FAIR-Coach: Biomechanically Informed Evidence Representation for Movement Diagnosis with Offline Squat-Based Evaluation

Monireh Vahdati

Thesis submitted to the University of Ottawa in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Electrical and Computer Engineering

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Thesis Supervisor: Abdulmotaleb El Saddik

© Monireh Vahdati, Ottawa, Canada, 2026

Abstract

Accurate classification is not sufficient for reliable movement diagnosis. The system must also identify and evaluate the biomechanical evidence underlying its decisions. Domain-shift experiments support this distinction: full or partial classifier collapse occurred in 77.8% of out-of-domain transfer cases.

A central gap is that changes in evidence representation often come with changes in reasoning complexity. Predictive accuracy alone does not establish whether a diagnosis relies on biomechanical evidence as intended.

This thesis investigates these questions using FAIR-Coach and a structured per-repetition representation called RepEvidence. The core empirical evaluation is offline and squat-based; Path A and live agentic interaction belong to the broader system architecture but are not part of the core diagnosis experiments. The evaluation compares generic, globally aggregated evidence with diagnosis-aligned biomechanical evidence and tests four relational reasoning configurations against corresponding non-relational controls. The primary comparison follows a pre-specified, frozen protocol on a held-out TEST split of five participants with paired bootstrap confidence intervals, followed by multi-seed and subject-level robustness analyses on the same TEST cohort.

Diagnosis-aligned evidence outperformed generic evidence at all seven seeds under an architecture-fixed Reasoner v2 comparison (mean $\Delta = +0.038$, SD$= 0.010$). A separate protocol-matched Concat+MLP rerun showed the same direction of effect (mean $\Delta = +0.052$, SD$= 0.010$), and the advantage was positive for all five TEST participants. Because the diagnosis-aligned configuration also adds ten measurements, these results do not isolate alignment alone, and absolute performance remained modest (TEST micro-F1 0.3435). In contrast, none of the four relational configurations showed a statistically confirmed advantage over their corresponding non-relational comparisons on the TEST split, and one performed significantly worse. In a corrected intervention-based evaluation of the final diagnosis model, only one of three tested labels satisfied the grounding criterion, so diagnosis-level grounding was not demonstrated. At the language layer, LoRA-DPO trained on template-constructed preference pairs improved diagnosis-to-language grounding under an automatic evaluation protocol on 40 held-out examples. Unsupported claims

decreased from 75% to 17.5%, evidence specificity increased from 75% to 92.5%, and consistency with the model's predicted diagnosis increased from 22.5% to 77.5%.

In this offline squat-diagnosis setting, the full diagnosis-aligned evidence configuration had a greater impact on predictive performance than increasing reasoning-architecture complexity. The results therefore support diagnosis-aligned evidence representation while also identifying clear limits to claims of diagnosis-level biomechanical grounding.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Abdulmotaleb El Saddik, for his guidance, support, and encouragement throughout my PhD journey. His mentorship has been invaluable to both my academic and personal growth. I am especially grateful for the trust he placed in me and for his constant encouragement to keep moving forward, even during the most challenging moments. His words of confidence, his thoughtful feedback, and his belief in my abilities gave me the motivation to persevere and continue growing as a researcher. This journey would not have been possible without his support and mentorship.

I would also like to sincerely thank my friend and research collaborator, Kamran Gholizadeh, for his support and companionship throughout this journey. Our collaborations led to several research publications and, more importantly, gave me many opportunities to learn, exchange ideas, and grow as a researcher. I am grateful for his help, encouragement, and the many discussions and experiences we shared along the way.

Finally, I would like to dedicate this thesis to my family, whose love, sacrifices, and unwavering support have been the foundation of this journey.

To my beloved father, whose physical presence I have deeply missed throughout this journey, but whose spirit has never left my side. His voice of encouragement, his unwavering belief in me, and the pride he always took in my efforts continue to live in my heart. Through the way he lived his life, he taught me the meaning of perseverance, strength, and never giving up. Although I can no longer share this moment with him in person, I carry his love, his lessons, and his belief in me in every step I take. This achievement is, in many ways, a continuation of the dreams and hopes he inspired in me, and I dedicate it to his memory with all my love.

To my beloved mother, whose love has been my quiet source of strength through every difficult moment. Her patience, kindness, and unwavering support have carried me through distance, challenges, and moments when I felt alone. Even from thousands of kilometers away, she has always found a way to remind me that I am never truly alone. Her love has been my sense of home, no matter where in the world I have been.

To my beloved sister and brother, who have always believed in me, encouraged me, and stood beside me throughout this journey. Their love, support, and constant encouragement have reminded me that I never had to walk this path alone.

To my family, this thesis is not only a reflection of my own journey, but also a tribute to the love, sacrifices, and unwavering faith that have carried me this far. This achievement belongs to all of you as much as it belongs to me.

Author Contributions

The author conceived, designed, implemented, and evaluated the research under the supervision of the thesis advisor. The author developed the FAIR-Coach framework, the structured RepEvidence representation, and the methodology for comparing generic and diagnosis-aligned biomechanical evidence.

The evidence-to-diagnosis experiments included non-relational baselines, relational-reasoning models, a subject-disjoint data-processing protocol, leakage safeguards, biomechanical reference construction, ablation studies, calibration analysis, grounding interventions, and statistical evaluation procedures. We conducted evaluations on the held-out TEST split and multi-seed robustness analyses for the closest relational comparisons.

The diagnosis-to-language grounding stage was developed to support constructing evidence-licensed preference pairs, training LoRA-DPO, and evaluating feedback faithfulness under contamination-controlled conditions. The complete FAIR-Coach pipeline integrated movement perception, structured evidence construction, diagnosis, reliability assessment, grounded feedback generation, logging, and reproducibility checks.

The author interpreted the results, formulated the research questions and contributions, and synthesised the findings with guidance and feedback from the thesis supervisor. Collaborative work that contributed to this thesis is identified and cited in the relevant chapters.

Declaration of Generative AI Use

During the revision of this doctoral thesis, generative artificial intelligence (AI) tools were used as supplementary resources to assist with text curation and the preparation and visual refinement of selected figures and diagrams. Their use is summarised as follows:

- **Claude Code (Anthropic):** Used to assist with reviewing and refining research-related materials, performing additional consistency checks, and supporting technical reliability assessments during the revision process.
- **ChatGPT (OpenAI):** Used for academic language polishing, improving text clarity and readability, and assisting with the generation and visual refinement of selected figures and diagrams based on the author's original concepts, instructions, and scientific requirements.
- **Grammarly:** Used for final proofreading, including grammar, spelling, punctuation, and language consistency checks.

All research concepts, methodologies, algorithms, implementations, experiments, analyses, and scientific contributions presented in this thesis were developed by the author. The use of AI tools was limited to supporting the revision and presentation of the thesis and did not replace the author's independent research, technical development, or scientific contributions.

All AI-assisted suggestions, revisions, and outputs were carefully reviewed, checked, and verified by the author before incorporation into the final thesis. The author retains full responsibility for the originality, accuracy, validity, and academic integrity of the thesis, including its research findings, interpretations, and conclusions.

Contents

Abstract **ii**

Acknowledgements **iv**

Author Contributions **vi**

Declaration of Generative AI Use **vii**

List of Figures **xii**

List of Tables **xv**

List of Abbreviations **xvii**

List of Symbols **xviii**

Nomenclature **xix**

1 Introduction **1**
- 1.1 Motivation and Background 1
- 1.2 Problem Statement and Research Gap 3
- 1.3 Research Questions and Hypotheses 6
- 1.4 Contributions and Scope 7
- 1.5 Scholarly Achievements 11
- 1.6 Thesis Organisation 14

2 Literature Review and Background **16**

2.1 Human Movement Analysis and Intelligent Coaching 16
2.2 Skeleton-Based Exercise Quality Assessment and Reliability 21
2.3 Biomechanical Evidence Representation 24
2.4 Evidence Fusion, Reasoning, and Grounded Feedback 27
2.5 Research Gap and Positioning of FAIR-Coach 30

3 FAIR-Coach Methodology **33**
3.1 Problem Formulation and System Overview 33
3.1.1 From Movement Observation to Evidence-Grounded Diagnosis . . . 33
3.1.2 FAIR-Coach Architecture 36
3.2 Movement Perception and Dual-Path Evidence Acquisition 38
3.2.1 Skeleton Representation and Reliability Gating 38
3.2.2 Continuous and Repetition-Level Evidence 39
3.3 Structured Biomechanical Evidence Representation 41
3.3.1 RepEvidence: Unified Per-Repetition Representation 41
3.3.2 Reference-Based and Diagnosis-Aligned Evidence 43
3.4 Evidence-to-Diagnosis Modelling 47
3.4.1 Structured Diagnosis Formulation 47
3.4.2 Non-Relational Fusion and Relational Reasoning 50
3.4.3 Model Comparison and Selection Protocol 53
3.5 Diagnosis Grounding and Grounded Feedback 55
3.5.1 Evidence-Grounding and Intervention Framework 55
3.5.2 Diagnosis-to-Language Feedback and Preference Optimisation . . . 59
3.6 Scope and Methodological Boundaries 61

4 Implementation and Evaluation **65**
4.1 Implementation and Experimental Protocol 66
4.1.1 Datasets and Their Experimental Roles 66
4.1.2 Subject-Disjoint Splits, Filtering, and Leakage Control 68
4.1.3 Reference Construction and Evidence Extraction 71
4.1.4 Model Training, Metrics, and Statistical Testing 75
4.2 Generic Evidence and Baseline Performance 77
4.2.1 Baseline Comparisons 77
4.2.2 Interpretation of Generic Reference Evidence 79
4.3 Diagnosis-Aligned Evidence and Relational Reasoning 80
4.3.1 Diagnosis-Aligned Evidence 80

4.3.2 Relational Versus Non-Relational Reasoning . . . 86
4.3.3 Final Diagnosis Model Selection . . . 90
4.4 Diagnosis Quality, Calibration, Ablation, and Grounding . . . 91
4.4.1 Predictive Performance and Calibration . . . 91
4.4.2 Evidence Ablation and Missing-Evidence Robustness . . . 93
4.4.3 Evidence-Grounding Interventions . . . 95
4.5 Diagnosis-to-Language Grounding and Final System Evaluation . . . 98
4.5.1 Language-Grounding Baseline and Preference Optimisation . . . 98
4.5.2 Final Pipeline Integration and Reliability Validation . . . 100
4.5.3 End-to-End Demonstration and Reproducibility Gate . . . 102
4.6 Overall Findings, Limitations, and Threats to Validity . . . 104
4.6.1 Findings with Respect to the Research Questions . . . 104
4.6.2 Limitations and Threats to Validity . . . 106
4.6.3 Chapter Summary . . . 115

5 Discussion and Synthesis 117
5.1 Chapter Overview . . . 117
5.2 Answering the Research Questions . . . 118
5.2.1 Evidence Representation and Diagnostic Performance . . . 118
5.2.2 The Limits of Relational Reasoning . . . 121
5.3 From Predictive Accuracy to Evidence-Grounding Evaluation . . . 124
5.4 Positioning FAIR-Coach Within Its Research Trajectory . . . 127
5.4.1 From Interaction to Structured Diagnosis . . . 127
5.4.2 Revisiting the Domain-Shift Motivation . . . 129
5.5 Implications for Evidence-Grounded Movement Diagnosis . . . 130
5.6 Methodological Trade-offs and Design Lessons . . . 132
5.7 Chapter Summary . . . 134

6 Conclusion 135
6.1 Chapter Overview . . . 135
6.2 Summary of the Research Problem and Approach . . . 135
6.3 Summary of Contributions . . . 137
6.4 Answers to the Research Questions . . . 139
6.5 Final Position on Relational Reasoning . . . 140
6.6 FAIR-Coach in Retrospect . . . 141
6.7 Limitations Carried Forward . . . 141

6.8 Future Work 144
6.9 Closing Statement 146

References **148**

APPENDICES **156**

A RepEvidence Schema and Evidence Groups **157**
A.1 RepEvidence Schema 157
A.2 Evidence-Group Definitions 157

B QEVD Subject Split **159**
B.1 Subject-Level Split 159

C Experimental Configuration **160**
C.1 Diagnosis-Model Configuration 160
C.2 LoRA-DPO Configuration 162
C.3 Evaluation Constants 163
C.4 Frozen Checkpoint Inventory 164

D Additional Results **165**

E Feedback Preference Construction **167**

List of Figures

2.1 Architecture of the multi-agent coaching framework introduced in [16]. Multimodal sensing and pose information are combined with short- and long-term memory and processed by specialised orchestration, recommendation, and feedback agents before the system returns voice/text feedback to the user. The figure summarises the multimodal, memory-augmented system foundation that precedes FAIR-Coach. 18

2.2 Real-time Digital Twin AI Fitness Coaching architecture from [17]. The system extends the earlier multi-agent framework with a web-based interaction loop, multimodal feature extraction, pose estimation, contextual memory, LLM-supported orchestration, and multimodal feedback for synchronous coaching. 19

2.3 Protocol-integrated multimodal coaching architecture based on [18]. The architecture extends prior coaching systems with A2A task delegation, remote agent executors, MCP-based tool/model access, multimodal output channels, and supporting observability and service components. 20

3.1 Overview of the FAIR-Coach architecture. Continuous and repetition-level movement evidence are organised into a structured biomechanical representation for diagnosis. Diagnosis is separated from downstream feedback generation, while orchestration and tool-access components provide supporting execution infrastructure. 37

4.1 QEVD data-processing and leakage-control protocol. Subject-disjoint splits are validated independently, while normalisation statistics and the biomechanical reference are derived from TRAIN only and reused unchanged for validation and TEST. Final evidence records are frozen after the corresponding leakage checks are satisfied. 70

4.2 Comparison of reference formulations and supplementary generic-deviation validation analysis. **(A)** The mean-curve reference provides greater clean-versus-error separation than the tested range-based variants. **(B)** A global-deviation analysis on EC3D (141 repetitions) yields AUC = 0.602 (95% CI $[0.501, 0.698]$). 74
4.3 Effect of evidence representation across two diagnosis architectures. VAL and TEST micro-F1 are shown for generic and diagnosis-aligned evidence using Concat+MLP and Reasoner v2. The annotated TEST comparison holds the Reasoner v2 architecture fixed and shows the original seed-42 improvement from diagnosis-aligned evidence ($\Delta = +0.039$, 95% CI [0.025, 0.053]). The direction was subsequently consistent across seven training seeds (42–48), with a mean $\Delta = +0.038$ and SD= 0.010; all seven seed-specific 95% confidence intervals excluded zero, as reported in Table 4.8. The Concat+MLP points shown here are the original historical runs and used different training-selection protocols; the protocol-matched Concat+MLP comparison is reported separately in Table 4.7. 85
4.4 Paired comparisons on the held-out TEST split evaluate relational mechanisms and evidence quality. In the first set of comparisons, the reasoning architecture is varied while the evidence representation remains fixed. In the final comparison, the evidence representation is varied with the Reasoner v2 architecture held constant. Filled markers indicate comparisons where the 95% confidence interval does not include zero, and open markers indicate intervals that do include zero. The results show that the coarse relational model performs significantly worse than the non-relational baseline. Comparisons involving the generic Concat+MLP use the original historical fixed-epoch 8.8 baseline and should therefore be interpreted with the baseline-training caveat described in Section 4.6.2. The only statistically confirmed improvement in the displayed seed-42 comparisons is observed when generic evidence is replaced with diagnosis-aligned evidence; this effect was consistent across seven training seeds (Table 4.8). 88

4.5 Calibration and per-label diagnostic performance on the locked TEST split (n = 1,052) for three generic-evidence models with persisted per-example predictions. **(A)** Ten-bin reliability diagrams compare mean predicted probability with observed empirical frequency; the dashed diagonal represents perfect calibration. **(B)** Per-label F1 across the eight diagnosis categories. The locked coarse relational model is not included in the per-example analysis because its TEST predictions are unavailable. Table 4.13 reports its aggregate results. 93
4.6 Ablation results for each evidence group were obtained using the locked TEST split. Filled markers indicate 95% bootstrap confidence intervals that do not include zero, while hollow markers indicate intervals that include zero. Knee and trunk ablations were evaluated only for fine-OFF and for 8.8 Concat+MLP, both under the frozen evaluation protocol. 95
4.7 End-to-end FAIR-Coach trace on a real, unmodified video (`SQUAT_516.mp4`), from raw evidence through reliability assessment, diagnosis, and preference-optimised feedback, concluding with a passed hard reproducibility gate. . . 103

List of Tables

3.1 Principal blocks of the generic `RepEvidence` representation used by the diagnosis experiments. . . . 42
3.2 TRAIN-derived positive-class weights used by the learned neural diagnosis models. . . . 49
3.3 Diagnosis-model families evaluated in FAIR-Coach. . . . 52
3.4 Corrected intervention directions used in the grounding re-evaluation. . . . 57

4.1 Datasets used in the final FAIR-Coach evaluation and their restricted roles. 67
4.2 Final subject-disjoint QEVD split and validator acceptance. . . . 68
4.3 Validator rejections by reason. . . . 69
4.4 Contradictory-label sensitivity analyses for the protocol-matched Concat+MLP comparison on TEST. . . . 72
4.5 Validation performance of the generic-evidence baselines. . . . 78
4.6 Original Concat+MLP results obtained under different historical training protocols. A protocol-matched comparison is reported in Table 4.7. . . . 80
4.7 Protocol-matched multi-seed comparison of generic and diagnosis-aligned Concat+MLP models on TEST. . . . 81
4.8 Seven-seed robustness analysis of the diagnosis-aligned versus generic evidence comparison using Reasoner v2 on TEST. . . . 82
4.9 Post hoc per-exercise TEST micro-F1 for the protocol-matched Concat+MLP comparison (seed 42). . . . 83
4.10 In the relational-reasoning experiments, conclusions come from paired bootstrap comparisons rather than relying solely on point estimates. . . . 86
4.11 Paired TEST micro-F1 differences between relational configurations and their non-relational controls (seed 42). . . . 86
4.12 Seven-seed follow-ups of the two borderline relational comparisons on TEST (paired Δ micro-F1 with 95% CI). . . . 89

4.13 Locked TEST diagnosis and calibration metrics ($n = 1{,}052$). Values for the diagnosis-aligned MLP marked † were computed post hoc from frozen TEST predictions and were not used for model selection. 91
4.14 TEST evidence ablations. Positive Δ indicates that removing the evidence reduced micro-F1. 94
4.15 Corrected evidence-grounding intervention results for the final diagnosis-aligned Concat+MLP model, with only generic evidence features perturbed. 96
4.16 Mean relevant-evidence Δp_ℓ across grounding evaluations ($n = 30$ cases per label). 97
4.17 Diagnosis-to-language grounding before and after LoRA-DPO on the held-out $n = 40$ evaluation set. 99
4.18 Final end-to-end demonstration on `SQUAT_516.mp4`. 102
4.19 Summary of the thesis research questions and the corresponding experimental evidence. 105
4.20 Per-participant TEST micro-F1 for the protocol-matched Concat+MLP comparison (seed 42). 108
4.21 Chronological record of substantive access to the TEST split. 110

A.1 Final `RepEvidence` input schema and TRAIN-set missingness. 158
A.2 Evidence-group definitions used in the relational-reasoning experiments. . . 158

B.1 QEVD subject-level TRAIN, validation, and TEST assignments. 159

C.1 Principal diagnosis-model configurations. 161
C.2 Final LoRA-DPO configuration. 162
C.3 Principal constants of the frozen evaluation protocol. 163
C.4 Final frozen checkpoint inventory. 164

D.1 Per-label F1 scores on the TEST split (threshold 0.5) for fine-OFF, fine-ON, the 8.8 Concat+MLP baseline, and the final diagnosis-aligned Concat+MLP (Diag. MLP). 165
D.2 Trainable parameters of the frozen diagnosis checkpoints. 166

E.1 Preference-pair construction used for diagnosis-to-language grounding. . . . 167

List of Abbreviations

A2A Agent-to-Agent xvi

AQA Action Quality Assessment xvi

AUC Area Under the Receiver Operating Characteristic Curve xvi

CI Confidence Interval xvi

DPO Direct Preference Optimization xvi

ECE Expected Calibration Error xvi

GCN Graph Convolutional Network xvi

LLM Large Language Model xvi

LoRA Low-Rank Adaptation xvi

MCP Model Context Protocol xvi

MLP Multilayer Perceptron xvi

ROM Range of Motion xvi

ST-GCN Spatial Temporal Graph Convolutional Network xvi

List of Symbols

$\alpha_{\ell g}$ Label-conditioned attention weight for diagnosis label ℓ and evidence group g xvi

Δp_{ℓ} Change in predicted probability for diagnosis label ℓ under an evidence intervention xvi

E_r Evidence representation for repetition r xvi

L Number of diagnosis labels; $L = 8$ in the core diagnosis task xvi

N Number of repetitions or evaluated examples xvi

$p_{i,\ell}$ Predicted probability for diagnosis label ℓ of example i xvi

$\hat{\mathbf{y}}_i$ Thresholded predicted multi-label diagnosis vector for example i xvi

$\mathbf{y}_i$ Ground-truth multi-label diagnosis vector for example i xvi

Nomenclature

`insufficient` Term used in two distinct senses: as a QEVD-derived diagnosis label and as a FAIR-Coach reliability state xvi

Diagnosis-aligned evidence Biomechanical evidence organised around features relevant to specific movement-diagnosis categories xvi

Evidence grounding Degree to which a diagnostic prediction responds appropriately to diagnostically relevant evidence while remaining stable to irrelevant evidence xvi

FAIR-Coach Coaching system whose name originates from the Flexible, Agentic, Intelligent, Real-Time lineage of prior work; the core empirical evaluation in this thesis is offline and squat-based xvi

Generic evidence Globally aggregated movement evidence that is not organised around features specific to individual diagnosis categories xvi

Language grounding Degree to which generated coaching feedback remains faithful to and consistent with an already-fixed StructuredDiagnosis xvi

Path A Continuous sliding-window evidence path used in the broader live FAIR-Coach architecture xvi

Path B Completed-repetition evidence path used in the core offline diagnosis experiments xvi

QEVD-FIT-COACH Primary benchmark dataset used for the core squat-based diagnosis experiments xvi

RepEvidence Structured repetition-level representation storing movement, biomechanical, reference, reliability, and contextual evidence together with explicit missingness information xvi

StructuredDiagnosis Structured output of the evidence-to-diagnosis stage containing the predicted movement diagnosis and associated confidence xvi

1

Introduction

1.1 Motivation and Background

Intelligent movement coaching requires more than recognising an exercise. A useful system must observe how a movement is performed, identify a meaningful execution problem, and communicate corrective feedback [1–3]. Recent work in human-motion understanding increasingly reflects this shift from action recognition toward fine-grained assessment [4–6]. Human-centric motion representations, pose-enhanced vision–language models, and action-quality assessment (AQA) methods increasingly model body structure, temporal phases, and task-specific semantics explicitly [7–10]. Semantic and rubric-informed AQA further demonstrate that assessment can benefit from explicit domain knowledge and structured evaluation criteria [5, 11, 12].

This distinction matters in fitness and rehabilitation, where repeated executions of the same exercise may vary in quality even when classified as the same action [5, 13, 14]. For this reason, a coaching system must separate movement identity, movement quality, and movement diagnosis [5, 13]. Benchmarks such as QEVD address this challenge by requiring

continuous user monitoring and timely feedback [15]. Recent work, including ExpertAF, extends video-based assessment to provide expert-style, actionable feedback [1, 3].

Earlier work in this research trajectory established the system-level foundations for this problem. A multimodal multi-agent coaching system integrated pose analysis, language interaction, memory, and specialised agents for personalised feedback [16]. Subsequent work extended this framework toward real-time multimodal coaching [17] and interoperable agent/tool coordination using Agent-to-Agent communication and related infrastructure [18]. These systems established increasingly capable coaching pipelines, but system-level interaction and orchestration alone do not establish whether the underlying movement diagnosis is supported by appropriate biomechanical evidence. The name FAIR-Coach refers to the *Flexible, Agentic, Intelligent, Real-Time* coaching-system lineage established across prior work. This use of the acronym FAIR is specific to the system name and is unrelated to the standard FAIR data principles of Findable, Accessible, Interoperable, and Reusable [19]. In this thesis, however, the core empirical evaluation is intentionally narrower: the diagnosis experiments are offline and squat-based, and Path A and live agentic interaction are implemented in the broader architecture but are not part of the core diagnosis evaluation.

This concern is also supported by a cross-domain benchmark conducted as part of this research [20]. Across nine GCN- and transformer-based models and four heterogeneous datasets, 77.8% of out-of-domain transfer experiments showed full or partial classifier collapse, and no evaluated architecture provided a consistent solution after statistical correction. These findings do not shift the thesis toward domain generalisation. Instead,

they provide empirical motivation for examining the limitations of an architecture-centric approach to movement understanding.

FAIR-Coach therefore shifts the scientific focus from "which model is more sophisticated?" to two linked questions: *what biomechanical evidence should support a movement diagnosis, and does explicit relational reasoning add value once that evidence is fixed?* Grounding analyses then examine whether diagnostic predictions respond to the intended evidence [21, 22], while the downstream language evaluation tests whether feedback remains faithful to a fixed diagnosis [23, 24]. This separation between *evidence*, *diagnosis*, and *communication* defines the thesis's final direction.

1.2 Problem Statement and Research Gap

Recent AQA and human-motion research suggests that coarse global representations may be insufficient for fine-grained assessment [10, 13]. FineParser emphasises localised spatial–temporal parsing [8]; vision–language AQA aligns domain terminology with multiple temporal scales [11]; and RICA^2 incorporates expert rubric structure and uncertainty into quality assessment [12]. Together, these studies support the principle that an assessment model should receive evidence that reflects the structure of the task being assessed.

However, two factors can be difficult to disentangle: *evidence representation* and *reasoning architecture*. More elaborate graph or attention mechanisms may change how evidence is combined [25–28], but greater architectural complexity does not establish that the un-

derlying evidence is diagnostically meaningful. Conversely, appropriately aligned evidence may allow a simpler fusion model to capture the information needed for diagnosis. FAIR-Coach therefore evaluates evidence representation and reasoning architecture separately.

The first gap concerns how evidence is represented. Generic measures, including overall deviation scores or classifier probabilities for an entire repetition, can summarise movement quality [29, 30] but may not identify the specific biomechanical characteristics associated with different error categories. Diagnosis-aligned evidence instead preserves spatial- or phase-specific characteristics relevant to particular diagnoses [31, 32].

The second gap concerns evaluation. Predictive accuracy alone does not demonstrate that a model relies on biomechanically meaningful evidence. Intervention-based interpretability research addresses this by modifying specific intermediate variables and measuring their effect on model predictions, rather than depending solely on post-hoc explanations [21, 22, 33]. In this work, FAIR-Coach implements this approach using evidence ablation and directed interventions on label-relevant evidence and non-primary-evidence control features.

The third gap concerns the language layer. Even with a fixed diagnosis, generated feedback can contain claims that are unsupported by or inconsistent with its grounding information [23, 24, 34, 35]. Preference optimisation provides a mechanism for adapting downstream responses toward preferred behaviour [36–39]. In FAIR-Coach, this adaptation is restricted to the language layer and does not alter the underlying movement diagnosis. Diagnosis grounding and language grounding are therefore evaluated separately.

The following problem statement summarises the main research questions this thesis addresses.

> **Thesis problem statement.** Movement-diagnosis systems may rely on aggregated evidence or classifier outputs that do not explicitly preserve the distinct biomechanical signatures of different movement errors. At the same time, introducing more complex reasoning architectures does not establish whether the underlying evidence is appropriate for diagnosis. This thesis examines how heterogeneous biomechanical evidence should be represented and evaluated for movement diagnosis, and whether explicit relational reasoning provides measurable benefit beyond simpler fusion of the same evidence. It further evaluates whether diagnostic predictions respond to intended biomechanical evidence and whether a fixed diagnosis can be communicated without unsupported language.

The experimental results support the value of diagnosis-aligned, literature-grounded evidence for diagnostic performance, while none of the tested relational architectures achieved a statistically confirmed advantage over corresponding non-relational fusion. The predefined intervention analysis did not establish intended biomechanical grounding under its defined criterion. With the diagnosis held fixed, lightweight preference optimisation improved language-grounding measures under the bounded automatic evaluation protocol. Together, these findings motivate the thesis's distinction between evidence quality, reasoning complexity, diagnosis grounding, and downstream communication.

1.3 Research Questions and Hypotheses

The thesis is organised around one central research question, addressed through two primary sub-questions.

> **Main RQ:** How should heterogeneous biomechanical evidence be represented and evaluated to support movement diagnosis and explicit evaluation of evidence grounding?

RQ1 – Evidence representation and diagnostic value. Does diagnosis-aligned, literature-grounded biomechanical evidence improve movement-diagnosis performance compared with generic, globally aggregated evidence on held-out subjects?

RQ2 – Relational reasoning. Does explicit relational reasoning over heterogeneous biomechanical evidence provide measurable benefit beyond corresponding non-relational fusion of the same evidence?

Grounding is treated as an evaluation dimension across the framework, rather than as a separate research question. It examines whether diagnostic predictions respond to the biomechanical evidence intended to support them.

A complementary objective is to evaluate whether lightweight preference optimisation can improve the faithfulness of coaching language while holding the structured diagnosis fixed.

The corresponding hypotheses are:

H1. Diagnosis-aligned, literature-grounded biomechanical evidence is expected to improve movement-diagnosis performance compared with generic, globally aggregated evidence under the evaluated conditions.

H2. Explicit relational modelling of heterogeneous evidence is expected to provide better performance than corresponding non-relational fusion of the same evidence under the evaluated conditions.

H3. Preference optimisation using evidence-licensed response pairs is expected to improve the faithfulness of language generated from a fixed diagnosis under the evaluated conditions.

The available evidence supports H1 within QEVD. H2 is not supported, as none of the tested relational architectures achieved a statistically confirmed improvement over their corresponding non-relational baselines on the held-out TEST split. The grounding evaluation did not demonstrate overall biomechanical grounding under the predefined intervention criterion. H3 is supported within the bounded automatic language-grounding evaluation. Detailed evidence is reported in Chapter 4.

