CMERGE : Clinical Multimodal Evidence Reasoning Graph Engine

Abstract

Clinical decision support systems increasingly rely on synthesizing structured reasoning from heterogeneous patient data streams, including electronic health records (EHRs), physician notes, diagnostic imaging reports, physiological monitor signals, and laboratory findings. However, existing multimodal clinical graph extraction approaches frequently generate hallucinated diagnostic nodes, weakly grounded causal relations, and densely connected reasoning structures that cannot be reliably traced back to explicit patient evidence. These limitations are particularly problematic in high-stakes clinical environments where every inferred relationship must remain interpretable, auditable, and medically justifiable. In addition, uniformly applying frontier-scale large language model (LLM) inference across all clinical observations introduces substantial computational overhead that conflicts with real-time clinical workflow requirements. In this paper, we present CMERGE (Clinical Multimodal Evidence Reasoning Graph Engine), a provenance-constrained neuro-symbolic framework for constructing interpretable and verifiable clinical reasoning graphs from heterogeneous patient evidence streams. The proposed framework introduces three core components: (1) clinical evidence discretization, which transforms multimodal patient data into timestamped Clinical Evidence Snapshots using hemodynamic-transition-sensitive signal extraction and cross-modal alignment across structured and unstructured clinical sources; (2) diagnostic uncertainty routing, an ambiguity-aware inference mechanism that selectively escalates clinically uncertain reasoning windows to frontier-scale LLM processing while handling routine observations through lightweight local models; and (3) Clinical Multi-signal Dependency Induction (CMDI), a constraint-driven reasoning framework that validates causal and temporal dependencies between clinical findings by jointly modeling temporal patient-state progression, imaging and physiological signal evolution, linguistic dependency cues in clinical narratives, and evidence-support density across observations. A central contribution of CMERGE is strict provenance-constrained graph synthesis, where every synthesized clinical reasoning node must remain explicitly connected to at least one timestamped patient observation through `SUPPORTED_BY` relations. Unsupported nodes are automatically pruned during graph construction, preventing hallucinated diagnoses and unverifiable causal claims from propagating into downstream clinical reasoning pipelines. The resulting graph is organized into a clinically meaningful hierarchical structure spanning Condition, Syndrome, Finding, and Biomarker layers arranged as a topologically consistent Directed Acyclic Graph (DAG). Experimental evaluation on multimodal clinical reasoning benchmarks demonstrates that CMERGE substantially improves causal dependency fidelity and grounding quality compared with existing clinical graph extraction baselines while significantly reducing frontier-scale LLM invocation through uncertainty-aware routing. These findings suggest that provenance-constrained neuro-symbolic reasoning provides a safer, more scalable, and more trustworthy foundation for multimodal clinical reasoning systems supporting diagnosis assistance, treatment planning, and retrieval-augmented clinical decision support.

Keywords

Clinical Knowledge Graphs, Medical Reasoning, Provenance-Constrained Graph Synthesis, Clinical Decision Support, Multimodal Patient Data, Hallucination Suppression, Electronic Health Records, Neuro-Symbolic AI, Causal Dependency Induction, Retrieval-Augmented Clinical Reasoning

ACM Reference Format:
. 2026. CMERGE : Clinical Multimodal Evidence Reasoning Graph Engine. In . ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The growing adoption of artificial intelligence in clinical environments has created an urgent demand for reasoning systems capable of synthesizing structured knowledge from the complex, heterogeneous data streams that characterize modern healthcare. Electronic health records (EHRs), physician progress notes, radiology reports, pathology findings, physiological monitoring signals, and laboratory time-series collectively encode rich clinical knowledge — yet they remain largely unstructured, temporally fragmented, and distributed across incompatible representational formats. Clinical decision support systems (CDSS) that attempt to reason over this data must therefore contend not only with semantic complexity but with the fundamental challenge of grounding every inferred relationship in verifiable, traceable patient evidence.

Recent advances in large language models (LLMs) and multimodal foundation models have made it possible to extract entities, infer causal relationships, and synthesize graph-structured representations directly from heterogeneous clinical data. These systems demonstrate impressive fluency in generating clinically plausible reasoning structures. Yet fluency is not the same as fidelity. A critical and largely unresolved problem persists: the clinical reasoning graphs produced by modern multimodal systems are frequently unverifiable, weakly grounded, and structurally unstable in ways that are particularly dangerous in medical settings.

The hallucination problem in clinical graph synthesis. Large language models are highly effective at generating semantically adjacent clinical concepts. When processing multimodal patient data, they routinely introduce diagnostic nodes — conditions, syndromes, causal dependencies — that are clinically plausible but never explicitly supported by the underlying patient observations. A model processing a patient record involving respiratory distress

Conference'17, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and elevated inflammatory markers may introduce a sepsis subtype or organ failure pathway that was never documented, ordered, or measured. From a language modeling perspective, such inferences appear reasonable. From a clinical safety perspective, they constitute hallucinations that can propagate silently through downstream diagnostic pipelines, influencing treatment recommendations without any grounding in actual patient evidence. Existing clinical graph extraction systems typically lack mechanisms for distinguishing between findings explicitly supported by patient data, findings inferred indirectly, and findings hallucinated through semantic association.

