Table 1. To evaluate its performance, we compare it against three baseline strategies: Stage-1 only, which relies solely on the low-cost symptom expert; Always-Upgrade, which invokes the complete expert set for all samples; and L2D, a traditional deferred-decision baseline. Table 1 reports Accuracy, Sensitivity, Specificity, Precision, F1-score, AUC, and Mean Cost. Sensitivity reflects positive-case detection, Specificity reflects correct exclusion of negative cases, Precision represents the proportion of predicted positive cases that are truly positive, and Mean Cost measures the average normalized expert-invocation proxy cost.

Table 1. Baseline comparison results. Panel (a) reports the primary performance and cost metrics; Panel (b) reports additional clinical diagnostic metrics.

(a) Primary performance and cost metrics

Method	Acc	F1-score	AUC	Mean Cost
Stage-1 only	0.7326 ± 0.0407	0.6874 ± 0.0663	0.8182 ± 0.0327	**0.5000 ± 0.0000**
Always-Upgrade	**0.8583 ± 0.0193**	**0.8368 ± 0.0179**	**0.9368 ± 0.0124**	3.7297 ± 0.0380
L2D	0.8327 ± 0.0212	0.8028 ± 0.0279	0.9143 ± 0.0181	3.1263 ± 0.1931
PACE-MoE (Ours)	0.8519 ± 0.0230	0.8255 ± 0.0301	0.9218 ± 0.0165	1.4836 ± 0.0152

(b) Additional clinical diagnostic metrics

Method	Sensitivity	Specificity	Precision
Stage-1 only	0.7556 ± 0.1236	0.7175 ± 0.0723	0.6398 ± 0.0504
Always-Upgrade	**0.9152 ± 0.0412**	0.8210 ± 0.0451	**0.7736 ± 0.0397**
L2D	0.8646 ± 0.0887	0.8117 ± 0.0662	0.7578 ± 0.0481
PACE-MoE (Ours)	0.8869 ± 0.0513	**0.8289 ± 0.0255**	0.7734 ± 0.0257

Note: Bold values indicate the best results for each metric. For Mean Cost, the lowest value is considered the best.

Stage-1 only yields the lowest Mean Cost, but its F1-score is 0.6874, indicating the limited diagnostic performance of relying solely on the symptom expert. Always-Upgrade achieves the highest F1-score and AUC; however, its Mean Cost reaches 3.7297, corresponding to a substantially higher normalized expert-invocation proxy cost. In comparison, PACE-MoE achieves an F1-score of 0.8255, close to that of Always-Upgrade, while reducing the normalized expert-invocation proxy cost from 3.7297 to 1.4836, a reduction of approximately 60.2%. As shown in Table 1b, PACE-MoE achieved the highest Specificity

among the four primary inference strategies. Its Precision was nearly identical to that of Always-Upgrade, whereas Always-Upgrade achieved the highest Sensitivity.

Compared to L2D, PACE-MoE achieves both a higher F1-score and a lower Mean Cost. This comparison suggests that the proposed framework differs from a binary upgrade decision by selecting expert plans adaptively according to the state of each sample.

Figure 2 further illustrates the relationship between diagnostic performance and normalized expert-invocation proxy cost across the evaluated methods. Stage-1 only is located in the low-cost, lower-performance region, whereas Always-Upgrade occupies the high-cost, high-performance region. PACE-MoE is located between these two operating points, combining a lower normalized expert-invocation proxy cost than Always-Upgrade with comparable diagnostic performance.

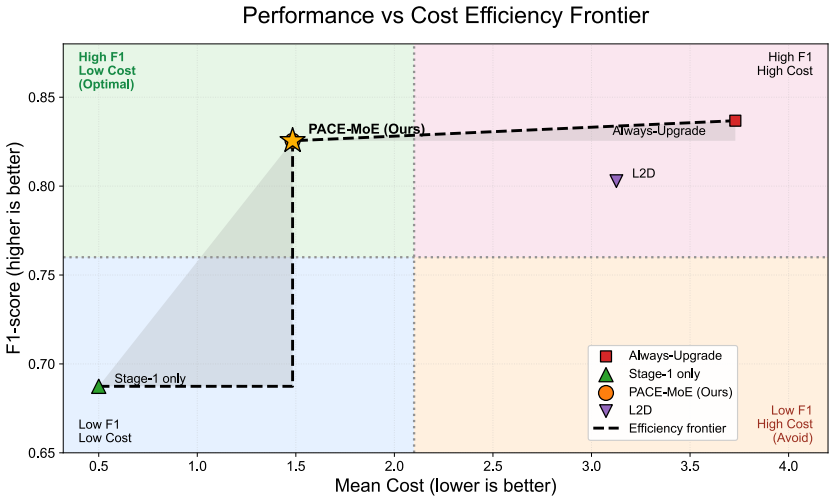


Figure 2. Diagnostic performance versus normalized expert-invocation proxy cost across the evaluated methods.