1.4 Contributions and Scope

This thesis makes four main contributions.

C1. Diagnosis-aligned organisation of established biomechanical evidence. FAIR-Coach organises established biomechanical measurements, including knee and hip angles, trunk inclination, knee–ankle offset, stance width, and left–right symmetry, into an evidence representation aligned with specific diagnosis categories. The individual measurements are not novel; related joint-angle and clinician-defined features have been used in prior exercise-assessment work [40, 41]. The contribution is their diagnosis-aligned organisation for multi-label diagnosis and its controlled comparison with generic evidence. In the pre-specified Reasoner v2 comparison, diagnosis-aligned evidence improved TEST micro-F1 at all seven training seeds (mean $\Delta = +0.038$). A protocol-matched Concat+MLP comparison showed the same direction, with mean micro-F1 increasing from 0.283 to 0.335 across seven seeds. Because these seed-level and later subject-level analyses reuse one five-participant TEST cohort, they are supporting analyses rather than independent replications. The comparison supports the full evidence design; it does not isolate alignment alone, because the diagnosis-aligned configuration also adds features and input dimensions.

C2. Controlled evaluation of explicit relational reasoning. Four relational configurations were compared with corresponding non-relational controls using frozen subject-level splits and staged evaluations on the held-out TEST split. None of the tested relational configurations achieved a statistically confirmed improvement over their corresponding non-relational controls. This result is limited to the evaluated configurations and does not establish that relational reasoning is ineffective in gen-

eral.

C3. Evidence-grounding evaluation for movement diagnosis. FAIR-Coach adapts intervention-based evaluation [21, 33] to movement diagnosis, testing whether diagnostic predictions respond to label-relevant evidence as intended, separately from predictive performance. In the corrected evaluation of the final diagnosis model, only one of three tested diagnoses satisfied the predefined grounding criterion when generic evidence was perturbed, and neither of the two labels that could be tested passed when the diagnosis-aligned features were also perturbed. Overall biomechanical grounding was therefore not demonstrated. The interventions modify extracted features rather than physically coherent movements, and their sign conventions are design assumptions rather than expert-validated mappings. The protocol remains a methodological contribution for testing this property.

C4. Diagnosis-to-language grounding through lightweight preference optimisation. The evaluation held the structured diagnosis fixed and trained a LoRA-DPO feedback model using constructed evidence-licensed preference pairs. On a 40-example held-out set with template-constructed preference pairs and automatic scoring, this approach improved measures of unsupported-claim avoidance and consistency with the fixed diagnosis, but did not eliminate unsupported claims. The diagnosis is fixed before generation and is not modified by the language model, whose role is limited to verbalising it. The result has not been validated through expert or user evaluation, and it does not establish that an LLM is necessary or superior to a

carefully designed template-based feedback system.

Taken together, the central finding is that, within the evaluated QEVD setting, how established biomechanical evidence was organised and evaluated mattered more for diagnostic performance than added relational complexity, and higher predictive performance did not by itself establish that diagnoses depended on the intended biomechanical evidence.

The claims in this thesis are limited to the evaluated setting. The final diagnosis experiments use an eight-category, multi-label taxonomy based on QEVD fine-grained annotations [15]. The evaluated data include accepted clips of basic squats and several squat-based exercise variants, whereas the generic reference representation is constructed from a basic-squat reference cohort to avoid averaging across distinct movement variants. Although the evidence-to-diagnosis framework is designed to be adaptable to other exercises by specifying appropriate biomechanical evidence, reference criteria, and diagnosis labels, this thesis evaluates it only within this squat-based setting. The main held-out TEST split contains 1,052 clips from only five independent participants, and several analyses reused this cohort; the results therefore characterise performance within QEVD and do not establish external generalisation. All results apply to the 6,310 of 13,013 clips accepted by the clip validator, whose acceptance rates differed across labels and subjects. The annotations also contain contradictory labels, such as 371 TRAIN clips labelled both *no obvious issue* and at least one error. Validation across additional exercises, participants, and rehabilitation tasks remains as future work.

The core diagnosis experiments use short, monocular 2D QEVD clips. Path A, the

continuous sliding-window branch of the broader live architecture, is not available in the core QEVD diagnosis records and is therefore not part of the empirical comparison. The evaluated diagnosis pipeline uses Path B together with biomechanical and reference-based evidence. Synchronised multi-view video and 3D kinematics are outside the evaluated setting; consequently, out-of-plane motion, absolute depth, and spinal curvature are not assessed directly. Responsible-joint and exact movement-phase outputs are not used as supervised targets because QEVD does not provide valid ground truth for them. All diagnosis experiments are conducted offline and address movement-quality assessment for coaching rather than longitudinal, clinical, or high-stakes decision-making. The evaluated models are not sufficiently calibrated for their predicted probabilities to be interpreted as dependable confidence estimates.

This work uses RTMPose, ST-GCN, LangGraph, A2A, MCP, logging, and user-interface components as supporting infrastructure. The main research claims address evidence representation, comparative diagnosis modelling, grounding evaluation, and controlled diagnosis-to-language communication.

1.5 Scholarly Achievements

This section lists publications and manuscripts that include the author of this thesis as an author or co-author. Entries are grouped by year, with first-author publications listed before co-authored publications.

1. Monica Vahdati, Kamran Gholizadeh HamlAbadi, Ali Etemad, and Abdulmotaleb El Saddik, "Skeleton-Based Exercise Quality Assessment Under Domain Shift: A Benchmark Study of Cross-Domain Failure Modes," accepted at *The 3rd International Workshop on Multimedia Computing for Health and Medicine (MCHM '26)*, co-located with ACM Multimedia 2026, Rio de Janeiro, Brazil, 2026, doi: 10.1145/3840471.3842592.

2. Monica Vahdati, Fedwa Laamarti, and Abdulmotaleb El Saddik, "Meta-Review of Wearable Devices for Healthcare in the Metaverse," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 21, no. 7, Article 203, 36 pages, 2025, doi: 10.1145/3705320.

3. Monica Vahdati, Kamran Gholizadeh HamlAbadi, Haiwei Dong, Basim Hafidh, and Abdulmotaleb El Saddik, "Real-Time 2D-to-3D LiDAR Reconstruction and Omniverse-Based Visualization for Scalable Point Cloud Applications," *Proceedings of the International Workshop on Application-driven Point Cloud Processing and 3D Vision (APP3DV '25)*, pp. 17–24, 2025, doi: 10.1145/3728486.3759213.

4. Monica Vahdati, Kamran Gholizadeh HamlAbadi, Fedwa Laamarti, and Abdulmotaleb El Saddik, "A Multi-Agent Digital Twin Framework for AI-Driven Fitness Coaching," *Proceedings of the 2025 ACM International Conference on Interactive Media Experiences (IMX '25)*, pp. 380–385, 2025, doi: 10.1145/3706370.3731651.

5. Monica Vahdati, Kamran Gholizadeh HamlAbadi, Fedwa Laamarti, Devansh Kumar, and Abdulmotaleb El Saddik, "Digital Twin AI Fitness Coach: An Intelligent Multi-Agent System for Personalized Exercise Guidance," *Proceedings of the 1st International*

Workshop on Multi-Sensorial Media and Applications (MSMA '25), pp. 18–26, 2025, doi: 10.1145/3728485.3759171. **Best Student Paper Award.**

6. Kamran Gholizadeh HamlAbadi, Monica Vahdati, and Abdulmotaleb El Saddik, "MedPhyGraph: Counterfactual Support-Graph Maintenance for Dynamic Built-Environment Digital Twins," *ECCV 2026 TwinWorld: 1st Workshop on Visual Intelligence for Built Environment Digital Twins*, 2026.

7. Kamran Gholizadeh HamlAbadi, Monica Vahdati, Ali Etemad, and Abdulmotaleb El Saddik, "Benchmarking the Physical Fidelity of Feed-Forward 3D Reconstruction for Physical AI," *The International Workshop on Industrial Physical AI (WIPAI '26)*, Rio de Janeiro, Brazil, 2026, 11 pages, doi: 10.1145/3841560.3841563.

8. Kamran Gholizadeh HamlAbadi, Monica Vahdati, and Abdulmotaleb El Saddik, "NemoBiopsy: Agentic Language-to-Action for Precision Robotic Liver Biopsy," submitted to the *AAAI 2027 Innovative Applications of Artificial Intelligence (IAAI)* track, 2026.

9. Kamran Gholizadeh HamlAbadi, Monica Vahdati, Haiwei Dong, and Abdulmotaleb El Saddik, "AI-Enhanced Creation of Digital Twins from iPhone LiDAR for Immersive XR Experiences in NVIDIA Omniverse," *Proceedings of the International Workshop on Intelligent Immersification in the Metaverse: AI-Driven Immersive Multimedia (I2M-MM '25)*, pp. 25–33, 2025, doi: 10.1145/3728487.3759232.

10. Kamran Gholizadeh HamlAbadi, Monica Vahdati, Fedwa Laamarti, and Abdulmotaleb

El Saddik, "Agent-to-Agent (A2A) Protocol Integrated Digital Twin System with AgentIQ for Multimodal AI Fitness Coaching and Personalized Well-Being," *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM '25)*, pp. 12483–12491, 2025, doi: 10.1145/3746027.3758176.

11. Y. E. Valdivieso et al., "The Potential of Olfactory Stimuli in Stress Reduction Through Virtual Reality," *2025 IEEE Medical Measurements & Applications (MeMeA)*, Chania, Greece, pp. 1–6, 2025, doi: 10.1109/MeMeA65319.2025.11068102.

1.6 Thesis Organisation

Chapter 2 reviews the literature on intelligent movement coaching, skeleton-based exercise quality assessment and reliability, biomechanical evidence representation, evidence fusion, grounding, and preference-based feedback alignment. It concludes by positioning FAIR-Coach around the distinction between evidence quality and reasoning complexity.

Chapter 3 presents the FAIR-Coach methodology, including the data protocol, movement-evidence extraction, generic and diagnosis-aligned evidence, relational and non-relational diagnosis models, the statistical evaluation protocol, grounding interventions, and diagnosis-to-language adaptation.

Chapter 4 presents the implementation and experimental results. It evaluates reference formulations, generic and diagnosis-aligned evidence, relational and non-relational diagnosis models, grounding behaviour, and downstream language adaptation using held-out

evaluation.

Chapter 5 interprets and synthesises the experimental findings. It examines why evidence representation produced stronger gains than explicit relational reasoning in the evaluated setting, distinguishes predictive performance from evidence grounding, situates FAIR-Coach within the broader research trajectory, and discusses methodological implications and design lessons.

Chapter 6 concludes the thesis by summarising the main contributions and addressing the research questions. It presents the final position on relational reasoning, discusses the main limitations, and outlines future work on joint Path A/Path B evaluation, controlled 3D evidence, improved grounding methods, independent human evaluation, and live longitudinal coaching.

2

Literature Review and Background

This chapter reviews the literature relevant to positioning FAIR-Coach. Rather than surveying all work on human-action analysis, the review is organised around four questions that motivate the thesis. These are: how human movement is represented for intelligent coaching; how exercise quality is assessed and how reliable those assessments are; how biomechanical evidence can be used for diagnosis; and how heterogeneous evidence and language feedback are evaluated for grounding.

2.1 Human Movement Analysis and Intelligent Coaching

Recent work in human movement analysis has shifted focus from action recognition to more detailed interpretation of movement execution [4–6]. Skeleton-based representations are commonly used because they preserve body structure and reduce reliance on appearance or background information [27, 42]. ST-GCN models human motion as a spatiotemporal graph, where joints are represented as nodes connected across time [25]. CTR-GCN extends

this approach by introducing adaptive refinement of the graph topology [26]. SkateFormer applies structured skeletal-temporal attention to capture multiple types of joint and temporal relationships [28]. These approaches illustrate the effectiveness of explicitly modelling spatial and temporal structure in human motion [27].

Building on these advances, subsequent work has incorporated information beyond skeleton-only recognition [43]. For example, PeVL integrates pose, video, and language to support fine-grained action recognition [7]. H-MoRe focuses on learning motion representations that emphasise human movement while reducing the influence of unrelated scene dynamics [9]. HuMoCon identifies motion concepts with semantic meaning by aligning multiple modalities [44]. While these methods address different tasks, their results indicate that the quality of the underlying representation influences the inferences downstream models can make.

Coaching requires the system to distinguish between different executions of the same action [5, 45]. For example, a squat can vary in depth, stance width, symmetry, or trunk posture without changing the action category. Recent AQA benchmarks likewise emphasise fine-grained assessment of execution quality within the same activity, including multimodal and detailed performance annotations [13, 14, 46]. Action recognition alone is therefore insufficient for coaching; the system must also assess movement quality and identify the basis for corrective feedback [2, 3].

This requirement also appears in the broader health and rehabilitation context discussed in earlier work [47]. A meta-review of wearable systems identified rehabilitation,

monitoring, and personalised healthcare as key application areas, and noted the difficulty of converting multimodal measurements into reliable decisions [48]. The initial system in this thesis's coaching lineage was a multi-agent framework that integrated pose analysis, multimodal interaction, memory components, and specialised agents for orchestration, recommendation, and feedback [16]. Figure 2.1 summarises the framework, which structures sensing and user context around a memory-augmented coaching core and provides multimodal feedback through a closed interaction loop.

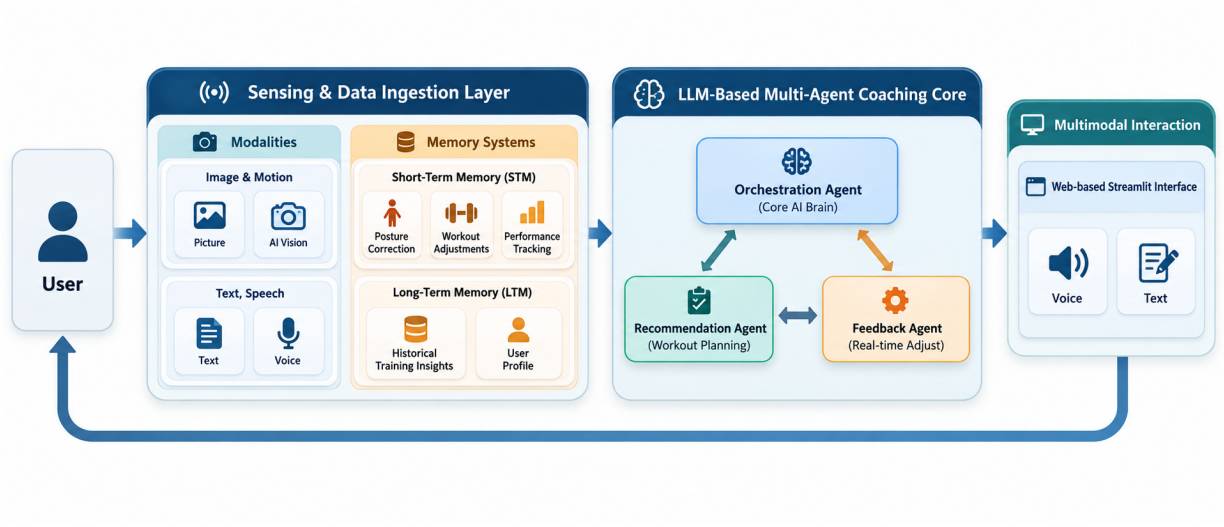


Figure 2.1: Architecture of the multi-agent coaching framework introduced in [16]. Multimodal sensing and pose information are combined with short- and long-term memory and processed by specialised orchestration, recommendation, and feedback agents before the system returns voice/text feedback to the user. The figure summarises the multimodal, memory-augmented system foundation that precedes FAIR-Coach.

The architecture was extended to support synchronous, real-time coaching [17]. The system uses a Streamlit-based interaction loop, live multimodal input, pose estimation,

contextual memory, and LLM-backed orchestration to provide low-latency feedback. As shown in Figure 2.2, user input passes through multimodal feature extraction and memory management, then through orchestration, recommendation, and feedback agents. The system then returns voice or text feedback through the same interactive loop. This extension shifts the system from primarily multimodal personalised coaching to continuous and responsive interaction.

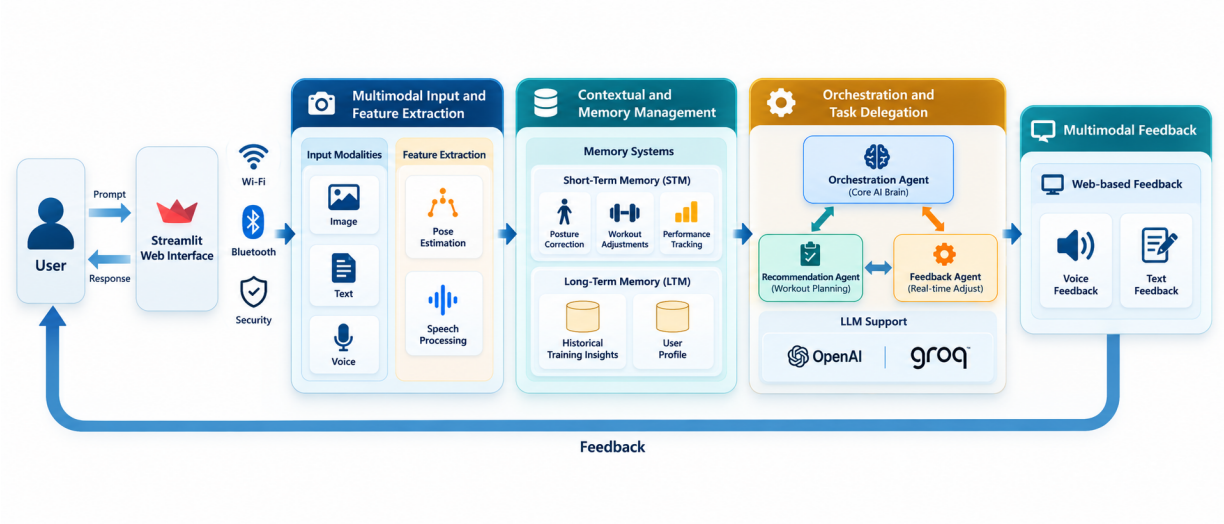


Figure 2.2: Real-time Digital Twin AI Fitness Coaching architecture from [17]. The system extends the earlier multi-agent framework with a web-based interaction loop, multimodal feature extraction, pose estimation, contextual memory, LLM-supported orchestration, and multimodal feedback for synchronous coaching.

A prior study examined a systems-level limitation that arises as the number of agents, models, tools, and output modalities increases. In such settings, tightly coupled orchestration becomes less scalable and more difficult to monitor [49–51]. The A2A-integrated architecture addresses this by introducing protocol-level agent communication, structured

task delegation, MCP-based access to tools and models, and observability-focused services [18]. Figure 2.3 illustrates the layered structure: an A2A orchestration layer processes multimodal input and contextual memory, remote agents use MCP to access external capabilities, and multimodal outputs provide user feedback. This architecture extends the earlier coaching pipeline to support a more interoperable agent-and-tool ecosystem.

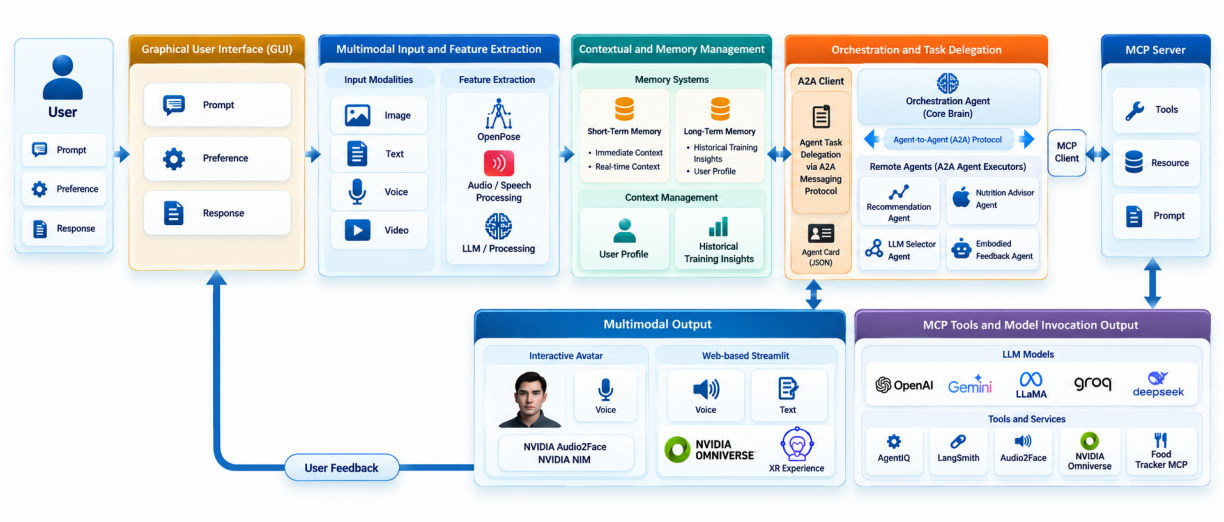


Figure 2.3: Protocol-integrated multimodal coaching architecture based on [18]. The architecture extends prior coaching systems with A2A task delegation, remote agent executors, MCP-based tool/model access, multimodal output channels, and supporting observability and service components.

Figures 2.1–2.3 illustrate the development of FAIR-Coach, beginning with multimodal and memory-augmented coaching, followed by real-time interaction, and then interoperable agent and tool coordination. These systems primarily provide infrastructure for intelligent interaction, not for biomechanical diagnosis. FAIR-Coach extends this infrastructure and shifts the focus to the evidence required for movement diagnosis and the methods used to

evaluate that evidence.

Recent work has examined how coaching systems can provide more situated and actionable feedback. QEVD models live fitness coaching as an asynchronous interaction problem, where the system observes activity, detects errors, and determines appropriate communication [15]. ExpertAF extends assessment by generating actionable feedback from video and pose data [1]. These approaches show that current systems connect perception with language [2, 3, 43, 47]. As a result, the reliability of both diagnosis and supporting evidence becomes central, since feedback is only useful when grounded in accurate information [23].

The methodology developed in this thesis is designed to apply to tasks beyond squat assessment. For a different exercise or rehabilitation task, the diagnosis taxonomy and relevant biomechanical evidence would differ [14, 52]. Still, the design principle remains the same: identify task-specific evidence, preserve it explicitly, evaluate its contribution to diagnosis, and keep it separate from verbalisation.This principle is evaluated only on squat-based QEVD data, and no empirical generalisation to other exercises or clinical settings is claimed.

2.2 Skeleton-Based Exercise Quality Assessment and Reliability

Action Quality Assessment (AQA) focuses on evaluating how well an action is performed rather than only identifying which action occurs. Early AQA research formulated

movement-quality assessment primarily through action-score estimation, with later work extending this direction through multitask learning and uncertainty-aware score modelling [53–56]. Recent systematic reviews show that AQA spans video-, skeleton-, and multimodal approaches across applications such as sports, skill assessment, and rehabilitation [57]. Recent methods increasingly emphasise more structured and fine-grained representations [10, 58, 59]. FineParser models fine-grained spatiotemporal structure [8], vision-language AQA aligns domain terminology with video representations at multiple levels [11], and RICA^2 incorporates expert rubrics and prediction uncertainty into assessment [12]. Together, these works motivate representations that better match the decomposition required by the assessment task.

In FAIR-Coach, exercise diagnosis presents a different challenge from conventional AQA scoring because the target is categorical and multi-label rather than a single scalar score. Individual exercises may include several distinct execution errors, each linked to specific movement features. As a result, methods that predict an overall quality score well do not always provide the information needed to identify or separate error categories [5, 13]. Recent exercise-specific systems have also moved beyond scalar quality scores toward identifying incorrect form and providing corrective guidance from pose or skeleton representations [29, 30].

Clinical movement-assessment literature similarly uses task-specific biomechanical measures and scoring criteria to distinguish movement-quality patterns, while also reporting substantial variation in their reliability and validity [60–62].

Reliability is an additional concern. A recent systematic review of markerless lower-extremity and trunk kinematics found that reliability and agreement vary across tasks, joints, planes of motion, and measurement systems [63, 64]. Markerless skeleton-based pipelines depend on estimated keypoints and may be sensitive to perturbations and pose-estimator changes. MaskCLR, for example, reports substantial degradation under skeleton perturbations and cross-estimator changes [65]. Multimodal video test-time adaptation research further shows that temporal and modality-specific domain shifts create distinct adaptation challenges [66]. In incomplete multimodal AQA, MCMoE further shows that performance can deteriorate when modalities expected during training are unavailable at inference [67]. These findings are relevant to FAIR-Coach because its movement evidence is heterogeneous and not always uniformly available [68, 69].

A domain-shift benchmark conducted as part of this research examined this problem directly for skeleton-based exercise-quality assessment [20]. The study evaluated nine GCN- and transformer-based models across four heterogeneous datasets under in-domain, out-of-domain, and mixed protocols. Although several models performed strongly within matched conditions, 77.8% of out-of-domain transfers exhibited full or partial classifier collapse, and architecture choice alone did not provide a consistent solution after statistical correction. The thesis therefore does not assume that a stronger backbone automatically yields more dependable movement understanding [45, 70].

Evaluation of movement diagnosis should consider not only aggregate accuracy or F1, but also class-specific performance, the presence or absence of relevant evidence, calibration, and the reasons for changes in predictions [12]. For this reason, the FAIR-Coach

evaluation strategy combines standard predictive metrics with evidence ablation and intervention to address these factors.

Predictive performance and confidence calibration are also distinct properties. Modern neural networks can achieve strong classification performance while producing probabilities that do not accurately reflect empirical correctness, motivating explicit calibration evaluation [71, 72]. This distinction is relevant to movement diagnosis because confidence scores may influence how strongly a diagnosis or feedback decision is interpreted.

2.3 Biomechanical Evidence Representation

Biomechanical movement assessment can draw on multiple kinematic observations rather than a single summary value, including joint angles, range of motion, symmetry, segment orientation, and phase-specific movement behaviour [31, 73]. When these measurements are used in machine-learning systems, an important representational choice is which aspects of the movement are preserved explicitly and which are aggregated into broader summaries.

Squat biomechanics provide a direct basis for selecting task-specific evidence. Squat depth and trunk position influence lower-extremity kinematics and the distribution of hip and knee loading [31, 74]. Stance width likewise affects lower-limb kinematics and three-dimensional hip moment demands [32]. These findings motivate the use of phase-specific joint angles, trunk orientation, lower-limb alignment, and stance-related measures rather than relying only on a global movement-deviation score.

Recent rehabilitation research has already used exercise-specific, clinically motivated skeletal features for movement assessment. Tang et al. extract features such as knee and hip flexion, trunk inclination, and foot symmetry, use them with prompt-engineered large language models for binary assessment of rehabilitation exercise quality and feedback generation, and report better assessment with these exercise-specific features than with raw joint inputs on UI-PRMD [40]. Liao et al. further review computational approaches for rehabilitation-exercise evaluation, including clinician-defined feature engineering and kinematic measurements for movement assessment [41]. Therefore, FAIR-Coach does not claim novelty in the individual biomechanical measurements themselves. Its contribution is the organisation and controlled evaluation of a full diagnosis-aligned evidence configuration for multi-label movement diagnosis, separately from reasoning architecture and language generation. Unlike these works, FAIR-Coach addresses eight-category multi-label diagnosis from monocular 2D in-the-wild video, trains evidence-to-diagnosis models, and compares evidence configurations under matched model controls, multiple training seeds, and subject-level analyses.

Reference-based assessment compares an observed movement with an explicit target or exemplar [2, 3]. Fine-grained AQA has used query–exemplar correspondence to compare semantically aligned action stages rather than only whole-video representations [59, 75]. In FAIR-Coach, the corresponding design question is how the reference itself should be represented: for example, as a mean trajectory or a distribution around that mean, with deviation measured globally, by joint, or at diagnosis-relevant movement phases.

As reviewed in Section 2.2, recent AQA and human-motion research suggests that the

structure of a representation, and the information it makes available to the downstream task, can materially affect both predictive performance and the granularity of assessment [8, 9, 58].

FAIR-Coach adopts this evidence-centric perspective by distinguishing *generic evidence* from *diagnosis-aligned evidence*. Generic evidence summarises overall movement behaviour, for example through whole-repetition classifier probabilities or global deviation scores. Such features can indicate that a repetition differs from a reference but may not identify why. Diagnosis-aligned evidence instead asks which measurable biomechanical signature should support a specific diagnosis and when in the movement that signature is most informative.

For squat assessment, this distinction has a biomechanical basis. Stance, trunk position, squat depth, and bilateral symmetry correspond to distinct aspects of squat kinematics [31, 32, 73]. Accordingly, diagnosis-aligned evidence can preserve these distinctions rather than collapsing them into a single global deviation measure.

Reference modelling forms one part of this question. A distribution-based reference may appear more realistic because acceptable movement varies across individuals. However, a richer reference does not necessarily provide better diagnosis if correct and incorrect repetitions overlap under the same global deviation measure. FAIR-Coach therefore evaluates reference formulation separately from diagnosis-aligned feature design. This distinction is important because it prevents a failure of one global reference statistic from being mistaken for evidence that biomechanics itself is uninformative.

This approach may also be applicable to other exercises and rehabilitation movements, although the relevant features would differ. In each case, task-specific evidence can be structured using clinically or biomechanically meaningful phases and variables. However, this thesis does not address empirical validation of this broader generalisation.

2.4 Evidence Fusion, Reasoning, and Grounded Feedback

Once heterogeneous evidence has been defined, it must be combined into a diagnosis. The simplest strategy concatenates evidence features and trains a nonlinear classifier. More structured alternatives use graphs, message passing, or attention to model explicit relationships among evidence groups [27]. Skeleton-recognition research provides strong examples of the value of relational modelling: CTR-GCN learns adaptive joint topology [26], while SkateFormer models multiple skeletal-temporal relation types [28]. Multimodal methods likewise use cross-source interaction to integrate pose, vision, and language [7, 43, 68]. AQA-specific work has also modelled explicit joint relations and relative comparisons between performances [76, 77].

These successes make relational reasoning a plausible hypothesis for movement diagnosis, but they do not establish that it is necessary. In action recognition, joint relationships are part of the raw signal required to recognise the action. In FAIR-Coach, the diagnosis model inputs are already derived evidence such as classifier probabilities, joint deviations,

or critical-frame measurements. A simple nonlinear fusion model may therefore capture much of the useful interaction without an explicit graph [78, 79]. For this reason, the thesis evaluates relational models against corresponding concatenation controls using the same evidence rather than assuming that additional structural complexity is beneficial.

