SHIFT-M3: Pre-fusion Alignment-based Consistency Screening for Multimodal ECG Record Integrity

Md Ashik Khan AASSHHIK98@GMAIL.COM
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India

Md Nahid Siddique MSIDD040@FIU.EDU
Knight Foundation School of Computing and Information Sciences
Florida International University
Miami, FL, USA

Abstract

Multimodal clinical AI typically assumes that the waveform, report, metadata, and downstream predictions attached to a record belong to the same patient. In practice, linkage failures can silently assemble individually plausible but cross-patient components, creating a safety problem that standard predictive models are not designed to detect. We study this problem as *multimodal record integrity triage*: given an assembled record, should its modalities be trusted to belong together? We introduce **SHIFT-M3**, a lightweight text-based pre-fusion screen that measures alignment-based consistency between two separately produced ECG text views: an LLM-generated interpretation and a clinical report summary. On 784,680 MEETI ECG records, SHIFT-M3 achieves **97.6%** TPR@5% FPR for full text-view swaps (AUROC 0.996), **90.3%** for partial swaps (AUROC 0.974), and **97.7%** for label-matched hard negatives (AUROC 0.996) with only 573,569 parameters. Compared with same-dataset lexical baselines, the gains are largest on partial swaps and hard negatives, suggesting that the model is learning more than surface overlap. We also introduce the **CMST** (Conflict-type Multimodal Stress Test) evaluation taxonomy, a three-seed stability study, a loss ablation, a temporal-tolerance sweep, and a shared-token masking control. The main remaining failure mode is longitudinal ambiguity: at the default operating point, same-patient cross-visit pairs still produce 87.0% Type-II false positives.

1. Introduction

Modern clinical ML systems increasingly fuse waveforms, reports, images, and metadata to improve downstream prediction or triage (Huang et al., 2020; Kline et al., 2022; Zhang et al., 2020; Lyu et al., 2023; Garriga et al., 2023). In practice, clinical information systems remain vulnerable to interoperability and documentation failures that can associate one patient's data with another's through clerical mistakes, workflow breakdowns, or software errors (Hyvämäki et al., 2023; Riplinger et al., 2020; Grannis et al., 2019, 2022). Standard benchmarks rarely test this failure mode: they assume the observed multimodal pair is valid. Yet a record-level linkage error can propagate into decision support, quality dashboards, or retrospective research cohorts. This problem sits adjacent to EHR data-quality assessment and patient matching, but what remains less studied is *content-level integrity checking* once

a multimodal record has already been assembled (Weiskopf and Weng, 2013; Kahn et al., 2012, 2016; Brown et al., 2013; Lewis et al., 2023; Riplinger et al., 2020; Li et al., 2022).

Most multimodal representation learning methods optimize fusion, retrieval, or alignment under the assumption that observed pairs are correct inputs (Saporta et al., 2024; Liu et al., 2024, 2025; Wang et al., 2025, 2023b; Yan et al., 2023). This leaves three gaps for integrity detection. First, **post-fusion detectors** are structurally disadvantaged once inconsistent components have already been merged, because downstream predictors are usually trained to map assembled records to outcomes rather than to audit whether the assembly itself is valid. Second, **identity-centric matching** depends on strong biometric signal; recent work shows this can be useful for ECG images (Sangha et al., 2024), but in our ECG-text setting identity signal alone is too weak. Third, many clinical workflows expose **text-rich but image-poor** records, so approaches that require heavy image pipelines are operationally brittle.

We therefore ask a different question: should the information have been fused at all? Our hypothesis is that contrastively trained alignment between *separately produced views* of the same event is a stronger integrity signal than latent identity matching. This reframes the task from "better multimodal prediction" to "pre-fusion consistency estimation."

We test this idea in the MEETI dataset (Zhang et al., 2026), where each ECG record contains two complementary text views: (i) an `LLM_Interpretation`, a detailed automated analysis, and (ii) a short clinical `report`. For a correctly linked record, both texts describe the same ECG. For a swapped record, the interpretation describes Patient A while the report describes Patient B, creating a detectable content mismatch. SHIFT-M3 measures that mismatch directly with a lightweight pre-fusion model. The resulting score is not a diagnosis; it is a triage signal that asks whether the assembled record warrants integrity review.

Our contributions are as follows:

1. A **problem formulation** of multimodal record integrity triage as pre-fusion alignment-based consistency estimation rather than standard multimodal prediction.
2. A **lightweight method**, SHIFT-M3, that detects full and partial text-view swaps on 784K ECG records with only 573,569 parameters.
3. The **CMST taxonomy** (Conflict-type Multimodal Stress Test), which separates full text-view swaps, partial swaps, hard negatives, and longitudinal pairs.
4. An **evaluation package** comprising same-dataset lexical baselines, frozen pretrained encoder baselines (MiniLM and Bio_ClinicalBERT), three-seed stability analysis, a temporal sweep, a full operating-point frontier, and a shared-token masking control extended to all stressors.

Generalizable Insights about Machine Learning in the Context of Healthcare

A cross-patient linkage error is undetectable from either text view in isolation: each document is individually plausible, yet a downstream model processing the assembled pair would attribute findings from two different patients to one, potentially generating incorrect risk predictions or clinical alerts. This study provides three generalizable insights for machine learning in healthcare. First, multimodal clinical pipelines should treat record assembly

as an explicit source of risk rather than assuming that paired modalities are valid inputs. Second, lightweight pre-fusion consistency screens can detect many cross-component linkage failures before downstream fusion or prediction occurs, even when each individual component appears clinically plausible. Third, content-sensitive integrity screening creates a predictable tradeoff with longitudinal tolerance: same-patient clinical change can resemble cross-patient inconsistency unless identity-aware or temporal context is incorporated. These findings suggest that integrity screening should be evaluated as a distinct stage in multimodal healthcare pipelines, especially in settings where silent record assembly errors can propagate into decision support, quality measurement, or retrospective cohorts.

2. Related Work

Most multimodal clinical AI work is organized around a downstream prediction task: combining images, EHR variables, waveforms, and text to improve classification, risk stratification, or triage (Huang et al., 2020; Kline et al., 2022). Concrete EHR systems follow the same pattern by fusing structured variables with clinical notes for mortality, readmission, or crisis prediction (Zhang et al., 2020; Lyu et al., 2023; Garriga et al., 2023). Recent general multimodal representation-learning methods likewise treat observed pairs as trustworthy supervision for fusion, retrieval, or alignment (Saporta et al., 2024; Chen et al., 2020). These methods improve prediction, but they do not test whether the pair itself is valid.