Table 1 reports the mean and standard deviation across the five test folds. For the confusion-matrix analysis, the patient-level out-of-fold predictions from the five mutually exclusive test folds were pooled, with each patient included once. This pooling yielded 439 true positives, 56 false negatives, 625 true negatives, and 129 false positives for PACE-MoE. Relative to Always-Upgrade, PACE-MoE reduced Mean Cost from 3.7297 to 1.4836, with 6 fewer false positives and 14 more false negatives. Figure 3A visualizes the confusion counts and row-normalized percentages for Stage-1 only, Always-Upgrade, L2D, and PACE-MoE.

Across the prespecified age bands of <8, 8 to <9, and ≥ 9 years, the Acc/F1-score values were 0.9231/0.9575, 0.7863/0.8447, and 0.8493/0.7683, respectively. In the <8-year subgroup, Specificity was estimated from only 12 negative cases and consequently had a wide 95% confidence interval. Complete age-stratified results with bootstrap 95% confidence intervals are provided in Table 2.

Post hoc analysis of the pooled out-of-fold predictions showed that Acc, Sensitivity, Specificity, Precision, F1-score, and AUC were 0.8299, 0.8973, 0.7647, 0.7868, 0.8384, and 0.8918 in female patients and 0.9283, 0.6111, 0.9502, 0.4583, 0.5238, and 0.8827 in male patients, respectively. The high male-subgroup accuracy reflected the predominance of negative cases, while the small number of positive male cases resulted in greater uncertainty for sensitivity, precision, and F1-score. Complete subgroup results with bootstrap 95% confidence intervals are provided in Table 2.

Values in Table 2 are reported as point estimates with bootstrap 95% confidence intervals in brackets.

Table 2. Post hoc sex- and age-stratified performance of PACE-MoE using pooled out-of-fold test predictions.

Subgroup	*n*	Pos./Neg.	Acc	Precision	Sensitivity	Specificity	F1-Score	AUC
Female	970	477/493	0.8299 [0.8062–0.8526]	0.7868 [0.7509–0.8193]	0.8973 [0.8701–0.9232]	0.7647 [0.7276–0.8012]	0.8384 [0.8131–0.8614]	0.8918 [0.8703–0.9121]
Male	279	18/261	0.9283 [0.8961–0.9570]	0.4583 [0.2609–0.6667]	0.6111 [0.3750–0.8333]	0.9502 [0.9213–0.9737]	0.5238 [0.3226–0.6897]	0.8827 [0.7638–0.9619]
<8 years	143	131/12	0.9231 [0.8741–0.9650]	0.9688 [0.9355–0.9924]	0.9466 [0.9055–0.9844]	0.6667 [0.3750–0.9167]	0.9575 [0.9297–0.9810]	0.9052 [0.7891–0.9835]
8 to <9 years	117	78/39	0.7863 [0.7094–0.8547]	0.8193 [0.7317–0.8974]	0.8718 [0.7901–0.9412]	0.6154 [0.4571–0.7647]	0.8447 [0.7778–0.8987]	0.7751 [0.6644–0.8714]
≥9 years	989	286/703	0.8493 [0.8261–0.8706]	0.6919 [0.6429–0.7387]	0.8636 [0.8219–0.9037]	0.8435 [0.8161–0.8696]	0.7683 [0.7305–0.8035]	0.9151 [0.8955–0.9336]

As shown in Figure 3B, ECE decreased from 0.0451 (95% CI: 0.0280–0.0668) to 0.0322 (0.0196–0.0544) after temperature scaling, whereas the Brier score remained similar (0.1489 [0.1389–0.1599] versus 0.1493 [0.1378–0.1617]). This indicates a modest reduction in binned confidence error rather than a consistent improvement across calibration measures. Because no universal ECE threshold defines good calibration, ECE was interpreted relative to its ideal value of zero.