A second consideration is whether a prediction relies on the intended evidence. Standard accuracy metrics indicate whether a model predicts the correct label, but do not address the basis for that prediction. Intervention-based interpretability methods test this aspect more directly [22, 80]. Steinmann et al. investigate learned interventions on concept bottlenecks [21], while Laguna et al. extend intervenability to pretrained black-box models [33]. Concept Bottleneck Models provide an earlier related formulation in which predicted concepts can be directly modified to observe their effect on the final prediction [81]. The underlying idea applies to movement diagnosis: if a biomechanical variable matters for a diagnosis, changing it in the expected direction should alter the diagnosis, whereas changes to non-primary evidence may have comparatively less impact.

FAIR-Coach therefore distinguishes three complementary evaluation ideas. *Ablation* asks whether removing an evidence source degrades performance. *Relevant-evidence intervention* asks whether the diagnosis responds appropriately when evidence supporting that diagnosis is strengthened or weakened. *Non-primary-evidence control intervention* asks whether the model remains comparatively stable when non-primary evidence is perturbed. Together, these tests provide a behavioural protocol for assessing grounding; predictive accuracy alone does not establish that a diagnosis is grounded in the intended evidence.

The final stage is language feedback. QEVD identifies the difficulty of delivering timely and appropriate coaching messages [15], while ExpertAF demonstrates actionable feedback generation from multimodal movement evidence [1]. Generative models can nevertheless produce claims that are unsupported or inconsistent with their grounding information [23, 24, 34, 35]. Grounded generation therefore requires evaluating consistency between generated content and the evidence that is intended to support it [82, 83]. For this reason, FAIR-Coach fixes the structured diagnosis before generating language feedback.

Preference optimisation is applied only at the downstream stage. Direct Preference Optimization (DPO) avoids explicit reward-model training by optimising preferred responses directly against rejected responses [36]. SimPO provides a subsequent reference-free preference-optimisation formulation that removes the need for a separate reference model [37, 84]. In FAIR-Coach, preference optimisation is used only to improve diagnosis-to-language faithfulness [39]: the preference pairs penalise unsupported or diagnosis-inconsistent feedback while the underlying diagnosis remains fixed.

This separation defines two types of grounding. Diagnosis grounding concerns whether the movement model relies on biomechanical evidence as intended. Language grounding refers to whether the generated feedback accurately reflects the diagnosis and the evidence used. The thesis examines these two aspects separately.

2.5 Research Gap and Positioning of FAIR-Coach

Recent literature demonstrates advances in movement representations, AQA models, multimodal coaching systems, and generative feedback mechanisms. Taken together, these studies motivate four gaps synthesised in this thesis, which form the basis for its research questions.

First, **evidence representation and reasoning complexity can be difficult to separate**. Recent methods often introduce richer representations together with new modelling mechanisms, including fine-grained parsing, cross-modal alignment, or rubric-informed reasoning [8, 11, 12, 58]. This makes it difficult to determine whether a gain arises from better information, a more sophisticated model, or both. FAIR-Coach compares the full Generic and Diagnosis-Aligned evidence configurations while separately comparing relational and non-relational models. These comparisons evaluate the configurations as a whole; they do not isolate the contribution of individual added features.

Second, **generic movement-quality evidence may be insufficient for multi-category diagnosis**. A substantial part of the AQA literature focuses on overall quality scoring or related aggregate assessment objectives [57, 68, 85]. Coaching diagnosis poses a different requirement because distinct error categories may have different spatial and phase-specific signatures. Recent fine-grained and rubric-informed AQA supports more structured assessment [8, 11–13]and exercise-specific features have been compared with raw joint inputs [40], but controlled comparisons between generic movement evidence and

diagnosis-aligned biomechanics for multi-label diagnosis, under matched model controls, remain limited.

Third, **predictive accuracy does not establish biomechanical grounding**. Standard metrics do not determine whether a model's prediction responds appropriately to the evidence that should support it. FAIR-Coach therefore complements accuracy and F1 with intervention-based tests inspired by recent work on model intervenability [21, 22, 33, 80].

Fourth, **diagnosis grounding and language grounding should be treated separately**. Recent coaching systems increasingly incorporate free-form feedback [1, 3, 15, 45, 47], while generated responses can contain claims that are unsupported by their grounding information [23, 24, 34]. FAIR-Coach therefore fixes the structured diagnosis before language generation and evaluates whether preference optimisation improves the faithfulness of its verbalisation without altering that diagnosis.

These gaps define the thesis's position. Earlier coaching work established how an intelligent system could perceive movement, maintain context, coordinate models, and communicate with users [16–18], following the system progression summarised in Figures 2.1–2.3. The domain-shift benchmark then showed that sophisticated skeleton classifiers can still fail under changing data conditions [20]. Together, these gaps position FAIR-Coach around the controlled comparison of evidence design and reasoning architecture, complemented by separate evaluations of diagnosis grounding and language grounding.

Chapter 3 operationalises this question through the FAIR-Coach evidence representation, controlled comparisons of diagnosis models, grounding interventions, and diagnosis-

to-language adaptation.

3

FAIR-Coach Methodology

This chapter presents the methodology used to address the thesis's main scientific questions. FAIR-Coach separates movement observation, structured biomechanical evidence, diagnosis, and language feedback so that evidence representation, reasoning architecture, diagnosis grounding, and downstream language grounding can be evaluated independently. Explicit relational reasoning is treated as a hypothesis and compared directly with corresponding non-relational fusion rather than assumed to improve performance [78, 79, 86]. Detailed implementation results, ablations, and statistical comparisons are reported in Chapter 4.

3.1 Problem Formulation and System Overview

3.1.1 From Movement Observation to Evidence-Grounded Diagnosis

Building on the coaching-system lineage reviewed in Chapter 2, FAIR-Coach shifts the methodological focus from interaction and orchestration [16–18] to whether movement di-

agnosis is supported by explicit biomechanical evidence. The domain-shift findings further motivate this shift [20]; accordingly, the method does not treat architectural complexity, model confidence, or fluent feedback as sufficient evidence of reliable movement understanding [23, 71].

A validated movement repetition is denoted by x_r, where r is the repetition index. FAIR-Coach maps each observed movement to a structured evidence record,

$$E_r = h(x_r), \tag{3.1}$$

where E_r is the structured evidence for repetition r, and $h(\cdot)$ represents pose extraction, temporal processing, repetition-level analysis, biomechanical feature construction, and evidence packaging. A diagnosis model f_θ then predicts an eight-dimensional multi-label probability vector,

$$\mathbf{p}_r = f_\theta(E_r), \qquad \mathbf{p}_r \in [0,1]^8. \tag{3.2}$$

where $\mathbf{p}_r$ contains the predicted probabilities for the eight diagnosis categories. The corresponding target is the multi-hot vector

$$\mathbf{y}_r \in \{0,1\}^8, \tag{3.3}$$

where each element indicates the presence or absence of one diagnosis label. A multi-label target is required because more than one movement issue may be annotated for the same repetition in QEVD [15].

The prediction is then materialised as a versioned `StructuredDiagnosis`,

$$D_r = \mathcal{S}(\mathbf{p}_r), \tag{3.4}$$

where D_r is the structured diagnosis for repetition r and $\mathcal{S}(\cdot)$ denotes the deterministic mapping from model probabilities to the stored diagnosis representation. The diagnosis is fixed before recommendations or free-form language are generated. The feedback model then receives the diagnosis together with only the evidence and context permitted for verbalisation:

$$t_r = g_\phi \left(D_r, E_r^{\text{licensed}}, c_r \right), \tag{3.5}$$

where g_ϕ is the feedback model with parameters ϕ, E_r^{licensed} is the subset of evidence permitted for communication, c_r denotes available session context, and t_r is the final coaching message. The feedback stage may alter wording, tone, or strategy, but it may not redefine D_r.

This formulation separates two grounding problems. Diagnosis grounding asks whether f_θ responds to biomechanical evidence as intended [21, 87], whereas language grounding asks whether g_ϕ communicates the fixed diagnosis without introducing unsupported information [23, 24]. FAIR-Coach evaluates these two properties independently.

3.1.2 FAIR-Coach Architecture

Figure 3.1 summarises the FAIR-Coach methodology. Video is converted into a smoothed skeletal representation, and movement evidence is collected at two temporal scales. Path A provides continuous window-level evidence in the live system, whereas Path B summarises completed repetitions using whole-repetition assessment, biomechanical analysis, and reference comparison. Because QEVD short clips do not provide the continuous session context required for Path A, the core evidence-to-diagnosis experiments use Path B together with reference-based and diagnosis-aligned biomechanical evidence.

These observations are organised into `RepEvidence`, which provides the structured per-repetition input to the diagnosis stage. Non-relational fusion and explicit relational reasoning are evaluated over the same evidence as alternative diagnosis mechanisms. Their output is stored as a `StructuredDiagnosis`, after which recommendation and language generation operate without redefining the diagnosis. LangGraph, A2A, MCP, logging, session state, and related interface components provide supporting execution and interoperability infrastructure rather than altering the scientific evidence-to-diagnosis comparison [18, 49].

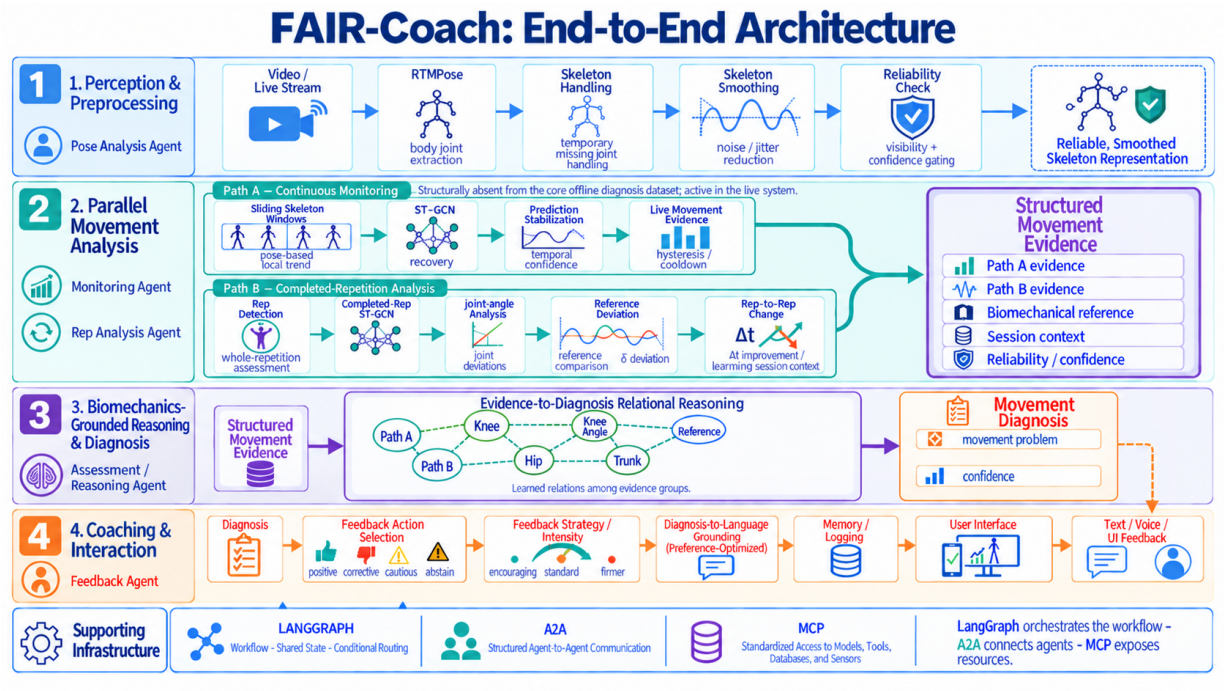


Figure 3.1: Overview of the FAIR-Coach architecture. Continuous and repetition-level movement evidence are organised into a structured biomechanical representation for diagnosis. Diagnosis is separated from downstream feedback generation, while orchestration and tool-access components provide supporting execution infrastructure.

The primary scientific focus is therefore on evidence representation, diagnosis modelling, grounding evaluation, and diagnosis-to-language faithfulness. Supporting components are necessary for end-to-end operation but are not treated as evidence of improved diagnostic grounding. FAIR-Coach also does not evaluate autonomous multi-agent decision-making [50, 51]: repetition-level execution is largely structured, with the principal evidence-driven conditional behaviour arising from reliability-gated feedback.

3.2 Movement Perception and Dual-Path Evidence Acquisition

3.2.1 Skeleton Representation and Reliability Gating

FAIR-Coach begins with 2D pose estimation using RTMPose in MMPose [88]. The pose estimator is treated as an upstream measurement component rather than a thesis contribution. It provides a consistent COCO-style skeleton [89] from which joint trajectories, repetition-level features, and reliability signals are derived.

Each frame provides keypoint coordinates and confidence values. Short gaps in keypoint availability are handled before temporal smoothing so isolated detector failures do not create discontinuous trajectories. Smoothing is applied before movement segmentation and biomechanical feature extraction to reduce short-term detector noise without inferring biomechanics across long missing intervals [63, 65].

A skeleton reliability gate determines whether the observed movement evidence is sufficiently trustworthy for coaching feedback. Here, reliability refers to input-evidence quality rather than diagnosis confidence: a model may be confident despite weak evidence, and a reliable skeleton does not imply a correct diagnosis [71, 72].

Reliability serves two methodological roles. Reliability indicators are stored in `RepEvidence` so that weak or unavailable evidence remains distinguishable from a valid numerical zero, and the feedback policy uses reliability as a communication gate. Low-

quality evidence leads to cautious or non-specific feedback, whereas sufficiently reliable evidence permits diagnosis-specific feedback [90]. Reliability does not alter the predicted movement category. The final reliability formula and its engineering validation are reported in Chapter 4; the methodological requirement here is to preserve and propagate uncertainty in the observed evidence.

3.2.2 Continuous and Repetition-Level Evidence

FAIR-Coach retains two complementary temporal views of movement. Path A captures continuous local evidence from the live stream, whereas Path B summarises a completed repetition. The two paths are retained because local movement behaviour and whole-repetition assessment represent different temporal scales.

Path A: continuous local evidence. Path A processes overlapping skeleton windows and aggregates the frozen window-level classifier outputs over time rather than treating any single window as a diagnosis. The resulting summaries provide evidence about localised movement irregularities and stable local events during live execution.

QEVD short clips do not provide the continuous session context required to reconstruct this sliding-window process [15]. Path-A fields are therefore structurally absent from QEVD-derived core diagnosis records and are marked as missing rather than as measured zeros [91]. Consequently, Path-A-only and learned Path-A/Path-B fusion are excluded from the controlled core diagnosis experiments. Path A remains part of the live FAIR-

Coach architecture, so the term "dual-path" refers to the system architecture rather than to the evidence evaluated statistically in the core diagnosis study.

Path B: completed-repetition evidence. Path B operates after a repetition has been detected or validated. In the live system, repetition boundaries are estimated from the temporal evolution of lower-body joint angles; in the offline QEVD pipeline, a short-clip validator first determines whether the clip contains a usable repetition before the frozen evidence extractor is applied. A clip is accepted only if it contains at least 10 reliable frames, reliable frames form at least half of the clip, the minimum knee angle reaches 115° or below, and the knee-angle range of motion is at least 40°. A frame is reliable when its knee angle is defined and its minimum keypoint confidence is at least 0.3.

For each accepted repetition, Path B provides whole-repetition classification outputs together with probabilities, confidence, and trajectory information used to derive range-of-motion and reference-deviation features. The Path-B classifier is a frozen ST-GCN [25]. Its training record was not preserved; reconstruction from the stored training arrays (614 sequences) indicates that it was trained on a pooled set of EC3D [92], REHAB24 [52], and a custom collection of 282 YouTube-derived squat sequences. None of these sources is part of QEVD, so no QEVD participant overlap is expected, although it was not separately checked for the YouTube-derived clips. Its outputs on QEVD are cross-domain predictions, and its own held-out accuracy was not recorded. Local and whole-repetition evidence can therefore expose different failure patterns: brief irregularities may not characterise an entire repetition, while a single repetition-level summary may obscure when a deviation occurred

[8, 58]. The live system retains both views, whereas the core diagnosis experiments remain restricted to the evidence available in the QEVD short-clip setting.

3.3 Structured Biomechanical Evidence Representation

3.3.1 RepEvidence: Unified Per-Repetition Representation

`RepEvidence` provides the structured interface between perception and diagnosis. Each validated repetition is stored as a versioned evidence record that can be inspected, normalised, serialised, and used consistently across live and offline diagnosis models [93].

The generic representation is grouped as

$$E_r = \left[E_r^A,\, E_r^B,\, E_r^{\text{bio}},\, E_r^{\text{rel}},\, E_r^{\text{ctx}},\, M_r\right], \tag{3.6}$$

where E_r^A contains Path-A evidence, E_r^B contains Path-B evidence, E_r^{bio} contains generic biomechanical/reference information, E_r^{rel} stores reliability signals, E_r^{ctx} stores available session context, and M_r explicitly represents missingness.

The frozen generic model-input schema contains 27 raw evidence dimensions before the missingness representation is applied. Its principal blocks are summarised in Table 3.1.

Table 3.1: Principal blocks of the generic `RepEvidence` representation used by the diagnosis experiments.

Block	**Representative contents**	**QEVD short-clip status**
Path A	stable/local window evidence	structurally missing
Path B	correct/incorrect probabilities, confidence, repetition assessment	available
Biomechanical	ROM, overall deviation, nine joint/segment deviations	available
Reliability	low-confidence and evidence-quality indicators	partially available
Context	repetition index and previous-repetition change when defined	limited / partly missing

Missingness is represented explicitly rather than encoded as numerical zero [67, 91, 94]. This distinction is necessary because Path A is unavailable in QEVD short clips, previous-repetition context may be absent for isolated clips, and some live reliability signals cannot be reconstructed offline. A zero in these positions would incorrectly imply that the quantity had been observed and measured as zero.

The serialised record also retains provenance information, including schema, source, extraction, reference, and model-version metadata. These fields are not necessarily diagnosis-model inputs, but they allow the evidence definition associated with each diagnosis to be reconstructed and reduce the risk of silent feature drift across experiments [93].

Continuous input features are standardised using statistics estimated only from the training set:

$$\tilde{x}_{r,k} = \frac{x_{r,k} - \mu_k^{\mathrm{TRAIN}}}{\sigma_k^{\mathrm{TRAIN}} + \epsilon}. \tag{3.7}$$

Here, $x_{r,k}$ is feature k for repetition r, $\tilde{x}_{r,k}$ is its standardised value, μ_k^{TRAIN} and σ_k^{TRAIN} are the TRAIN mean and standard deviation for feature k, and ϵ is a small numerical-stability constant. The TRAIN statistics are fixed and reused without modification for validation, TEST, and final inference. Reference construction follows the same TRAIN-only principle to avoid indirect data leakage.

3.3.2 Reference-Based and Diagnosis-Aligned Evidence

The methodology separates two questions: how acceptable movement is summarised through a biomechanical reference, and which biomechanical variables are exposed to the multi-category diagnosis model.

Reference-based evidence. The original reference pipeline uses nine joint and segment angle trajectories from accepted TRAIN repetitions whose QEVD annotations contain both `squats` and `no_obvious_issue` [3, 59, 75]. This produces a dataset-specific reference cohort of $N_{\text{ref}} = 529$ basic-squat repetitions, each temporally normalised to $T = 100$ samples. The reference is therefore a descriptive baseline for this dataset, rather than a universal or clinical definition of an ideal squat.

The original selection rule did not require `no_obvious_issue` to be the only label. Consequently, 333 of the 529 reference repetitions (62.95%) also carry at least one movement-error label. This contamination is a limitation of the original reference construction. A sensitivity analysis therefore rebuilds the reference from the 196 TRAIN repetitions carry-

ing only `no_obvious_issue`; the corresponding results are reported in Chapter 4.

For trajectory j, the mean reference is

$$\mu_j(t) = \frac{1}{N_{\text{ref}}} \sum_{i=1}^{N_{\text{ref}}} \theta_{i,j}(t), \qquad N_{\text{ref}} = 529, \tag{3.8}$$

where t is the normalised time index, i indexes reference repetitions, j indexes the nine joint or segment trajectories, and $\theta_{i,j}(t)$ is the angle value for reference repetition i.

Per-time-point variability is also computed as

$$s_j(t) = \sqrt{\frac{1}{N_{\text{ref}} - 1} \sum_{i=1}^{N_{\text{ref}}} (\theta_{i,j}(t) - \mu_j(t))^2}, \tag{3.9}$$

where $s_j(t)$ is the standard deviation around the dataset reference at time t. This variability is retained for reference analysis, but the final generic deviation feature uses the mean trajectory only. Across the nine trajectories and 100 time points, $s_j(t)$ has a mean of 26.8° (range 6.2°–46.1°), indicating substantial spread around the mean reference. Because the reference pools participants with different body proportions and movement strategies, a population-mean trajectory may blur individual variation; a morphology-conditioned reference is left for future work.

A repetition-specific deviation is computed as

$$d_{r,j} = \frac{1}{T} \sum_{t=1}^{T} \left| \theta_{r,j}(t) - \mu_j(t) \right|, \tag{3.10}$$

where $\theta_{r,j}(t)$ is trajectory j for repetition r, and $d_{r,j}$ is its mean absolute deviation from the reference trajectory. The generic whole-repetition deviation is

$$\delta_r = \frac{1}{9} \sum_{j=1}^{9} d_{r,j}. \tag{3.11}$$

A distribution-aware reference formulation was also evaluated separately. Reference deviation remains generic evidence: it indicates similarity to this dataset-specific reference but does not by itself determine the appropriate diagnosis.

Diagnosis-aligned evidence. The diagnosis-aligned representation instead exposes evidence based on the biomechanical structure of the diagnosis taxonomy. This design is consistent with fine-grained and rubric-informed AQA findings showing that diagnostically relevant information may be localised to particular body regions and movement stages [8, 11, 12, 58].

Ten additional features are extracted from the available 2D skeleton:

- left and right knee angles at the deepest squat frame;
- left and right hip angles at the deepest squat frame [31];
- trunk inclination at the deepest frame [31, 74];
- variance of trunk inclination across the repetition, computed as an engineered summary statistic of the trunk trajectory;

- left and right knee–ankle horizontal offsets at the deepest frame [31, 74];
- stance-width-to-hip-width ratio [32]; and
- a left–right symmetry difference [73].

The deepest frame is a reproducible heuristic for depth- and posture-related measurements rather than a ground-truth movement-phase annotation. Because the COCO-17 keypoint representation does not include a toe keypoint [89], the ankle is used as a proxy for the knee–foot horizontal relationship. These choices preserve interpretability without implying measurements or annotations unavailable in the source data.

The two evidence regimes used in the controlled comparisons are therefore

$$E_r^{\text{generic}} = \texttt{RepEvidence}_r^{\text{generic}}, \tag{3.12}$$

$$E_r^{\text{aligned}} = \left[E_r^{\text{generic}}, E_r^{\text{diag}}\right], \tag{3.13}$$

where E_r^{generic} is the generic evidence representation for repetition r, E_r^{diag} contains the ten diagnosis-aligned features, and E_r^{aligned} is their combined representation. This design enables relational and non-relational architectures to be compared using the same evidence within each condition. However, the Generic and Diagnosis-Aligned conditions differ in both evidence content and input dimensionality: the aligned representation adds ten measurements to the 27-feature generic representation. Consequently, the RQ1 comparison evaluates the full Diagnosis-Aligned configuration rather than isolating the causal contribution of alignment or of any individual added feature. Feature-ranking, ten-feature-only,

and matched-dimensional control experiments are not part of the present study.

3.4 Evidence-to-Diagnosis Modelling

3.4.1 Structured Diagnosis Formulation

The diagnosis task is defined over eight QEVD-derived categories:

> `no_obvious_issue`, `shallow`, `knees_over_toes`, `back_not_straight`, `narrow`, `wide`, `over_90_degrees`, and `insufficient`.

Here, `insufficient` denotes the QEVD category described as "User is doing the exercise insufficiently"; the source dataset provides no more specific biomechanical criterion for this label, and it is distinct from the FAIR-Coach reliability state with the same name.

Because QEVD contains co-occurring annotations, diagnosis targets are represented as multi-hot vectors rather than converted to single-label classification [15]. Contradictory annotations are retained as given in the main experiments: 371 TRAIN, 100 validation, and 97 TEST clips are labelled both `no_obvious_issue` and at least one error. Sensitivity analyses that remove these clips are reported in Chapter 4. For each repetition r and label ℓ, the diagnosis model produces a logit $z_{r,\ell}$ and probability

$$p_{r,\ell} = \sigma(z_{r,\ell}). \tag{3.14}$$

Here, $z_{r,\ell}$ is the model logit, $\sigma(\cdot)$ is the sigmoid function, and $p_{r,\ell}$ is the predicted probability for label ℓ, that is, the ℓ-th element of $\mathbf{p}_r$.

Learned neural diagnosis models use weighted binary cross-entropy with logits to address label imbalance. For each label ℓ, the positive-class weight is calculated only from the accepted TRAIN repetitions:

$$w_\ell^+ = \frac{n_\ell^-}{n_\ell^+}, \tag{3.15}$$

where n_ℓ^+ and n_ℓ^- denote the numbers of positive and negative TRAIN examples for label ℓ, respectively. The implemented loss uses the `pos_weight` argument of `BCEWithLogitsLoss` with default mean reduction:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{NK} \sum_{i=1}^{N} \sum_{\ell=1}^{K} \left[w_\ell^+ y_{i,\ell} \log \sigma(z_{i,\ell}) + (1 - y_{i,\ell}) \log \left(1 - \sigma(z_{i,\ell})\right) \right], \tag{3.16}$$

where N is the number of TRAIN repetitions in a batch, $K = 8$ is the number of diagnosis labels, $z_{i,\ell}$ is the logit for repetition i and label ℓ, $\sigma(\cdot)$ is the sigmoid function, and $y_{i,\ell} \in \{0, 1\}$ is the corresponding ground-truth indicator.

For the original accepted TRAIN split ($N_{\mathrm{TRAIN}} = 4{,}175$), the per-label positive counts and resulting positive-class weights were as follows:

Table 3.2: TRAIN-derived positive-class weights used by the learned neural diagnosis models.

Label	n_ℓ^+	n_ℓ^-	w_ℓ^+
`no_obvious_issue`	1,257	2,918	2.321
`shallow`	279	3,896	13.964
`knees_over_toes`	589	3,586	6.088
`back_not_straight`	192	3,983	20.745
`narrow`	273	3,902	14.293
`wide`	379	3,796	10.016
`over_90_degrees`	334	3,841	11.500
`insufficient`	64	4,111	64.234

The weights are computed from TRAIN data only and are not estimated from validation or TEST labels. The same weighting principle is used for the learned neural diagnosis models; the logistic-regression baselines use their separate balanced sample-weighting procedure.

Predicted labels and probabilities are stored in a `StructuredDiagnosis` together with the model and version metadata required for reproducibility. Supporting evidence may be attached separately for evaluation or language generation, but the stored diagnosis is not modified after this stage.

FAIR-Coach does not introduce responsible-joint or exact movement-phase targets as supervised outputs because validated ground truth for these targets is unavailable in the core dataset. The methodology therefore avoids creating richer pseudo-labels that cannot be independently validated.

3.4.2 Non-Relational Fusion and Relational Reasoning

The model comparisons separate representation quality from reasoning complexity. Direct comparisons use identical subject splits and, when the relational mechanism is tested, the same underlying evidence inputs.

Non-relational fusion. The principal non-relational model concatenates the available evidence into one feature vector and applies a multilayer perceptron:

$$\mathbf{h}_r = \phi\left(W_1\tilde{\mathbf{x}}_r + b_1\right), \qquad \mathbf{z}_r = W_2\mathbf{h}_r + b_2, \tag{3.17}$$

where $\tilde{\mathbf{x}}_r$ is the normalised evidence-and-mask representation for repetition r, W_1 and W_2 are learned weight matrices, b_1 and b_2 are bias vectors, $\phi(\cdot)$ is the hidden activation function, $\mathbf{h}_r$ is the hidden representation, and $\mathbf{z}_r$ contains the output logits. The same formulation is evaluated with generic and diagnosis-aligned evidence. Simpler rule and linear models are retained for context, while Concat+MLP serves as the principal non-relational control [78, 79].

Relational reasoning. Relational reasoning is evaluated through several architectures rather than a single graph formulation [25–27]. Relations are defined among evidence groups; the diagnosis models do not process temporal sequences directly because temporal information has already been summarised into per-repetition evidence.

The first formulation uses a coarse five-group decomposition:

$$\{\text{Path A}, \text{Path B}, \text{biomechanical}, \text{reliability}, \text{context}\}.$$

Each group is encoded separately before message passing between group representations.

The second formulation uses a finer eight-group decomposition:

$$\{\text{Path A}, \text{Path B}, \text{reference}, \text{knee}, \text{hip}, \text{trunk}, \text{reliability}, \text{context}\}.$$

This representation exposes more specific cross-evidence relations, including knee–hip, knee–reference, and hip–trunk interactions. Relations-ON and relations-OFF use the same group encoders and differ only in whether inter-group message passing is enabled. In the QEVD short-clip setting, Path A is structurally unavailable, the repetition-count field is fixed at one, and several reliability and context fields have limited observed variation. Consequently, only a subset of the nominal evidence groups carries substantial movement-specific information in the evaluated records. The relational experiments therefore assess the chosen architectures under this sparse-node condition; a non-significant relational advantage should not be interpreted as evidence that relational reasoning would be unhelpful when richer, dynamic, and more heterogeneous evidence is available.

The generic update is

$$\mathbf{h}_g^{(l+1)} = \mathcal{U}\left(\mathbf{h}_g^{(l)}, \sum_{j \in \mathcal{N}(g)} \Psi\left(\mathbf{h}_g^{(l)}, \mathbf{h}_j^{(l)}\right)\right), \tag{3.18}$$

where g denotes an evidence group, l is the message-passing layer, $\mathbf{h}_g^{(l)}$ is the representation of group g at layer l, $\mathcal{N}(g)$ is the set of neighbouring evidence groups, j indexes those neighbours, $\Psi(\cdot)$ computes pairwise messages, and $\mathcal{U}(\cdot)$ updates the group representation.