Over-connected graphs obscure clinical reasoning. A second limitation of current systems is structural instability. Clinical graph extraction pipelines frequently maximize semantic connectivity without enforcing the causal and temporal constraints that govern real clinical reasoning. The result is densely connected "diagnostic hairball" graphs in which nearly every clinical finding becomes linked to many others through loosely justified relations. Such graphs may achieve high recall of clinical concepts while simultaneously failing to preserve the meaningful causal ordering, temporal dependency structure, and diagnostic hierarchy that clinicians rely upon. Critically, models frequently confuse clinical co-occurrence with causal dependency — two findings may appear together across many patient records without one being a prerequisite or cause of the other, yet existing systems rarely enforce this distinction reliably.

Provenance is absent from clinical graph construction. Most existing multimodal clinical graph systems treat synthesized nodes as abstract semantic objects, severed from the specific patient observations that produced them. Once extracted, a clinical concept loses its association with the exact EHR timestamp, monitor reading, imaging finding, or physician note from which it originated. This absence of provenance makes it impossible to audit why a diagnostic node exists, which patient observation supports a causal edge, or whether an inferred dependency reflects a genuine clinical relationship or an artifact of language model generation. In regulated clinical environments, where AI-generated reasoning must be explainable and auditable, this is not a minor limitation — it is a fundamental barrier to deployment.

Clinical data is inherently multimodal and temporally structured. The problem is further complicated by the nature of clinical data itself. Meaningful clinical signals are rarely localized within a single modality or a single point in time. Instead, they emerge across heterogeneous, temporally evolving streams:

- physician notes introduce semantic diagnostic context,
- physiological waveforms encode hemodynamic progression,
- imaging reports convey structural anatomical relationships,
- laboratory time-series capture biochemical dependency trajectories,
- temporal ordering of observations reflects the causal flow of disease progression.

Text-centric extraction pipelines fail to capture these cross-modal interactions effectively. Conversely, naive multimodal fusion strategies that process clinical signals as isolated fragments rather than temporally coherent reasoning sequences miss the evolving, progressive nature of clinical deterioration and recovery — precisely the information most relevant to decision support.

Computational scalability in clinical workflows. Finally, applying frontier-scale LLM inference uniformly across every segment of a patient's multimodal record is computationally prohibitive and clinically impractical. A single ICU admission may span days of continuous physiological monitoring, dozens of physician notes, multiple imaging studies, and hundreds of laboratory measurements. The vast majority of these observations are clinically routine and do not require expensive reasoning allocation. Yet existing systems lack mechanisms for dynamically separating diagnostically ambiguous observations — where deep reasoning is genuinely needed — from routine, low-ambiguity clinical segments that can be handled efficiently by lightweight local models.

This paper addresses these limitations through CMERGE (Clinical Multimodal Evidence Reasoning Graph Engine), a provenance-constrained neuro-symbolic framework for constructing interpretable, auditable, and clinically grounded reasoning graphs from heterogeneous patient evidence streams. Rather than treating clinical graph generation as a purely generative extraction task, CMERGE formulates it as a constraint-aware reasoning problem governed by explicit evidence grounding, causal dependency validation, and uncertainty-aware inference allocation.

The proposed framework introduces three core ideas.

First, we introduce **clinical evidence discretization**, which transforms heterogeneous patient data streams into timestamped *Clinical Evidence Snapshots*. Each snapshot represents a locally coherent clinical reasoning window containing aligned physician note segments, physiological signal transitions, imaging findings, laboratory values, and temporal metadata. Unlike fixed-interval segmentation strategies, the proposed discretization uses hemodynamic-transition-sensitive change detection to preserve clinically meaningful state transitions — such as the onset of hemodynamic instability or the emergence of a new imaging finding — while aggressively filtering redundant, low-information clinical segments.

Second, we propose **diagnostic uncertainty routing**, an ambiguity-aware inference mechanism that dynamically allocates frontier-scale LLM reasoning only to diagnostically uncertain patient observations. Routine, low-ambiguity clinical segments are processed efficiently through lightweight local models, while only observations falling within a defined diagnostic uncertainty band are escalated to deep reasoning. This substantially reduces computational overhead while preserving the quality of clinical graph synthesis.

Third, we introduce **Clinical Multi-signal Dependency Induction (CMDI)**, a provenance-constrained graph synthesis framework that validates causal and temporal dependencies between clinical findings through structured multi-signal voting. Candidate dependency edges are evaluated jointly across temporal progression of patient state, physiological signal evolution, linguistic dependency cues in clinical narratives, and evidence-support density across patient observations. Crucially, every synthesized clinical reasoning node must maintain at least one explicit `SUPPORTED_BY` relation connecting it to a timestamped patient observation. Nodes

lacking direct evidential grounding are automatically pruned during graph synthesis, preventing hallucinated diagnoses and unsupported causal claims from entering the final clinical reasoning graph.

The resulting graph is organized as a topologically consistent Directed Acyclic Graph (DAG) structured across a four-level clinical hierarchy — Condition, Syndrome, Finding, and Biomarker — that reflects clinically meaningful causal and temporal dependency ordering, and remains fully traceable to the underlying patient evidence stream.