AI-enabled ECG research has expanded from single-task classification toward broader representation learning and multimodal report-aware modeling (Attia et al., 2021). Recent ECG-language methods align ECG signals with reports for zero-shot classification, arbitrary-lead representation learning, or report generation (Liu et al., 2024; Wang et al., 2025; Liu et al., 2025; Wan et al., 2025). These systems show that ECGs and associated text contain rich shared clinical content, which directly motivates our paired-view formulation. However, they still presume that the interpretation and report presented during training correspond to the same event. Our goal is different: not to predict from paired modalities or generate text, but to test whether two separately produced descriptions should be trusted as belonging to the same record.

Adjacent medical vision-language work studies how to generate or verbalize reports from correctly paired clinical inputs. Recent papers emphasize fine-grained image-report alignment, low-label self-training, style control, and longitudinal context for report generation (Wang et al., 2023b,a; Yan et al., 2023; Dalla Serra et al., 2023). This literature treats cross-modal agreement as useful structure, but still assumes the pair is trustworthy. SHIFT-M3 targets the missing pre-fusion screening step. Clinical informatics has separately developed strong frameworks for data quality assessment and record linkage. Foundational work formalizes data-quality dimensions for EHR reuse and multisite research (Weiskopf and Weng, 2013; Kahn et al., 2012, 2016; Bian et al., 2020; Brown et al., 2013; Lewis et al., 2023). Parallel patient-matching studies evaluate deterministic, probabilistic, and machine learning approaches, including standardization-sensitive matching, adaptive Fellegi-Sunter variants, and manual audits of real-world linkage quality (Riplinger et al., 2020; Eggerth et al., 2019; Grannis et al., 2019; Li et al., 2022; Grannis et al., 2022; Gupta et al., 2024). Patient-safety studies further show that interoperability and documentation failures remain an active source of clinical risk (Hyvämäki et al., 2023). At the modality level, identity-centric contrastive

learning can exploit patient-specific biometric signal when the channel is strong enough, as shown for ECG images (Sangha et al., 2024).

These literatures do not directly address the post-assembly question studied here: once a multimodal record exists, can its internal content be screened for cross-patient inconsistency? SHIFT-M3 sits between multimodal representation learning and clinical data-quality assessment. It leverages redundancy between two views of the same event, but uses that redundancy for integrity screening rather than prediction. Its distinguishing contribution is to recast multimodal consistency as a *pre-fusion integrity screening* problem and to evaluate it under explicit record-level stress tests rather than standard downstream prediction metrics.

3. Method

3.1. Problem Formulation

Each record is represented as $r = (t_A, t_B)$, where t_A is an LLM-derived ECG interpretation and t_B is the clinical report summary. The task is to learn a score $s(t_A, t_B) \in [0, 1]$ such that higher values indicate likely cross-patient mismatch. Screening operates at the level of a single patient event: the question is whether the two text views assembled into one record describe the same clinical event, not whether a patient's identity is consistent across separate visits.

We evaluate four stressors with the **CMST** taxonomy shown in Figure 1:

1. **Type-I (Full Text-view Swap):** replace t_B with another patient's report, creating a complete mismatch between the two text views while leaving each view individually plausible.
2. **Type-II (Temporal):** pair records from the same patient but different visits.
3. **Type-III (Partial Swap):** replace half of the interpretation TF-IDF features in t_A with features from another patient's interpretation, creating a mixed record in which only part of the evidence is inconsistent. This operation is performed in TF-IDF feature space and does not correspond to readable mixed text. Unlike real partial corruptions, which typically consist of coherent but misattributed readable text (such as copied-forward documentation or split-record merges), this substitution operates at the level of abstract feature overlap. Type-III is therefore a representation-space stress test that probes the model's sensitivity to partial feature-level mismatch.
4. **Hard Negative:** pair with a different patient sharing the same weak abnormal/normal label, where that auxiliary label is derived by keyword matching over the report text rather than by expert adjudication.

Types I, III, and Hard Negative are failures to detect; Type II is a legitimate longitudinal pair that should *not* be flagged. All four stressors are constructed algorithmically from clean MEETI records. No documented real-world linkage errors are included in the evaluation, and the degree to which these synthetic stressors reflect actual EHR assembly failures remains an open empirical question.

3.2. SHIFT-M3 Architecture

SHIFT-M3 is a jointly trained dual-stream text encoder, analogous in design to BGE-M3 (Chen et al., 2025), with separate TF-IDF vocabularies for interpretations and reports (500 and 325 features, respectively). The vocabulary sizes are motivated by a coverage analysis on the training split: clinical reports (5–20 tokens per document) reach 99.9% TF-IDF coverage at 325 features, and LLM interpretations reach 75.1% coverage at 500 features. To verify sufficiency, we trained SHIFT-M3 with LLM vocabulary size $\in \{500, 1000, 1500\}$: Type-I AUROC was 0.996 in all cases and TPR@5% varied by less than 0.06 pp, confirming that 500 features captures the full discriminative vocabulary. An identity stream produces normalized embeddings z_A^{id}, z_B^{id} and defines a conflict score via cosine distance:

$$s(t_A, t_B) = \frac{1 - \cos(z_A^{id}, z_B^{id})}{2}. \tag{1}$$

SHIFT-M3 contains separate auxiliary task and identity streams. The conflict score is computed exclusively from the identity stream. Each MLP follows a Linear(d_{in}, 256) $\rightarrow$ ReLU $\rightarrow$ BatchNorm $\rightarrow$ Dropout(0.2) $\rightarrow$ Linear(256, d_{out}) design, where the identity stream maps to a 128-dimensional embedding (d_{out}=128). The evaluated model has exactly 573,569 trainable parameters.

3.3. Training Objective

Training combines an auxiliary classification loss, a contrastive loss, a temporal positive loss, and an explicit swap margin loss:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_{ctr}\mathcal{L}_{ctr} + \lambda_{temp}\mathcal{L}_{temp} + \mathcal{L}_{swap}. \tag{2}$$

For a batch of size B, let $d_{ij} = 1 - \cos(z_{A,i}^{id}, z_{B,j}^{id})$ and $[x]_+ = \max(x, 0)$. The contrastive loss uses this unscaled cosine distance, whereas reported conflict scores use the scaled score s in Eq. (1). The task loss is binary cross-entropy between the auxiliary task-head output and the weak abnormal/normal label. The contrastive term uses diagonal pairs as positives and the closest off-diagonal report in the batch as the mined negative:

$$\mathcal{L}_{ctr} = \frac{1}{B}\sum_i d_{ii}^2 + \frac{1}{B}\sum_i \left[m - \min_{j \neq i} d_{ij}\right]_+^2. \tag{3}$$

The temporal term treats same-subject, different-visit pairs as positives and penalizes high conflict scores, $\mathcal{L}_{temp} = K^{-1}\sum_{k=1}^{K} s(t_{A,i_k}, t_{B,j_k})$, where i_k and j_k share a subject identifier but correspond to different ECG records. If a batch contains fewer than two eligible temporal pairs, this term is set to zero. Finally, the swap term applies an in-batch cyclic report permutation π and pushes swapped pairs above the margin:

$$\mathcal{L}_{swap} = \frac{1}{B}\sum_i \left[m - s(t_{A,i}, t_{B,\pi(i)})\right]_+. \tag{4}$$

We set $\lambda_{ctr} = 1.0$, $\lambda_{temp} = 0.5$, and swap margin $m = 0.5$ as the default configuration. The contrastive term brings matched pairs together while pushing hard negatives apart; the temporal term reduces penalties on same-patient cross-visit pairs; and the swap term explicitly pushes synthetic swaps toward higher conflict scores.