The third relational family uses label-conditioned attention. Here, $\mathbf{h}_g$ denotes an encoded evidence group and $\mathbf{q}_\ell$ is a learned query for diagnosis label ℓ. The label-specific representation is

$$\mathbf{c}_\ell = \sum_g \alpha_{\ell g} \mathbf{h}_g, \qquad \alpha_{\ell g} = \mathrm{softmax}_g \left(\frac{\mathbf{q}_\ell^\top \mathbf{k}_g}{\sqrt{d}} \right), \tag{3.19}$$

where $\mathbf{k}_g$ is the key representation for evidence group g, $\alpha_{\ell g}$ is the attention weight assigned by label ℓ to group g, d is the key/query dimensionality, and $\mathbf{c}_\ell$ is the resulting label-specific evidence representation. This allows different diagnoses to weight evidence groups differently.

Table 3.3 summarises the controlled architecture families.

Table 3.3: Diagnosis-model families evaluated in FAIR-Coach.

Family	**Evidence organisation**	**Purpose**
Concat+MLP	one concatenated vector	non-relational learned fusion baseline
Coarse message passing	5 evidence groups	test relations at broad evidence-source level
Fine message passing	8 evidence groups	test relations among diagnosis-relevant evidence groups
Label-conditioned attention	label queries over encoded groups	test diagnosis-specific relational aggregation

These architectures provide controlled tests of whether explicit relational structure adds

value when the underlying evidence is held fixed; relational benefit is evaluated rather than assumed [86].

3.4.3 Model Comparison and Selection Protocol

The main diagnosis experiments use the QEVD short-clip corpus with a fixed subject-disjoint split: 21 subjects for TRAIN, four for validation, and five for TEST. After short-clip validation, the corresponding datasets contain 4,175 TRAIN, 1,083 validation, and 1,052 TEST repetitions. Subjects do not overlap across splits. Reference examples and normalisation statistics are derived exclusively from TRAIN and then applied unchanged to validation and TEST [93].

Micro-F1 is the primary comparative metric for multi-label diagnosis, with macro-F1 and class-specific results used to expose behaviour on less frequent labels. Calibration and robustness are evaluated separately [71, 72]. Checkpoint selection and freezing use TRAIN and validation data only, and each checkpoint is locked before its TEST evaluation. However, the diagnosis-aligned Concat+MLP was designated as the final live diagnosis backend after TEST results for the frozen diagnosis-aligned models had been observed, although the diagnosis-aligned Reasoner v2 had the higher validation micro-F1. The TEST cohort therefore also informed final system selection.

For a fixed model comparison, paired bootstrap resampling draws the same clip-level prediction pairs from the locked TEST set with replacement ($B = 1{,}000$, seed 42). This procedure evaluates the uncertainty of the performance difference conditional on the fixed

TEST cohort; it does not create new TRAIN, validation, or TEST partitions. Because the 1,052 TEST repetitions come from only five independent participants, clip-level confidence intervals alone do not establish uncertainty for unseen people.

The robustness analysis therefore additionally reports per-subject results, a two-level bootstrap that resamples subjects and then clips within subjects (B = 10,000, seed 42), and six-fold subject-disjoint cross-validation. In the cross-validation analysis, each fold holds out five subjects, rebuilds the reference from that fold's TRAIN subjects using the original reference-selection rule, and retrains the matched Generic and Diagnosis-Aligned Concat+MLP models using seeds 42–46. Repeated training seeds assess sensitivity to random initialisation; they do not constitute independent TEST cohorts or external replication.

A relational architecture is credited with a measurable benefit only when its paired improvement over the corresponding non-relational baseline has a 95% confidence interval excluding zero in the beneficial direction on the locked TEST split. Sequential analyses on the same TEST cohort are interpreted as supporting analyses within QEVD, rather than as independent confirmations. Complete numerical results, confidence intervals, and per-architecture comparisons are reported in Chapter 4. The relational-reasoning conclusions are limited to the FAIR-Coach evidence representation, diagnosis taxonomy, and data regime evaluated in this thesis [86].

3.5 Diagnosis Grounding and Grounded Feedback

3.5.1 Evidence-Grounding and Intervention Framework

Predictive performance and biomechanical grounding are evaluated as distinct properties. A model may predict the correct label while depending on evidence in a way that is inconsistent with the intended biomechanical interpretation [87, 95]. FAIR-Coach therefore complements predictive metrics with evidence ablation and controlled interventions.

Evidence ablation. An evidence group is removed or masked while the remaining input is held unchanged. The resulting performance change indicates whether that group is important to the learned representation. Because evidence groups may be correlated, however, ablation does not establish that a single feature causally determines a diagnosis [22, 95].

Relevant-evidence intervention. For diagnosis label ℓ, E_ℓ^{rel} denotes the evidence defined as relevant to that label. Here, this notation refers to label-relevant evidence and is distinct from the reliability evidence block defined earlier. The intervention applies a controlled perturbation [87]:

$$E'_r = \mathrm{do}\left(E_\ell^{\mathrm{rel}} \leftarrow \widetilde{E}_\ell^{\mathrm{rel}}\right),$$

where E'_r is the intervened evidence record for repetition r and $\widetilde{E}^{\text{rel}}_\ell$ is the perturbed version of the label-relevant evidence. The corresponding probability change is

$$\Delta p_\ell = f_\theta(E'_r)_\ell - f_\theta(E_r)_\ell. \tag{3.20}$$

Here, Δp_ℓ is the change in predicted probability for label ℓ, and $f_\theta(\cdot)_\ell$ denotes the diagnosis model output for that label. The original intervention protocol, including the evidence groups, the 0.5 magnitude, and the decision thresholds, was frozen before the first TEST grounding evaluation and applied +0.5 to every relevant feature. After that evaluation, the sign conventions were revised for a corrected re-evaluation: ROM is decreased rather than increased for `shallow`, the Path-B control perturbation is made directionally coherent, and the global reference deviation is recomputed whenever joint-deviation features are perturbed. These corrected specifications were fixed in code immediately before the corrected run and are therefore not preregistered. Table 3.4 records the corrected feature-level directions and their intended interpretations; these directions are design assumptions and have not been validated by biomechanics experts. All changes are applied in standardised feature space.

Table 3.4: Corrected intervention directions used in the grounding re-evaluation.

Diagnosis	**Primary evidence intervention**	**Intended interpretation**
`shallow`	ROM −0.5; global reference deviation +0.5	Less range of motion and greater deviation from the reference indicate a shallower repetition.
`knees_over_toes`	Left/right knee-deviation features +0.5; for the diagnosis-aligned sensitivity condition, deep knee angles −0.5, knee–ankle offsets +0.5, and left–right symmetry difference +0.5	Greater knee deviation, increased knee displacement relative to the ankle, greater knee flexion, and greater asymmetry indicate stronger knee-travel evidence.
`back_not_straight`	Spine and shoulder-deviation features +0.5; for the diagnosis-aligned sensitivity condition, trunk inclination and its variance +0.5	Greater trunk-related deviation, forward inclination, or trunk variability indicates stronger evidence of non-straight back posture.

For the final diagnosis-aligned Concat+MLP, two conditions are evaluated with the same frozen checkpoint. In the first, only the 27 generic evidence features are perturbed and the ten diagnosis-aligned features are left unchanged. In the diagnosis-aligned sensitivity condition, label-matched diagnosis-aligned features are also perturbed. `shallow` is not evaluated in the sensitivity condition because no diagnosis-aligned feature is mapped to it, and it is counted as not passing.

All relevant-evidence interventions use a fixed magnitude of 0.5 standard deviations. A 1.0-standard-deviation shift was considered and rejected before TEST evaluation because it exceeded the observed TRAIN+VAL support for several features. Perturbed values are clamped to the observed TRAIN+VAL range. These constraints keep individual values

within their empirical ranges, but the intervention modifies extracted features rather than generating a physically coherent full-body pose or trajectory.

For each diagnosis, the comparison intervention perturbs evidence that is not directly label-defining while the primary evidence remains unchanged. For `shallow`, the comparison features are Path B quality indicators; for `knees_over_toes`, they are spine and shoulder deviations; and for `back_not_straight`, they are knee deviations. These features are not assumed to be statistically independent or biomechanically irrelevant. They serve only as non-primary-evidence controls, because movement features may co-occur in realistic repetitions.

Non-primary-evidence control intervention. This intervention alters non-primary evidence while keeping the primary evidence unchanged. The corresponding Δp_ℓ is computed analogously to Eq. (3.20). Diagnostic predictions are expected to remain comparatively stable, as assessed by changes in predicted probabilities and the frequency of label changes.

For each label, 30 TEST cases are sampled (seed 42), and all confidence intervals use a paired bootstrap ($B = 1{,}000$, seed 42). A label passes only if three conditions hold. First, the mean relevant-evidence Δp_ℓ is at least 0.05, its 95% confidence interval excludes zero on the positive side, and it exceeds the corresponding response of the Generic Concat+MLP under the same perturbation, with the confidence interval of the paired difference also excluding zero on the positive side. Second, the mean absolute non-primary-evidence $|\Delta p_\ell|$ is below 0.05 and the non-primary-evidence flip rate at threshold 0.5 is at most 0.10.

Third, the per-case contrast $|\Delta p_\ell^{\text{rel}}| - |\Delta p_\ell^{\text{ctrl}}|$ has a 95% confidence interval excluding zero on the positive side. Grounding is considered demonstrated only if at least two of the three tested labels pass. These thresholds were fixed before the first TEST grounding evaluation; no separate formal derivation for the specific 0.05 and 0.10 values was recorded.

The framework follows intervention-based interpretability, which evaluates behaviour under controlled evidence changes rather than relying only on post-hoc importance [21, 33]. FAIR-Coach therefore distinguishes:

1. predictive accuracy: whether a diagnosis is correct;
2. evidence necessity: whether removing an evidence source harms performance; and
3. evidence grounding: whether the prediction responds appropriately to directed evidence changes.

The thresholds and the two-of-three decision rule were fixed before the first TEST grounding evaluation and were not changed in the corrected re-evaluation. The resulting evaluation is reported in Chapter 4.

3.5.2 Diagnosis-to-Language Feedback and Preference Optimisation

The language stage operates only after `StructuredDiagnosis` has been fixed. Diagnosis confidence, session context, and reliability may influence the response strategy, but they

do not modify the diagnosis itself. The reliability gate from Section 3.2.1 therefore controls whether feedback is cautious or diagnosis-specific [90].

The language model receives the immutable diagnosis, a restricted evidence summary, and optional context rather than the complete `RepEvidence` record. This separation constrains language generation to evidence suitable for communication and distinguishes what the system has diagnosed from how that diagnosis is expressed [23, 82].

After evaluating the baseline language model without task-specific preference optimisation, we applied Direct Preference Optimization (DPO) [36] using a lightweight LoRA adapter [96]. DPO is used here to prefer evidence-faithful formulations rather than to learn a new biomechanical decision policy; the underlying diagnosis remains fixed [39, 97].

Preference pairs are constructed automatically from the diagnosis and its associated evidence [39, 98, 99]. Chosen responses follow the stored diagnosis without introducing unsupported information, while rejected responses represent:

- selection of an incorrect diagnostic theme;
- generic praise that omits the supported correction; and
- a claim unsupported by the diagnosis or evidence record.

The final preference dataset contains 420 constructed pairs and does not overlap with the held-out language-grounding evaluation set. The preferences are not based on human judgements; chosen responses are template-constructed from the evidence-license map.

This design limits the conclusions to faithfulness under the defined protocol rather than general coaching-language quality. The final DPO adapter is evaluated under the held-out language-grounding protocol reported in Chapter 4.

3.6 Scope and Methodological Boundaries

FAIR-Coach is evaluated within a narrower scope than a general autonomous coaching platform. The following boundaries define the claims supported by the implemented and validated methodology.

The core diagnosis experiments are squat-based and use an eight-category QEVD-derived multi-label taxonomy. Although the methodology could be adapted by redefining task-specific evidence, references, and labels, this thesis does not establish cross-exercise or clinical generalisation.

All diagnosis claims are conditioned on the validator-accepted subset of QEVD. Of the 13,013 original records, 6,310 passed the pose-confidence, depth, and range-of-motion validation filters. This selection may favour clips with more observable movement information and may introduce selection bias. Acceptance rates differed by diagnosis label (from 18% for `insufficient` to 77% for `wide` in TRAIN) and by participant (22–65% in TRAIN, 35–56% in validation, and 42–54% in TEST); these rates are descriptive and do not identify causes. The results should therefore not be generalised to rejected or low-quality recordings.

The accepted diagnosis records include basic squats (about 35% of accepted clips) as well as squat jumps, squat kicks, squat jacks, squat punches, and squat jabs. In contrast, the reference cohort is constructed only from basic-squat repetitions. A post hoc per-exercise analysis is reported in Chapter 4; it does not establish equal diagnostic performance across exercise variations.

The perception pipeline relies on 2D COCO-17 pose data, so viewpoint, occlusion, and pose-estimation errors may affect the derived evidence [63, 64]. No multi-view capture or 3D reconstruction is evaluated. Consequently, out-of-plane movement, depth, and some aspects of trunk and joint alignment cannot be directly recovered from the available observations. Where a toe landmark would be biomechanically preferable, the ankle is used as a proxy rather than inferring an unavailable measurement. Similarly, the deepest frame is a reproducible heuristic for critical-phase evidence rather than a ground-truth phase annotation. Responsible-joint outputs are not supervised because such labels are unavailable in the core dataset.

Path A is implemented in the broader live-system architecture but is absent from the QEVD short-clip diagnosis records. The core experiments therefore evaluate Path B together with the available biomechanical, reference, reliability, context, and diagnosis-aligned evidence. They do not evaluate learned Path-A/Path-B fusion, continuous session dynamics, or the benefit of Path A for diagnosis.

Checkpoint selection and freezing use TRAIN and validation data before access to the TEST split, but the final backend designation was made after TEST results had been

observed (Section 3.4.3). The TEST set contains repetitions from only five independent participants. Additional seed-based analyses and subject-disjoint cross-validation assess sensitivity to training initialisation and participant composition, but they do not create external validation cohorts. Thus, the reported results provide evidence within QEVD rather than independent replication or population-level generalisation [93].

The grounding protocol is deliberately stricter than predictive evaluation. A strong F1 score is not treated as evidence of biomechanical understanding. The intervention experiments test whether predictions respond in the intended direction to primary evidence and remain comparatively stable under non-primary-evidence control interventions. These interventions modify standardised extracted features, rather than generating physically coherent pose sequences. Failure to satisfy the predefined criterion is retained as a methodological boundary.

Diagnosis-to-language grounding is evaluated on a bounded held-out set using rule- and theme-based automatic metrics. The comparison evaluates the off-the-shelf language model against its DPO-adapted version while holding the structured diagnosis fixed. Although templates were used to construct the preference pairs, no template-based feedback system was evaluated as a baseline. Therefore, the results do not establish that an LLM is necessary or superior to a carefully designed template system. The chosen responses are template-constructed, and the evaluation does not replace expert human assessment [100]. The language results support improved faithfulness only under the defined protocol, not a complete evaluation of coaching-language quality.

LangGraph-style orchestration, A2A/MCP-compatible tool access, logging, and session state provide supporting infrastructure and build on earlier coaching systems [16–18]. They are not treated as the principal methodological contribution, and FAIR-Coach is not evaluated as a fully autonomous multi-agent decision-making system.

The evaluated user feedback is text-only. Visual feedback, such as pose overlays, trajectory comparison, or highlighted movement phases, and auditory feedback were not implemented or evaluated. Whether multimodal feedback improves user comprehension, motivation, or movement correction remains future work.

Finally, FAIR-Coach is intended for movement-quality assessment in coaching, not medical diagnosis. The core evaluation is offline and does not establish clinical validity, longitudinal outcomes, live-user efficacy, or safety in high-stakes settings. Within these boundaries, the methodology evaluates diagnosis-relevant evidence representations, controlled relational reasoning, intervention-based diagnosis grounding, and diagnosis-conditioned language generation.

4

Implementation and Evaluation

This chapter evaluates the FAIR-Coach methodology introduced in Chapter 3. We present implementation choices, the experimental protocol, quantitative results, and interpretation separately. The evaluation addresses the two thesis research questions: whether diagnosis-aligned biomechanical evidence improves movement diagnosis compared with generic, globally aggregated evidence, and whether explicit relational reasoning adds value over corresponding non-relational fusion. It also evaluates diagnosis-level evidence grounding and downstream language faithfulness as additional objectives.

The main experiments use the QEVD-FIT-COACH short-clip corpus with a frozen subject-disjoint protocol [15, 93]. Other datasets are included only for specific validation or integration steps. Core diagnosis conclusions in this chapter use the final frozen experimental configuration. Results from earlier splits and preference-pair runs with contamination are excluded. Earlier training protocols and reliability formulas are reported only where needed to document a controlled comparison or an implementation correction.

4.1 Implementation and Experimental Protocol

4.1.1 Datasets and Their Experimental Roles

QEVD-FIT-COACH short clips are the primary dataset for evidence extraction, diagnosis training, baseline evaluation, relational-reasoning experiments, diagnosis-grounding analysis, and language-grounding evaluation [15]. QEVD was selected because it combines scale, subject diversity, in-the-wild video, and coaching-style movement-error labels. These characteristics are important for the thesis because the target problem is not generic action recognition but multi-category movement diagnosis under realistic visual variation.

The remaining datasets serve more limited roles. Reconstruction from the stored training arrays indicates that EC3D [92], REHAB24 [52], and the custom squat collection (282 YouTube-derived sequences) were pooled to train the frozen Path-B ST-GCN classifier [25], whose outputs enter every diagnosis model as generic evidence. EC3D was also used in a supplementary deviation-validation analysis (Section 4.1.3). None of these datasets is used to train or evaluate the diagnosis models. The custom squat collection also supplies the unmodified video for the final end-to-end demonstration, while personal recordings are used only in integration and regression experiments and do not affect the final diagnosis metrics. We considered QEVD long-range videos for continuous Path-A or session-context extraction but excluded them because of mixed exercise content, inconsistent repetition alignment, and the lack of a suitable per-repetition multi-category diagnosis taxonomy.

Table 4.1 summarises the final experimental roles.

Table 4.1: Datasets used in the final FAIR-Coach evaluation and their restricted roles.

Dataset	**Final role**	**Status**
QEVD short clips	Primary diagnosis dataset: evidence extraction, baselines, relational reasoning, diagnosis quality, grounding, and language-grounding evaluation	Final
QEVD long videos	Investigated for Path A/session context; rejected because of mixed activities, weak alignment, and incompatible labels	Not used
EC3D	Path-B ST-GCN training data; supplementary deviation-validation analysis	Path-B training
custom_squat	Path-B ST-GCN training data (reconstructed); final real-video end-to-end demonstration	Path-B training; demo
REHAB24	Path-B ST-GCN training data	Path-B training
Personal recordings	Regression testing and live reliability scans	Integration only

The Path-B training record was not preserved. The stored training arrays contain 614 sequences, and their class counts (281 incorrect, 333 correct) equal the combined counts of EC3D (141 sequences), REHAB24 (191), and the custom squat collection (282); no other combination of the four available collections matches. This reconstruction is indirect: it does not identify individual samples, and the classifier's training seed and held-out accuracy were not recorded. None of these collections is part of QEVD, so no QEVD participant overlap is expected, although it was not separately checked for the YouTube-derived clips. The classifier's outputs on QEVD are cross-domain predictions.

4.1.2 Subject-Disjoint Splits, Filtering, and Leakage Control

All core diagnosis experiments use the frozen subject split in `subject_splits.json` version 2.0. The split was randomised with seed 42 rather than constructed sequentially by subject identifier. It contains 21 TRAIN subjects, four validation subjects, and five TEST subjects. Subjects 0028 and 0029 remain in TEST in accordance with the original benchmark-design intent.

The initial split includes 8,765 TRAIN clips, 2,141 validation clips, and 2,107 TEST clips. Applying the same frozen validator to each split results in accepted diagnosis datasets of 4,175 TRAIN clips, 1,083 validation clips, and 1,052 TEST clips. The corresponding acceptance rates are 47.6%, 50.6%, and 49.9%. Table 4.2 summarises these final counts.

Table 4.2: Final subject-disjoint QEVD split and validator acceptance.

Split	**Subjects**	**Pre-validation clips**	**Accepted clips**	**Acceptance**
TRAIN	21	8,765	4,175	47.6%
Validation	4	2,141	1,083	50.6%
TEST	5	2,107	1,052	49.9%
Total	30	13,013	6,310	—

Each rejected clip is attributed to one of three validator checks (Table 4.3); insufficient depth accounts for most rejections. Acceptance also varied by diagnosis label, from 18.1% (`insufficient`) to 76.7% (`wide`) in TRAIN, from 27.5% to 83.0% in validation, and from 20.7% to 79.5% in TEST, and by participant, from 22.0% to 65.4% in TRAIN, 35.3% to

56.1% in validation, and 42.2% to 53.8% in TEST. These rates are descriptive. The validator may therefore favour some labels and participants, and all results apply to accepted clips only. Rejected inputs are not passed to the diagnosis model, and system behaviour on them was not evaluated.

Table 4.3: Validator rejections by reason.

Split	Rejected	Insufficient depth	Low confidence	Insufficient ROM
TRAIN	4,590	3,823	753	14
Validation	1,058	917	139	2
TEST	1,055	910	142	3

Basic squats form 34.8% of accepted clips (2,193 of 6,310); the remainder are squat kicks (1,423), squat jumps (1,391), squat jacks (666), squat punches (618), and squat jabs (19).

We systematically checked the dataset split for subject overlap, duplicate clips, preprocessing leakage, and reference contamination. We found no subject overlap between TRAIN, validation, and TEST. No duplicate clip identifiers were present within TRAIN, and none of the validation clips overlapped with the 529 clips used to construct the biomechanical reference.

One earlier candidate reference had inherited a subject overlap from a pre-existing QEVD split, with subject 0028 appearing on both sides. That candidate was discarded. The reference was rebuilt from the thesis-specific subject split before use in any final experiment. This correction is important because a biomechanical reference is part of the evidence pipeline and can otherwise become an indirect leakage path [93].

The validated TRAIN set includes 371 of 4,175 examples (8.9%) where the labels contain a contradiction: 'no obvious issue' is assigned alongside at least one error category. These cases are kept in the dataset using the original multi-hot label format and are not converted to single-label examples. Validation and TEST contain 100 and 97 such clips, respectively. Sensitivity analyses that remove them are reported in Section 4.1.3.

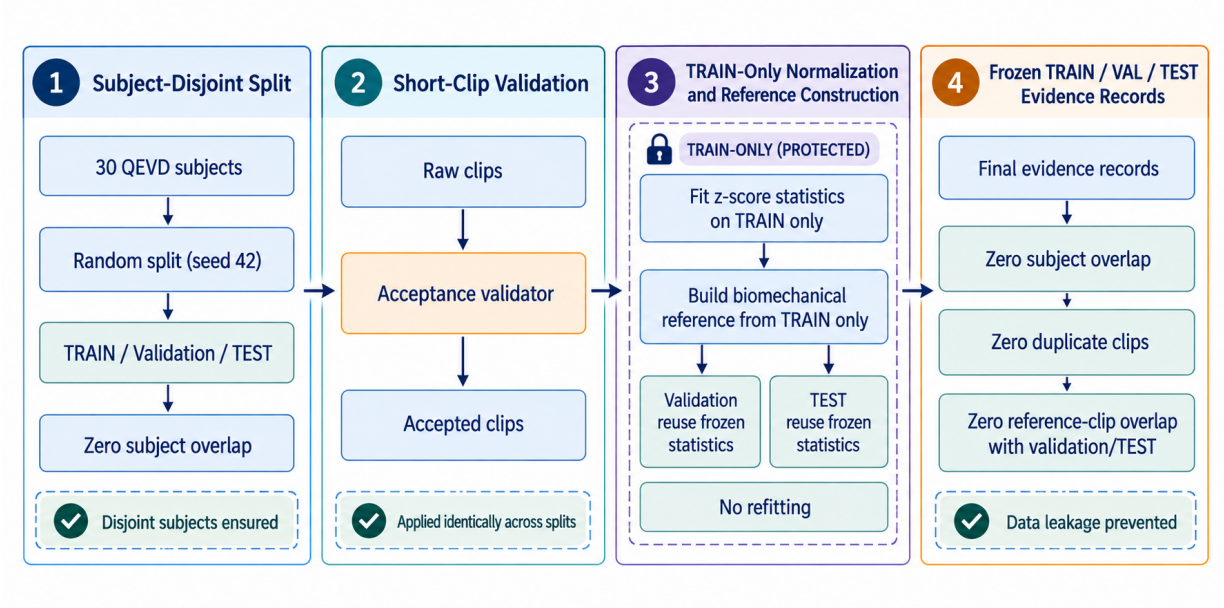


Figure 4.1: QEVD data-processing and leakage-control protocol. Subject-disjoint splits are validated independently, while normalisation statistics and the biomechanical reference are derived from TRAIN only and reused unchanged for validation and TEST. Final evidence records are frozen after the corresponding leakage checks are satisfied.

Figure 4.1 is intended to summarise the data flow and the points at which leakage is explicitly prevented.

4.1.3 Reference Construction and Evidence Extraction

We constructed the generic biomechanical reference from 529 accepted TRAIN basic-squat repetitions whose QEVD annotations contain both `squats` and `no_obvious_issue`. Each repetition includes nine joint or segment trajectories, temporally normalised to a length of 100. This reference is therefore a descriptive baseline for this dataset, rather than a universal or clinical definition of an ideal squat. Two reference formulations were implemented: a mean trajectory for each joint and time point, and a range-based formulation that retained per-time-point dispersion.

The original selection rule did not require `no_obvious_issue` to be the only label. Consequently, 333 of the 529 reference repetitions (62.95%) also carry at least one movement-error label. This is a limitation of the original reference construction. A sensitivity analysis therefore rebuilt the reference from the 196 TRAIN repetitions carrying only `no_obvious_issue`.

The mean-curve reference is

$$\mu_j(t) = \frac{1}{N_{\text{ref}}} \sum_{i=1}^{N_{\text{ref}}} \theta_{i,j}(t), \qquad N_{\text{ref}} = 529, \tag{4.1}$$

and the joint-level deviation for repetition r is

$$d_{r,j} = \frac{1}{100} \sum_{t=1}^{100} \left| \theta_{r,j}(t) - \mu_j(t) \right|. \tag{4.2}$$

The overall generic deviation is calculated as the mean of the nine joint or segment deviations.

Two post hoc sensitivity variants examined the contradictory labels (Table 4.4). Variant A removes all contradictory clips from every split (3,804 TRAIN, 983 validation, and 955 TEST clips) but keeps the original 529-clip reference; Variant B additionally rebuilds the reference from the 196 clean clips. Normalisation statistics were refitted on the cleaned TRAIN split in both variants. At seed 42, the diagnosis-aligned advantage was +0.041 in Variant A and +0.040 in Variant B, compared with +0.042 in the original configuration. A later follow-up over seeds 43–48, which was not part of the frozen sensitivity rule, gave positive differences at all seven seeds in both variants; in Variant A, the seed-48 confidence interval reached zero at its lower bound. Absolute micro-F1 is lower in both variants, which are evaluated on the cleaned 955-clip TEST split. These analyses reuse the same TEST participants and are not independent confirmations.

Table 4.4: Contradictory-label sensitivity analyses for the protocol-matched Concat+MLP comparison on TEST.

Config.	**Ref.**	**Generic**	**Diag.**	**Δ (seed 42)**	**95% CI**	**7-seed Δ (SD)**
Original	529	0.3018	0.3435	+0.0417	[0.0257, 0.0575]	+0.0522 (0.0104)
Variant A	529	0.2610	0.3017	+0.0407	[0.0271, 0.0541]	+0.0417 (0.0142)
Variant B	196	0.2630	0.3025	+0.0395	[0.0252, 0.0531]	+0.0513 (0.0129)

We compared the range-based and mean-curve references on QEVD validation clips (189 clean and 323 with an error label). The clean-versus-error mean gap for this reference ranged from +0.28° to +0.57° across the tested variants. In comparison, the mean-curve

formulation produced a $+1.31^{\circ}$ gap. Because the more complex formulation did not improve separation, we used the simpler curve formulation as the final reference. Panel A of Figure 4.2 summarises this comparison.

A supplementary analysis tested whether a global deviation score separates correct from incorrect repetitions on EC3D [92] (141 repetitions: 41 correct and 100 incorrect), using the mean absolute deviation from the EC3D correct-repetition mean over nine joints and 100 time points. The AUC was 0.602 (95% bootstrap CI $[0.501, 0.698]$; $B = 1{,}000$, seed 42), Because the lower bound lies just above 0.5, global deviation discriminated only weakly above chance. This analysis uses a different dataset and reference from the QEVD diagnosis experiments and is reported only as supporting evidence that a single global deviation score has limited discriminative value. Panel B of Figure 4.2 visualises this result.