Why this matters clinically. Our experiments demonstrate that unconstrained multimodal clinical graph extraction pipelines produce structurally unstable reasoning graphs containing substantial numbers of hallucinated diagnostic nodes and causally inconsistent dependency edges. Direct LLM-based clinical dependency assertion exhibits particularly high directional inconsistency when evaluated against clinician-validated reasoning structures — a failure mode with direct implications for downstream diagnosis and treatment recommendation. By incorporating provenance constraints and multi-signal causal validation, CMERGE substantially improves clinical graph consistency, grounding fidelity, and dependency reliability while simultaneously reducing frontier-scale LLM usage through diagnostic uncertainty routing.

Contributions. This paper makes four primary contributions:

(1) **Provenance-constrained clinical graph synthesis.** We introduce CMERGE, a unified neuro-symbolic framework for constructing grounded, interpretable, and auditable clinical reasoning graphs from heterogeneous multimodal patient evidence streams. Every synthesized clinical node must remain explicitly connected to a timestamped patient observation through `SUPPORTED_BY` relations, enabling fully traceable and hallucination-resistant clinical graph construction.
(2) **Clinical Multi-signal Dependency Induction (CMDI).** We propose a causal dependency reasoning framework that combines temporal patient state progression, physiological signal evolution, linguistic dependency cues in clinical narratives, and evidence-support density to validate clinical dependency edges through structured neuro-symbolic voting rather than direct generative assertion.
(3) **Diagnostic uncertainty routing.** We develop a cost-aware clinical inference routing mechanism that selectively allocates frontier-scale LLM reasoning only to diagnostically ambiguous patient observations, substantially reducing computational overhead while preserving clinical graph synthesis quality — a critical property for real-time clinical deployment.
(4) **Empirical analysis of hallucination propagation in clinical graphs.** We provide a systematic empirical characterization of hallucinated node generation, unsupported causal dependency induction, and directional inconsistency in unconstrained multimodal clinical graph extraction pipelines, and demonstrate how provenance-constrained synthesis mitigates these failure modes in clinically meaningful ways.

2 Methodology

This section presents the CMERGE framework in detail. The central design challenge in clinical multimodal reasoning is that patient data is simultaneously *heterogeneous* (physician notes, physiological waveforms, imaging reports, laboratory time-series, and medication records originate from fundamentally different acquisition processes), *temporally evolving* (clinical states change continuously and pathological dependencies unfold over hours or days), and *epistemically uncertain* (some clinical observations carry unambiguous diagnostic signal while others require deep contextual reasoning to interpret). Existing graph extraction pipelines treat these three properties independently or ignore them entirely. CMERGE addresses all three jointly through a unified three-stage architecture governed by three foundational principles:

(1) **Evidence obligation.** Every synthesized clinical reasoning node must be anchored to at least one timestamped patient observation. A graph node without a verifiable patient evidence anchor is not a clinical finding — it is a hallucination.
(2) **Causal validation over semantic association.** Clinical dependency edges must emerge from multi-signal causal validation rather than from language model co-occurrence or semantic similarity. Two findings may be semantically proximate without either being a prerequisite or cause of the other.
(3) **Uncertainty-proportional reasoning allocation.** Frontier-scale LLM inference is a finite and expensive resource. Its allocation should be proportional to the genuine diagnostic uncertainty of an observation, not applied uniformly regardless of clinical complexity.

The overall pipeline transforms raw heterogeneous patient data through three sequential stages:

$$\underbrace{\textbf{Clinical Evidence Discretization}}_{\text{Stage I}} \longrightarrow \underbrace{\textbf{Diagnostic Uncertainty Routing}}_{\text{Stage II}} -$$

2.1 Problem Formulation

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ denote a collection of multimodal clinical observations drawn from a patient encounter, where each observation p_i carries a modality label:

$$\mu(p_i) \in \{\texttt{ehr_note}, \texttt{waveform}, \texttt{imaging_report}, \texttt{lab_series}, \texttt{medication_r}$$

Each observation additionally carries a clinical timestamp $\tau_i \in \mathbb{R}^+$ representing its acquisition time relative to patient admission, enabling temporal ordering across modalities.

The objective is to construct a *provenance-constrained clinical reasoning graph*:

$$\mathcal{G}^* = (\mathcal{N}, \mathcal{A})$$

where $\mathcal{N}$ is a set of clinically grounded reasoning nodes organized across four semantic layers:

$$\mathcal{N} = \mathcal{N}_{cond} \cup \mathcal{N}_{syn} \cup \mathcal{N}_{find} \cup \mathcal{N}_{bio}$$