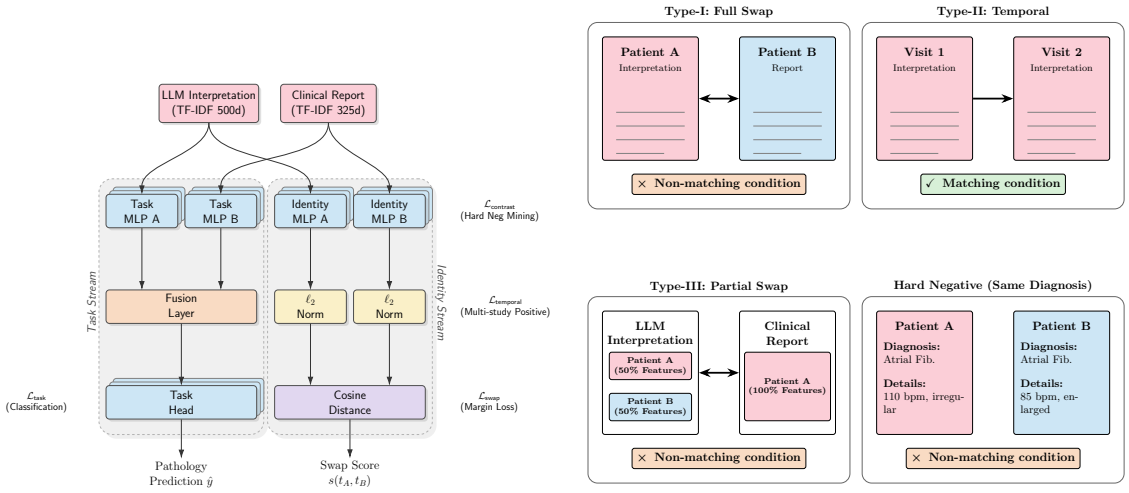

Figure 1: Method overview. On the left, SHIFT-M3 encodes the interpretation and report with separate modality-specific streams and produces a pre-fusion conflict score before any downstream multimodal model is invoked. On the right, the CMST taxonomy decomposes evaluation into full text-view swaps, partial swaps, hard negatives, and same-patient temporal controls, making it possible to separate cross-patient mismatch detection from longitudinal tolerance.

3.4. Dataset and Setup

MEETI is a publicly available dataset of 784,680 ECG records derived from MIMIC-IV-ECG (Zhang et al., 2026; Johnson et al., 2023); we use it as released without modification. The `LLM_Interpretation` field is generated by GPT-4o from extracted ECG parameters together with the associated clinical report (Zhang et al., 2026); the `report` field is the independently released clinical report text from the source ECG record. We therefore treat the two fields as separately produced textual views of the same ECG event, but not as independent expert-adjudicated diagnoses. Because both views share a common derivation from the clinical report, perfect cross-view independence cannot be assumed. For cross-patient swap stressors, Patient A's interpretation is paired with Patient B's independent record, breaking paired-view consistency; post-masking Type-I AUROC of 0.960 confirms that the alignment signal reflects clinical content beyond shared stylistic overlap. We acknowledge this derivation dependency as a limitation and note it in Section 6. We use a patient-level 80/10/10 split to prevent subject leakage (Table 1). Training uses Adam, learning rate 3×10^{-4}, batch size 128, 10 epochs, and 128-dimensional embeddings.

4. Results

4.1. Main Performance and Same-dataset Baselines

Table 3 compares SHIFT-M3 with same-dataset lexical baselines under a common 5% clean-pair false-positive budget. The lexical baselines are as follows: *word cosine* and *char*

Table 1: Patient-level split in the MEETI dataset pipeline. Splitting by subject prevents leakage of repeated visits across train, validation, and test.

Split	Records	Subjects	Multi-Visit
Train	629,386	128,477	82,414
Validation	76,742	16,059	10,342
Test	78,552	16,061	10,245
Total	**784,680**	**160,597**	**103,001**

cosine compute cosine similarity over TF-IDF vectors with no learned parameters; *LogReg* fits approximately ten hand-crafted pairwise similarity features using fewer than 50 weights. SHIFT-M3 reaches 97.6% Type-I TPR, 90.3% Type-III TPR, and 97.7% hard-negative TPR, outperforming all lexical baselines on every cross-patient stressor. The strongest baseline is a logistic regression over hand-crafted similarity features, but it still trails substantially on both full text-view and partial swaps. This difference matters because the partial-swap setting is the closest proxy to subtle record corruption: simple overlap can detect many obvious complete text-view swaps, yet it degrades sharply once only part of the evidence is inconsistent. For example, word TF-IDF cosine reaches Type-I AUROC 0.965but only Type-III AUROC 0.798, whereas SHIFT-M3 maintains Type-III AUROC 0.974. The temporal column shows a different ordering: the lexical baselines achieve substantially lower Type-II FPR because they are less sensitive to clinically meaningful semantic drift across visits. The empirical picture is therefore not that SHIFT-M3 is uniformly better, but that it trades temporal specificity for much stronger sensitivity to cross-patient corruption.

Table 2: Frozen pretrained encoder baselines versus SHIFT-M3 on the held-out test set. Both encoders score pairs via cosine distance between independently encoded embeddings without fine-tuning. Lower Type-II FPR in frozen models reflects weak conflict sensitivity across all pairs, not better longitudinal tolerance. Best value per column in bold; higher is better for AUROC, lower is better for Type-II FPR.