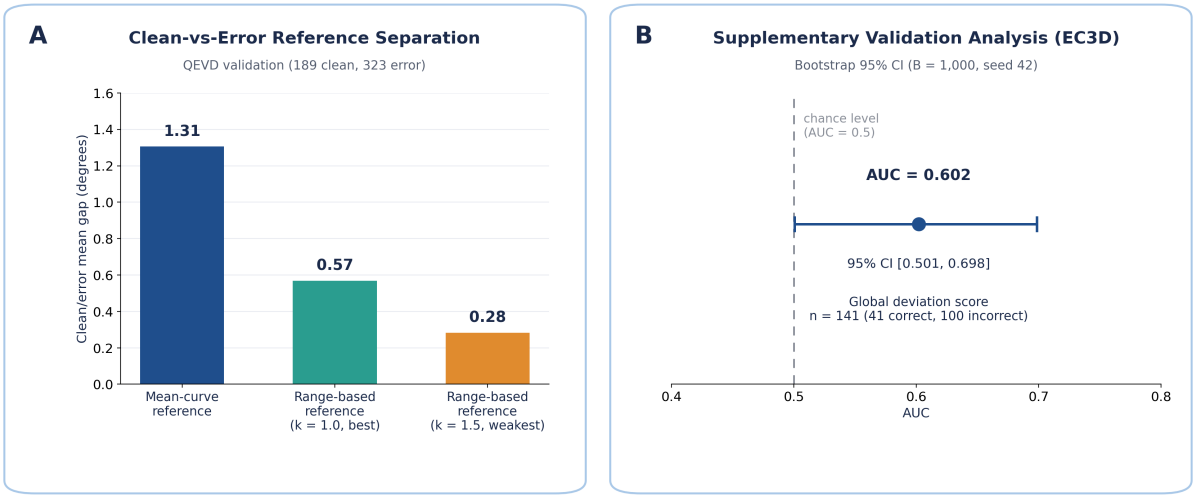


Figure 4.2: Comparison of reference formulations and supplementary generic-deviation validation analysis. **(A)** The mean-curve reference provides greater clean-versus-error separation than the tested range-based variants. **(B)** A global-deviation analysis on EC3D (141 repetitions) yields AUC $= 0.602$ (95% CI $[0.501, 0.698]$).

The final evidence extractor converts each validated repetition into a single `RepEvidence` record. The generic schema includes 27 raw evidence dimensions and a missingness mask. Path A defines three window-count features, but none appear in any QEVD short-clip record. Path B provides correct and incorrect probabilities, classification confidence, a normalised posture score, and the binary repetition assessment. The biomechanical block includes ROM, the global deviation score, and nine per-joint deviations. Reliability and context fields are included, with missing values explicitly marked [91].

The diagnosis-aligned extension introduces ten spatial features, most measured at the deepest frame: bilateral knee angles at the deepest frame, bilateral hip angles at the

deepest frame, trunk inclination at the deepest frame, trunk-inclination variance, bilateral knee–ankle horizontal offsets at the deepest frame, stance-width-to-hip-width ratio, and left–right symmetry difference [31, 32, 73, 74]. When evaluated individually, these diagnosis-aligned features show limited discriminative ability, with univariate AUCs ranging from 0.42 to 0.55. For comparison, the strongest generic single feature achieved an AUC of 0.65. The evaluation therefore focuses on their combined contribution rather than on individual univariate performance.

All z-score normalisation statistics are fitted on TRAIN only and reused unchanged for validation and TEST. Validation and TEST are never used to recompute normalisation parameters [93].

4.1.4 Model Training, Metrics, and Statistical Testing

The diagnosis task is eight-label multi-label classification. For N repetitions and $L = 8$ labels, $y_{i,\ell} \in \{0, 1\}$ denotes the ground-truth value for label ℓ of repetition i, $p_{i,\ell} \in [0, 1]$ denotes the corresponding predicted probability, and $\hat{y}_{i,\ell}$ denotes the thresholded binary prediction. Subset accuracy is

$$\text{SubsetAcc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I} \left[\hat{\mathbf{y}}_i = \mathbf{y}_i \right]. \tag{4.3}$$

Hamming accuracy is reported as the proportion of correctly predicted binary label

decisions:

$$\text{HammingAcc} = 1 - \frac{1}{NL}\sum_{i=1}^{N}\sum_{\ell=1}^{L}\mathbb{I}\left[\hat{y}_{i,\ell} \neq y_{i,\ell}\right]. \tag{4.4}$$

Any-hit accuracy measures whether at least one predicted label intersects the ground-truth set, where $\hat{Y}_i = \{\ell : \hat{y}_{i,\ell} = 1\}$ is the predicted label set and $Y_i = \{\ell : y_{i,\ell} = 1\}$ is the ground-truth label set:

$$\text{AnyHit} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[\hat{Y}_i \cap Y_i \neq \varnothing\right]. \tag{4.5}$$

Micro-F1 aggregates true positives, false positives, and false negatives over all labels:

$$F_1^{\text{micro}} = \frac{2TP}{2TP + FP + FN}, \tag{4.6}$$

whereas macro-F1 averages the label-wise F1 values:

$$F_1^{\text{macro}} = \frac{1}{L}\sum_{\ell=1}^{L} F_{1,\ell}. \tag{4.7}$$

Here TP, FP, and FN are the true-positive, false-positive, and false-negative counts aggregated over all $N \times L$ label decisions, and $F_{1,\ell}$ is the F1 score computed separately for label ℓ.

Calibration is evaluated with the Brier score,

$$\text{Brier} = \frac{1}{NL}\sum_{i=1}^{N}\sum_{\ell=1}^{L}\left(p_{i,\ell} - y_{i,\ell}\right)^2, \tag{4.8}$$

and label-averaged Expected Calibration Error (ECE),

$$\mathrm{ECE} = \frac{1}{L}\sum_{\ell=1}^{L}\sum_{m=1}^{M}\frac{|B_{m,\ell}|}{N}\left|\mathrm{acc}(B_{m,\ell}) - \mathrm{conf}(B_{m,\ell})\right|, \tag{4.9}$$

where $B_{m,\ell}$ contains the examples whose predicted probability for label ℓ falls in bin m, $M = 10$ is the number of equal-width bins, $\mathrm{acc}(B_{m,\ell})$ is the observed positive frequency, and $\mathrm{conf}(B_{m,\ell})$ is the mean predicted probability in that bin. ECE is computed separately for each label and then averaged over the eight labels. Bins are half-open except for the final bin, which includes probability one.

Paired model comparisons are conducted using 1,000 bootstrap resamples with seed 42. For each resample, we use the same example indices for both models, ensuring the resulting distribution reflects paired performance differences rather than independent scores. An improvement is attributed to a relational model only if the 95% confidence interval of the paired difference excludes zero in the beneficial direction on held-out TEST split. Improvements observed on validation data are not considered confirmatory evidence.

4.2 Generic Evidence and Baseline Performance

4.2.1 Baseline Comparisons

In the baseline sequence, the available evidence is varied while the target task remains fixed. Validation results for $n = 1{,}083$ accepted clips are shown in Table 4.5.

Table 4.5: Validation performance of the generic-evidence baselines.

Model	Subset	Hamming	Any-hit	Micro-F1	Macro-F1
Rule-based, all 1,083	0.243	0.844	0.191	0.234	0.086
Rule-based, taxonomy-fair subset (865)	0.150	0.819	0.239	0.249	0.090
Path-B-only logistic regression	0.020	0.467	0.367	0.188	0.188
Reference-only logistic regression	0.104	0.617	0.428	0.260	0.234
Path-B + reference logistic regression	0.110	0.618	0.413	0.255	0.230
Generic evidence concat+MLP	0.124	0.652	0.423	**0.278**	**0.247**

The rule baseline has a taxonomy mismatch: 20.1% of its predicted categories have no counterpart in the frozen QEVD eight-class taxonomy. For transparency, Table 4.5 therefore reports both the full-set result and the recomputed taxonomy-fair subset. The mismatch also explains why the full-set subset accuracy (0.243) exceeds any-hit accuracy (0.191). On the 218 clips whose predicted category has no counterpart, the mapped prediction is empty. Of these, 133 also have no label in the eight-class taxonomy, so they count as exact matches for subset accuracy but can never count as any-hit. This count is derived from the stored aggregate metrics (263 exact matches on all 1,083 clips versus 130 on the 865-clip subset).

The heuristic Path-A/Path-B fusion baseline cannot be applied to QEVD because Path A is not present in any of the short-clip records. As a result, none of the 1,083 validation examples achieve a reliability outcome other than `insufficient`, the lowest reliability state, under this heuristic. For this reason, we do not include Path-A-only and learned Path-A/Path-B baselines in the evaluation. These cases do not provide the required in-

puts. This exclusion reflects a defined scope boundary rather than an omitted experiment. Direct LLM, chain-of-thought, tree-of-thought, and graph-of-thought baselines [101–103] were considered as possible extensions in the original plan, but were not implemented after the main relational-reasoning comparison was completed using the concatenation baseline.

Among learned generic-evidence models, the original concatenation MLP establishes the strongest validation micro-F1 at 0.278 and reaches 0.2901 on the TEST split. This historical baseline used fixed-epoch training and serves as the principal generic non-relational baseline for the original relational-reasoning experiments. A subsequent protocol-matched comparison with the diagnosis-aligned Concat+MLP is reported in Section 4.3.1.

The baseline results show that no single generic evidence source consistently outperforms the others. Reference-only logistic regression achieves higher validation performance than Path-B-only logistic regression, but including both sources in a linear model does not improve over the reference-only model. Among the generic baselines, the nonlinear MLP performs best, suggesting that interactions among evidence fields can improve results even without an explicit relational structure [78, 79].

4.2.2 Interpretation of Generic Reference Evidence

The reference experiments show that improving the statistical description of acceptable movement does not improve performance on its own. We implemented and tested a distribution-aware range formulation, but it did not improve clean-versus-error separation compared with the mean trajectory. The supplementary EC3D analysis likewise indicates

only weak discriminability for a globally aggregated deviation score.

These findings shifted the emphasis from *how the reference is summarised* to *what evidence is exposed to the diagnosis task*. This distinction is important because the negative reference result does not imply that biomechanics is uninformative. It implies that a single globally aggregated distance from a reference can reach a representational ceiling when the target consists of multiple biomechanically distinct diagnosis categories [31, 58].

4.3 Diagnosis-Aligned Evidence and Relational Reasoning

4.3.1 Diagnosis-Aligned Evidence

The diagnosis-aligned representation extends the generic evidence by including ten literature-motivated spatial features, most measured at the deepest frame as a heuristic for the critical movement phase [31, 32, 73, 74]. Table 4.6 shows the main non-relational models evaluated using these features.

Table 4.6: Original Concat+MLP results obtained under different historical training protocols. A protocol-matched comparison is reported in Table 4.7.

Evidence / model	**VAL micro-F1**	**TEST micro-F1**
Generic evidence, Concat+MLP (original fixed-epoch)	0.278	0.2901
Diagnosis-aligned evidence, Concat+MLP (VAL-selected checkpoint)	**0.339**	**0.3435**

Because the original generic Concat+MLP used fixed-epoch training whereas the diagnosis-aligned Concat+MLP used validation-based early stopping and threshold selection over $\{0.4, 0.5, 0.6\}$, which selected 0.5, we repeated the comparison under identical training and model-selection conditions. In the protocol-matched comparison, both models used the same early-stopping rule and a fixed decision threshold of 0.5, together with identical optimisation settings, data splits, and training seeds. Across seven seeds (42–48), diagnosis-aligned evidence improved TEST micro-F1 at every seed, with a mean paired improvement of +0.052 and an across-seed standard deviation of 0.010. This matched comparison shows that the diagnosis-aligned improvement is not explained by the earlier difference in training protocol (Table 4.7). It was run after the original TEST evaluations and is a supporting robustness analysis, not the pre-specified primary comparison.

Table 4.7: Protocol-matched multi-seed comparison of generic and diagnosis-aligned Concat+MLP models on TEST.

Seed	**Generic F1**	**Diag. F1**	ΔF_1	**95% CI**	**Excludes 0?**
42	0.3018	0.3435	+0.0417	[0.0257, 0.0575]	Yes
43	0.2673	0.3361	+0.0688	[0.0508, 0.0863]	Yes
44	0.2864	0.3362	+0.0498	[0.0331, 0.0657]	Yes
45	0.2814	0.3331	+0.0514	[0.0371, 0.0657]	Yes
46	0.2772	0.3372	+0.0599	[0.0474, 0.0722]	Yes
47	0.2907	0.3461	+0.0552	[0.0397, 0.0705]	Yes
48	0.2735	0.3123	+0.0386	[0.0250, 0.0526]	Yes
Across-seed mean ΔF_1			+0.0522		
Across-seed SD			0.0104		

For the default diagnosis-aligned Concat+MLP configuration, whose checkpoint was

selected on validation data, the seed-42 TEST micro-F1 is 0.3435. This value is not a claim of the highest TEST score across all models or seed reruns, and the TEST cohort also informed the backend designation (Section 4.3.3). The pre-specified primary evidence comparison held the architecture fixed using Reasoner v2. At the original seed (42), replacing generic evidence with diagnosis-aligned evidence increased TEST micro-F1 by +0.039 (95% CI $[0.025, 0.053]$). The direction was consistent across seven training seeds (42–48): every seed showed a positive improvement with a 95% confidence interval excluding zero, with a mean $\Delta = +0.038$ and SD$= 0.010$ across seeds (Table 4.8).

Table 4.8: Seven-seed robustness analysis of the diagnosis-aligned versus generic evidence comparison using Reasoner v2 on TEST.

Seed	**Diag. F1**	**Generic F1**	ΔF_1	**95% CI**	**Excludes 0?**
42	0.3322	0.2932	+0.0392	[0.0246, 0.0532]	Yes
43	0.3183	0.2946	+0.0234	[0.0096, 0.0380]	Yes
44	0.3361	0.2915	+0.0442	[0.0300, 0.0586]	Yes
45	0.3438	0.2918	+0.0520	[0.0368, 0.0671]	Yes
46	0.3260	0.2828	+0.0431	[0.0293, 0.0569]	Yes
47	0.3262	0.2955	+0.0307	[0.0160, 0.0443]	Yes
48	0.3238	0.2917	+0.0317	[0.0167, 0.0468]	Yes
Across-seed mean ΔF_1			+0.038		
Across-seed SD			0.010		

A post hoc per-exercise breakdown of the seed-42 protocol-matched comparison (Table 4.9) shows a positive or zero difference for every exercise type. The gain was largest for squat kicks and basic squats and near zero for squat punches; squat jabs had only eight TEST clips, with micro-F1 of zero for both models. Absolute performance was much higher for basic squats, the only exercise type used to build the reference, than for the

other variants; this analysis does not identify the cause.

Table 4.9: Post hoc per-exercise TEST micro-F1 for the protocol-matched Concat+MLP comparison (seed 42).

Exercise	**TEST clips**	**Generic**	**Diag.**	**Δ**
Basic squats	379	0.4069	0.4450	+0.0381
Squat kick	228	0.2286	0.2727	+0.0441
Squat jump	209	0.2409	0.2693	+0.0284
Squat punch	114	0.0873	0.0902	+0.0029
Squat jacks	114	0.0444	0.0794	+0.0349
Squat jabs	8	0.0000	0.0000	+0.0000

Two additional literature-motivated features, a pelvis-relative trunk angle at the deepest frame and a dynamic knee–ankle offset averaged over the descent [31, 74], were evaluated together on validation data only. Adding both to the diagnosis-aligned Reasoner v2 changed validation micro-F1 from 0.3433 to 0.3353 (paired $\Delta = -0.0084$, 95% CI $[-0.0227, +0.0053]$). Because the difference was not distinguishable from zero and its point estimate was negative, the features were not retained and were not evaluated on TEST.

An exploratory appearance-feature extension added 32 principal components of EfficientNet [104] features, fitted on TRAIN only, to the 27 generic features. On validation, it reached micro-F1 0.564 at threshold 0.5 (0.574 at a validation-selected threshold of 0.6), compared with 0.278 for the generic Concat+MLP, and EfficientNet features alone reached 0.584 at threshold 0.6. Leakage checks confirmed TRAIN-only principal-component fitting and no subject or clip overlap. A subject-identification probe trained and tested within the four validation participants did not support a subject-identity shortcut (accuracy 0.444 ver-

sus a majority baseline of 0.435). Because appearance features alone matched or exceeded the combined model, this gain does not arise from the biomechanical evidence studied in this thesis. Its source was not established, the extension was not evaluated on TEST, and it lies outside the scope of the research questions.

The results indicate that improvement is associated with the combined representation rather than with any single diagnosis-aligned scalar. Individual diagnosis-aligned features achieve univariate AUCs between 0.42 and 0.55. Their contribution becomes apparent when multiple diagnosis-aligned measurements are provided together to the diagnosis model.

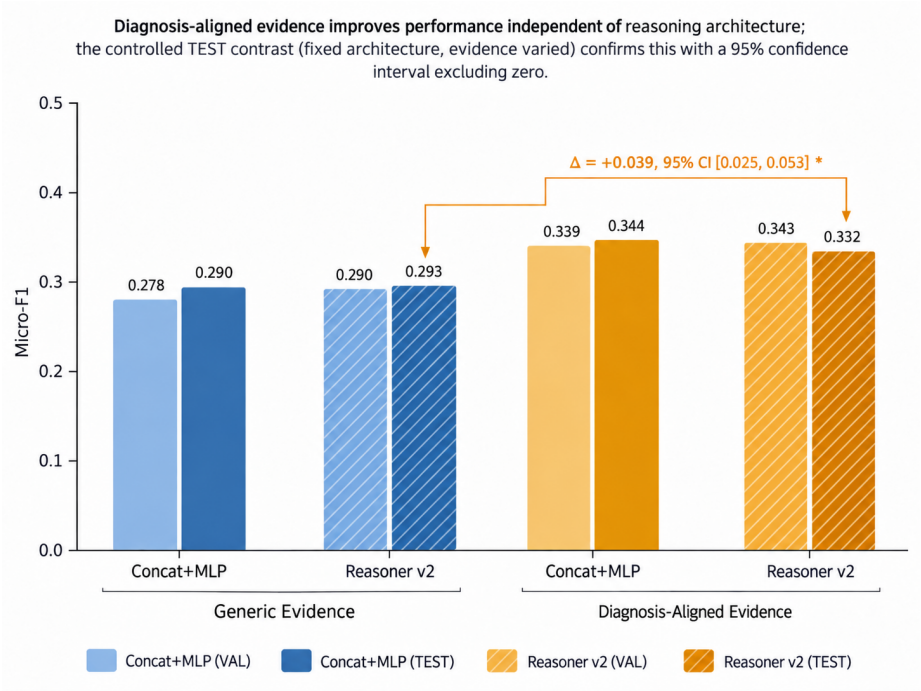


Figure 4.3: Effect of evidence representation across two diagnosis architectures. VAL and TEST micro-F1 are shown for generic and diagnosis-aligned evidence using Concat+MLP and Reasoner v2. The annotated TEST comparison holds the Reasoner v2 architecture fixed and shows the original seed-42 improvement from diagnosis-aligned evidence ($\Delta = +0.039$, 95% CI [0.025, 0.053]). The direction was subsequently consistent across seven training seeds (42–48), with a mean $\Delta = +0.038$ and SD= 0.010; all seven seed-specific 95% confidence intervals excluded zero, as reported in Table 4.8. The Concat+MLP points shown here are the original historical runs and used different training-selection protocols; the protocol-matched Concat+MLP comparison is reported separately in Table 4.7.

Figure 4.3 highlights the performance difference between generic and diagnosis-aligned evidence across the two diagnosis architectures.

4.3.2 Relational Versus Non-Relational Reasoning

We tested four relational configurations against corresponding non-relational controls. Table 4.10 reports the frozen validation and TEST results.

Table 4.10: In the relational-reasoning experiments, conclusions come from paired bootstrap comparisons rather than relying solely on point estimates.

Architecture	**Evidence**	**VAL micro-F1**	**TEST micro-F1**
Coarse 5-node message passing	Generic	0.2557	0.2648
Fine 8-node, relations ON	Generic	0.2911	0.2969
Fine 8-node, relations OFF	Generic	0.2860	0.2905
Label-conditioned attention (Reasoner v2)	Generic	0.2898	0.2932
Label-conditioned attention (Reasoner v2)	Diagnosis-aligned	0.3433	0.3322
Diagnosis-aligned concat+MLP	Diagnosis-aligned	0.339	**0.3435**

Table 4.11 reports the paired TEST differences (B = 1,000, seed 42). The Generic Concat+MLP control is the original fixed-epoch baseline.

Table 4.11: Paired TEST micro-F1 differences between relational configurations and their non-relational controls (seed 42).

Comparison	**Mean Δ**	**95% CI**	**Excludes 0?**
Coarse 5-node − Generic Concat+MLP	−0.0251	[−0.0387, −0.0096]	Yes (worse)
Fine relations-ON − relations-OFF	+0.0063	[−0.0053, +0.0190]	No
Fine relations-ON − Generic Concat+MLP	+0.0068	[−0.0063, +0.0201]	No
Generic Reasoner v2 − Generic Concat+MLP	+0.0032	[−0.0080, +0.0143]	No
Diag. Reasoner v2 − Diag. Concat+MLP	−0.0110	[−0.0241, +0.0024]	No

The coarse five-node message-passing model performs worse than the generic con-

cat+MLP on TEST.

$$\Delta F_1^{\text{micro}} = -0.0251, \qquad 95\%\ \text{CI} = [-0.0387, -0.0096]. \tag{4.10}$$

The confidence interval lies entirely below zero, which supports the presence of a detrimental relation mechanism in this configuration [78]. This result does not indicate a null effect.

The fine relations-ON model achieves a validation score of 0.2911, which is higher than the generic MLP baseline at 0.278. On the test split, however, this advantage is not statistically supported, as the relevant confidence intervals include zero when compared to both the generic MLP and the relations-OFF control. A similar pattern is observed for generic Reasoner v2. While it improves over the generic MLP on validation, the TEST-split contrasts do not provide evidence of a significant difference.

We evaluated the diagnosis-aligned Reasoner v2 to isolate the effects of evidence quality and relational reasoning at the strongest evidence level. At the original seed (42), compared to the generic-evidence Reasoner v2, the diagnosis-aligned model increased TEST micro-F1 by +0.039 (95% CI: [0.025, 0.053]). This evidence-design effect was subsequently consistent across seven training seeds, with a mean $\Delta = +0.038$ and SD= 0.010 (Table 4.8). In contrast, when compared to the diagnosis-aligned concat+MLP, the relational model is lower by -0.011 (95% CI: [-0.024, $+0.002$]). Because this interval includes zero, the relational and non-relational models differ by no statistically significant amount when evidence quality is controlled. Figure 4.4 presents the paired TEST results and distinguishes

the effects of reasoning architecture from those of evidence quality.

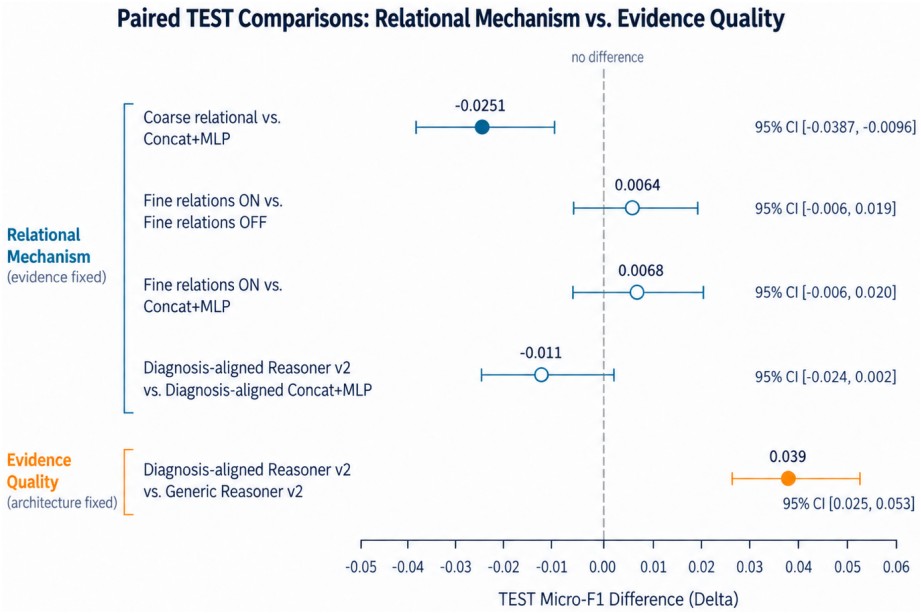


Figure 4.4: Paired comparisons on the held-out TEST split evaluate relational mechanisms and evidence quality. In the first set of comparisons, the reasoning architecture is varied while the evidence representation remains fixed. In the final comparison, the evidence representation is varied with the Reasoner v2 architecture held constant. Filled markers indicate comparisons where the 95% confidence interval does not include zero, and open markers indicate intervals that do include zero. The results show that the coarse relational model performs significantly worse than the non-relational baseline. Comparisons involving the generic Concat+MLP use the original historical fixed-epoch 8.8 baseline and should therefore be interpreted with the baseline-training caveat described in Section 4.6.2. The only statistically confirmed improvement in the displayed seed-42 comparisons is observed when generic evidence is replaced with diagnosis-aligned evidence; this effect was consistent across seven training seeds (Table 4.8).

The results do not support a benefit for explicit relational mechanisms in this setting. Across coarse message passing, fine-grained message passing, generic label-conditioned attention, and diagnosis-aligned label-conditioned attention, none of the relational approaches achieves a statistically significant improvement over the corresponding

non-relational baseline on the held-out TEST split. In one comparison, the relational mechanism performs significantly worse, as shown in Figure 4.4 and Table 4.11.

A second methodological observation concerns the validation-to-TEST pattern. Three relational configurations showed promising results during validation, but these findings were not statistically supported on the TEST split. This repeated pattern indicates that conclusions should not be based solely on validation point estimates or a single data split [86, 93].

Seven-seed follow-ups were run for the two borderline fine-model comparisons (Table 4.12). Relations-ON versus relations-OFF varied in sign across seeds: five seeds were positive, one was significantly positive, and one was significantly negative (mean $\Delta = +0.003$). Relations-ON versus the fixed-epoch Generic Concat+MLP was positive at five seeds, significantly positive at three, and significantly negative at one (mean $\Delta = +0.010$). Neither comparison is consistent across seeds, and these follow-ups do not establish a relational benefit.

Table 4.12: Seven-seed follow-ups of the two borderline relational comparisons on TEST (paired Δ micro-F1 with 95% CI).

Seed	**Relations-ON − OFF**	**Relations-ON − Generic Concat+MLP**
42	$+0.0064\ [-0.0053, +0.0190]$	$+0.0068\ [-0.0063, +0.0201]$
43	$+0.0316\ [+0.0197, +0.0445]$	$+0.0183\ [+0.0066, +0.0299]$
44	$+0.0020\ [-0.0093, +0.0133]$	$+0.0195\ [+0.0071, +0.0317]$
45	$+0.0007\ [-0.0088, +0.0109]$	$+0.0113\ [-0.0005, +0.0239]$
46	$-0.0078\ [-0.0172, +0.0026]$	$-0.0002\ [-0.0138, +0.0144]$
47	$-0.0198\ [-0.0313, -0.0088]$	$-0.0175\ [-0.0292, -0.0052]$
48	$+0.0071\ [-0.0040, +0.0174]$	$+0.0288\ [+0.0165, +0.0413]$
Mean	+0.003	+0.010

4.3.3 Final Diagnosis Model Selection

The default diagnosis backend is the diagnosis-aligned Concat+MLP configuration. Its checkpoint was selected on validation data (micro-F1 0.3393) and frozen before TEST access. However, the choice of this configuration as the default backend is recorded only in the final experiment manifest, written after the TEST results for both frozen diagnosis-aligned models had been observed. The diagnosis-aligned Reasoner v2 had the higher validation micro-F1 (0.3433) but the lower TEST micro-F1 (0.3322 versus 0.3435), and their paired TEST difference was not significant. The TEST cohort therefore informed the final backend choice, and the TEST micro-F1 of 0.3435 is not an untouched estimate for the selected system.

The default deployment choice favours non-relational fusion of the full diagnosis-aligned evidence while avoiding additional relational-model complexity. Alternative evaluated checkpoints remain available behind a configuration switch for reproducibility. This should not be interpreted as a general rejection of relational modelling [27, 28]; it is a bounded conclusion for the evidence representation, diagnosis taxonomy, and data regime evaluated in this thesis.

4.4 Diagnosis Quality, Calibration, Ablation, and Grounding

4.4.1 Predictive Performance and Calibration

Table 4.13 reports locked TEST results for the principal generic-evidence models and post-hoc calibration metrics for the final diagnosis-aligned MLP.

Table 4.13: Locked TEST diagnosis and calibration metrics ($n = 1{,}052$). Values for the diagnosis-aligned MLP marked † were computed post hoc from frozen TEST predictions and were not used for model selection.

Model	Subset	Hamming	Any-hit	Micro-F1	Macro-F1	Brier	ECE
fine-OFF	0.129	0.698	0.378	0.290	0.260	0.194	0.304
fine-ON	0.126	0.666	0.422	0.297	0.262	0.201	0.308
Generic Concat+MLP	0.121	0.667	0.410	0.290	0.259	0.203	0.316
Locked coarse 5-node	0.191	0.641	0.384	0.265	0.215	0.222	0.357
Diagnosis-aligned MLP	–	–	–	0.344	$0.314^{\dagger}$	$0.171^{\dagger}$	$0.256^{\dagger}$

Among the four generic-evidence models, fine-ON achieves the highest micro-F1, while fine-OFF obtains the lowest Brier score and ECE. The locked coarse model has the highest subset accuracy but the lowest micro-F1 and macro-F1 in this group. The final diagnosis-aligned MLP has lower post-hoc Brier score (0.1706) and ECE (0.2557) than the generic-model comparators. These lower values are descriptive only: they were calculated after model selection from frozen TEST predictions and do not establish a statistically confirmed calibration advantage. Its probabilities should therefore not yet be treated as validated confidence estimates for high-stakes use.

In addition to diagnostic accuracy, we evaluated the calibration of confidence estimates. Figure 4.5 reports reliability curves and per-label diagnostic performance for three generic-evidence models with persisted per-example TEST predictions. The figure does not include the final diagnosis-aligned MLP; its Brier and ECE values are instead reported as post-hoc aggregate metrics in Table 4.13. Panel A shows that all three reliability curves are below the perfect-calibration line across most of the confidence range. Mean predicted probabilities near 0.55 correspond to observed frequencies of 0.15 to 0.20, which is consistent with the large ECE values reported in Table 4.13 [71, 72].

Panel B presents per-label F1 scores for the eight movement-error categories. All three models show similar class-wise profiles, with higher performance in some categories and lower in others. This provides more detail than the aggregate micro- and macro-F1 scores. The locked coarse relational model is not included because its per-example TEST predictions were not saved. We report its aggregate results separately in Table 4.13.