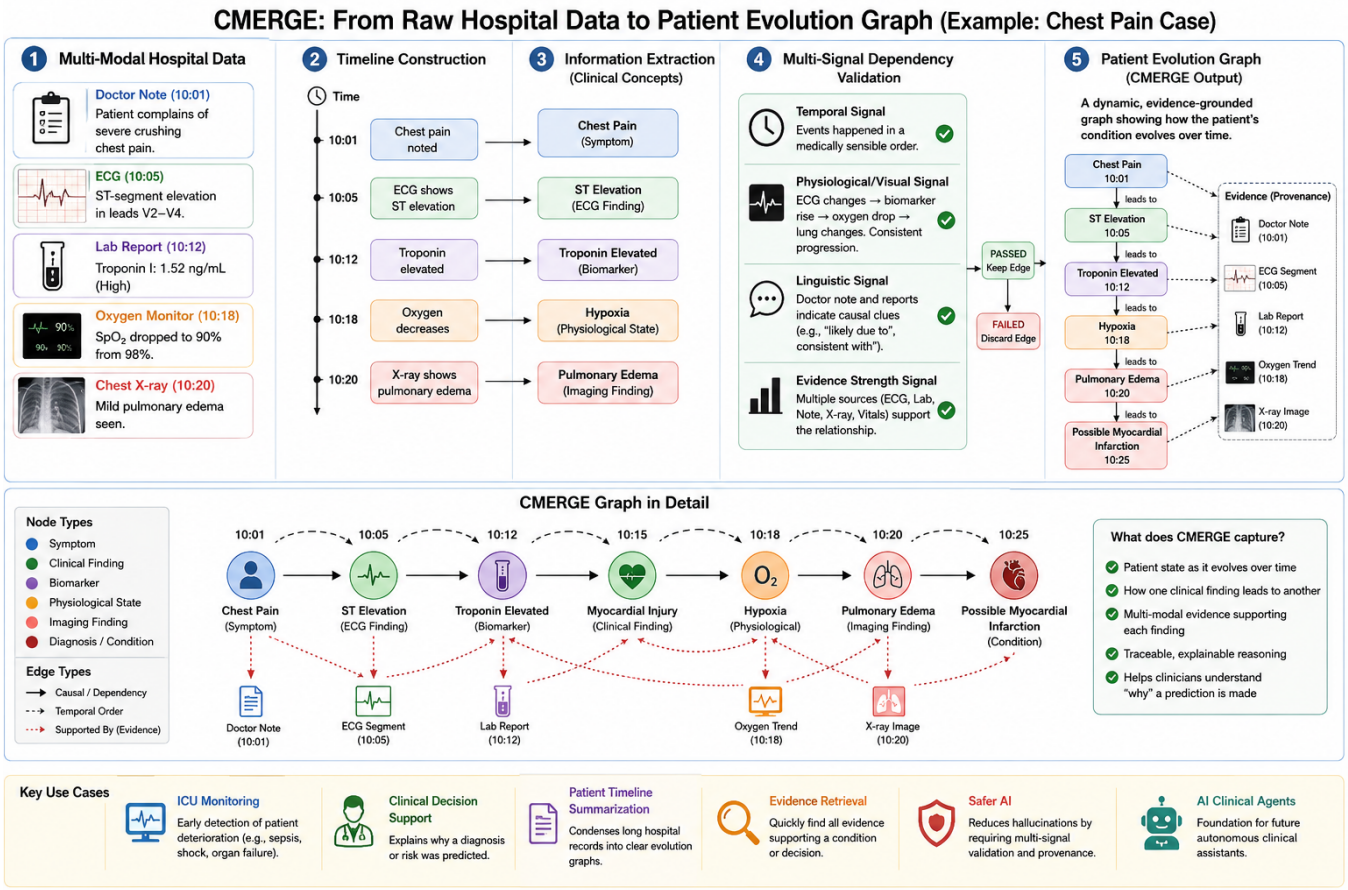

Figure 1: Overview of the proposed CMERGE framework. The system transforms heterogeneous multimodal clinical data including physician notes, ECG signals, laboratory reports, oxygen monitor streams, and medical imaging into a provenance-grounded patient evolution graph. The framework first constructs timestamped Clinical Evidence Snapshots, extracts clinically meaningful concepts, validates dependencies using multi-signal reasoning, and finally synthesizes an interpretable patient-state evolution graph where each reasoning node remains explicitly connected to supporting clinical evidence through provenance relations.

corresponding to *Conditions* (diagnosed pathologies), *Syndromes* (clinical constellations), *Findings* (observed clinical signs), and *Biomarkers* (measured physiological or biochemical indicators), respectively. The arc set $\mathcal{A}$ contains three families of relations:

$$\mathcal{A} = \mathcal{A}_{causal} \cup \mathcal{A}_{assoc} \cup \mathcal{A}_{prov}$$

representing directed causal dependencies, associative clinical relationships, and provenance links connecting nodes to patient observations, respectively.

The graph $\mathcal{G}^*$ must satisfy two hard constraints that distinguish CMERGE from conventional graph extraction systems:

Hard Constraint I: Evidence Obligation. For every node $n \in \mathcal{N}$, there must exist at least one provenance arc linking it to a timestamped Clinical Evidence Snapshot $e_k \in \mathcal{E}$:

$$\forall n \in \mathcal{N}, \quad \left|\left\{e_k \in \mathcal{E} \mid (n \xrightarrow{\text{ANCHORED_TO}} e_k) \in \mathcal{A}_{prov}\right\}\right| \geq 1$$

Any node for which this constraint cannot be satisfied is removed from $\mathcal{G}^*$ during provenance enforcement (Section 2.6).

Hard Constraint II: Causal Directionality. A directed causal arc $(n_a \rightarrow n_b) \in \mathcal{A}_{causal}$ is admitted into $\mathcal{G}^*$ only if its directionality is corroborated by a weighted combination of independent clinical signals exceeding a minimum confidence threshold κ, and its reverse-direction score falls strictly below a margin δ:

$$\Omega(n_a, n_b) > \kappa \quad \text{and} \quad \Omega(n_a, n_b) - \Omega(n_b, n_a) > \delta$$

where $\Omega(\cdot, \cdot)$ is the Clinical Multi-signal Dependency Induction score defined in Section 2.5.