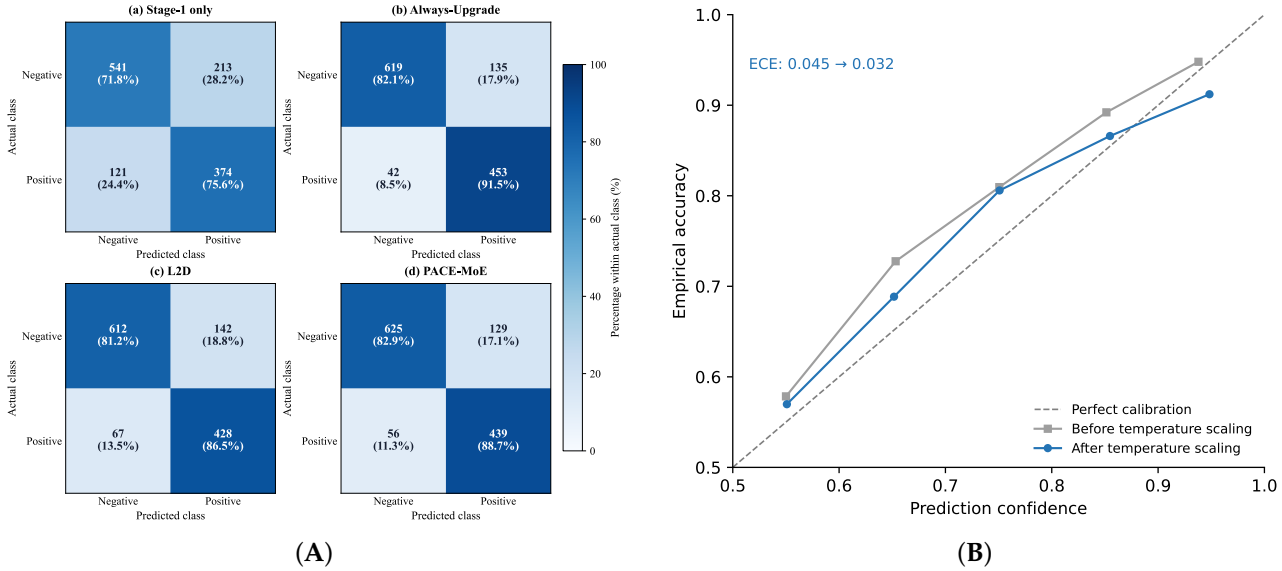


Figure 3. Classification-error and calibration analysis. (**A**) Confusion matrices for the primary inference strategies, showing counts and row-normalized percentages. (**B**) Reliability curves before and after temperature scaling using the pooled out-of-fold test predictions.

4.2. Comparative Analysis Under Comparable Cost Conditions

Because the average invocation costs vary across dynamic inference methods, we compare the competing methods under matched or near-matched Mean Cost conditions in Table 3. In addition to Accuracy, F1-score, AUC, and Mean Cost, Sensitivity, Specificity, and Precision are reported to provide a more complete assessment of clinical screening performance.

As shown in Table 3, PACE-MoE achieved the highest reported mean Accuracy, F1-score, Specificity, and Precision among the adaptive comparators. Joint-AFA-like achieved the highest AUC at a similar Mean Cost, while CWCF-like achieved the highest Sensitivity but lower Accuracy, Specificity, Precision, and F1-score. SimMLM-adapted had the lowest Mean Cost but also the lowest Accuracy, Precision, and F1-score, whereas DynMM-adapted produced lower Accuracy and F1-score under the evaluated settings. These results indicate complementary operating characteristics rather than uniform superiority across all metrics.

Table 3. Comparison with classical and recent baselines under matched or near-matched cost conditions.

Method	Acc	F1-Score	Sensitivity	Specificity	Precision	AUC	Mean Cost
Joint-AFA-like [10]	0.8415	0.8176	0.8949	0.8064	0.7574	**0.9264**	1.4712
CWCF-like [11]	0.7967	0.7870	**0.9434**	0.7004	0.6774	0.8690	1.3939
DynMM-adapted [17]	0.8159	0.7873	0.8586	0.7878	0.7358	0.9009	1.7538
SimMLM-adapted [18]	0.7542	0.7224	0.8040	0.7215	0.6607	0.8382	**1.2523**
PACE-MoE (Ours)	**0.8519**	**0.8255**	0.8869	**0.8289**	**0.7734**	0.9218	1.4836

Note: Bold values indicate the best results for each metric. For Mean Cost, the lowest value is considered the best.

Paired fold-bootstrap analyses were conducted using the shared cross-validation splits. The F1-score differences between PACE-MoE and CWCF-like, DynMM-adapted, SimMLM-adapted, and Joint-AFA-like were 0.0385 (95% CI: 0.0256–0.0494), 0.0382 (0.0062–0.0784), 0.1031 (0.0899–0.1166), and 0.0079 (−0.0131 to 0.0308), respectively. The intervals for CWCF-like, DynMM-adapted, and SimMLM-adapted excluded zero, whereas the interval for Joint-AFA-like included zero. Because only five folds were available as resampling units, these intervals were interpreted as exploratory measures of cross-validation uncertainty rather than definitive significance tests.

4.3. Ablation Study Analysis

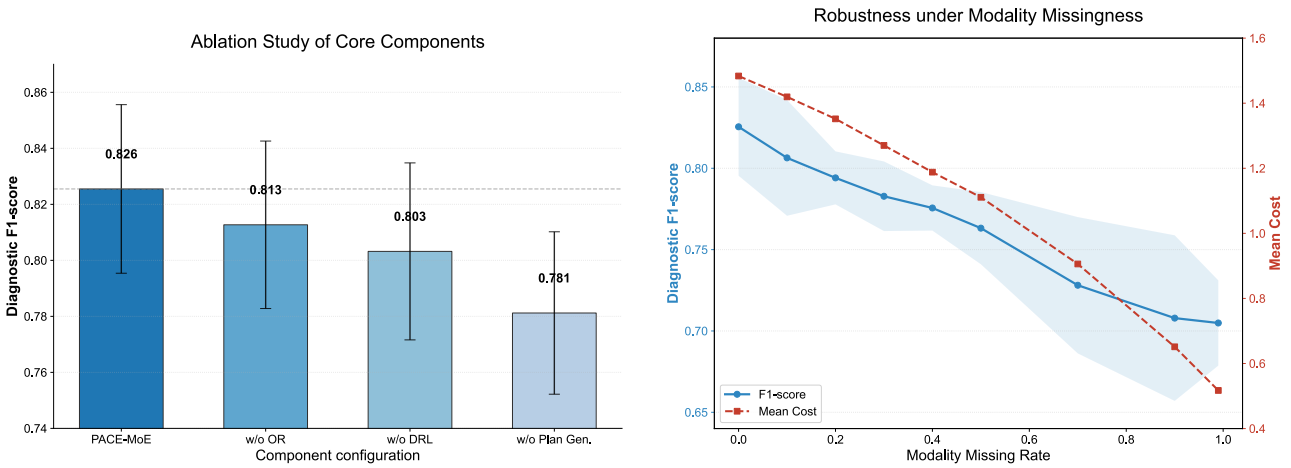
To assess the effects of the core components within PACE-MoE, we conduct ablation studies on candidate plan generation, OR-based constraint filtering, and resource-aware plan scoring.