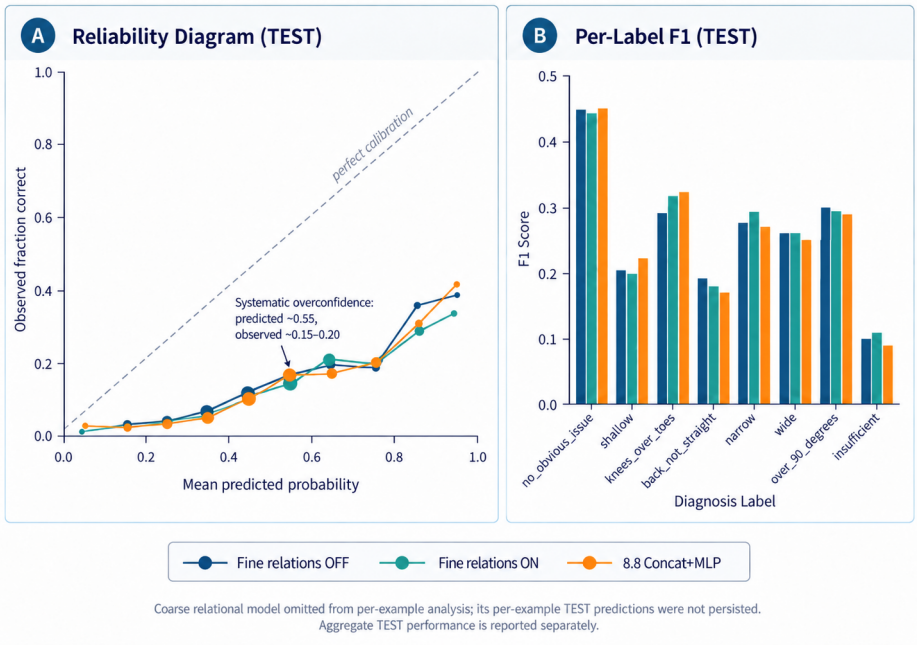


Figure 4.5: Calibration and per-label diagnostic performance on the locked TEST split (n = 1,052) for three generic-evidence models with persisted per-example predictions. **(A)** Ten-bin reliability diagrams compare mean predicted probability with observed empirical frequency; the dashed diagonal represents perfect calibration. **(B)** Per-label F1 across the eight diagnosis categories. The locked coarse relational model is not included in the per-example analysis because its TEST predictions are unavailable. Table 4.13 reports its aggregate results.

4.4.2 Evidence Ablation and Missing-Evidence Robustness

Ablation is measured as the difference in micro-F1 between the full model and the same model with a specific evidence group excluded.

$$\Delta F_1^{\text{micro}} = F_{1,\text{full}}^{\text{micro}} - F_{1,\text{ablated}}^{\text{micro}}. \tag{4.11}$$

Positive values therefore indicate a performance loss after ablation.

Table 4.14 reports the frozen quantitative ablations for Path B, reference evidence, and reliability.

Table 4.14: TEST evidence ablations. Positive Δ indicates that removing the evidence reduced micro-F1.

Ablation	**Model**	Δ	**95% CI**	**Excludes 0?**
Path B	fine-OFF	+0.019	[0.007, 0.030]	Yes
Path B	Generic concat+MLP	+0.011	[−0.000, 0.022]	No
Reference	fine-OFF	+0.004	[−0.007, 0.014]	No
Reference	fine-ON	+0.025	[0.015, 0.034]	Yes
Reference	Generic concat+MLP	+0.012	[0.004, 0.020]	Yes
Reliability	fine-OFF / fine-ON / MLP	small, mixed sign	all include 0	No

The ablation results indicate that the importance of each evidence group varies across model families. Path B contributes significantly to fine-OFF, but not to the generic MLP. Reference evidence is significant for fine-ON and the generic MLP, but not for fine-OFF. Reliability evidence does not show a consistent or statistically significant contribution in these settings.

Evidence redundancy may account for the observed architecture-dependent behaviour. The group-correlation analysis found reference–knee correlations up to $r = 0.736$, which suggests that related geometric information is present in multiple evidence groups. This overlap means that removing one evidence group can sometimes be offset by another [22, 95]. Figure 4.6 shows that knee and trunk ablations produce the largest confirmed performance reductions for fine-OFF and the generic MLP, the two architectures evaluated with these evidence groups under the frozen protocol.

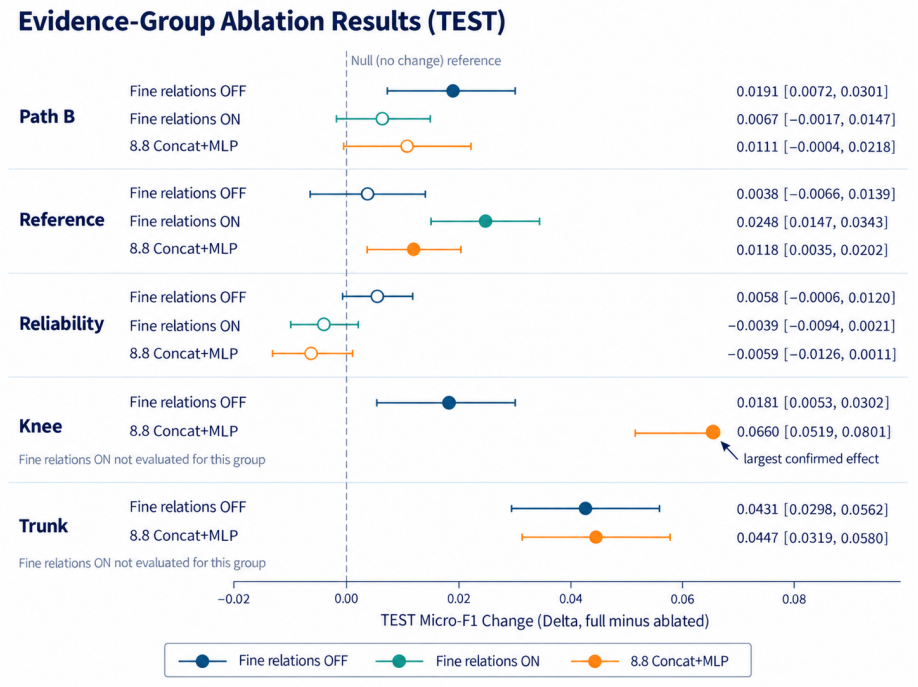


Figure 4.6: Ablation results for each evidence group were obtained using the locked TEST split. Filled markers indicate 95% bootstrap confidence intervals that do not include zero, while hollow markers indicate intervals that include zero. Knee and trunk ablations were evaluated only for fine-OFF and for 8.8 Concat+MLP, both under the frozen evaluation protocol.

4.4.3 Evidence-Grounding Interventions

Ablation tests whether evidence contributes to model performance, but does not indicate whether the model's response aligns with the intended direction. To address this, FAIR-Coach uses the intervention protocol described in chapter 3 [21, 87]. For diagnosis label ℓ, the relevant-evidence response is summarised as

$$\Delta p_\ell = p_\ell(E^{\text{intervened}}) - p_\ell(E^{\text{original}}). \tag{4.12}$$

Here $p_\ell(E)$ denotes the model-predicted probability for diagnosis label ℓ given evidence representation E; it is the single-example counterpart of $p_{i,\ell}$ defined above. The grounding rule and its thresholds are defined in Section 3.5.1. The original protocol was frozen before the first TEST evaluation on September 7 and applied to the generic fine-OFF model. The corrected sign conventions and the evaluation of the final diagnosis-aligned model were added afterwards and run on October 17; they were fixed in code immediately before that run and are not preregistered.

Table 4.15 reports the corrected grounding evaluation for the final diagnosis-aligned concat+MLP model ($n = 30$ cases per label).

Table 4.15: Corrected evidence-grounding intervention results for the final diagnosis-aligned Concat+MLP model, with only generic evidence features perturbed.

Label	**Primary-evidence response**	**Non-primary-evidence control stability**	**Overall**
`knees_over_toes`	+0.0065, 95% CI [−0.001, 0.014]; does not meet the primary-evidence criterion	Passes	Fail
`shallow`	+0.0622, 95% CI [0.054, 0.071]; passes in the expected direction	Passes	Pass
`back_not_straight`	−0.0760, 95% CI [−0.092, −0.059]; statistically confirmed wrong-direction response	Fails	Fail

Only `shallow` satisfies all conditions of the predefined grounding rule. `knees_over_toes`

does not show a sufficiently supported primary-evidence response, while `back_not_straight` responds in the opposite direction and also fails the non-primary-evidence control condition. Because only one of the three labels passes, grounding is not demonstrated under the overall rule.

Table 4.16 places this result alongside the other grounding evaluations. In the original evaluation of the generic fine-OFF model, no label passed, and after the sign correction fine-OFF still passed none. When the label-matched diagnosis-aligned features of the final model were also perturbed, no label passed: the `knees_over_toes` response rose to +0.038 but remained below the 0.05 threshold without a significant contrast, and the `back_not_straight` response became more strongly negative (−0.160). `shallow` could not be tested in this condition. The final-model result therefore depends on which features are perturbed, and perturbing the diagnosis-aligned features did not produce grounded behaviour for the tested labels.

Table 4.16: Mean relevant-evidence Δp_ℓ across grounding evaluations ($n = 30$ cases per label).

Model / condition	`knees`	`shallow`	`back`	**Passing**
Fine-OFF, original protocol	+0.026	−0.011	−0.103	0/3
Fine-OFF, corrected	+0.023	+0.021	−0.108	0/3
Final model, generic features perturbed	+0.007	+0.062	−0.076	1/3
Final model, aligned features also perturbed	+0.038	n/a	−0.160	0/3
Generic Concat+MLP (comparator)	+0.084	+0.029	−0.056	not assessed

The interventions modify standardised extracted feature values by 0.5 standard deviations and clamp them to observed TRAIN+VAL ranges. They do not generate physically

coherent poses or video trajectories. Thus, this evaluation tests whether predictions respond as intended to controlled feature-space changes; it does not establish causal reliance on biomechanically feasible movement changes.

The grounding results indicate that predictive performance and evidence-grounded behaviour are distinct properties [87, 95]. Although the diagnosis-aligned representation improves predictive performance, the final-model intervention results do not demonstrate consistently grounded diagnostic behaviour.

4.5 Diagnosis-to-Language Grounding and Final System Evaluation

4.5.1 Language-Grounding Baseline and Preference Optimisation

The downstream language experiment uses a fixed diagnosis as input. The language model does not make or revise the movement diagnosis; its role is limited to expressing the fixed diagnosis and evidence in feedback language. The baseline model, `Qwen3.8-27B` [105], is evaluated in zero-shot mode on 40 held-out validation examples without tool use or agentic actions.

Before preference optimisation, unsupported claims occur in 30 of 40 examples (75%), evidence specificity is 75%, and own-predicted-label consistency is 9 of 40 examples (22.5%)

[23].

The final preference dataset consists of 420 constructed pairs, all disjoint from the 40-example evaluation set. Rejected responses represent three distinct failure modes: incorrect diagnostic theme, generic praise, and unsupported claims. The LoRA-DPO configuration [36, 96] applies `Qwen3.8-27B` with rank $r = 8$, $\alpha = 16$, q/k/v/o projection targets, a learning rate of 5×10^{-5}, $\beta = 0.1$, and 156 optimisation steps. Training exposure matches their effective epoch fraction, controlling for differences in optimisation exposure during the final comparison. An earlier preference file containing 540 pairs overlapped between the training and evaluation sets and is therefore excluded from all thesis results [93]. Only the contamination-corrected and exposure-matched 420-pair run is used in the reported analyses.

Table 4.17 reports the held-out, contamination-corrected result.

Table 4.17: Diagnosis-to-language grounding before and after LoRA-DPO on the held-out $n = 40$ evaluation set.

Metric	**Before DPO**	**After DPO**
Unsupported-claim rate	75.0%	**17.5%**
Evidence specificity	75.0%	**92.5%**
Own-predicted-label consistency	22.5%	**77.5%**

Preference optimisation reduces unsupported claims by 57.5 percentage points, increases evidence specificity by 17.5 points, and increases own-label consistency by 55.0 points. Because the diagnosis is fixed before the language stage, these changes are attributable to the verbalisation layer rather than to a change in movement diagnosis [39].

The claim's scope is limited to the defined evaluation conditions. This experiment compares zero-shot Qwen with LoRA-DPO Qwen; it does not evaluate a deterministic template-based feedback baseline. Therefore, it does not establish that an LLM is necessary, computationally justified, or superior to retrieving carefully designed evidence-licensed templates for this task.

The preference pairs are constructed using templates based on the same evidence-license map as the automatic grounding metrics. The adapter may therefore learn both template style and evidence-faithful preference. The evaluation set includes 40 examples, and the metrics rely on rule-based or keyword-theme criteria rather than expert human judgements [100]. The results indicate improvement within the specified automatic language-grounding protocol, but do not establish overall coaching-language quality or user benefit.

4.5.2 Final Pipeline Integration and Reliability Validation

The integrated system is configured with `diag-concat-MLP` as the default diagnosis backend and the Phase-12c DPO LoRA adapter as the default feedback backend. For reproducibility, `reasoner_v2` and the rule-based path can still be selected through configuration switches.

The reliability score combines three components. With n_w Path-A windows, n_{inc} of them classified incorrect, and incorrect rate $\rho = n_{\text{inc}}/n_w$, the agreement rate is $a = \rho$ if Path B labels the repetition INCORRECT and $a = 1 - \rho$ otherwise; window coverage is $c = \min(n_w/5, 1)$; and path agreement g is 1 if Path B is INCORRECT with $\rho \geq 0.5$ or

CORRECT with $\rho < 0.5$, and 0 otherwise. The score is

$$R = 0.33\,a + 0.33\,c + 0.34\,g, \tag{4.13}$$

clipped to $[0, 1]$. A repetition is labelled strong for $R \geq 0.75$, partial for $R \geq 0.45$, conflicting for $R \geq 0.20$, and insufficient otherwise. These weights and thresholds are fixed engineering choices. In Version 3, n_{inc} counts stable incorrect-window states when available.

A reliability implementation correction was evaluated retrospectively. Earlier scans counted only event-based `incorrect_windows`, whereas Version 3 uses stable incorrect-window states when available. This correction changes the reliability calculation but does not alter the Path B assessment label.

The original scans cover 315 videos and 375 repetitions. However, only the 237 `custom_squat` repetitions had cached keypoints and could be rescored under Version 3; the remaining 138 demonstration repetitions were not rescored. Within the rescorable `custom_squat` subset, the number of strong-reliability repetitions increased from 126 under the old formula to 188 under Version 3.

Among the 89 `custom_squat` repetitions labelled INCORRECT by Path B, none had strong reliability under the old event-based formula. Under Version 3, 55 INCORRECT repetitions achieved strong reliability and emitted negative feedback. The Path B labels themselves remained unchanged. Because the `custom_squat` sequences were probably part of the Path-B training data (Section 4.1.1), the Path-B labels on these repetitions are

likely in-sample. This analysis therefore checks the reliability calculation only; it does not estimate Path-B accuracy.

For `SQUAT_516.mp4`, 23 of 25 windows were stably classified as incorrect. The reliability score consequently changed from 0.3564 (conflicting) under the old event-based formula to 0.9736 (strong) under Version 3. This correction is implementation validation rather than a thesis contribution, but it matters because the reliability policy controls whether diagnosis-specific language can reach the user [90].

4.5.3 End-to-End Demonstration and Reproducibility Gate

We conducted the end-to-end demonstration using the unmodified clip `SQUAT_516.mp4`. The pipeline generated 25 Path-A windows, of which 23 were identified as incorrect. The final reliability score was 0.974, Path B classified the repetition as INCORRECT, and the diagnosis backend returned `over_90_degrees`. The DPO feedback backend then produced deterministic feedback from the frozen inference path.

Table 4.18 records the frozen demonstration outputs.

Table 4.18: Final end-to-end demonstration on `SQUAT_516.mp4`.

Field	**Frozen output**
Path-A windows	25 total; 23 stable-incorrect
Reliability	0.974, strong
Path-B assessment	INCORRECT
Diagnosis	`over_90_degrees`
Feedback backend	DPO LoRA
Reproducibility gate	PASS

This trace provides a qualitative, single-example demonstration of the complete pipeline operating end-to-end on a real, unmodified video. It is not a statistical evaluation, and its interpretation is limited to this example. The final reproducibility gate disables stochastic generation (`do_sample=False`) and reconstructs the inference path using only the frozen artefacts. The reliability score, diagnosis, and feedback text are reproduced exactly, confirming that the gate passes.

Figure 4.7 summarises this end-to-end trace from evidence extraction to the final reproducibility check.

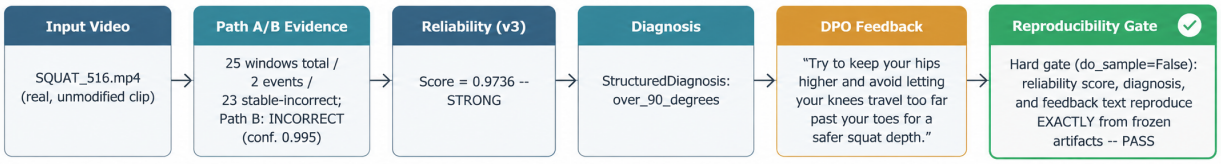


Figure 4.7: End-to-end FAIR-Coach trace on a real, unmodified video (`SQUAT_516.mp4`), from raw evidence through reliability assessment, diagnosis, and preference-optimised feedback, concluding with a passed hard reproducibility gate.

The final manifest contains the checkpoint hashes, split version, thresholds, and runtime versions used in the experiments. The frozen runtime consists of Python 3.10.20, PyTorch 2.5.1+cu121, Transformers 5.17.0, and PEFT 0.20.0. Since no version-control repository is available, reproducibility depends on recorded artefact hashes and timestamps rather than commit history [93].

4.6 Overall Findings, Limitations, and Threats to Validity

4.6.1 Findings with Respect to the Research Questions

Table 4.19 summarises the principal experimental conclusions.

Table 4.19: Summary of the thesis research questions and the corresponding experimental evidence.

RQ	Question	Experimental conclusion
RQ1	Does diagnosis-aligned, literature-grounded biomechanical evidence improve movement-diagnosis performance compared with generic, globally aggregated evidence on held-out subjects?	Yes, for the full diagnosis-aligned configuration. Diagnosis-aligned evidence improved TEST performance across all seven training seeds under the architecture-fixed Reasoner v2 comparison (mean $\Delta = +0.038$, SD= 0.010). A separate protocol-matched Concat+MLP comparison showed the same direction across seven seeds (mean $\Delta = +0.052$, SD= 0.010). Because this configuration retains generic evidence and adds ten extra measurements, these comparisons do not isolate alignment alone as the cause of the improvement.
RQ2	Does explicit relational reasoning over heterogeneous biomechanical evidence provide measurable benefit beyond corresponding non-relational fusion of the same evidence?	No confirmed benefit. None of the tested relational mechanisms produced a statistically confirmed TEST improvement over its corresponding non-relational comparison, and the coarse relational model performed significantly worse.

Taken together, the results support a specific scientific interpretation. The strongest improvement observed in this thesis comes from redesigning the evidence representation so that the diagnosis model receives diagnosis-relevant biomechanical information. Increas-

ing reasoning complexity after the evidence representation is fixed does not produce a corresponding statistically confirmed gain.

This result does not imply that the final diagnosis model is fully biomechanically grounded. In the final-model intervention evaluation, `shallow` satisfies the predefined grounding rule. However, `knees_over_toes` does not show a sufficiently supported primary-evidence response, and `back_not_straight` responds in the wrong direction. Because only one of three labels passes, overall grounding is not demonstrated. The resulting conclusion is therefore twofold: diagnosis-aligned evidence improves predictive performance, but predictive improvement alone does not establish correct biomechanical reasoning.

At the language layer, lightweight preference optimisation reduces unsupported claims and improves diagnosis-consistent verbalisation under the specified automatic evaluation protocol. These results concern the communication layer after the diagnosis is fixed. They do not establish that an LLM is necessary or superior to a carefully designed template-based feedback system [39, 97].

4.6.2 Limitations and Threats to Validity

Data and label validity. The validated TRAIN split contains an 8.9% contradiction rate under the frozen eight-class taxonomy. Multi-hot labels preserve these source annotations rather than resolving them heuristically. The rule-based baseline also exhibits a 20.1% taxonomy mismatch, which is why both full-set and taxonomy-fair results are reported.

A further limitation concerns the original 529-repetition biomechanical reference. Although it is constructed from TRAIN only and therefore does not introduce validation or TEST leakage, 333 repetitions (62.95%) also carry at least one movement-error label. The clean-reference sensitivity analysis shows that the diagnosis-aligned advantage remains under a 196-repetition clean reference, but the original reference should not be interpreted as a universal or clinically ideal squat trajectory.

Perception and evidence availability. Path A is structurally unavailable in all QEVD short-clip diagnosis records. The relational-reasoning findings therefore apply to Path B combined with biomechanical and reference evidence, rather than to learned Path-A/Path-B fusion. The final evidence representation is based on 2D pose, and camera viewpoint remains an uncontrolled confound for lateral measures such as stance width and knee–ankle offset [63, 64]. In addition, COCO-17 does not include a toe keypoint [89], so ankle position is used as the available proxy.

Statistical scope. The held-out TEST split contains 1,052 clips from only five independent participants. The principal paired-bootstrap analyses resample clips ($B = 1{,}000$); therefore, their confidence intervals quantify clip-level variation and should not be interpreted as evidence of broad participant-level generalisation.

Supplementary subject-aware analyses were conducted for the main evidence-design comparison. A two-level bootstrap resampled participants and then clips within participants ($B = 10{,}000$), and the seed-42 diagnosis-aligned advantage remained positive (95%

CI [+0.023, +0.067]). At the individual-participant level, all five TEST participants showed a positive effect (from +0.028 to +0.073; Table 4.20), although the corresponding two-sided sign test is not significant with only five participants ($p = 0.0625$). A six-fold subject-level cross-validation, in which each fold rebuilt the reference from its own TRAIN subjects using the original selection rule, produced positive mean differences in all six folds (mean $\Delta = +0.049$, SD 0.007; two-sided sign test $p = 0.031$). Averaged over folds, micro-F1 increased from 0.277 to 0.326. With a larger reference that included all `no_obvious_issue` clips of each fold's TRAIN subjects, the mean difference was +0.046. Because the folds share training participants, the sign test is approximate.

Table 4.20: Per-participant TEST micro-F1 for the protocol-matched Concat+MLP comparison (seed 42).

Participant	**Clips**	**Generic**	**Diag.**	**Δ**
0009	147	0.3002	0.3727	+0.0726
0012	168	0.3229	0.3523	+0.0294
0023	413	0.3232	0.3516	+0.0284
0028	133	0.2528	0.2973	+0.0445
0029	191	0.2671	0.3211	+0.0540
All	1,052	0.3018	0.3435	+0.0417

The primary model-development runs used training seed 42. Seven training seeds (42–48) were used to assess initialisation sensitivity for the main architecture-fixed Reasoner v2 evidence-design comparison. These seeds do not represent independent participant cohorts. Diagnosis-aligned evidence improved TEST micro-F1 at every seed, with a mean paired improvement of +0.038 and an across-seed standard deviation of 0.010. These analyses strengthen the evidence-design conclusion, but external validation with more participants

remains necessary.

The original generic Concat+MLP used fixed-epoch training without validation-based early stopping, whereas the diagnosis-aligned Concat+MLP used validation-based model selection. The evidence comparison was therefore repeated under identical training conditions and the same fixed decision threshold across seven seeds. The diagnosis-aligned advantage remained positive at every seed (mean $\Delta = +0.052$, SD$= 0.010$; Table 4.7).

The same original 8.8 Concat+MLP baseline was also used as a comparison point in several relational-reasoning experiments. The difference in training protocol should therefore also be considered when interpreting those comparisons. Because a less-optimised baseline is, if anything, an easier target for a competing model to exceed, this asymmetry is unlikely to have disadvantaged the relational-reasoning conclusions; however, this was not independently re-verified for those specific comparisons. The matched fine-ON versus fine-OFF comparison is unaffected by this baseline-training difference.

The diagnosis-aligned Reasoner v2 evaluation constituted the third distinct access to the held-out TEST split in the broader experimental sequence. Although the models and the two planned comparisons were fixed before this access, no formal correction was applied for repeated access to the TEST split or for multiple comparisons across the broader sequence [93]. Accordingly, this result is interpreted as corroborating evidence within the overall experimental pattern rather than as a standalone confirmatory finding independent of the preceding TEST evaluations.

Table 4.21 provides a chronological record of the substantive evaluations on the TEST

split across the experimental sequence.

Table 4.21: Chronological record of substantive access to the TEST split.

Date	**Evaluation**	**Fixed before TEST access**	**Interpretation**
Sept. 7	Four principal models: diagnostic performance, ablations, and grounding intervention	Full evaluation protocol frozen and mechanically preflight-verified	Initial predefined TEST evaluation
Sept. 8	Reasoner v2 with generic evidence vs. the 8.8 baseline	Architecture, training procedure, and VAL-based pass criterion	Second TEST access
Sept. 10	Diagnosis-aligned Reasoner v2 vs. diagnosis-aligned MLP and generic Reasoner v2	Both models trained and selected using VAL; two comparisons specified before TEST access	Third TEST access
Sept. 16	Seven-seed robustness analysis for two borderline comparisons	Original seed-42 results and conclusions already recorded before the follow-up	Post hoc robustness replication; no model-selection changes

Continued on next page

Table 4.21: Chronological record of substantive access to the TEST split (continued).

Date	**Evaluation**	**Fixed before TEST access**	**Interpretation**
Sept. 21	Multi-seed replication of diagnosis-aligned vs. generic Reasoner v2 (six additional seeds; seven seeds total, 42–48)	Original seed-42 comparison, models, evidence definitions, and evaluation protocol already fixed	Follow-up robustness analysis; positive across all seven seed reruns, which are not independent participant cohorts
Sept. 22	Protocol-matched generic vs. diagnosis-aligned Concat+MLP comparison across seven seeds	Architecture, training conditions, early-stopping rule, fixed decision threshold, data splits, and seeds matched between conditions	Follow-up robustness analysis; diagnosis-aligned advantage positive under matched training conditions across the seven seed reruns

Continued on next page

Table 4.21: Chronological record of substantive access to the TEST split (continued).

Date	**Evaluation**	**Fixed before TEST access**	**Interpretation**
Sept. 23	Contradictory-label sensitivity, Variant A: cleaned splits with the original 529-clip biomechanical reference	Cleaning rule, model comparison, seed-42 protocol, and normalisation procedure were frozen before execution	Post-hoc sensitivity analysis; not independently confirmatory
Sept. 23	Contradictory-label sensitivity, Variant B: cleaned splits with the rebuilt 196-clip biomechanical reference	Variant definition, reference-rebuild procedure, model comparison, and seed-42 protocol were frozen before execution	Post-hoc sensitivity analysis; not independently confirmatory
Sept. 23	Multi-seed robustness follow-up for contradictory-label sensitivity Variants A and B (seeds 43–48; seven seeds total)	Seed-42 sensitivity results and the original frozen escalation rule were already recorded; preprocessing, references, normalisation, and evaluation protocol were held fixed	Separate post-hoc robustness follow-up for consistency with the thesis-wide multi-seed evaluation standard

Continued on next page

Table 4.21: Chronological record of substantive access to the TEST split (continued).

Date	**Evaluation**	**Fixed before TEST access**	**Interpretation**
Oct. 17	Per-participant TEST analysis and two-level participant/clip bootstrap of the protocol-matched Concat+MLP comparison	Frozen seed-42 checkpoints and stored TEST predictions; no retraining	Post hoc participant-level robustness analysis; same five participants
Oct. 17	Six-fold subject cross-validation over all 30 participants, including the five TEST participants	Reference-selection rule, model recipe, and seeds 42–46	Post hoc robustness analysis; folds share training participants
Oct. 17	Corrected grounding re-evaluation of fine-OFF and the final diagnosis-aligned Concat+MLP	Corrected sign conventions fixed in code immediately before the run; thresholds and decision rule unchanged	Post hoc corrected evaluation; not preregistered
Oct. 17–30	Per-exercise breakdown and calibration of the final model from stored TEST predictions	Frozen predictions; no retraining	Descriptive post hoc analyses

Grounding evaluation. The diagnosis-grounding protocol evaluates only three labels under the predefined intervention rule. Although this provides stronger evidence than predictive accuracy alone, it does not establish causal biomechanical reasoning in a formal causal model [87]. The intervention directions are design assumptions that have not been validated by biomechanics experts. The language-grounding evaluation is limited to 40 held-out examples and relies on rule- and keyword-based metrics [100]. In addition, the preferred DPO responses are template-constructed rather than independently human-written, so part of the measured improvement may reflect adaptation to the template style as well as improved grounding.

Reproducibility. The project does not have a version-control repository. Artefact provenance therefore relies on file hashes, timestamps, split manifests, configuration files, and frozen model checkpoints rather than commit history [93]. The final deterministic reproducibility gate increases confidence that the frozen pipeline can be reconstructed consistently, but it does not provide the same development-history traceability as commit-based version control. In particular, no training log, seed, or held-out accuracy was kept for the Path-B classifier, so its training data were reconstructed from stored arrays (Section 4.1.1) and it cannot be retrained from the stored records.

Application scope. The thesis evaluates movement diagnosis, not medical diagnosis. Its primary objective is to evaluate evidence-representation methodology rather than to establish deployment-grade diagnostic accuracy; accordingly, the post-selection TEST micro-F1

of 0.3435 for the default diagnosis-aligned concat+MLP model should be interpreted as a comparative result within the evaluated experimental setting, not as a claim of practical diagnostic performance. All principal experiments are offline, and no claim of longitudinal or live end-user efficacy is made. The framework is designed to be adaptable to other exercises, but empirical generalisation beyond squat assessment remains future work.

4.6.3 Chapter Summary

The implementation and evaluation lead to four main conclusions. First, a richer statistical reference does not resolve the limitations of globally aggregated deviation; under the tested criterion, the simpler mean-curve reference remains preferable. Second, the full diagnosis-aligned evidence configuration produces the most consistent improvement across the evaluated comparisons in movement diagnosis. Third, explicit relational reasoning does not provide a statistically measurable advantage over corresponding non-relational fusion in the evaluated setting, and one relational configuration performs significantly worse. Fourth, diagnostic accuracy does not guarantee evidence-grounded behaviour. When the diagnosis is held fixed, LoRA-DPO improves the specified automatic diagnosis-to-language metrics, but this result does not establish that an LLM is necessary or superior to deterministic template-based feedback.