Model	Type-I AUROC	Type-III AUROC	Hard Neg AUROC	Type-II FPR@5%
MiniLM (frozen)	0.729	0.527	0.731	9.9%
Bio_ClinicalBERT (frozen)	0.555	0.500	0.556	**7.3%**
SHIFT-M3	**0.996**	**0.974**	**0.996**	87.0%

Comparison with frozen pretrained encoders: To assess whether supervised contrastive training adds value beyond off-the-shelf text representations, we evaluated two frozen contextual encoders as additional baselines: MiniLM (Wang et al., 2020) and Bio_ClinicalBERT (Alsentzer et al., 2019), each scoring pairs via cosine distance between independently encoded LLM interpretation and report embeddings without any fine-tuning. Table 2 reports results on the three cross-patient stressors. Both frozen encoders fall sub-

stantially below SHIFT-M3 across all stressors, with Bio_ClinicalBERT performing near chance on Type-III and Hard Negatives. Their lower Type-II FPR is not evidence of better longitudinal tolerance: it reflects uniformly weak conflict sensitivity across all pair types. These results demonstrate that task-specific contrastive training over TF-IDF features is a more effective strategy for record integrity screening than frozen clinical embedding cosine similarity.

Table 3: Held-out CMST performance. SHIFT-M3 improves substantially over same-dataset lexical baselines on full text-view swaps, partial swaps, and hard negatives. Type-II is a false-positive rate, so lower is better. Best value per column in bold.

(a) Recall at 5% clean-pair FPR

Model	T-I	T-III	HN
Word cosine	84.4	24.4	84.1
Char cosine	45.7	13.6	46.0
LogReg	89.1	32.5	89.2
SHIFT-M3	**97.6**	**90.3**	**97.7**

(b) Ranking quality and temporal error

Model	T-I AUC	T-III AUC	HN AUC	T-II FPR
Word cosine	0.965	0.798	0.965	50.8
Char cosine	0.904	0.729	0.903	**26.0**
LogReg	0.973	0.824	0.973	58.1
SHIFT-M3	**0.996**	**0.974**	**0.996**	87.0

Figures 2 and 3 complement Table 3 with a visual breakdown of the per-stressor margins and the score distributions that underlie them. The gap is largest on partial swaps, where lexical overlap degrades sharply but SHIFT-M3 remains consistent across seeds.

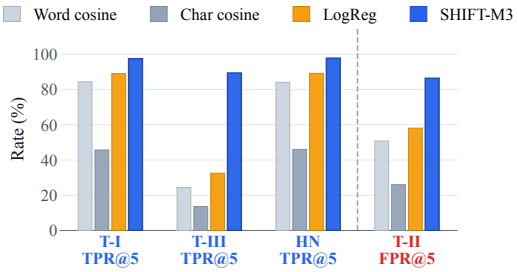

Figure 2: Performance at the 5% clean-pair FPR budget. SHIFT-M3 outperforms all lexical baselines on every cross-patient stressor, with the largest margin on partial swaps, whereas same-patient longitudinal pairs remain the dominant residual error source.

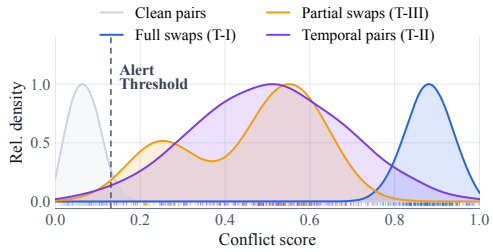

Figure 3: Conflict-score densities by pair type. Clean pairs and full swaps are well separated, whereas partial swaps and temporal pairs occupy the intermediate region near the alert threshold.

Figure 4 confirms these rankings across the full operating range. SHIFT-M3 achieves the highest AUROC on all three cross-patient stressors, with the widest margin on partial swaps (0.974 vs. 0.824 for the best lexical baseline).

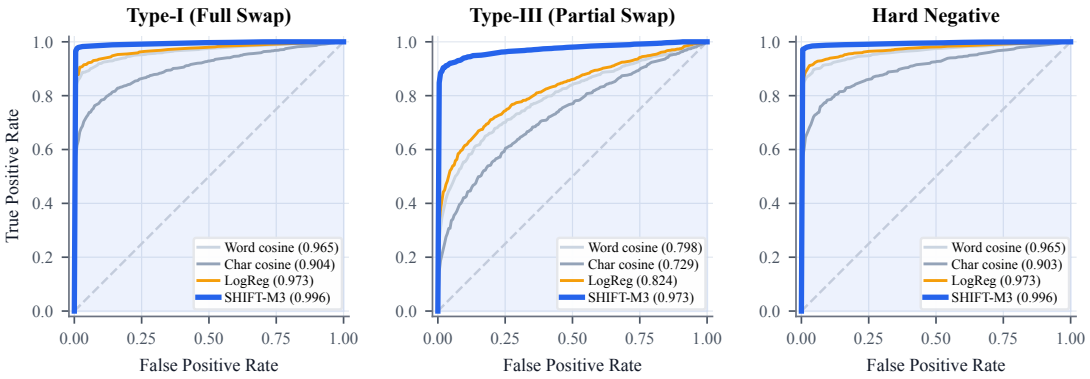

Figure 4: ROC curves across the three cross-patient stressors. SHIFT-M3 (blue) consistently hugs the upper-left corner, with the largest margin on partial swaps (center panel, AUROC 0.974 vs. 0.824). Each curve is annotated with its AUROC. The dashed diagonal represents random chance.

4.2. Stability Across Seeds

Table 4 reports all three completed seeds, and Table 5 summarizes the mean and standard deviation. Variance is minimal for the main swap-detection metrics: Type-I AUROC remains within 10^{-4} and Type-III AUROC within roughly 2×10^{-3} across runs. The larger spread in Type-II FPR is informative rather than alarming, because longitudinal pairs sit near the conceptual boundary between legitimate change and suspicious inconsistency.

Table 4: Per-seed held-out metrics from the three-run stability analysis. The main detection metrics vary little across seeds, indicating that the reported gains are not artifacts of a single run.

Seed	Type-I TPR@5	Type-III TPR@5	Hard Neg TPR@5	Type-II FPR@5	Type-I AUROC	Type-III AUROC	Hard Neg AUROC
42	97.55	89.40	97.83	86.42	0.9962	0.9725	0.9964
43	97.66	91.41	97.81	87.86	0.9963	0.9770	0.9961
44	97.47	90.02	97.57	86.56	0.9961	0.9737	0.9960

Table 5: Three-seed stability summary. Small standard deviations on Type-I, Type-III, and hard-negative metrics indicate stable optimization.