As observed in Table 4 and Figure 4A, removing any key module leads to a decline in overall classification performance. Upon removing the OR constraint filtering module, the F1-score drops to 0.8127, while the Mean Cost remains close to the full model. This indicates that the primary contribution of the OR module is not cost reduction, but rather the filtering of infeasible or unsuitable expert combinations before plan scoring. Removing the DRL routing module reduces the F1-score to 0.8032 and the AUC to 0.9048; although the Mean Cost decreases to 1.2838, the accompanying performance degradation indicates that the alternative routing strategy does not preserve the same balance between diagnostic performance and resource consumption. Removing the candidate plan generation module drops the F1-score to 0.7812 and the Recall to 0.8020, representing the largest performance decline among the ablated variants and underscoring the importance of a diverse candidate plan space for subsequent filtering and scoring. These three modules correspond to proposing candidate plans, filtering infeasible plans, and selecting the highest-scoring feasible plan, respectively, and together they constitute the core mechanism underlying the dynamic inference capability of PACE-MoE.

Table 4. Ablation studies of the proposed components in PACE-MoE.

Proposed Components			Metrics				
OR	DRL	PG	Acc	F1-Score	Rec	AUC	Mean Cost
✓	✓	✓	**0.8519**	**0.8255**	0.8869	**0.9218**	1.4836
	✓	✓	0.8399	0.8127	0.8747	0.9167	1.4788
✓		✓	0.8255	0.8032	**0.8990**	0.9048	**1.2838**
✓	✓		0.8231	0.7812	0.8020	0.9027	1.5380

Note: ✓ indicates inclusion; blank indicates removal. Bold values indicate the best results.



(**A**) (**B**)

Figure 4. Ablation and robustness analysis. (**A**) Ablation studies of core modules in PACE-MoE. (**B**) Changes in model performance and mean cost under different modality missingness ratios.

Furthermore, to evaluate the impact of the routing network architecture on dynamic expert selection, we substitute the DRL router in PACE-MoE with Multi-Layer Perceptron (MLP) and Gated Linear Unit (GLU) structures, keeping candidate plan generation, OR constraint filtering, output fusion, and other experimental settings identical. The results are reported in Table 5.

Table 5. Experimental results of router architecture substitution.

Method	Acc	F1-Score	Rec	AUC	Mean Cost
MLP Router	0.8463	0.8204	0.8869	0.9196	1.4836
GLU Router	0.8471	0.8228	**0.8929**	**0.9220**	1.4836
PACE-MoE (Ours)	**0.8519**	**0.8255**	0.8869	0.9218	1.4836

Note: Bold values indicate the best result for each metric among the three routing configurations.

As shown in Table 5, the three routing configurations had the same Mean Cost of 1.4836 and produced closely matched results. PACE-MoE achieved the highest reported mean Accuracy and F1-score, whereas the GLU router achieved slightly higher mean Recall and AUC. The small differences indicate that the candidate-generation and feasibility-filtering pipeline can operate with different learned plan scorers. PACE-MoE uses the contextual policy-gradient router as its primary implementation because it provides a stochastic-policy formulation for optimizing immediate utility over discrete feasible plans.

5. Discussion

PACE-MoE implements sample-adaptive review by selecting a patient-specific plan according to modality availability, Stage-1 confidence, and expert-cost information. Budget awareness is imposed through feasibility constraints and the normalized expert-invocation proxy cost used in plan filtering and scoring. Accordingly, the lower Mean Cost relative to Always-Upgrade represents a trade-off between diagnostic performance and resource use within this predefined proxy-cost framework and does not imply an equivalent reduction in clinical examination expenditure.

5.1. Resource Configuration and Deployment Efficiency Analysis

Table 6 displays the performance and system resource overhead of PACE-MoE under various resource configurations. It is important to clarify that the Small (S), Medium (M), and Large (L) tier experiments were conducted in identical hardware environments rather

than on different GPUs or deployment devices; the primary distinctions among these three tiers lie in the candidate plan scale and the expert search scope. Specifically, the S-tier utilizes a smaller candidate plan scale, the M-tier moderately expands the search scope, and the L-tier corresponds to our primary baseline configuration.