Overall, the results support the central empirical position of this thesis: within the evaluated movement-diagnosis setting, the full diagnosis-aligned evidence configuration produced more consistent predictive gains than the tested relational architectures. Because

this configuration also adds ten measurements, the results do not isolate alignment alone as the cause of the gain. Reliable grounding at the diagnosis layer remains an open problem.

5

Discussion and Synthesis

5.1 Chapter Overview

Chapter 4 established the empirical results of FAIR-Coach. This chapter interprets those findings in relation to the two research questions and the central problem of how heterogeneous biomechanical evidence should be represented and evaluated to support movement diagnosis and assess evidence grounding. The discussion focuses on why diagnosis-aligned evidence had a stronger effect than increased reasoning complexity, what the relational results imply within the evaluated setting, and why predictive performance must be distinguished from evidence grounding. It then connects these findings to the broader research trajectory and domain-shift motivation and derives the main methodological implications and design lessons. Detailed experimental results remain in Chapter 4 and are repeated here only when needed for interpretation.

5.2 Answering the Research Questions

5.2.1 Evidence Representation and Diagnostic Performance

RQ1 asks whether diagnosis-aligned biomechanical evidence improves movement-diagnosis performance relative to generic, globally aggregated evidence. The experiments support this conclusion, but the more important scientific question is why the change in representation mattered.

The generic representation primarily describes how much a repetition differs from a reference through whole-repetition classification, range-of-motion information, reference deviation, and per-joint summaries. This is useful for characterising overall movement difference, but the diagnosis task requires distinguishing error categories with different spatial and phase-dependent signatures [31, 58]. A shallow squat, excessive knee translation, trunk deviation, and stance-width error are therefore not simply different magnitudes of one generic deviation signal [31, 32, 74].

Diagnosis-aligned evidence addresses this mismatch by exposing measurements that correspond more directly to the distinctions required by the target labels, including deepest-frame joint angles, trunk behaviour, spatial knee–ankle relations, and stance or symmetry information [31, 32, 73]. The resulting improvement suggests that the full diagnosis-aligned configuration provided diagnosis-relevant information not adequately captured by the generic representation. Because this configuration also adds ten measurements, the

result does not establish whether the performance advantage comes from diagnosis alignment, additional evidence quantity, or their combination [8, 12].

The earlier reference experiments support the same interpretation. A range-based reference provided a richer description of acceptable movement than a single mean trajectory, yet it did not improve clean-versus-error separation. Increasing the statistical richness of the same generic deviation signal was therefore insufficient. Improvement was observed with the full diagnosis-aligned configuration, which reorganised evidence around diagnosis-relevant biomechanical distinctions and added ten measurements. The comparison does not attribute the gain to either of these changes alone.

A supplementary EC3D analysis points in the same direction: a global deviation score separated correct from incorrect repetitions only weakly above chance (AUC 0.602, 95% CI $[0.501, 0.698]$). Because this analysis uses a different dataset and reference, it supports the interpretation only indirectly.

Importantly, the diagnosis-aligned features were not individually strong discriminators; their univariate AUC values remained modest. Their benefit appeared when multiple diagnosis-relevant measurements were available together. The positive result should therefore not be interpreted as the discovery of a single decisive biomechanical feature, but as evidence that the organisation of multiple complementary measurements matters for multi-category diagnosis.

The performance advantage of the full diagnosis-aligned configuration was consistent across the matched comparisons. With Reasoner v2 held fixed, diagnosis-aligned evi-

dence improved TEST performance across seven initialisation seeds (mean $\Delta = +0.038$, SD= 0.010), and the protocol-matched Concat+MLP comparison showed the same direction across seven seeds (mean $\Delta = +0.052$, SD= 0.010). These seed reruns do not constitute independent participant cohorts. The advantage was also positive for all five TEST participants and in all six folds of a subject-level cross-validation, although five TEST participants remain a small basis for generalisation. A post-hoc cleaned-data sensitivity analysis also preserved the diagnosis-aligned advantage when contradictory source annotations were removed, both with the original training reference and with a rebuilt reference. This supports robustness to that specific label inconsistency without implying that QEVD label quality is fully resolved.

RQ1 therefore supports a specific, scoped conclusion: the full diagnosis-aligned evidence configuration provides the diagnosis model with biomechanical distinctions that are more useful for the evaluated movement-diagnosis task [9, 58]. Within the data regime, evidence definitions, and model families evaluated in this thesis, this configuration was more consequential than increasing reasoning complexity. Because it retains the generic evidence and adds ten additional measurements, however, the comparisons do not isolate alignment alone as the cause of the improvement. This does not establish a universal priority of representation over architecture, but it shows that additional model sophistication cannot compensate for evidence that does not adequately encode the distinctions required by the target diagnosis.

The absolute performance of the best diagnosis model remains modest (post-selection TEST micro-F1 0.3435), and the results do not support practical or deployment-level

diagnosis. Several features of the evaluated setting are consistent with this ceiling, although none was isolated experimentally. The source labels contain contradictions, with 8.9% of TRAIN clips labelled both `no_obvious_issue` and an error. Some labels are rare, for example 64 TRAIN positives for `insufficient`. Monocular 2D pose cannot recover out-of-plane motion. In addition, accepted clips mix basic squats with five squat variants, and per-exercise micro-F1 was much lower for the variants than for basic squats, the only exercise used to build the reference. Directions suggested by these observations include cleaner or expert-adjudicated labels, more participants per label, exercise-specific references or models, and multi-view or 3D evidence. These are hypotheses for future work rather than tested remedies.

5.2.2 The Limits of Relational Reasoning

RQ2 asked whether explicit relational reasoning over heterogeneous biomechanical evidence provides measurable value beyond corresponding non-relational fusion. The evaluated experiments did not demonstrate such a benefit. Importantly, this conclusion does not rest on a single unsuccessful architecture. Relational reasoning was tested through coarse and fine-grained message passing, relations-ON versus relations-OFF controls, and label-conditioned attention, under both generic and diagnosis-aligned evidence. Across these formulations, no relational mechanism produced a statistically confirmed advantage over the appropriate non-relational comparison on held-out TEST split, and one configuration performed significantly worse.

This breadth narrows the interpretation of the negative result. The evidence does not support the general claim that relational learning is ineffective [25–28]. Rather, within the evidence definitions, architectures, and data regime tested here, explicit relational structure did not provide measurable value beyond simpler fusion. This is therefore a boundary finding for the evaluated setting, not a universal conclusion about relational reasoning.

One plausible explanation is partial redundancy among the evidence groups. Reference–knee correlations reached $r = 0.736$, indicating that related geometric information was available through multiple evidence blocks. The architecture-dependent ablation results are consistent with the same possibility: the contribution of Path B, reference, and reliability evidence varied across model families rather than revealing one universally indispensable source. If several groups encode overlapping aspects of the same movement, an explicit relation mechanism may have limited additional structure to exploit [78, 79]. In addition, Path A was absent and several reliability and context fields varied little in the QEVD records, so fewer evidence groups carried movement-specific information than the nominal graph structure suggests. Seven-seed follow-ups of the two borderline fine-model comparisons were inconsistent in sign across seeds and did not change this conclusion.

This explanation remains tentative. Correlation between evidence groups does not establish that redundancy caused the relational models to fail, and no causal experiment in this thesis isolates that mechanism [87]. Redundancy should therefore be treated as an interpretation consistent with the observed correlations and ablations, rather than as an established cause.

A broader methodological lesson concerns validation discipline. Several relational models showed promising validation differences that did not remain statistically supported on a held-out TEST split. This repeated VAL-promising/TEST-null pattern illustrates the risk of interpreting small architecture gains from validation results alone, where model-selection noise or sampling variation may appear as improvement [86].

The predefined decision rule was therefore important: a relational model was credited with an improvement only when its paired bootstrap confidence interval excluded zero in the beneficial direction on held-out TEST split. Applying this rule prevented validation optimism from becoming a thesis claim and made held-out TEST evidence, rather than architectural complexity, the basis for interpretation [93]. Because the same TEST cohort was used repeatedly, these comparisons are supporting evidence within QEVD rather than independent confirmations.

The RQ2 conclusion also has an important scope boundary. Path A was structurally absent from the QEVD short-clip diagnosis records [15], so these experiments concern relational reasoning over Path B together with biomechanical and reference evidence. They do not evaluate learned Path-A/Path-B fusion and therefore do not establish whether relational modelling would help in a dataset where both evidence paths were jointly available.

The negative result is nevertheless scientifically informative because it rules out the tested relational mechanisms as the explanation for the main diagnostic improvement. Once diagnosis-aligned evidence was available, a more elaborate relational architecture did not produce a confirmed held-out gain over non-relational fusion. The diagnosis-aligned

Concat+MLP was retained as the default diagnosis model. This designation was made after TEST results for the frozen diagnosis-aligned models had been observed, so the TEST cohort informed the final choice. This decision does not treat relational reasoning as excluded a priori.

5.3 From Predictive Accuracy to Evidence-Grounding Evaluation

Improved diagnostic performance is necessary but not sufficient to establish evidence grounding. A model may predict the correct label while depending on evidence that is unstable or directionally inconsistent with the intended biomechanics. The intervention experiments directly exposed this distinction. Of the three tested labels, only `shallow` satisfied the full predefined grounding criterion. `knees_over_toes` changed by +0.0065 with a 95% confidence interval of [−0.001, 0.014] and did not meet the primary-evidence criterion. `back_not_straight` changed in the wrong direction (−0.0760, 95% CI [−0.092, −0.059]) and also failed the non-primary-evidence control condition. Because only one of the three labels passed, overall diagnosis grounding was not demonstrated. The interventions modify standardised extracted features rather than generating physically coherent poses or video trajectories. They therefore test controlled feature-space responses, not causal reliance on physically feasible movement changes.

When the label-matched diagnosis-aligned features were also perturbed, no tested label

passed: the `knees_over_toes` response increased but stayed below the threshold, and the `back_not_straight` response became more strongly negative. The generic fine-OFF model passed no label in either the original or the corrected evaluation.

The experiments do not identify why grounding failed, but several explanations are consistent with the design. First, the intervention directions are design assumptions; for example, in 2D a larger trunk-angle deviation does not necessarily mean a less straight back. Second, the interventions change features independently, whereas real movements change correlated features together; reference and knee evidence correlate up to $r = 0.736$, so the model may rely on proxies rather than the intended variables. Third, the labels are noisy and some are rare, for example 192 TRAIN positives for `back_not_straight`. Possible improvements include expert-validated sign conventions, perturbing physically coherent poses and recomputing all dependent features, and training objectives that encourage the intended evidence dependence. These remain untested.

This result places an important boundary around the positive evidence-representation finding. Diagnosis-aligned evidence improved predictive performance, but the intervention study did not demonstrate that the resulting predictions depend on the tested evidence in the intended biomechanical direction. Prediction and grounding therefore answer different questions, and improvement in the former does not establish the latter.

Calibration provides a further, distinct view of reliability. The principal diagnosis models were overconfident across much of the confidence range (ECE 0.30–0.36 for the generic-evidence models and 0.256, computed post hoc, for the final model), but calibration and

grounding measure different properties: calibration asks whether predicted probabilities correspond to empirical frequencies [71, 72], whereas grounding asks whether predictions respond appropriately to primary biomechanical evidence and non-primary-evidence control interventions [21, 33]. Neither should be treated as evidence for the other. Together, these results show why aggregate predictive performance alone is insufficient to characterise diagnosis reliability.

The language layer produced a different outcome. With the diagnosis fixed before generation, LoRA-DPO [36, 96] reduced unsupported content and improved evidence specificity and consistency with the predicted diagnosis under the defined held-out automatic evaluation protocol [39], which uses 40 examples, template-constructed preference pairs, and rule-based scoring. This task is structurally different from diagnosis grounding: the diagnosis model maps heterogeneous movement evidence to labels, whereas the language model verbalises an already-fixed `StructuredDiagnosis` under an explicit evidence-license constraint. The language-grounding result was therefore more favourable under the present protocol, but the experiments do not establish a general causal explanation for this difference because the two layers differ in representation, supervision, and evaluation. No deterministic template-based feedback baseline was evaluated. Therefore, these results do not establish that deploying an LLM is necessary, computationally justified, or superior to retrieving carefully designed evidence-licensed templates.

This separation is essential for interpreting the DPO result. More faithful language cannot repair an incorrect or ungrounded diagnosis [23]. The improvement should therefore be understood as increased communication fidelity after the diagnostic decision has been

fixed, not as evidence that the underlying biomechanical reasoning has been corrected.

Taken together, the grounding experiments refine the meaning of diagnosis quality in FAIR-Coach. Diagnosis-aligned evidence improves the ability to distinguish movement labels, while diagnosis-level grounding remains unresolved under the intervention criterion. Downstream language grounding can improve once the diagnosis is fixed, but that progress belongs to a different layer. The positive evidence-design result and the negative diagnosis-grounding result are therefore complementary rather than contradictory: one concerns predictive discrimination, while the other concerns whether predictions depend on evidence in the intended way.

5.4 Positioning FAIR-Coach Within Its Research Trajectory

5.4.1 From Interaction to Structured Diagnosis

The research trajectory leading to FAIR-Coach began with questions of interactive capability rather than evidence-grounded diagnosis. Earlier work established multimodal multi-agent coaching, real-time webcam-based interaction, and interoperable communication through agent and tool interfaces [16–18]. These systems demonstrated that an intelligent coach could perceive, interact, and coordinate components, but they did not address the scientific question that became central to this thesis: what biomechanical

evidence supports a specific movement diagnosis?

FAIR-Coach therefore represents a shift in the level of inquiry rather than simply an accumulation of system features. The thesis moves from external coaching capability toward internal diagnostic accountability. `RepEvidence` makes the evidence between perception and diagnosis explicit, while `StructuredDiagnosis` separates the diagnostic decision from its linguistic expression. These boundaries make it possible to evaluate evidence representation, diagnosis behaviour, diagnosis grounding, and language grounding as distinct scientific problems.

This framing also defines the role of the agent-oriented infrastructure. LangGraph, A2A, MCP, pose estimation, logging, and interface components enable the end-to-end workflow [49, 50], but their presence is not treated as evidence that movement diagnosis is better grounded. The final repetition-level execution path is largely structured rather than autonomously deliberative, and the thesis does not evaluate FAIR-Coach as an autonomous multi-agent decision-making system [51]. Its scientific contribution instead lies in making the evidence-to-diagnosis pathway explicit and experimentally testable.

Viewed across the broader trajectory, the key progression is therefore from external capability to internal accountability. Earlier systems established that coaching interaction could be implemented; FAIR-Coach asks whether the movement interpretation produced within that interaction can withstand scientific scrutiny. This shift, rather than the addition of further software components, defines the role of FAIR-Coach in the research programme.

5.4.2 Revisiting the Domain-Shift Motivation

The domain-shift benchmark provided an important motivation for this thesis by showing that strong within-domain skeleton-model performance does not guarantee reliable cross-domain behaviour [65, 70]. Across the tested transfer settings, 77.8% of out-of-domain experiments showed full or partial classifier collapse [20]. This result motivated scepticism toward treating the output of a single movement classifier as sufficient evidence of reliable movement understanding.

FAIR-Coach responds to this concern by making biomechanical evidence explicit and separating evidence representation from diagnosis. The completed experiments, however, refine the original motivation. Simply adding more evidence sources or introducing explicit relations between them did not guarantee better diagnosis. Evidence groups contributed differently across model families, and the tested relational mechanisms did not produce a confirmed advantage over simpler fusion. The more defensible conclusion is therefore that performance improved when the full diagnosis-aligned configuration combined generic evidence with additional diagnosis-relevant measurements. Because this configuration changes both the content and the number of measurements, the experiment does not separate their individual effects.

The final results add a second caution. Even when diagnosis-aligned evidence improved predictive performance, the intervention study did not demonstrate that the diagnosis model responded to that evidence in the intended biomechanical direction. Reliability therefore cannot be reduced either to classifier accuracy or to the presence of multiple

evidence sources. It requires separate evaluation of what information is represented, how it affects prediction, and whether the resulting evidence dependence is consistent with the intended interpretation.

The domain-shift motivation is thus retained but made more specific by the thesis results. The earlier study motivated moving beyond unquestioned dependence on a single classifier; FAIR-Coach shows that the alternative must itself be evaluated critically. Within the setting studied here, the strongest empirical gain came from aligning evidence with diagnosis-relevant biomechanics rather than from adding model complexity, while reliable diagnosis-level grounding remained unresolved. Later reliability-policy corrections in the integrated system are implementation-specific and are not treated as evidence for, or a replication of, the earlier domain-shift result.

5.5 Implications for Evidence-Grounded Movement Diagnosis

The completed experiments support several methodological implications for movement-diagnosis systems. These conclusions are specific to the evidence, models, datasets, and evaluation setting studied in this thesis rather than universal claims about human-movement analysis.

First, evidence design should be treated as a primary modelling decision rather than as a preprocessing detail. The most consistent improvement occurred with the full diagnosis-

aligned evidence configuration, not with additional reasoning complexity. Because this configuration retains generic evidence and adds ten measurements, the comparison does not isolate alignment alone as the source of the gain. It nevertheless shows that architecture search may be premature when the available evidence does not encode the distinctions required by the target diagnosis [86]. Complexity should therefore be justified by demonstrated value rather than assumed to compensate for a weak representation.

Second, the boundary between evidence, diagnosis, and communication should be made explicit. FAIR-Coach's structured intermediate representations make it possible to distinguish failures in evidence extraction, evidence-to-diagnosis mapping, and diagnosis-to-language generation. This separation is scientifically important because a fluent feedback sentence can otherwise obscure an incorrect or weakly grounded diagnosis [23, 35].

Third, predictive performance should be complemented by direct tests of evidence dependence [21, 80, 87]. The intervention experiments showed that improved classification does not establish that predictions respond to biomechanical evidence in the intended direction. Evidence-grounding evaluation therefore addresses a different question from accuracy and should be treated as a separate requirement when biomechanical justification is part of the system claim.

Fourth, controlled negative results and held-out evaluation are informative. The relational experiments narrowed the explanation of the positive diagnosis result by showing that the tested relational mechanisms did not provide a confirmed advantage over corresponding non-relational controls. The recurring VAL-promising/TEST-null pattern further

demonstrates why small architectural improvements should not be promoted to scientific claims without held-out confirmation [86, 93].

Fifth, grounding should be treated as layered. Diagnosis-level grounding remained unresolved under the intervention criterion, whereas diagnosis-to-language grounding improved once the diagnosis was fixed [24, 39]. Improved verbal faithfulness therefore does not make the diagnosis itself more trustworthy. The same principle should apply to new evidence sources: predictive gain alone is insufficient without testing whether they produce the intended evidence-dependent behaviour.

Taken together, these implications favour evaluating evidence representation, diagnosis behaviour, grounding, and communication as distinct layers before treating their integration as evidence of overall reliability. FAIR-Coach does not resolve all of these layers, but it makes their successes and failures separable enough to be evaluated rather than hidden within a single accuracy score or plausible coaching response.

5.6 Methodological Trade-offs and Design Lessons

Several design choices changed as the experiments progressed because greater complexity did not consistently produce greater scientific value. Across these decisions, the recurring principle was to retain a more elaborate formulation only when it improved the target behaviour under controlled evaluation or enabled a capability unavailable to the simpler alternative.

The biomechanical-reference experiments illustrate this distinction. A range-based reference was statistically richer than a single mean trajectory, but it did not improve held-out clean-versus-error separation. The later diagnosis-aligned results showed why this distinction matters: richer description of the same generic deviation is different from reorganising the evidence around diagnosis-relevant biomechanics. The methodological lesson is therefore not that simpler representations are inherently preferable, but that task relevance matters more than complexity for its own sake.

The relational experiments led to the same principle at the model level. Explicit relations were initially plausible because the evidence sources were heterogeneous, but the tested relational formulations did not establish a held-out advantage over corresponding non-relational controls [78]. The diagnosis-aligned Concat+MLP was therefore retained as the default configuration; this designation was made after TEST results had been observed. Relational structure was not excluded in advance. Model complexity was treated as a hypothesis to justify empirically rather than as a contribution by itself.

A similar caution governed exploratory extensions. EfficientNet [104] appearance features produced a large validation gain, but appearance features alone performed at least as well as the combined model, so the gain did not arise from the biomechanical evidence studied here. Its source could not be established, and the extension was not evaluated on TEST. Possible monocular 3D extensions [42, 106] were also left outside the final diagnosis study because their additional scientific value had not been isolated from potential redundancy with the existing 2D evidence. Neither decision establishes that appearance or 3D evidence is unhelpful; instead, they reflect a requirement that additional modalities enter

the thesis only with a clear hypothesis and evaluation plan.

Preserving these attribution boundaries is itself an important methodological outcome of the thesis.

5.7 Chapter Summary

This chapter interpreted the FAIR-Coach results around the distinction between evidence design, reasoning complexity, and grounding. The full diagnosis-aligned biomechanical evidence configuration improved movement diagnosis in the evaluated setting, whereas the tested relational mechanisms provided no statistically confirmed advantage over corresponding non-relational fusion. Because the diagnosis-aligned configuration retains generic evidence and adds ten measurements, this result does not isolate alignment alone as the cause of the improvement. Diagnosis-level grounding remained unresolved under the predefined intervention criterion, while diagnosis-to-language grounding improved under the defined automatic evaluation protocol with the diagnosis fixed. Together, these findings shift the emphasis from system complexity toward explicit, task-aligned evidence and careful evaluation of what predictive improvements do and do not establish. Chapter 6 summarises the contributions, limitations, and future research directions.

6

Conclusion

6.1 Chapter Overview

This chapter closes the thesis by summarising the research problem, the FAIR-Coach approach, the principal contributions, and the final answers to the research questions. It also states the final position on relational reasoning, identifies the main limitations that bound the thesis claims, and outlines the future work that follows from them. Rather than re-arguing the evidence presented in Chapters 4 and 5, the chapter states what the thesis has established, what remains unresolved, and how the final system should be understood.

6.2 Summary of the Research Problem and Approach

This thesis addressed a central problem in intelligent movement coaching: successful classification does not by itself establish which biomechanical evidence supports a diagnosis. Generic, globally aggregated movement-quality measures can describe deviation from an

acceptable reference without necessarily exposing the distinctions required to separate specific movement errors. The research problem was therefore framed around how biomechanical evidence should be represented, mapped to diagnosis, and evaluated for grounding.

FAIR-Coach addressed this problem by making the evidence-to-diagnosis pathway explicit and by separating evidence representation from reasoning architecture. The diagnosis experiments compared generic with diagnosis-aligned biomechanical evidence while independently testing non-relational and relational reasoning. Diagnosis grounding was then evaluated separately from downstream language grounding, with feedback generated only after the diagnostic decision had been fixed.

The resulting thesis position is specific. The full diagnosis-aligned evidence configuration produced the most consistent improvement in movement diagnosis across the evaluated comparisons, whereas the tested relational mechanisms did not provide a statistically confirmed advantage over corresponding non-relational controls. Because this configuration retains generic evidence and adds ten measurements, the result does not isolate alignment alone as the cause of the improvement. The grounding interventions showed that better prediction did not establish correct biomechanical evidence dependence, while preference optimisation improved diagnosis-to-language faithfulness under the defined automatic evaluation protocol with the diagnosis fixed. Reliability must therefore be evaluated as a layered problem rather than inferred from predictive performance alone.

6.3 Summary of Contributions

The thesis makes four principal contributions:

1. **C1 — Diagnosis-aligned organisation of established biomechanical evidence:** FAIR-Coach organises established biomechanical measurements into an evidence representation aligned with specific diagnosis categories; the individual measurements are not novel [40, 41]. In the pre-specified Reasoner v2 comparison, the full diagnosis-aligned configuration improved TEST micro-F1 at all seven seeds (mean $\Delta = +0.038$, SD$= 0.010$), and the protocol-matched Concat+MLP comparison showed the same direction (mean $\Delta = +0.052$, SD$= 0.010$). The advantage was also positive for all five TEST participants and in all six cross-validation folds. These analyses reuse one TEST cohort and are not independent replications. Because the configuration retains generic evidence and adds ten measurements, the results do not isolate alignment alone as the cause of the improvement.

2. **C2 — Controlled empirical test of relational reasoning:** Controlled evaluations of four relational configurations (Tables 4.11 and 4.12) showed that the tested relational mechanisms did not provide a statistically confirmed advantage over corresponding non-relational controls in the evaluated setting. The contribution is the controlled negative result and its scope, rather than a winning relational architecture.

3. **C3 — Evidence-grounding intervention protocol:** Intervention-based evaluation was adapted to movement diagnosis to test responses to primary biomechan-

ical evidence and non-primary-evidence control interventions in feature space. For the final diagnosis-aligned Concat+MLP, only `shallow` satisfied the full criterion when generic evidence was perturbed; `knees_over_toes` did not meet the primary-evidence requirement, and `back_not_straight` responded in the wrong direction. Neither of the two labels that could be tested passed when the diagnosis-aligned features were also perturbed. The intervention directions are design assumptions, not expert-validated mappings. Thus, predictive improvement did not establish overall diagnosis-level grounding.

4. **C4 — Diagnosis-to-language grounding through preference optimisation:** A lightweight LoRA-DPO procedure using programmatically constructed, evidence-licensed preference pairs improved diagnosis-to-language faithfulness under the defined automatic evaluation protocol (40 held-out examples, rule-based scoring), with the diagnosis fixed before language generation, but did not eliminate unsupported claims. This contribution is limited to the communication layer. Because no deterministic template-based feedback baseline was evaluated, it does not establish that an LLM is necessary or superior to retrieving carefully designed evidence-licensed templates.

Together, these contributions include both supported improvements and scientifically informative boundaries. C1 and C4 show bounded improvements, whereas C2 and C3 show respectively that the tested relational mechanisms did not explain the main diagnostic gain and that predictive improvement did not establish diagnosis-level grounding. This balance

reflects the thesis's emphasis on conclusions supported by controlled evidence.

6.4 Answers to the Research Questions

RQ1: Does diagnosis-aligned biomechanical evidence improve movement-diagnosis performance compared with generic, globally aggregated evidence? Yes, for the full diagnosis-aligned evidence configuration. It consistently improved diagnostic performance over generic evidence in the evaluated setting, with a positive advantage across the matched multi-seed comparisons. Because this configuration retains generic evidence and adds ten measurements, the result does not isolate alignment alone as the cause of the improvement.

RQ2: Does explicit relational reasoning provide measurable benefit beyond corresponding non-relational fusion of the same evidence? No statistically confirmed benefit was demonstrated in the evaluated setting. Across the tested relational formulations, explicit relational reasoning did not outperform the corresponding non-relational comparisons with statistical confirmation, and one tested configuration performed significantly worse.

6.5 Final Position on Relational Reasoning

Relational reasoning was a genuine thesis hypothesis evaluated through controlled, multi-architecture held-out comparisons (Tables 4.11 and 4.12). These evaluations did not demonstrate a statistically confirmed advantage over corresponding non-relational controls in the evidence and task setting studied here. This controlled negative result is scientifically informative because it rules out the tested reasoning mechanisms as the source of the principal diagnostic improvement.

The diagnosis-aligned Concat+MLP was retained as the default diagnosis model. This designation was made after TEST results had been observed, so the TEST cohort informed the final choice. The contribution therefore lies not in preserving relational complexity, but in testing the hypothesis rigorously enough to establish a clear empirical boundary.

This conclusion remains deliberately scoped. The thesis does not claim that relational reasoning is generally ineffective or that relation-aware models would be unnecessary under different datasets, evidence representations, or tasks. It concludes only that the relational mechanisms evaluated here did not provide a statistically confirmed benefit for the movement-diagnosis problem as instantiated in this thesis.

6.6 FAIR-Coach in Retrospect

The research trajectory leading to FAIR-Coach moved from enabling interaction and orchestration toward making the system's internal movement interpretation explicit and scientifically testable. Earlier work established multimodal coaching, real-time operation, and interoperable agent and tool communication [16–18]. FAIR-Coach shifted the central question from whether an intelligent coach could operate to whether the biomechanical evidence supporting its diagnosis could be represented explicitly and evaluated for grounding.

In retrospect, the scientific value of FAIR-Coach lies in separating evidence, diagnosis, diagnosis grounding, and feedback so that success or failure can be attributed to the appropriate stage. The supporting perception and orchestration infrastructure enables end-to-end operation, but the thesis claims attach to these experimentally testable boundaries rather than to the infrastructure itself. This transition from external capability to internal accountability defines FAIR-Coach's role within the broader research trajectory.

6.7 Limitations Carried Forward

Diagnosis-level grounding remains unresolved. Only `shallow` satisfied the predefined intervention criterion. `knees_over_toes` did not meet the primary-evidence condition, and `back_not_straight` responded in the wrong direction, and no tested label passed when the diagnosis-aligned features were also perturbed. The thesis therefore does not es-

tablish overall diagnosis-level grounding for the final model, despite the predictive benefit of the full diagnosis-aligned configuration. This remains the principal scientific boundary on claims of trustworthy biomechanical reasoning.

Path A is absent from the core diagnosis dataset. Path A is structurally unavailable in the QEVD short-clip records used for the core diagnosis experiments [15]. The RQ2 conclusion is therefore limited to relational reasoning over Path B together with biomechanical and reference evidence and does not establish how learned relational fusion would behave when genuine Path-A and Path-B evidence are jointly available.

Multi-seed coverage remains targeted. The main evidence-design result was consistent across seven initialisation seeds with both Reasoner v2 and protocol-matched Concat+MLP. These seed reruns are not independent participant cohorts. Two borderline relational comparisons were also evaluated across seven seeds without a stable advantage. Other model, ablation, and grounding experiments were not systematically retrained across multiple seeds.

Language-grounding evaluation remains narrowly bounded. The final language-grounding result is based on 40 held-out examples, rule- and keyword-based metrics, and template-constructed preference data rather than independently authored human preferences or expert judgement. The demonstrated improvement should therefore be interpreted within the defined automatic evaluation protocol rather than as a complete evaluation of

coaching-language quality. No deterministic template-based feedback baseline was evaluated; therefore, the result does not establish that an LLM is necessary or superior to carefully designed evidence-licensed templates.