Metric	Mean	Std
Type-I TPR@5	97.56	0.09
Type-III TPR@5	90.28	1.03
Hard Neg TPR@5	97.73	0.15
Type-II FPR@5	86.95	0.79
Type-I AUROC	0.9962	0.0001
Type-III AUROC	0.9744	0.0023
Hard Neg AUROC	0.9962	0.0002

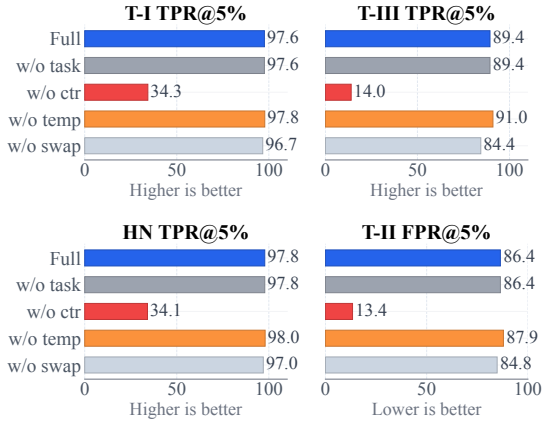

Figure 5: Loss ablation across mismatch conditions. The contrastive term provides the primary detection signal, whereas the swap and temporal terms specifically tune partial-swap sensitivity and longitudinal tolerance.

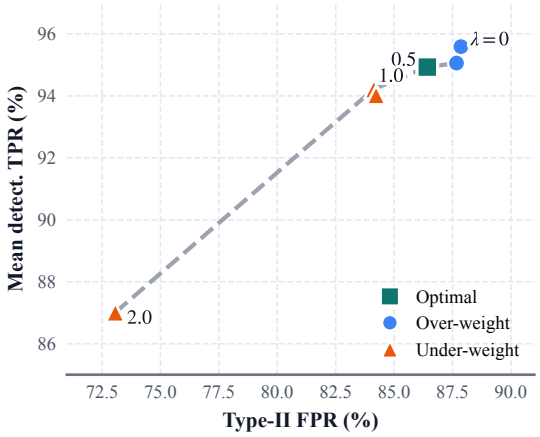

Figure 6: Temporal-weight sweep. The tradeoff frontier demonstrates that moderate temporal weighting ($\lambda_{temp} = 0.5$) optimally balances same-patient longitudinal tolerance against cross-patient detection.

4.3. Loss Ablation

Table 6 and Figure 5 report the full ablation results. Removing the contrastive term causes a pronounced collapse across all swap types, showing that aligned representation learning is the primary mechanism behind the detector. Removing the explicit swap loss also degrades performance, especially on partial swaps, which indicates that synthetic conflict supervision is not redundant with the contrastive objective alone. By contrast, removing the temporal term slightly improves raw detection but worsens Type-II false positives, confirming that its role is to trade some sensitivity for better longitudinal tolerance. Setting the auxiliary task-loss weight to zero produced identical conflict-detection metrics, as expected because the task and identity streams have disjoint parameters. Accordingly, the reported conflict scores are optimized through the contrastive, temporal, and swap objectives. The reported results retain the evaluated 573,569-parameter architecture.

4.4. Temporal Tolerance Sweep

The temporal-weight sweep exposes a real tradeoff rather than a trivial tuning oversight. Increasing the temporal weight improves same-patient tolerance, but too much temporal encouragement suppresses true swap sensitivity. Table 8 lists all completed sweep runs and Figure 6 plots the resulting tradeoff frontier; together they show a clear pattern: the highest raw detection appears when the temporal loss is removed, whereas aggressive temporal weighting lowers Type-II FPR at a substantial cost to Type-I and Type-III recall. The

Table 6: Loss ablation. The contrastive and swap losses provide most of the integrity signal, whereas the temporal term mainly moderates false positives on same-patient longitudinal pairs. The task-loss variant matches Full because the task and identity streams share no parameters.

Variant	Type-I TPR@5	Type-III TPR@5	Hard Neg TPR@5	Type-II FPR@5	Type-I AUROC	Type-III AUROC	Hard Neg AUROC
Full	97.55	89.40	97.83	86.42	0.9962	0.9725	0.9964
No task loss	97.55	89.40	97.83	86.42	0.9962	0.9725	0.9964
No contrastive loss	34.28	14.01	34.12	13.37	0.8445	0.6634	0.8443
No temporal loss	97.76	91.03	97.98	87.86	0.9964	0.9760	0.9967
No swap loss	96.73	84.40	97.04	84.79	0.9945	0.9575	0.9949

most defensible operating point in the current grid is therefore the compromise setting $\lambda_{temp} = 1.0$ with swap margin 0.5, which reduces Type-II FPR relative to the default without eroding detection as sharply as the aggressive setting. We recommend the default setting ($\lambda_{temp} = 0.5$) for first-linkage screening contexts, where cross-patient swap detection is the primary objective, and the compromise setting ($\lambda_{temp} = 1.0$) for high-revisit cardiology departments, where repeated same-patient ECGs are common and reducing alert burden is a practical priority.

Full operating-point frontier: To clarify whether any threshold setting can resolve longitudinal ambiguity, Table 7 reports the full operating-point frontier at clean-pair FPR budgets from 1% to 20%. Type-II FPR ranges from 77.2% at a 1% budget to 90.3% at 20%, while Type-I TPR remains above 96% throughout. No threshold choice resolves the longitudinal false-positive problem: tightening the threshold reduces both cross-patient recall and same-patient errors simultaneously, because genuine longitudinal change and cross-patient swaps produce similar text divergence patterns when only text content is available.

Table 7: Operating-point frontier across clean-pair FPR budgets. Type-III results use the text-splice variant. No threshold setting resolves longitudinal ambiguity: Type-II FPR remains elevated across the full range.

Clean-pair FPR budget	Type-I TPR@thr	Type-III TPR@thr	Hard Neg TPR@thr	Type-II FPR@thr
1%	96.6%	63.0%	96.8%	77.2%
2%	97.0%	69.0%	97.3%	80.2%
5%	97.7%	73.7%	97.8%	82.8%
10%	98.5%	76.6%	98.4%	85.4%
20%	99.7%	84.3%	99.5%	90.3%

Table 8: Completed temporal-weight sweep runs. Larger temporal weight lowers Type-II false positives, but the gain comes at the cost of weaker cross-patient detection.