Table 6. Performance and resource overhead under different resource configurations.

Setting	Acc	F1-Score	Rec	AUC	Mean Cost	Time/Patient
S	0.8311	0.7992	0.8485	0.9133	1.4972	**0.1861**
M	0.8479	0.8167	0.8586	0.9198	1.4868	0.1938
L	**0.8519**	**0.8255**	**0.8869**	**0.9218**	**1.4836**	0.2079

Note: Bold values indicate the best result for each metric; the lowest Time/Patient is highlighted.

As the resource configuration scales from the S-tier to the L-tier, the F1-score increases from 0.7992 to 0.8255, indicating that a broader candidate plan search is associated with improved diagnostic performance. Meanwhile, the Mean Cost remains closely matched across all tiers and even decreases slightly from 1.4972 to 1.4836, indicating that the expanded candidate search does not substantially increase the average expert invocation cost. The measured end-to-end inference latency was 0.1861, 0.1938, and 0.2079 s per patient for the S-, M-, and L-tier configurations, respectively; for the primary L-tier configuration, the peak CUDA memory allocated by PyTorch was 0.993 GB. Under the evaluated hardware and software environment, expanding the candidate-plan search scope introduced a modest increase in measured latency while improving the diagnostic-performance point estimates.

5.2. Robustness to Missing Modalities

To evaluate the stability of PACE-MoE under incomplete optional modalities, we assessed its performance across varying missingness ratios. Synthetic missingness was applied during testing to the laboratory, ultrasound, and X-ray inputs, while clinical text was retained for all cases. Table 7 reports the results at key missingness ratios.

Table 7. Robustness results of PACE-MoE under different modality missingness ratios.

Missing Rate	Acc	F1-Score	Rec	AUC	Mean Cost
0.000	**0.8519**	**0.8255**	**0.8869**	**0.9218**	**1.4836**
0.100	0.8335	0.8064	0.8727	0.9144	1.4195
0.300	0.8223	0.7828	0.8121	0.8969	1.2702
0.500	0.7958	0.7632	0.8303	0.8794	1.1109
0.900	0.7350	0.7079	0.8202	0.8300	0.6513
0.990	0.7294	0.7049	0.8162	0.8213	0.5172

Note: Bold values indicate the results under 0% missingness.

As shown in Table 7 and Figure 4B, diagnostic performance declined gradually as optional-modality missingness increased, while Mean Cost decreased because fewer additional experts remained executable. At missingness ratios of 0.10, 0.30, and 0.50, the F1-scores were 0.8064, 0.7828, and 0.7632, with Mean Costs of 1.4195, 1.2702, and 1.1109, respectively. At the extreme ratios of 0.90 and 0.99, the F1-scores were 0.7079 and 0.7049, while Mean Cost decreased to 0.6513 and 0.5172. Clinical text was retained in all cases; therefore, these extreme settings represent near-complete absence of the optional laboratory and imaging inputs rather than absence of all evidence.

To place the robustness results in context, the primary baseline strategies and adaptive competing methods were evaluated under the same optional-modality missingness protocol at 0%, 50%, and 90% missingness. The corresponding F1-score and Mean Cost results are reported in Table 8.

Table 8. Comparison of F1-score and Mean Cost under additional optional-modality missingness.

Method	F1 (0%)	F1 (50%)	F1 (90%)	Cost (0%)	Cost (50%)	Cost (90%)
Stage-1 only	0.6874	0.6874	0.6874	**0.5000**	**0.5000**	**0.5000**
Always-Upgrade	**0.8368**	0.7731	0.6923	3.7297	1.6613	0.6918
L2D	0.8028	0.7717	0.7075	3.1263	1.4432	0.6181
Joint-AFA-like	0.8176	**0.7745**	0.7184	1.4712	1.1361	0.6521
CWCF-like	0.7870	0.7705	**0.7228**	1.3939	1.1806	0.6181
DynMM-adapted	0.7873	0.7573	0.7150	1.7538	1.2218	0.6589
SimMLM-adapted	0.7224	0.6923	0.6794	1.2523	1.0529	0.6473
PACE-MoE	0.8255	0.7632	0.7079	1.4836	1.1109	0.6513

Note: Bold values indicate the best result for each metric at each missingness level. Higher F1-scores and lower costs are considered better.

Stage-1 only produced identical results across the three missingness levels because it relies exclusively on clinical text, which was retained for all patients.

Under the original clinical availability pattern, PACE-MoE achieved the highest Acc and F1-score among the adaptive comparators, with an F1-score of 0.8255. At 50% additional optional-modality missingness, Joint-AFA-like achieved an F1-score of 0.7745, compared with 0.7632 for PACE-MoE. At 90% missingness, Joint-AFA-like, CWCF-like, and DynMM-adapted achieved F1-scores of 0.7184, 0.7228, and 0.7150, respectively, compared with 0.7079 for PACE-MoE. Under severe evidence depletion, the feasible plan space becomes smaller, and the evaluated acquisition strategies produced comparable performance at these operating points.