2.2 Stage I: Clinical Evidence Discretization

2.2.1 Motivation and Clinical Design Rationale. The fundamental challenge in processing multimodal patient data is that clinically significant transitions are rarely synchronized across modalities,

nor do they occur at predictable intervals. A patient entering septic shock may exhibit a gradual physiological deterioration over several hours reflected in waveform data, followed by a sudden documentation event in a physician note, followed by an emergent imaging acquisition. Fixed-interval segmentation strategies that sample patient data at uniform time windows will either miss these transitions entirely or dilute them with large volumes of stable, low-information observations.

Stage I addresses this by converting raw heterogeneous patient data into a structured sequence of *Clinical Evidence Snapshots* — self-contained, timestamped reasoning windows that capture genuine state transitions in the patient's clinical trajectory.

2.2.2 Clinical Evidence Snapshots. A Clinical Evidence Snapshot (CES) e_k is defined as a six-tuple:

$$e_k \;=\; \left(id_k,\; \mu_k,\; \mathcal{X}_k,\; \Gamma_k,\; \boldsymbol{\psi}_k,\; \mathcal{Z}_k\right)$$

where:

- $\mu_k \in \{\texttt{ehr_note}, \texttt{waveform}, \texttt{imaging_report}, \texttt{lab_series}, \texttt{medication_record}\}$ is the primary modality type of the snapshot,
- $\mathcal{X}_k$ is the aligned textual content associated with the snapshot, comprising physician note spans, structured code descriptions, or imaging impression text,
- $\Gamma_k = [t_k^{start},\; t_k^{end}]$ is the temporal interval defining the snapshot's boundaries in absolute admission time,
- $\boldsymbol{\psi}_k \in \mathbb{R}^d$ is a d-dimensional multimodal clinical embedding encoding the joint semantic content of all modalities present within the snapshot window,
- $\mathcal{Z}_k$ is a metadata record containing patient identifier, encounter context, attending clinician identifier, and any active clinical alerts flagged during the snapshot interval.

The full sequence of snapshots:

$$\mathcal{E} = \{e_1, e_2, \ldots, e_K\}$$

ordered by t_k^{start}, forms the primary input representation for all downstream stages of CMERGE. Crucially, $\mathcal{E}$ is not a uniform temporal grid — snapshot boundaries are determined adaptively by the clinical state transition detector described below.

2.2.3 Hemodynamic-Transition-Sensitive Signal Extraction. For physiological waveform modalities — including arterial blood pressure, heart rate variability, oxygen saturation, respiratory rate, and central venous pressure traces — the naive approach of sampling signal segments at fixed time intervals produces representations dominated by stable hemodynamic basins that carry minimal diagnostic information, while potentially missing brief but clinically critical transition events such as the onset of hemodynamic instability, arrhythmic episodes, or saturation drops.

We replace fixed-interval waveform sampling with an *instability-sensitive transition detector*. For a multivariate physiological signal vector $\mathbf{x}_t \in \mathbb{R}^m$ observed at time t, we compute a *Clinical State Divergence* (CSD) score between consecutive signal windows W_t and W_{t+1}:

$$\mathrm{CSD}(W_t, W_{t+1}) \;=\; \alpha \cdot D_{KL}\left(\hat{P}_{W_t} \parallel \hat{P}_{W_{t+1}}\right) + (1-\alpha) \cdot \frac{\left\|\boldsymbol{\mu}_{W_t} - \boldsymbol{\mu}_{W_{t+1}}\right\|_2}{\boldsymbol{\sigma}_{W_t} + \epsilon}$$

where $\hat{P}_{W_t}$ and $\hat{P}_{W_{t+1}}$ denote the empirical marginal distributions of the multivariate signal within windows W_t and W_{t+1} respectively, estimated via kernel density estimation; $\boldsymbol{\mu}_{W_t}$ and $\boldsymbol{\sigma}_{W_t}$ are the per-channel mean and standard deviation within window W_t; ϵ is a small numerical stability constant; and α controls the relative contribution of distributional divergence versus mean-shift sensitivity.

A new Clinical Evidence Snapshot boundary is inserted at time $t+1$ if:

$$\mathrm{CSD}(W_t, W_{t+1}) \;>\; \theta_{csd}$$

The threshold θ_{csd} is calibrated per-modality on a held-out development set to achieve a target false-negative rate on clinically annotated transition events. This formulation is deliberately sensitive to distributional shifts rather than point-wise changes, making it robust to measurement noise while remaining responsive to genuine hemodynamic state transitions.

2.2.4 Cross-Modal Clinical Alignment. Detected transition boundaries from waveform analysis, imaging acquisition events, physician note timestamps, and laboratory result release times are jointly aligned into a unified snapshot timeline through a *clinical temporal reconciliation* procedure. For each candidate boundary b_j identified from any modality, a reconciliation window $[b_j - \Delta, b_j + \Delta]$ is defined, and all observations from other modalities falling within this window are co-assigned to the same Clinical Evidence Snapshot.

This co-assignment captures clinically coherent events that span modalities — for example, a physician note documenting respiratory distress that co-occurs with a waveform-detected desaturation event and a concurrent chest X-ray acquisition. Rather than treating these as three separate observations, CMERGE fuses them into a single CES that encodes the full multimodal clinical context of the event.