Config	λ_{temp}	Margin	Type-I TPR@5	Type-III TPR@5	Hard Neg TPR@5	Type-II FPR@5	Type-I AUROC	Type-III AUROC
No temporal	0.0	0.5	97.76	91.03	97.98	87.86	0.9964	0.9760
Low temporal	0.25	0.5	97.66	90.64	97.89	87.67	0.9963	0.9747
Default	0.5	0.5	97.6	90.3	97.7	87.0	0.996	0.974
Compromise	1.0	0.5	97.21	89.13	97.32	84.13	0.995	0.970
Margin 0.6	1.0	0.6	97.39	88.62	97.58	84.23	0.9957	0.9693
Aggressive	2.0	0.5	92.79	77.44	93.06	73.07	0.984	0.936

4.5. Shortcut Control and Audit Export

The shared-token masking control tests whether the model succeeds mainly because the two text views share obvious tokens. Masking shared tokens causes a substantial but not catastrophic drop, which shows that lexical overlap contributes signal but does not fully explain the result. We extend the masking analysis beyond Type-I to cover the text-splice Type-III variant and Hard Negatives, addressing a key reviewer request. For the text-splice Type-III stressor (source first half + donor second half), performance drops substantially under masking (AUROC 0.927 → 0.796; TPR@5% 73.7% → 31.5%), confirming that partial-swap detection is more lexically dependent than full-swap detection. In contrast, Hard Negative detection remains strong post-masking (AUROC 0.996 → 0.967), indicating that the model learns beyond surface overlap for the label-matched case. The table caption distinguishes the text-splice Type-III variant used here from the feature-space substitution reported in Table 3, whose AUROC of 0.974 should not be directly compared. Table 9 summarizes the masking control metrics and the audit table sizes. To support later qualitative analysis, we also exported three audit tables of 250 examples each at the main 5% clean-pair threshold.

Table 9: Compact control summaries shown side by side. Left: shared-token masking control. Right: audit tables exported at the main 5% clean-pair threshold.

(a) Shared-token masking control

Setting	Thr.	T-I TPR@5	T-I AUC	Clean mean
Original	0.131	97.6	0.996	0.067
Masked shared tokens	0.452	83.3	0.960	0.212

(b) Audit table sizes at the 5% FPR threshold

Table	Rows
Clean FP	250
Type-I FN	250
Type-II FP	250

5. Qualitative Error Analysis

To understand the practical failure modes of our text-based alignment approach, we extracted the top-scoring false positives (clean records flagged as swaps) and lowest-scoring false negatives (swaps flagged as clean) at the 5% FPR threshold. Table 10 illustrates two representative edge cases.

In the False Positive case (Table 10a), the LLM generates a granular, step-by-step reasoning trace that notes minor deviations (*bradycardia*, *fluctuations*) before ultimately concluding the ECG is normal. The clinical report, however, skips the intermediate noise and directly asserts a *Normal ECG*. The model penalizes this difference in granularity, incorrectly flagging the clean pair as a content mismatch.

Conversely, the False Negative case (Table 10b) highlights a fundamental vulnerability of any strictly text-based sentinel. Here, the report from Case B was paired with the interpretation from Case A. Because both cases had similar findings (*Sinus rhythm with PVCs*), the two text views were nearly identical. The text-text contrastive model successfully recognized the lexical equivalence between the views, but in doing so, failed to detect the underlying cross-patient mismatch. Resolving such clinically similar collisions requires independent verification of identity attributes such as demographics, which is outside the scope of the present text-only model.

Table 10: Representative false positive and false negative cases at the 5% FPR threshold. Terms contributing to model conflict are highlighted in red; terms consistent across both views are highlighted in teal.

(a) False Positive: Clean Pair Flagged as Swap	
LLM Interp.	The ECG image analysis begins with the global features revealing a heart rate of 40 bpm, which suggests bradycardia... PR interval measurements vary significantly... large fluctuations in P and QRS amplitudes... conclusion stands with a sinus rhythm without prominent abnormalities, aligning with the representation of a normal ECG.
Clin. Report	Sinus rhythm, Normal ECG
(b) False Negative: Full Text-view Swap Flagged as Clean Match	
Case A LLM	...overall heart rate is 81 bpm, suggesting a normal sinus rhythm. However... indicates the presence of premature ventricular contractions (PVCs), aligning with the finding of a borderline ECG.
Case B Report	Sinus rhythm with PVC(s), Borderline ECG

6. Discussion

The central finding is that, in this setting, record integrity is better captured by contrastively trained alignment between *separately produced views* of the same event than by generic lexical similarity or latent identity matching alone. A compact text-text model with task-specific contrastive training substantially outperforms both same-dataset lexical baselines and frozen pretrained encoders on the harder stressors despite using simpler features. When the detailed LLM interpretation and short clinical report diverge substantially, that disagreement is often a stronger integrity cue than a single global similarity score. This is consistent with recent ECG-language and medical report-generation work, which shows that paired clinical views encode dense shared semantics beyond surface token overlap (Liu et al., 2024; Wang et al., 2025, 2023b; Yan et al., 2023). The main remaining tradeoff is between content sensitivity and temporal tolerance.

The Type-II FPR of 87.0% is not merely a tuning issue. In a typical cardiology department where a substantial fraction of ECG records are longitudinal re-visits, this rate

would generate a large alert burden (e.g., roughly 864 false alerts per 1,000 same-patient cross-visit pairs screened), motivating the staged deployment model described below. A patient can be stable in identity yet genuinely inconsistent in content across visits because new ischemia, conduction changes, or treatment effects alter the ECG narrative. Without an explicit identity variable, SHIFT-M3 interprets strong content drift as suspicious. This is also why the lexical baselines appear better on Type-II: they are less responsive to semantic change overall, which helps temporal tolerance but hurts corruption detection on the harder cross-patient stressors. SHIFT-M3 is therefore better suited to first-pass screening than final adjudication for longitudinal records.

The temporal sweep moderates this tension only partially, and the loss ablation supports the same interpretation from another angle: contrastive alignment and explicit swap supervision carry most of the integrity signal, whereas the temporal term mainly controls how aggressively the model treats legitimate clinical change. That limitation echoes a broader theme in longitudinal clinical modeling: temporal context is informative, but it cannot be treated as noise when the underlying patient state is genuinely evolving (Dalla Serra et al., 2023; Attia et al., 2021).

The practical deployment model is therefore staged: SHIFT-M3 should be used as a *first-pass* integrity filter whose outputs are routed to lightweight verification rather than treated as automatic rejection decisions. The method is equally applicable for post-market surveillance of deployed clinical decision support systems (Kahn et al., 2012): a screen that flags suspicious assembled records a posteriori can support ongoing monitoring of multimodal pipelines after deployment, not only at the point of record creation. The next research step is not merely a larger encoder, but an identity-aware consistency model that can distinguish "different patient" from "same patient, clinically changed." That future direction is well aligned with the informatics literature on data quality and patient matching, where automated screening is most useful when it feeds a broader verification workflow rather than acting as a fully autonomous decision rule (Kahn et al., 2012; Lewis et al., 2023; Gupta et al., 2024). The present study also has clear limitations: it relies on two text views, uses TF-IDF rather than contextual encoders (though frozen pretrained encoders perform substantially worse in the present evaluation), evaluates on a single dataset (MEETI/MIMIC-IV-ECG) with only synthetic linkage errors, and shares a view-derivation dependency because the LLM interpretation is generated from the clinical report. The auxiliary classification objective is disjoint from the identity stream used for conflict detection. The reported results retain the evaluated architecture, and conflict inference uses only the identity stream. Conflict-score calibration is not evaluated in the present work; a calibrated score would allow direct mismatch-probability interpretation and is an explicit direction for future work. The present evaluation is confined to a single dataset (MEETI/MIMIC-IV-ECG), where the TF-IDF vocabulary, threshold calibration, and stressor construction are all corpus-specific, and generalizability to ECG recordings from other institutions or acquisition systems remains to be established.