5.3. Sensitivity Analysis of the Review Fusion Coefficient

In the two-stage inference workflow of PACE-MoE, the first stage produces a baseline assessment derived from low-cost symptom data, and the second stage conducts an evidence review utilizing an expert combination selected via dynamic routing. The final prediction requires fusing the first-stage baseline screening with the second-stage review evidence. To analyze the impact of this blending mechanism on overall model performance, we conduct a sensitivity analysis on the review fusion coefficient α.

As shown in Table 9, performance remained stable when α ranged from 0.5 to 0.8, with only small differences in the diagnostic metrics. Although $\alpha = 0.5$ yielded slightly higher point estimates for Accuracy, F1-score, and Recall, the primary setting of $\alpha = 0.8$ achieved a marginally higher AUC. Performance decreased at $\alpha = 0.2$ and $\alpha = 1.0$, suggesting that the final prediction benefited from retaining contributions from both stages. Mean Cost remained unchanged because α affected only the final logit-space fusion rather than expert-plan selection. The original primary setting of $\alpha = 0.8$ was retained throughout the main experiments to maintain a consistent evaluation configuration.

Table 9. Sensitivity analysis of the review fusion coefficient α.

α	Acc	F1-Score	Rec	AUC	Mean Cost
0.2	0.7654	0.7502	0.8889	0.8720	1.4836
0.5	**0.8591**	**0.8356**	**0.9030**	0.9215	1.4836
0.8	0.8519	0.8255	0.8869	**0.9218**	1.4836
1.0	0.8303	0.8065	0.8848	0.9074	1.4836

Note: Bold values indicate the best result for each metric across the evaluated fusion coefficients.

5.4. Misclassification Analysis

Figure 5 illustrates the expert-level decision behavior of PACE-MoE on four representative misclassified samples. For a clearer comparison, we selected four representative

failures with different expert participation and routing patterns. The results show that these errors are not simply caused by all experts failing simultaneously; rather, they can arise from conflicts between expert predictions and the routing weights assigned to them.