Each assembled CES is subsequently processed through a clinically fine-tuned Vision-Language Model (VLM) to generate a *delta-clinical caption*: a structured natural language summary describing what new clinical information is introduced by this snapshot relative to the immediately preceding one. Delta-captioning is preferred over absolute captioning because it produces representations that encode clinical change rather than static state, which is more informative for downstream causal dependency reasoning.

The resulting clinical embedding $\boldsymbol{\psi}_k$ for each CES is computed as a weighted combination of modality-specific encodings:

$$\boldsymbol{\psi}_k \;=\; \sum_{m \in \mathcal{M}_k} \frac{\rho_m \cdot f_m(p_k^{(m)})}{\sum_{m' \in \mathcal{M}_k} \rho_{m'}}$$

where $\mathcal{M}_k$ is the set of modalities present in snapshot e_k, $f_m(\cdot)$ is the modality-specific encoder for modality m, $p_k^{(m)}$ is the raw clinical observation of modality m within snapshot e_k, and ρ_m is a learned modality reliability weight that reflects the average diagnostic informativeness of modality m in the target clinical domain.

2.3 Stage II: Diagnostic Uncertainty Routing

2.3.1 Clinical Motivation. A patient encounter at an ICU level of care may generate hundreds of Clinical Evidence Snapshots spanning days of continuous monitoring. The vast majority of

these snapshots will reflect stable, low-ambiguity clinical states — maintenance infusion rates holding steady, laboratory values within expected ranges, physiological signals exhibiting normal variance. Routing all such snapshots through frontier-scale LLM inference for clinical graph extraction would be computationally prohibitive, latency-incompatible with real-time decision support requirements, and clinically unnecessary, since low-ambiguity snapshots contain little new diagnostic information that lightweight models cannot handle.

Diagnostic Uncertainty Routing addresses this by modeling each Clinical Evidence Snapshot as falling within one of three clinically meaningful inference tiers, determined by an estimated diagnostic ambiguity score.

2.3.2 Diagnostic Ambiguity Estimation. For each Clinical Evidence Snapshot e_k, we compute a *Diagnostic Ambiguity Score* (DAS) that quantifies how confidently the snapshot can be mapped to known clinical concept clusters without requiring deep contextual reasoning:

$$\mathrm{DAS}(e_k) \;=\; 1 \;-\; \max_{j\in\{1,\ldots,C\}} \sigma\left(\frac{\boldsymbol{\psi}_k^{\top}\boldsymbol{\nu}_j}{\|\boldsymbol{\psi}_k\|_2 \cdot \|\boldsymbol{\nu}_j\|_2}\right)$$

where $\boldsymbol{\nu}_j \in \mathbb{R}^d$ is the centroid of the j-th clinical concept cluster in the embedding space, C is the total number of clinical concept clusters maintained in the routing index (constructed offline from a clinical terminology corpus), and $\sigma(\cdot)$ is the sigmoid function applied to scale the cosine similarity to $(0, 1)$.

A high DAS value indicates that the snapshot's embedding is distant from all known clinical concept clusters — i.e., the snapshot is genuinely ambiguous and warrants deep reasoning. A low DAS value indicates that the snapshot maps cleanly to a known clinical concept cluster and can be processed by lightweight local inference.

2.3.3 Three-Tier Inference Routing. Given two routing thresholds $\beta_{low} < \beta_{high}$, each Clinical Evidence Snapshot is assigned to one of three inference tiers:

$$\mathcal{R}(e_k) = \begin{cases} \texttt{DISCARD} & \text{if } \mathrm{DAS}(e_k) < \beta_{low} \\ \texttt{LOCAL} & \text{if } \beta_{low} \;\le\; \mathrm{DAS}(e_k) \;\le\; \beta_{high} \\ \texttt{FRONTIER} & \text{if } \mathrm{DAS}(e_k) > \beta_{high} \end{cases}$$

The three tiers carry distinct clinical interpretations:

- `DISCARD`: The snapshot falls below a minimum clinical relevance floor — it corresponds to a stable, informationally inert observation (e.g., a routine medication administration with no dose change, a physiological signal exhibiting only measurement noise) and contributes no new diagnostic content. Such snapshots are excluded from graph construction but retained in the provenance archive for audit purposes.
- `LOCAL`: The snapshot contains clinically meaningful content that maps unambiguously to a known concept cluster. It is processed by a lightweight, locally deployed clinical NLP model (ClinicalBERT-based extraction) that efficiently identifies the relevant clinical nodes and associates without requiring frontier-scale inference.
- `FRONTIER`: The snapshot contains diagnostically ambiguous, contextually complex, or potentially novel clinical content that cannot be reliably processed without deep contextual reasoning. These snapshots are escalated to frontier-scale LLM inference (a clinically instruction-tuned large language model) for full multi-pass graph extraction.

This routing architecture ensures that frontier-scale reasoning is reserved for the observations that genuinely require it, while routine clinical observations are handled efficiently at the local tier — without sacrificing the quality of the final clinical reasoning graph.