Type-III as a representation-space stress test: The Type-III stressor is a controlled representation-space probe rather than a simulation of clinically realistic partial record corruption. Real-world partial corruptions (including copied-forward documentation, split-record merges, and report amendments) manifest as coherent readable text that is internally

consistent yet attributed to the wrong patient. Type-III, by contrast, substitutes half of the TF-IDF feature vector with features drawn from a different patient, producing a hybrid representation with no readable-text analogue. The reported 90.3% TPR should therefore be read as sensitivity to feature-level partial overlap under this controlled substitution, not as robustness to the full range of realistic partial record failures.

A more faithful approximation of readable partial corruption is the text-splice Type-III variant used in the masking analysis, in which the source interpretation's first half is concatenated with a donor interpretation's second half. Under shared-token masking, this variant declines from AUROC 0.927 to 0.796 and from 73.7% to 31.5% TPR@5%, indicating that partial-swap detection is more dependent on lexical overlap than full-swap detection and that partial corruptions lacking shared surface tokens would require identity-aware signals beyond the scope of the present model.

7. Conclusion

This study frames linkage failure as a *multimodal record integrity* problem rather than a downstream prediction problem. On MEETI, SHIFT-M3 shows that contrastively trained alignment between two separately produced ECG text views is sufficient to detect full text-view swaps, partial swaps, and hard negatives at strong operating points with a compact pre-fusion model. Across same-dataset baselines, seed-stability analysis, loss ablation, and shared-token masking, the evidence is consistent: the strongest signal comes from cross-view contrastive alignment rather than lexical overlap alone.

The main limitation is longitudinal ambiguity. Records from the same patient can change meaningfully over time, and the high Type-II false-positive rate shows that content inconsistency is not yet the same as identity mismatch. For that reason, SHIFT-M3 is best viewed as a first-pass integrity screen that prioritizes suspicious records for verification, not as a fully autonomous adjudicator. The next step is an identity-aware temporal consistency model that can separate "different patient" from "same patient, clinically changed." More broadly, the results suggest that integrity checking should become a standard stage before multimodal fusion in clinical AI pipelines, especially in settings where record assembly errors can silently propagate downstream.

References

E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 72–78, 2019. doi: 10.18653/v1/W19-1909.

Z. I. Attia, D. M. Harmon, E. R. Behr, and P. A. Friedman. Application of artificial intelligence to the electrocardiogram. *European Heart Journal*, 42(46):4717–4730, 2021. doi: 10.1093/eurheartj/ehab649.

J. Bian, T. Lyu, A. Loiacono, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era

of real-world data. *Journal of the American Medical Informatics Association*, 27(12): 1999–2010, 2020. doi: 10.1093/jamia/ocaa245.

J. S. Brown, M. G. Kahn, and S. Toh. Data quality assessment for comparative effectiveness research in distributed data networks. *Medical Care*, 51(8 Suppl 3):S22–S29, 2013. doi: 10.1097/MLR.0b013e31829b1e2c.

J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2025. doi: 10.48550/arXiv.2402.03216.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pages 1597–1607, 2020.

F. Dalla Serra, C. Wang, F. Deligianni, J. Dalton, and A. O'Neil. Controllable chest X-ray report generation from longitudinal representations. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 4891–4904, 2023. doi: 10.18653/v1/2023.findings-emnlp.325.

A. Eggerth, D. Hayn, K. Kreiner, S. Veeranki, H. Traninger, R. Modre-Osprian, and G. Schreier. Patient record linkage for data quality assessment based on time series matching. *Studies in Health Technology and Informatics*, 260:210–217, 2019.

R. Garriga, T. S. Buda, J. Guerreiro, J. Omaña Iglesias, I. Estella Aguerri, and A. Matić. Combining clinical notes with structured electronic health records enhances the prediction of mental health crises. *Cell Reports Medicine*, 4(11):101260, 2023. doi: 10.1016/j.xcrm.2023.101260.

S. J. Grannis, H. Xu, J. R. Vest, S. Kasthurirathne, N. Bo, B. Moscovitch, R. Torkzadeh, and J. Rising. Evaluating the effect of data standardization and validation on patient matching accuracy. *Journal of the American Medical Informatics Association*, 26(5): 447–456, 2019. doi: 10.1093/jamia/ocz056.

S. J. Grannis, J. L. Williams, S. Kasthurirathne, M. Murray, and H. Xu. Evaluation of real-world referential and probabilistic patient matching to advance patient identification strategy. *Journal of the American Medical Informatics Association*, 29(8):1409–1415, 2022. doi: 10.1093/jamia/ocac068.

A. K. Gupta, H. Xu, X. Li, J. R. Vest, and S. J. Grannis. Manual evaluation of record linkage algorithm performance in four real-world datasets. *Applied Clinical Informatics*, 15(3):620–628, 2024. doi: 10.1055/a-2291-1391.

S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *npj Digital Medicine*, 3:136, 2020.

P. Hyvämäki, S. Sneck, M. Meriläinen, et al. Interorganizational health information exchange-related patient safety incidents: A descriptive register-based qualitative study. *International Journal of Medical Informatics*, 174:105045, 2023.

A. E. W. Johnson, L. Bulgarelli, L. Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, and J. F. Steiner. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical Care*, 50(Suppl):S21–S29, 2012. doi: 10.1097/MLR.0b013e318257dd67.

M. G. Kahn, T. J. Callahan, J. Barnard, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Washington, DC)*, 4(1):1244, 2016. doi: 10.13063/2327-9214.1244.

A. Kline, H. Wang, Y. Li, et al. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5:171, 2022.

A. E. Lewis, N. Weiskopf, Z. B. Abrams, R. Foraker, A. M. Lai, P. R. O. Payne, and A. Gupta. Electronic health record data quality assessment and tools: A systematic review. *Journal of the American Medical Informatics Association*, 30(10):1730–1740, 2023. doi: 10.1093/jamia/ocad120.