Small and repeatedly used TEST cohort. The TEST split contains only five participants and was accessed repeatedly, including for the final backend designation. Seed reruns, subject-level analyses, and cross-validation are supporting analyses within QEVD, not external validation.

Data selection and label quality. All results apply to the 6,310 of 13,013 clips accepted by the validator, whose acceptance rates differed across labels and participants. The source labels contain contradictions, and the original reference included clips that also carried error labels.

Modest diagnostic performance and calibration. The final model reached a post-selection TEST micro-F1 of 0.3435 and remained poorly calibrated (post hoc ECE 0.256). The thesis therefore makes no claim of practical or deployment-level diagnostic accuracy.

Exercise scope and real-time use. The core evaluation is offline and squat-based. Performance was much lower for squat variants than for basic squats, and transfer to other exercises would require new evidence definitions, references, and labels. The live pipeline runs end to end, but its diagnostic accuracy in real-time use was not evaluated.

Evidence design not isolated. No ten-feature-only, leave-one-feature-out, or matched-dimension control was run, so the contribution of alignment itself, as distinct from added measurements, remains untested.

Path-B classifier provenance. The Path-B ST-GCN is frozen, but its training log, seed, and held-out accuracy were not preserved, and its training data (EC3D, REHAB24, and a custom YouTube-derived collection) were reconstructed from stored arrays. All diagnosis models use its outputs as generic evidence, so the reported diagnosis results cannot be fully reproduced without the frozen checkpoint.

6.8 Future Work

1. **Joint Path-A/Path-B datasets:** Future work should construct or adopt datasets that preserve genuine continuous Path-A context together with completed-repetition Path-B evidence and compatible per-repetition diagnosis labels, enabling learned fusion that could not be evaluated with the current QEVD short clips [5, 15].

2. **Broader multi-seed replication:** The main evidence-design comparison was consistent across seven seeds under two model architectures, and two borderline relational comparisons were also stress-tested. Multi-seed evaluation should be extended to the remaining ablation and grounding experiments [93].

3. **Controlled 3D evidence evaluation:** Given the redundancy observed among cur-

rent evidence groups, including a reference–knee correlation of up to $r = 0.736$, future 3D experiments [42, 106] should test whether new measurements contribute independent diagnostic information rather than assuming that additional geometric dimensions will improve performance.

4. **Stronger diagnosis-grounding methods:** Because only one of the three tested labels satisfied the predefined grounding criterion, future diagnosis models should be evaluated with physically coherent interventions and primary-evidence versus non-primary-evidence control tests, so that predictive gains are accompanied by direct tests of evidence dependence [80, 87].

5. **Independent human preference and feedback evaluation:** Future language-grounding studies should complement the current template-constructed preferences and automatic metrics with independently authored preference data and expert or human evaluation of feedback faithfulness, specificity, and grounding [97, 100].

6. **Live and longitudinal evaluation:** Because the principal scientific claims were evaluated offline and no longitudinal end-user study was conducted, future work should evaluate FAIR-Coach in sustained live use to examine diagnosis, reliability, and grounded feedback over repeated real interactions [47].

7. **Isolating the effect of alignment:** Ten-feature-only, leave-one-feature-out, and matched-dimension control experiments should determine whether the gain comes from diagnosis alignment itself or from additional measurements.

8. **Larger and external participant cohorts:** Evaluation on more participants, ideally an untouched external cohort, is needed to estimate performance for unseen people.

9. **Exercise-specific and morphology-aware references:** References built per exercise variant, or conditioned on body proportions, should be compared with the current population-mean basic-squat reference.

10. **Expert-validated grounding interventions:** Intervention directions should be validated by biomechanics experts and applied to physically coherent poses with all dependent features recomputed.

11. **Template baseline and multimodal feedback:** A deterministic template-based feedback baseline should be compared with the LLM, and visual or auditory feedback, such as pose overlays or highlighted movement phases, should be evaluated for user comprehension and movement correction.

6.9 Closing Statement

The central contribution of this thesis is not a single winning model, but the controlled empirical separation of evidence design, reasoning complexity, diagnosis grounding, and language grounding. In the offline squat-diagnosis setting studied here, the full diagnosis-aligned evidence configuration produced more consistent predictive gains than increasing the complexity of the tested reasoning mechanisms. Because this configuration retains

generic evidence and adds ten measurements, the result does not isolate alignment alone as the cause of the gain. Diagnosis-level grounding remained unresolved. The transferable lesson is methodological: for any exercise, the evidence should be defined around the target diagnoses, compared against generic evidence under matched controls, and tested for evidence dependence separately from accuracy, rather than assuming that greater model complexity or higher accuracy implies grounded reasoning. By treating both positive and negative findings as scientific outcomes, the thesis provides a clearer basis for future work on evidence-grounded movement coaching.

References

1. Ashutosh, K., Nagarajan, T., Pavlakos, G., Kitani, K. & Grauman, K. ExpertAF: Expert Actionable Feedback from Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 13582–13594 (June 2025).
2. Fang, Q., Tang, C., Tekin, B. & Yang, Y. Cigtime: Corrective instruction generation through inverse motion editing. *Advances in Neural Information Processing Systems* **37,** 102011–102035 (2024).
3. Yeh, W.-H. *et al.* Coachme: Decoding sport elements with a reference-based coaching instruction generation model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 29126–29151 (2025).
4. Grauman, K. *et al.* Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 19383–19400 (2024).
5. Li, Y.-M. *et al.* Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In *European conference on computer vision* 363–382 (2024).
6. Zhang, S. *et al.* Narrative action evaluation with prompt-guided multimodal interaction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 18430–18439 (2024).
7. Zhang, H., Leong, M. C., Li, L. & Lin, W. PeVL: Pose-Enhanced Vision-Language Model for Fine-Grained Human Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 18857–18867 (June 2024).
8. Xu, J., Yin, S., Zhao, G., Wang, Z. & Peng, Y. FineParser: A Fine-grained Spatio-temporal Action Parser for Human-centric Action Quality Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14628–14637 (June 2024).
9. Huang, Z., Liu, X. & Kong, Y. H-MoRe: Learning Human-centric Motion Representation for Action Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 22702–22713 (June 2025).
10. Zhou, K. *et al.* CoFInAl: Enhancing Action Quality Assessment with Coarse-to-Fine Instruction Alignment. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)* 1771–1779 (2024).
11. Xu, H. *et al.* Vision-Language Action Knowledge Learning for Semantic-Aware Action Quality Assessment. In *Computer Vision – ECCV 2024* **15100,** 423–440 (Springer, 2024).
12. Majeedi, A., Gajjala, V. R., GNVV, S. S. S. N. & Li, Y. RICA^2: Rubric-Informed, Calibrated Assessment of Actions. In *Computer Vision – ECCV 2024* **15121,** 143–161 (Springer, 2024).

13. Dong, L., Wang, W., Qiao, Y. & Sun, X. Lucidaction: A hierarchical and multi-model dataset for comprehensive action quality assessment. *Advances in neural information processing systems* **37,** 96468–96482 (2024).

14. Li, J. *et al.* Finerehab: A multi-modality and multi-task dataset for rehabilitation analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 3184–3193 (2024).

15. Panchal, S. *et al.* What to Say and When to Say it: Live Fitness Coaching as a Testbed for Situated Interaction. In *Advances in Neural Information Processing Systems* **37.** Datasets and Benchmarks Track, 75853–75882 (2024).

16. Vahdati, M. M., Gholizadeh HamlAbadi, K., Laamarti, F. & El Saddik, A. A Multi-Agent Digital Twin Framework for AI-Driven Fitness Coaching. In *Proceedings of the 2025 ACM International Conference on Interactive Media Experiences (IMX '25)* 380–385 (Association for Computing Machinery, New York, NY, USA, 2025).

17. Vahdati, M., Gholizadeh HamlAbadi, K., Laamarti, F., Kumar, D. & El Saddik, A. Digital Twin AI Fitness Coach: An Intelligent Multi-Agent System for Personalized Exercise Guidance. In *Proceedings of the 1st International Workshop on Multi-Sensorial Media and Applications (MSMA '25)* 18–26 (Association for Computing Machinery, New York, NY, USA, 2025).

18. Gholizadeh HamlAbadi, K., Vahdati, M., Laamarti, F. & El Saddik, A. Agent-to-Agent (A2A) Protocol Integrated Digital Twin System with AgentIQ for Multimodal AI Fitness Coaching and Personalized Well-Being. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)* 12483–12491 (Association for Computing Machinery, New York, NY, USA, 2025).

19. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3,** 160018 (2016).

20. Vahdati, M., Gholizadeh HamlAbadi, K., Etemad, A. & El Saddik, A. Skeleton-Based Exercise Quality Assessment Under Domain Shift: A Benchmark Study of Cross-Domain Failure Modes. In *Proceedings of the ACM Multimedia 2026 Workshop on Multimodal Computing for Health and Medicine (MCHM '26)* (Association for Computing Machinery, Rio de Janeiro, Brazil, 2026).

21. Steinmann, D., Stammer, W., Friedrich, F. & Kersting, K. Learning to Intervene on Concept Bottlenecks. In *Proceedings of the 41st International Conference on Machine Learning* **235,** 46556–46571 (PMLR, 2024).

22. Zhang, F. & Nanda, N. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations (ICLR)* (2024).

23. Huang, L. *et al.* A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM transactions on information systems* **43,** 1–55 (2025).

24. Niu, C. *et al.* Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 10862–10878 (2024).

25. Yan, S., Xiong, Y. & Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence* **32,** 7444–7452 (2018).

26. Chen, Y. *et al.* Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 13359–13368 (Oct. 2021).

27. Zhou, Y. *et al.* Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2049–2058 (2024).

28. Do, J. & Kim, M. SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition. In *Computer Vision – ECCV 2024* **15099,** 401–420 (Springer, 2024).

29. Rao, P., Asha, C. & Rao, P. R. Real-time posture correction of squat exercise: a deep learning approach for performance analysis and error correction. *IEEE Access* **13,** 39557–39571 (2025).

30. Wang, X. Spatiotemporal skeleton based learning for automatic exercise quality assessment and real time posture correction. *Scientific Reports* **16,** 21663 (2026).

31. Straub, R. K. & Powers, C. M. A biomechanical review of the squat exercise: implications for clinical practice. *International Journal of Sports Physical Therapy* **19,** 490–501 (2024).

32. Hopkins, J. E., Hopkins, C. E. & Chiu, L. Z. Greater squat stance width alters three-dimensional hip moment demands. *Journal of Biomechanics* **177,** 112391 (2024).

33. Marcinkevičs, R., Laguna, S., Vandenhirtz, M. & Vogt, J. E. Beyond Concept Bottleneck Models: How to Make Black Boxes Intervenable? In *Advances in Neural Information Processing Systems* **37** (2024).

34. West, P., Quirk, C., Galley, M. & Choi, Y. Probing factually grounded content transfer with factual ablation. In *Findings of the Association for Computational Linguistics: ACL 2022* 3732–3746 (2022).

35. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630,** 625–630 (2024).

36. Rafailov, R. *et al.* Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In *Advances in Neural Information Processing Systems* **36** (2023).

37. Meng, Y., Xia, M. & Chen, D. SimPO: Simple Preference Optimization with a Reference-Free Reward. In *Advances in Neural Information Processing Systems* **37** (2024).

38. Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D. & Kiela, D. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning* (2024).

39. Tian, K., Mitchell, E., Yao, H., Manning, C. & Finn, C. Fine-tuning language models for factuality. In *International Conference on Learning Representations* **2024,** 44851–44866 (2024).

40. Tang, J., Abedi, A., Colella, T. J. & Khan, S. S. Rehabilitation exercise quality assessment and feedback generation using large language models with prompt engineering. In *International Joint Conference on Artificial Intelligence* 60–75 (2025).

41. Liao, Y., Vakanski, A., Xian, M., Paul, D. & Baker, R. A review of computational approaches for evaluation of rehabilitation exercises. *Computers in biology and medicine* **119,** 103687 (2020).

42. Khirodkar, R. *et al.* Sapiens: Foundation for human vision models. In *European Conference on Computer Vision* 206–228 (2024).

43. Feng, Y. *et al.* Chatpose: Chatting about 3d human pose. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2093–2103 (2024).

44. Fang, Q., Tang, C., Tekin, B., Ma, S. & Yang, Y. HuMoCon: Concept Discovery for Human Motion Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 7179–7190 (June 2025).

45. Ashutosh, K. & Grauman, K. Learning skill-attributes for transferable assessment in video. *Advances in Neural Information Processing Systems* **38,** 160403–160430 (2025).

46. Pan, Y., Zhang, C. & Bertasius, G. Basket: A large-scale video dataset for fine-grained skill estimation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 28952–28962 (2025).

47. Khasentino, J. *et al.* A personal health large language model for sleep and fitness coaching. *Nature Medicine* **31,** 3394–3403 (2025).

48. Vahdati, M. M., Laamarti, F. & El Saddik, A. Meta-Review of Wearable Devices for Healthcare in the Metaverse. *ACM Transactions on Multimedia Computing, Communications, and Applications* **21** (July 2025).

49. Hong, S. *et al.* MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations* **2024,** 23247–23275 (2024).

50. Xi, Z. *et al.* The Rise and Potential of Large Language Model Based Agents: A Survey. *Science China Information Sciences* **68,** 121101 (2025).

51. Hu, S., Lu, C. & Clune, J. Automated design of agentic systems. In *International Conference on Learning Representations* **2025,** 21344–21377 (2025).

52. Černek, A., Sedmidubsky, J. & Budikova, P. REHAB24-6: Physical therapy dataset for analyzing pose estimation methods. In *International Conference on Similarity Search and Applications* 18–33 (2024).

53. Pirsiavash, H., Vondrick, C. & Torralba, A. Assessing the quality of actions. In *European conference on computer vision* 556–571 (2014).

54. Parmar, P. & Morris, B. T. Learning to score olympic events. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* 76–84 (2017).

55. Parmar, P. & Morris, B. T. What and how well you performed? a multitask learning approach to action quality assessment. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 304–313 (2019).

56. Tang, Y. *et al.* Uncertainty-aware score distribution learning for action quality assessment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 9836–9845 (2020).

57. Liu, J. *et al.* Vision-based human action quality assessment: A systematic review. *Expert Systems with applications* **263,** 125642 (2025).

58. Xu, J., Yin, S. & Peng, Y. Human-centric fine-grained action quality assessment. *IEEE transactions on pattern analysis and machine intelligence* **47,** 6242–6255 (2025).

59. Xu, J., Rao, Y., Zhou, J. & Lu, J. Procedure-aware action quality assessment: Datasets and performance evaluation. *International journal of computer vision* **132,** 6069–6090 (2024).

60. Warner, M. B. *et al.* A systematic review of the discriminating biomechanical parameters during the single leg squat. *Physical Therapy in Sport* **36,** 78–91 (2019).

61. Lally, E. M., Ericksen, H. & Earl-Boehm, J. Measurement properties of clinically accessible movement assessment tools for analyzing single-leg squats and step-downs: a systematic review. *Journal of Sport Rehabilitation* **31,** 476–489 (2022).

62. Gomes, D. A. *et al.* Are visual assessments of the single-leg squat valid to be used in clinical practice? A systematic review of measurement properties based on the COSMIN guideline. *Physical Therapy in Sport* **63,** 118–125 (2023).

63. Yoma, M., Llurda-Almuzara, L., Herrington, L. & Jones, R. Reliability and validity of lower extremity and trunk kinematics measured with markerless motion capture during sports-related and functional tasks: A systematic review. *Journal of Sports Sciences* **43,** 1703–1730 (2025).

64. Lima, Y. L., Collings, T., Hall, M., Bourne, M. N. & Diamond, L. E. Validity and reliability of trunk and lower-limb kinematics during squatting, hopping, jumping and side-stepping using OpenCap markerless motion capture application. *Journal of sports sciences* **42,** 1847–1858 (2024).

65. Abdelfattah, M., Hassan, M. & Alahi, A. MaskCLR: Attention-Guided Contrastive Learning for Robust Action Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 18678–18687 (June 2024).

66. Xiong, B., Yang, X., Song, Y., Wang, Y. & Xu, C. Modality-Collaborative Test-Time Adaptation for Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 26732–26741 (June 2024).

67. Xu, H. *et al.* MCMoE: Completing Missing Modalities with Mixture of Experts for Incomplete Multimodal Action Quality Assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence* **40,** 11241–11249 (2026).

68. Zeng, L.-A. & Zheng, W.-S. Multimodal action quality assessment. *IEEE Transactions on Image Processing* **33,** 1600–1613 (2024).

69. Xu, H. *et al.* Language-guided audio-visual learning for long-term sports assessment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 23967–23977 (2025).

70. Zhou, K., Shum, H. P., Li, F. W., Zhang, X. & Liang, X. Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. *IEEE transactions on image processing* **34,** 3718–3732 (2025).

71. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning* 1321–1330 (2017).

72. Minderer, M. *et al.* Revisiting the calibration of modern neural networks. *Advances in neural information processing systems* **34,** 15682–15694 (2021).

73. Johnson, C. C. *et al.* Hip kinematics in healthy adults during gait and squatting: sex differences and asymmetry revealed through dynamic biplane radiography. *Journal of Biomechanics* **143,** 111280 (2022).

74. Graber, K. A., Halverstadt, A. L., Gill, S. V., Kulkarni, V. S. & Lewis, C. L. The effect of trunk and shank position on the hip-to-knee moment ratio in a bilateral squat. *Physical Therapy in Sport* **61,** 102–107 (2023).

75. Xu, J. *et al.* Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2949–2958 (2022).

76. Pan, J.-H., Gao, J. & Zheng, W.-S. Action assessment by joint relation graphs. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 6330–6339 (2019).

77. Yu, X., Rao, Y., Zhao, W., Lu, J. & Zhou, J. Group-aware contrastive regression for action quality assessment. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 7899–7908 (2021).

78. Bechler-Speicher, M., Amos, I., Gilad-Bachrach, R. & Globerson, A. Graph Neural Networks Use Graphs When They Shouldn't. In *Proceedings of the 41st International Conference on Machine Learning (ICML)* **235,** 3284–3304 (2024).

79. Bechler-Speicher, M., Globerson, A. & Gilad-Bachrach, R. Tree-g: Decision trees contesting graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* **38,** 11032–11042 (2024).

80. Vandenhirtz, M., Laguna, S., Marcinkevičs, R. & Vogt, J. E. Stochastic concept bottleneck models. *Advances in Neural Information Processing Systems* **37,** 51787–51810 (2024).

81. Koh, P. W. *et al.* Concept bottleneck models. In *International conference on machine learning* 5338–5348 (2020).

82. Dhuliawala, S. *et al.* Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024* 3563–3578 (2024).

83. Luo, R., Xue, Z. S., Dimakis, A. & Grauman, K. When thinking drifts: Evidential grounding for robust video reasoning. *Advances in Neural Information Processing Systems* **38,** 83696–83727 (2025).

84. Hong, J., Lee, N. & Thorne, J. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* 11170–11189 (2024).

85. Xu, H. *et al.* Quality-Guided Vision-Language Learning for Long-Term Action Quality Assessment. *IEEE Transactions on Multimedia* **27,** 7326–7339 (2025).

86. Bechler-Speicher, M. *et al.* Position: Graph Learning Will Lose Relevance Due to Poor Benchmarks. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)* **267,** 81067–81089 (2025).

87. Geiger, A. *et al.* Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research* **26,** 1–64 (2025).

88. Jiang, T. *et al.* RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv preprint arXiv:2303.07399* (2023).

89. Lin, T.-Y. *et al.* Microsoft coco: Common objects in context. In *European conference on computer vision* 740–755 (2014).

90. Wen, B. *et al.* Know Your Limits: A Survey of Abstention in Large Language Models. *Transactions of the Association for Computational Linguistics* **13,** 529–556 (2025).

91. Le Morvan, M. & Varoquaux, G. Imputation for Prediction: Beware of Diminishing Returns. In *International Conference on Learning Representations (ICLR)* (2025).

92. Zhao, Z. *et al.* 3d pose based feedback for physical exercises. In *Asian conference on computer vision* 189–205 (2022).

93. Kapoor, S. *et al.* REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances* **10,** eadk3452 (2024).

94. Xu, W., Jiang, H. & Liang, X. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia* 438–446 (2024).

95. Makelov, A., Lange, G., Geiger, A. & Nanda, N. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations* **2024,** 26486–26515 (2024).

96. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)* (2022).

97. Yu, T. *et al.* Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 13807–13816 (2024).

98. Lee, H. *et al.* RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)* **235,** 26874–26901 (2024).

99. Yuan, W. *et al.* Self-Rewarding Language Models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)* **235,** 57905–57923 (2024).

100. Wang, P. *et al.* Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* 9440–9450 (2024).

101. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35,** 24824–24837 (2022).

102. Yao, S. *et al.* Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* **36,** 11809–11822 (2023).

103. Besta, M. *et al.* Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence* **38,** 17682–17690 (2024).

104. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* 6105–6114 (2019).

105. Qwen Team. *Qwen3.8-27B* Hugging Face model repository. Model identifier: Qwen/Qwen3.8-27B. 2026.

106. Shin, S., Kim, J., Halilaj, E. & Black, M. J. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2070–2080 (2024).

APPENDICES

A

RepEvidence Schema and Evidence Groups

A.1 RepEvidence Schema

One completed, validated repetition corresponds to one `RepEvidence` record. The frozen generic input contains 27 raw dimensions together with an explicit missing-evidence mask.

Z-score normalisation is fitted on TRAIN only and reused unchanged for validation and TEST.

A.2 Evidence-Group Definitions

Table A.1: Final `RepEvidence` input schema and TRAIN-set missingness.

Block	**Fields**	**Missingness (TRAIN)**
Path A	3 window counts	100% missing (structurally absent in QEVD short clips)
Path B	`prob_correct`, `prob_incorrect`, `classification_confidence`, `posture_score_norm`, `is_incorrect`	0% missing
Biomechanical	ROM, `deviation_score`, 9 per-joint deviations	0% missing
Reliability	`low_confidence`, `confidence_score`, 4-way reliability one-hot	`confidence_score` and one-hot: 100% missing; `low_confidence`: 0% missing
Context	Δt, `rep_count`	Δt: 100% missing; `rep_count`: 0% missing but constant (always 1)

Table A.2: Evidence-group definitions used in the relational-reasoning experiments.

Grouping	**Evidence groups**
Coarse 5-group split	`path_a`, `path_b`, `biomechanical`, `reliability`, `context`
Fine 8-group split	`path_a`, `path_b`, `reference`, `knee`, `hip`, `trunk`, `reliability`, `context`

B

QEVD Subject Split

B.1 Subject-Level Split

The frozen QEVD subject split is defined in `subject_splits.json`, version 2.0, using seed 42. Exact subject assignments are provided here for reproducibility; aggregate split and validation statistics are reported in Chapter 4.

Table B.1: QEVD subject-level TRAIN, validation, and TEST assignments.

Split	**Subject IDs**
TRAIN	0000, 0001, 0002, 0003, 0004, 0005, 0007, 0008, 0010, 0011, 0013, 0015, 0016, 0017, 0018, 0019, 0020, 0022, 0025, 0026, 0027
Validation	0006, 0014, 0021, 0024
TEST	0009, 0012, 0023, 0028, 0029

C

Experimental Configuration

C.1 Diagnosis-Model Configuration

Unless otherwise stated, learned diagnosis models use `BCEWithLogitsLoss` with TRAIN-derived per-label class weights, Adam optimisation, learning rate 1×10^{-3}, and batch size 64. Table C.1 summarises the settings that differ across the principal models.

Table C.1: Principal diagnosis-model configurations.

Model	**Input / structure**	**Training and selection**
Generic Concat+MLP	54-d evidence-and-mask input; hidden dimension 32	60 fixed epochs; no validation-based early stopping; decision threshold 0.5
Diagnosis-aligned Concat+MLP	74-d evidence-and-mask input; hidden dimension 32	Maximum 80 epochs; patience 12; validation micro-F1 at 0.5; validation threshold grid $\{0.4, 0.5, 0.6\}$, selecting 0.5
Coarse relational	Five evidence groups with message passing	Maximum 80 epochs; patience 12; learning rate selected from $\{10^{-3}, 3\times 10^{-4}\}$; final threshold 0.5
Fine relational ON/OFF	Eight evidence groups, with message passing enabled or disabled	Maximum 80 epochs; patience 12; validation micro-F1 at 0.5; final threshold 0.5
Reasoner v2	Label-conditioned attention over evidence groups; evaluated with generic and diagnosis-aligned evidence	Maximum 80 epochs; patience 12; validation micro-F1 at 0.5; final threshold 0.5

The protocol-matched generic-versus-diagnosis-aligned Concat+MLP robustness comparison used seeds 42–48, identical early stopping and optimisation settings, fixed subject splits, and a fixed decision threshold of 0.5. Other multi-seed follow-ups likewise kept the architecture, evidence definition, and dataset split fixed within each comparison.

C.2 LoRA-DPO Configuration

Table C.2: Final LoRA-DPO configuration.

Parameter	**Specification**
Base model	`Qwen/Qwen3.8-27B`
LoRA rank / alpha / dropout	8 / 16 / 0.05
Target modules	`q`, `k`, `v`, and `o` projections
Learning rate	5×10^{-5}
DPO beta	0.1
Batch / gradient accumulation	1 / 4
Training steps	156
Preference pairs	420 template-constructed pairs, disjoint from the 40-example held-out evaluation set
Generation	Greedy decoding (`do_sample=False`)

C.3 Evaluation Constants

Table C.3: Principal constants of the frozen evaluation protocol.

Parameter	**Specification**
Final TEST diagnosis threshold	0.5
Validation threshold candidates where selection was used	$\{0.4, 0.5, 0.6\}$
Paired bootstrap resamples	1,000
Bootstrap seed	42
Two-level bootstrap resamples	10,000 (participants, then clips within participants)
Reliability thresholds	strong ≥ 0.75; partial ≥ 0.45; conflicting ≥ 0.20
Clip validator	min. knee angle $\leq 115°$; knee ROM $\geq 40°$; keypoint confidence ≥ 0.3; ≥ 10 reliable frames covering $\geq 50\%$ of the clip
Grounding intervention	0.5 SD, clamped to the TRAIN+VAL range; 30 cases per label (seed 42)
Grounding thresholds	relevant $\Delta p \geq 0.05$; non-primary $\lvert\Delta p\rvert < 0.05$; flip rate ≤ 0.10 at threshold 0.5; at least 2 of 3 labels must pass

C.4 Frozen Checkpoint Inventory

Table C.4: Final frozen checkpoint inventory.

Model	**Checkpoint file**	**SHA-256 prefix**
8.8 Concat+MLP	`phase8_8_concat_mlp_frozen.pt`	`34377d32...`
Fine 8-node, relations OFF	`fine8_relations_off_frozen.pt`	`4aa53db4...`
Fine 8-node, relations ON	`fine8_relations_on_frozen.pt`	`3b161959...`
Reasoner v2, generic evidence	`reasoner_v2_frozen.pt`	`63254200...`
Diagnosis-aligned Concat+MLP	`diag_concat_mlp_27plus10_frozen.pt`	`16e2ab06...`
Diagnosis-aligned Reasoner v2	`reasoner_v2_diag_spatial_frozen.pt`	`b57acae0...`
Locked coarse 5-node relational	`relational_reasoner_locked.pt`	`30c6b08d...`
Phase 12c DPO LoRA adapter	`phase12c_dpo_lora/adapter_model.safetensors`	`7ca01b0f...`

D

Additional Results

The principal evidence-ablation results and confidence intervals are reported in Chapter 4. This appendix reports per-label TEST scores that provide additional detail beyond the aggregate results. The final diagnosis-aligned Concat+MLP has high recall but low precision for every label (micro-averaged precision 0.232 and recall 0.666), indicating that it predicts many labels per clip at threshold 0.5.

Table D.1: Per-label F1 scores on the TEST split (threshold 0.5) for fine-OFF, fine-ON, the 8.8 Concat+MLP baseline, and the final diagnosis-aligned Concat+MLP (Diag. MLP).

Label	**fine-OFF**	**fine-ON**	**8.8**	**Diag. MLP**
`no_obvious_issue`	0.442	0.439	0.444	0.481
`shallow`	0.205	0.203	0.223	0.258
`knees_over_toes`	0.292	0.319	0.322	0.353
`back_not_straight`	0.194	0.179	0.172	0.172
`narrow`	0.278	0.290	0.272	0.349
`wide`	0.263	0.263	0.254	0.416
`over_90_degrees`	0.301	0.295	0.291	0.339
`insufficient`	0.102	0.110	0.089	0.141

In the QEVD short-clip records, not all nominal groups carry observed, varying evidence (Table A.1): `path_a` is entirely missing, the `reliability` group observes only

`low_confidence`, and the `context` group contains a missing Δt and a constant `rep_count`. The relational comparisons therefore operate with fewer informative groups than the nominal five or eight. Table D.2 reports the trainable parameter counts of the frozen checkpoints.

Table D.2: Trainable parameters of the frozen diagnosis checkpoints.

Model	**Parameters**
Generic Concat+MLP (8.8)	2,024
Diagnosis-aligned Concat+MLP	2,664
Coarse 5-node relational	3,513
Fine 8-node, relations OFF	3,576
Fine 8-node, relations ON	4,377
Reasoner v2, generic evidence	3,432
Reasoner v2, diagnosis-aligned evidence	3,752

E

Feedback Preference Construction

The LoRA-DPO preference set was constructed from the fixed `StructuredDiagnosis` and its evidence-license map. Chosen responses were template-constructed from licensed evidence, while rejected responses represented one of three controlled failure modes.

Table E.1: Preference-pair construction used for diagnosis-to-language grounding.

Component	**Construction**
Chosen response	Template-constructed text licensed by the fixed diagnosis and supporting evidence.
`wrong-theme`	A coaching theme licensed for a different predicted label.
`generic-praise`	Generic positive feedback without the corrective theme required by the predicted error.
`unsupported-claim`	A coaching theme not licensed by any predicted label for the example.

The held-out diagnosis-to-language results are reported in Chapter 4.