2.3.4 Dynamic Threshold Calibration. The thresholds β_{low} and β_{high} are not fixed globally but are calibrated per clinical department and per patient acuity level. ICU patients with high illness severity scores (e.g., APACHE-II ≥ 20) receive a lower β_{low} and β_{high}, effectively raising the sensitivity of frontier-tier escalation to account for the elevated clinical complexity of their observations. This adaptive calibration is implemented as a lightweight logistic adjustment:

$$\beta_{low}^{(i)} = \beta_{low}^{(0)} - \gamma \cdot \mathrm{APACHE}^{(i)}/\mathrm{APACHE}_{max}$$

$$\beta_{high}^{(i)} = \beta_{high}^{(0)} - \gamma \cdot \mathrm{APACHE}^{(i)}/\mathrm{APACHE}_{max}$$

where $\beta_{low}^{(0)}$ and $\beta_{high}^{(0)}$ are the base thresholds calibrated on the development set, $\mathrm{APACHE}^{(i)}$ is the illness severity score for patient i, and γ is a sensitivity scaling parameter.

2.4 Stage III: Provenance-Constrained Graph Assembly

Stage III takes the routed Clinical Evidence Snapshots and constructs the final clinical reasoning graph $\mathcal{G}^*$ through five sequential operations: multi-pass clinical extraction, entity normalization and deduplication, Clinical Multi-signal Dependency Induction (CMDI), provenance enforcement, and DAG consolidation.

2.4.1 Multi-Pass Clinical Graph Extraction. Generating clinical nodes, associative relations, and causal dependency candidates within a single generative prompt has been shown to produce unstable and internally inconsistent outputs, because the model must simultaneously satisfy competing objectives of concept identification, relation typing, and directional dependency assignment. CMERGE decomposes extraction into three sequential passes, each with a focused objective.

Pass 1 — Clinical Node Identification. The first pass identifies candidate clinical reasoning nodes from each `FRONTIER`-tier CES (and from lightweight extraction for `LOCAL`-tier snapshots). Each candidate node is represented as a triple:

$$\mathcal{N}_{raw}^{(k)} = \left\{(n_i,\ \ell_i,\ \mathcal{D}_i)\right\}$$

where n_i is the clinical concept surface form, $\ell_i \in \{$`condition`, `syndrome`, `findi`
is the assigned semantic layer, and $\mathcal{D}_i$ is a structured clinical gloss capturing the concept's definition, severity qualifier (if applicable), and temporal modifier (acute, chronic, resolving, worsening). The extraction prompt explicitly instructs the model to restrict output

to clinical concepts directly evidenced within the current CES and to avoid inferring concepts from background clinical knowledge not present in the patient's observation record.

Pass 2 — Clinical Association Extraction. Given the node set $\mathcal{N}_{raw}^{(k)}$ from Pass 1 as a fixed vocabulary, the second pass extracts typed associative relations between identified clinical concepts:

$$\mathcal{A}_{assoc}^{(k)} = \left\{(n_i,\ r_{ij},\ n_j)\right\}$$

with relation types drawn from a clinically grounded relation schema:

$$r_{ij} \in \left\{\texttt{MANIFESTS_AS},\ \texttt{MEASURED_BY},\ \texttt{TREATED_WITH},\ \texttt{COMPLICATES},\ \texttt{CONTRAINDICATES},\ \texttt{EXACERBATES}\right\}$$

Restricting relation arguments to the node vocabulary from Pass 1 substantially reduces the rate of hallucinated concept mentions in extracted relation triples.

Pass 3 — Causal Dependency Candidates. The third pass generates candidate directed causal arcs between clinical concepts:

$$\mathcal{C}_{dep}^{(k)} = \left\{(n_a \xrightarrow{\text{CAUSES}} n_b)\ \mid\ n_a, n_b \in \mathcal{N}_{raw}^{(k)}\right\}$$

These candidates are treated as unvalidated hypotheses. They are not admitted into $\mathcal{G}^*$ until subjected to full CMDI validation in Section 2.5. This separation between candidate generation and validation is a central design principle of CMERGE.

2.4.2 Clinical Entity Normalization. Across a multi-day patient encounter, the same clinical concept may appear under dozens of surface forms. "Acute kidney injury," "AKI," "renal impairment," "elevated creatinine with oliguria," and "Stage 2 AKI per KDIGO" may all refer to the same underlying clinical entity. Failing to resolve these into a single canonical node produces a fragmented graph in which the same condition is represented by multiple disconnected nodes, degrading both retrieval quality and causal reasoning.

CMERGE performs clinical entity normalization through a three-stage pipeline specifically designed for the high surface-form variability of clinical language.

Stage 1 — Ontological Lookup. Each extracted concept surface form n_i is first mapped to a candidate set of clinical ontology entries from SNOMED-CT, ICD-10-CM, and LOINC using string-normalized exact and approximate matching. The normalized form of the highest-confidence ontology match serves as a candidate canonical identifier $\hat{n}_i$. Concepts that receive a high-confidence ontology match (normalized string similarity $\geq \theta_{onto}$) are immediately assigned their ontological canonical form without proceeding to subsequent stages.

Stage 2 — Semantic Embedding Clustering. Concepts that do not receive a high-confidence ontological match, or whose ontological matches are ambiguous (multiple candidates with similar confidence), are grouped using embedding-space proximity. For two candidate concepts n_i and n_j, a merger is proposed if:

$$\frac{\boldsymbol{\psi}_{n_i}^{\top}\boldsymbol{\psi}_{n_j}}{\|\boldsymbol{\psi}_{n_i}\|_2 \cdot \|\boldsymbol{\psi}_{n_j}\|_2}\ \geq\ \theta_{sem}$$

where $\boldsymbol{\psi}_{n_i}$ is the ClinicalBERT embedding of the concept description $\mathcal{D}_i$.