X. Li, H. Xu, and S. Grannis. The data-adaptive Fellegi-Sunter model for probabilistic record linkage: Algorithm development and validation for incorporating missing data and field selection. *Journal of Medical Internet Research*, 24(9):e33775, 2022. doi: 10.2196/33775.

C. Liu, Z. Wan, C. Ouyang, A. Shah, W. Bai, and R. Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pages 31949–31963, 2024. arXiv:2403.06659.

C. Liu, C. Ouyang, Z. Wan, H. Wang, W. Bai, and R. Arcucci. Knowledge-enhanced multimodal ECG representation learning with arbitrary-lead inputs. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 7298–7316, 2025. doi: 10.18653/v1/2025.findings-emnlp.385.

W. Lyu, X. Dong, R. Wong, S. Zheng, K. Abell-Hart, F. Wang, and C. Chen. A multimodal transformer: Fusing clinical notes with structured EHR data for interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings*, pages 719–728, 2023.

L. Riplinger, E. Plesons, P. Sirsat, S. Thompson, and S. Maitra. Patient identification techniques – approaches, implications, and findings. *Yearbook of Medical Informatics*, 29 (1):81–86, 2020. doi: 10.1055/s-0040-1701998.

V. Sangha, A. Khunte, G. Holste, B. J. Mortazavi, Z. Wang, E. K. Oikonomou, and R. Khera. Biometric contrastive learning for data-efficient deep learning from electrocardiographic images. *Journal of the American Medical Informatics Association*, 31(4):855–865, 2024. doi: 10.1093/jamia/ocae002.

A. Saporta, A. Puli, M. Goldstein, and R. Ranganath. Contrasting with Symile: Simple model-agnostic representation learning for unlimited modalities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2411.01053.

Z. Wan, C. Liu, X. Wang, C. Tao, H. Shen, J. Xiong, R. Arcucci, H. Yao, and M. Zhang. MEIT: Multimodal electrocardiogram instruction tuning on large language models for report generation. In *Findings of the Association for Computational Linguistics: ACL*, pages 14510–14527, 2025. doi: 10.18653/v1/2025.findings-acl.749.

F. Wang, J. Xu, and L. Yu. From token to rhythm: A multi-scale approach for ECG-language pretraining. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, pages 65059–65074, 2025.

S. Wang, Z. Liu, and B. Peng. A self-training framework for automated medical report generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 16443–16449, 2023a. doi: 10.18653/v1/2023.emnlp-main.1024.

S. Wang, B. Peng, Y. Liu, and Q. Peng. Fine-grained medical vision-language representation learning for radiology report generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 15949–15956, 2023b. doi: 10.18653/v1/2023.emnlp-main.989.

W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 5776–5788, 2020.

N. G. Weiskopf and C. Weng. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013. doi: 10.1136/amiajnl-2011-000681.

B. Yan, R. Liu, D. Kuo, S. Adithan, E. P. Reis, S. Kwak, V. K. Venugopal, C. P. O'Connell, A. Saenz, P. Rajpurkar, and M. Moor. Style-aware radiology report generation with RadGraph and few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 14676–14688, 2023. doi: 10.18653/v1/2023.findings-emnlp.977.

D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang. Combining structured and unstructured data for predictive models: A deep learning approach. *BMC Medical Informatics and Decision Making*, 20(1):280, 2020. doi: 10.1186/s12911-020-01297-6.

D. Zhang, X. Lan, S. Geng, Q. Zhao, S. Fan, M. Feng, and S. Hong. MEETI: A multimodal ECG dataset from MIMIC-IV-ECG with signals, images, features and interpretations. *Scientific Data*, 2026. doi: 10.1038/s41597-026-06796-1.

Appendix A: Replication Details

A.1 Dataset

MEETI is publicly available via PhysioNet (Zhang et al., 2026). Records are split 80/10/10 at the subject level, assigning all visits of a given patient to the same partition (Table 1).

A.2 Feature Extraction

Separate unigram TF-IDF vocabularies are fit on the training split only, using 500 interpretation features and 325 report features. Standard English stop words are removed, and the vocabularies are applied to validation and test sets without refitting. Feature vectors are L2-normalized before input to the MLP streams.

A.3 Architecture

Each identity MLP stream follows Linear(d_{in}, 256) $\rightarrow$ ReLU $\rightarrow$ BatchNorm $\rightarrow$ Dropout(0.2) $\rightarrow$ Linear(256, 128), with d_{in}=500 for the interpretation stream and d_{in}=325 for the report stream. Its output embeddings are L2-normalized. The auxiliary task stream uses parallel modality-specific projections ($d_{in} \rightarrow 256 \rightarrow 128$), followed by a classifier ($256 \rightarrow 64 \rightarrow 1$). The task stream shares no parameters with the identity stream and is not used at inference: the conflict score is computed solely from the identity-stream embeddings, and the loss ablation in Table 6 confirms that zeroing the task loss leaves every conflict metric unchanged. It is retained here only to describe the architecture as trained. Total parameters: 573,569 including the auxiliary task stream, of which only the identity stream is active at screening time.

A.4 Training Configuration

Table 11 lists all hyperparameters. Three random seeds (42, 43, 44) are used for the stability analysis reported in Table 4. All experiments were run on a single NVIDIA Tesla T4 GPU (16 GB).

Table 11: Hyperparameter configuration used for all reported experiments.

Hyperparameter	**Value**
Optimizer	Adam
β_1 / β_2 / ϵ	0.9 / 0.999 / 10^{-8}
Learning rate	3×10^{-4}
Batch size	128
Epochs	10
Embedding dim	128
λ_{ctr}	1.0
λ_{temp}	0.5
Swap margin m	0.5
Dropout	0.2
Random seeds	42, 43, 44

A.5 Operating-Point Threshold

The default 5% FPR operating point corresponds to the 95th percentile of clean-pair conflict scores on the validation split. The frontier in Table 7 sweeps from the 99th percentile (1% FPR) to the 80th (20% FPR).

A.6 Stressor Construction

All stressors are constructed from held-out test records; no stressor pair appears in training.

- *Type-I:* t_B is replaced by a randomly drawn different-subject report.
- *Type-II:* two records from the same subject at different visits are paired.
- *Type-III:* 250 of 500 interpretation feature dimensions are replaced by the corresponding dimensions from a randomly drawn different-subject interpretation (feature-space substitution, not text splicing).
- *Hard Negative:* each record is paired with a different-subject record sharing the same coarse keyword label (normal vs. abnormal).
- *Text-splice Type-III* (masking analysis only): the first half of the source interpretation text is concatenated with the second half of a donor interpretation.

A.7 Shared-Token Masking

The shared-token mask is computed as the intersection of non-zero vocabulary indices in t_A and t_B. Those indices are zeroed in both feature vectors before conflict-score computation.