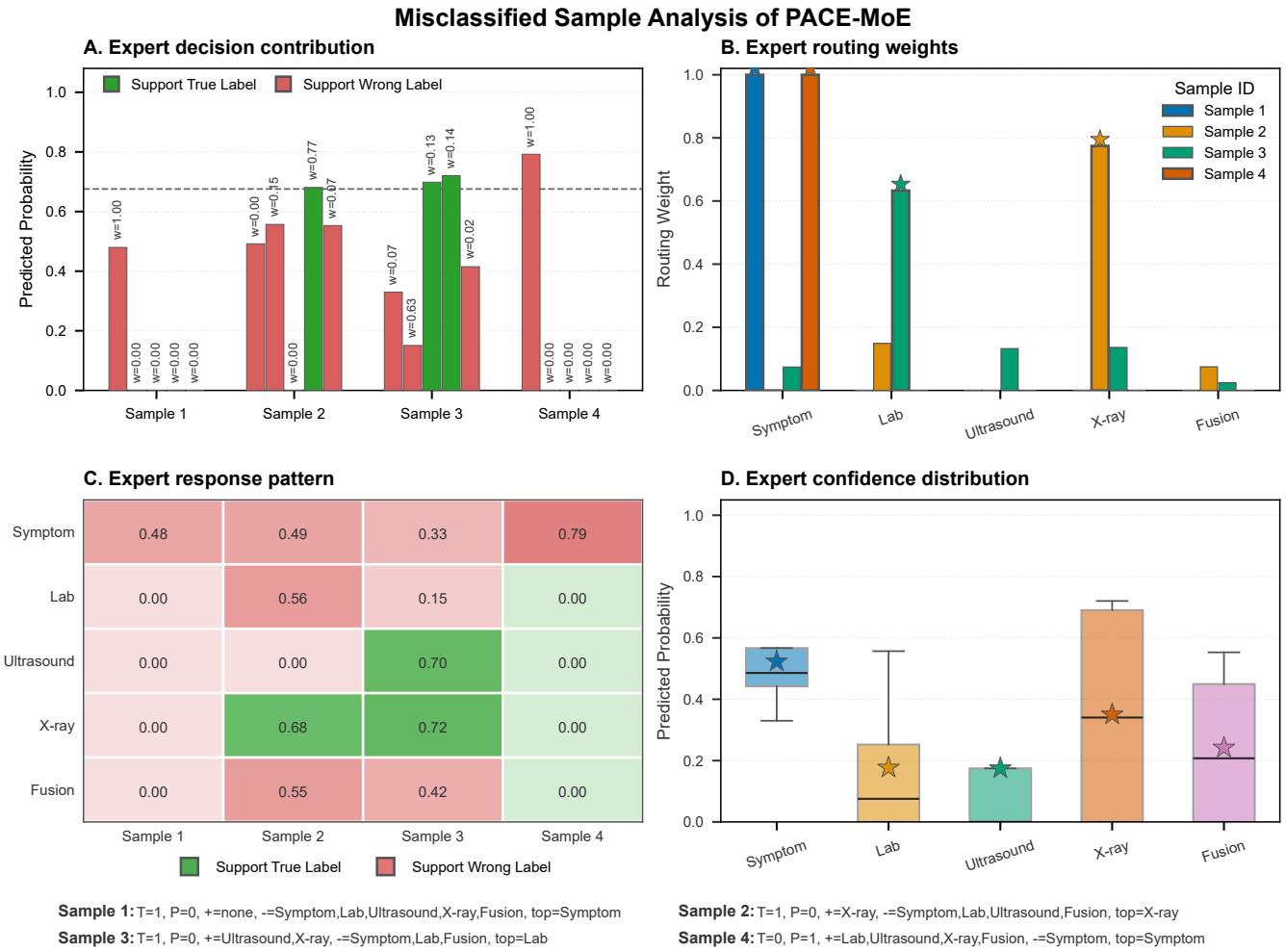


Figure 5. Expert-level decision analysis on misclassified samples of PACE-MoE. (**A**) Expert decision contribution showing predicted probabilities for true and incorrect labels. (**B**) Routing weight distribution of modality experts for four misclassified samples. (**C**) Expert response pattern matrix of per-sample predicted probabilities across modalities. (**D**) Predicted confidence distribution of different modality experts for misclassified samples.

In several cases, some experts assigned relatively high probabilities to the true label, while the routing mechanism assigned greater weights to experts supporting the incorrect label, ultimately biasing the final prediction toward the wrong class. Other cases show that the selected routing weights were concentrated on only a subset of experts, while the remaining experts provided either weak or conflicting evidence. These patterns indicate that misclassification may result from an imbalance between expert-level evidence and routing allocation rather than from uniformly incorrect predictions across all experts.

The expert response pattern further reveals substantial heterogeneity across the four samples. In some cases, a few experts provided useful corrective signals that received relatively limited routing weight; in others, positive and negative evidence coexisted across multiple experts. The expert confidence distributions in Figure 5D also show considerable variation across modalities, indicating that the confidence levels of individual experts are not necessarily consistent. Overall, these failure cases suggest that preserving complemen-

tary expert evidence while improving the router's ability to identify the most discriminative signals may help further improve the robustness of PACE-MoE.

6. Conclusions

This paper addresses the practical challenges of multimodal computer-aided diagnosis for pubertal development, where incomplete modality availability, uneven examination coverage, and heterogeneous expert invocation costs make fixed multimodal inference difficult to deploy. To address these issues, we proposed PACE-MoE, a two-stage dynamic mixture-of-experts framework that separates low-cost symptom screening from adaptive evidence review. After the initial symptom-based assessment, all samples proceed to Stage 2, where candidate expert plans are generated through structural masking and stochastic combinatorial sampling, filtered according to modality availability and resource constraints, and scored by a resource-aware DRL router to select the highest-scoring feasible plan. The predictions from the two stages are then integrated through logit-space fusion.

Experiments on 1249 puberty-related clinical cases show that PACE-MoE reduces the normalized expert-invocation proxy cost by approximately 60.2% compared with full-expert invocation while maintaining comparable diagnostic performance. Under matched or near-matched cost conditions, PACE-MoE achieves higher Accuracy and F1-score than the adapted baseline methods, while Joint-AFA-like obtains a higher AUC at a slightly lower Mean Cost. Ablation studies support the contributions of candidate plan generation, OR-based feasibility filtering, and resource-aware routing to the overall framework. In the robustness evaluation, diagnostic performance declined gradually as optional-modality missingness increased, while Mean Cost decreased with the number of executable additional experts.

The present study has several limitations. The experiments are based on retrospective data from a single medical center, and the expert invocation costs are represented by normalized proxy values rather than directly measured clinical resource consumption; therefore, the reported reduction in Mean Cost does not demonstrate an equivalent reduction in clinical examination expenditure. The limited number of positive male cases resulted in greater uncertainty in the male-subgroup estimates and restricts their generalizability. In addition, the current framework focuses on static binary classification and does not model the temporal dynamics of pubertal development. Future work will therefore investigate external multi-center validation, more realistic cost modeling, and longitudinal multimodal diagnosis. Overall, PACE-MoE provides a sample-adaptive expert-plan selection strategy for multimodal clinical diagnosis under incomplete evidence and constrained inference resources.

Author Contributions: Conceptualization, H.Z., Y.F., L.H. and L.L.; methodology, Y.F. and Y.B.; software, Y.F.; validation, X.M., L.H., L.L. and X.W.; formal analysis, Y.F.; investigation, X.M.; resources, L.L.; data curation, X.W.; writing—original draft preparation, Y.F.; writing—review and editing, H.Z., Y.B. and X.W.; visualization, Y.F. and X.M.; supervision, H.Z.; funding acquisition, X.W. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Liaoning Province Doctoral Research Start-up Fund Program (Grant No. 2025-BS-0604), in part by the Frontier Technology Project for Future Industries of Liaoning Province (Grant No. 2025JH2/101330122), and in part by the General Project of the Basic Scientific Research Project of Liaoning Provincial Department of Education (Grant No. LJ212410144025).

Institutional Review Board Statement: All procedures involving human participants in this study were approved by the Medical Ethics Committee of China Medical University (Ethics Committee Approval No. [2026] 157; 28 April 2026). The research was conducted ethically in strict accordance with the ethical standards of the World Medical Association Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to patient privacy protection and the constraints of the ethical approval protocol of China Medical University.

Acknowledgments: OpenAI ChatGPT 5.6 was used solely for English language polishing, as the authors are non-native English speakers. The authors reviewed and approved all revisions and take full responsibility for the content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Carel, J.C.; Léger, J. Precocious Puberty. *N. Engl. J. Med.* **2008**, *358*, 2366–2377. https://doi.org/10.1056/NEJMcp0800459.
2. Latronico, A.C.; Brito, V.N.; Carel, J.C. Causes, diagnosis, and treatment of central precocious puberty. *Lancet Diabetes Endocrinol.* **2016**, *4*, 265–274. https://doi.org/10.1016/S2213-8587(15)00380-0.
3. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. https://doi.org/10.1038/nature21056.
4. Johnson, A.E.W.; Pollard, T.J.; Shen, L.; Lehman, L.w.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.A.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. https://doi.org/10.1038/sdata.2016.35.
5. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. https://doi.org/10.1109/TPAMI.2018.2798607.
6. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7.
7. Neverova, N.; Wolf, C.; Taylor, G.; Nebout, F. ModDrop: Adaptive Multi-Modal Gesture Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1692–1706. https://doi.org/10.1109/TPAMI.2015.2461544.
8. Havaei, M.; Guizard, N.; Chapados, N.; Bengio, Y. HeMIS: Hetero-Modal Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2016, Athens, Greece, 17–21 October 2016; pp. 469–477. https://doi.org/10.1007/978-3-319-46723-8_54.
9. Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; Peng, X. SMIL: Multimodal Learning with Severely Missing Modality. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 2302–2310. https://doi.org/10.1609/aaai.v35i3.16330.
10. Shim, H.; Hwang, S.J.; Yang, E. Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 3–8 December 2018; Volume 31, pp. 1368–1378.
11. Janisch, J.; Pevný, T.; Lisý, V. Classification with Costly Features Using Deep Reinforcement Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 3959–3966. https://doi.org/10.1609/aaai.v33i01.33013959.
12. Graves, A. Adaptive Computation Time for Recurrent Neural Networks. *arXiv* **2016**, arXiv:1603.08983.
13. Teerapittayanon, S.; McDanel, B.; Kung, H.T. BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2464–2469. https://doi.org/10.1109/ICPR.2016.7900006.
14. Huang, G.; Chen, D.; Li, T.; Wu, F.; van der Maaten, L.; Weinberger, K.Q. Multi-Scale Dense Networks for Resource Efficient Image Classification. *arXiv* **2017**, arXiv:1703.09844.
15. Wang, X.; Yu, F.; Dou, Z.Y.; Darrell, T.; Gonzalez, J.E. SkipNet: Learning Dynamic Routing in Convolutional Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 409–424.
16. Wu, Z.; Nagarajan, T.; Kumar, A.; Rennie, S.; Davis, L.S.; Grauman, K.; Feris, R. BlockDrop: Dynamic Inference Paths in Residual Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8817–8826.
17. Xue, Z.; Marculescu, R. Dynamic Multimodal Fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 2575–2584. https://doi.org/10.1109/CVPRW59228.2023.00256.

18. Li, S.; Chen, C.; Han, J. SimMLM: A Simple Framework for Multi-Modal Learning with Missing Modality. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025; pp. 24068–24077. https://doi.org/10.1109/ICCV51701.2025.02231.
19. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive Mixtures of Local Experts. *Neural Comput.* **1991**, *3*, 79–87. https://doi.org/10.1162/neco.1991.3.1.79.
20. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv* **2017**, arXiv:1701.06538.
21. Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; Chen, Z. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. *arXiv* **2020**, arXiv:2006.16668.
22. Fedus, W.; Zoph, B.; Shazeer, N. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *J. Mach. Learn. Res.* **2022**, *23*, 1–39.
23. Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; Houlsby, N. Scaling Vision with Sparse Mixture of Experts. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–14 December 2021; Volume 34, pp. 8583–8595.
24. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1321–1330.
25. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016; Volume 48, pp. 1050–1059.
26. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 6402–6413.
27. Geifman, Y.; El-Yaniv, R. Selective Classification for Deep Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 4878–4887.
28. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; Volume 139, pp. 8748–8763.
29. Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI* **2025**, *2*, AIoa2400640. https://doi.org/10.1056/AIoa2400640.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008.
31. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
32. Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.H.; Jindi, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 72–78. https://doi.org/10.18653/v1/W19-1909.
33. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2020**, *36*, 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.
34. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc. HEALTH* **2021**, *3*, 2:1–2:23. https://doi.org/10.1145/3458754.
35. Williams, R.J. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.* **1992**, *8*, 229–256. https://doi.org/10.1007/BF00992696.
36. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.
37. Nan, F.; Saligrama, V. Adaptive Classification for Prediction Under a Budget. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 4727–4737.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.