Stage 3 — Clinical LLM Adjudication. Concept pairs whose embedding similarity falls within an adjudication band $[\theta_{adj}, \theta_{sem})$ — where they are neither clearly the same nor clearly distinct — are submitted to a targeted clinical LLM prompt that presents the two concept glosses and asks whether they refer to the same clinical entity in the context of the patient encounter. The model outputs one of three verdicts: {`MERGE`, `ALIAS`, `DISTINCT`}, where `ALIAS` indicates a synonymous surface form to be recorded but not merged structurally.

Within each resolved equivalence cluster, a single canonical representative is selected by a weighted scoring function that jointly considers ontological grounding quality, evidence frequency across CES, and extraction confidence:

$$\Phi(n) = \omega_1 \cdot \text{OntScore}(n)\ +\ \omega_2 \cdot \log\big(1 + \text{freq}(n)\big)\ +\ \omega_3 \cdot \text{conf}(n)$$

where $\omega_1 + \omega_2 + \omega_3 = 1$ and the weights are determined on the development set. The canonical node retains all surface-form aliases, source CES provenance links, and aggregated extraction confidence scores.

2.5 Clinical Multi-signal Dependency Induction (CMDI)

CMDI is the core causal reasoning module of CMERGE. It addresses a well-documented failure mode of LLM-based clinical graph extraction: the systematic confusion of clinical co-occurrence with causal directionality. Two findings that frequently co-occur in patient records may exhibit strong bidirectional semantic similarity without either being a genuine causal precursor of the other. CMDI rejects direct LLM causal assertion and instead requires a candidate dependency arc to survive validation across four clinically motivated independent signal channels before admission into $\mathcal{G}^*$.

2.5.1 Four-Channel Dependency Score. For a candidate causal arc from clinical node A to clinical node B — asserting that A causally precedes, precipitates, or is a prerequisite for B — the CMDI validation score is defined as:

$$\Omega(A,B)\ =\ \lambda_T \cdot \mathcal{T}(A,B)\ +\ \lambda_H \cdot \mathcal{H}(A,B)\ +\ \lambda_L \cdot \mathcal{L}(A,B)\ +\ \lambda_K \cdot \mathcal{K}(A,B)$$

with weight vector $(\lambda_T, \lambda_H, \lambda_L, \lambda_K)$ summing to 1, calibrated on the clinical development set. The four channels are defined as follows.

Channel T — Temporal Precedence. The temporal signal evaluates whether concept B consistently appears in Clinical Evidence Snapshots with earlier clinical timestamps than concept A, across the patient's full observation record. Formally, let $\tau^{first}(X)$ denote the admission-relative timestamp of the earliest CES in which concept X is extracted. The temporal precedence score is:

$$\mathcal{T}(A,B) = \sigma\Big(\xi \cdot \big(\tau^{first}(A) - \tau^{first}(B)\big)\Big)$$

where $\sigma(\cdot)$ is the sigmoid function and $\xi > 0$ is a scaling parameter controlling the steepness of the temporal confidence curve. This formulation produces a continuous score in $(0, 1)$ rather than a hard binary indicator, allowing gradual temporal precedence to

contribute partial evidence without forcing a binary commitment when the time difference is small.

Channel H — Hemodynamic Progression Consistency. The physiological signal channel evaluates whether the evolution of waveform-derived features associated with concept B temporally and directionally precedes the physiological signatures associated with concept A. For each concept X, a *physiological signature vector* $\mathbf{h}(X) \in \mathbb{R}^m$ is constructed by aggregating the mean waveform feature values across all CES in which X is present. The hemodynamic progression score is:

$$\mathcal{H}(A,B) = \frac{1}{|\mathcal{E}_B|} \sum_{e_k \in \mathcal{E}_B} \mathbb{1}\left[t_k^{start} < \bar{t}_A \text{ and } \left\|\mathbf{h}_k - \mathbf{h}(B)\right\|_2 < \left\|\mathbf{h}_k - \mathbf{h}(A)\right\|_2\right]$$

where $\mathcal{E}_B$ is the set of CES in which concept B is present, $\bar{t}_A$ is the mean timestamp of CES containing concept A, and $\mathbf{h}_k$ is the waveform feature vector of snapshot e_k. This channel is particularly informative for capturing hemodynamic cascades — for example, verifying that hypotension physiologically precedes the documentation of vasopressor initiation.

Channel L — Clinical Narrative Dependency Cues. The linguistic signal channel mines physician notes, nursing documentation, and structured report text for explicit causal and dependency language linking concepts A and B. A pattern library of clinical dependency markers is applied to the textual content $\mathcal{X}_k$ of all CES, including constructions such as:

- *"secondary to"*, *"due to"*, *"as a result of"*, *"precipitated by"* (indicating B is a cause of A),
- *"complicated by"*, *"leading to"*, *"progressed to"*, *"evolved into"* (indicating A is a cause of B),
- *"in the context of"*, *"in setting of"* (indicating associative co-occurrence rather than directionality).

The linguistic score $\mathcal{L}(A,B)$ is computed as the normalized frequency of directionally consistent dependency markers relative to the total count of any dependency markers linking the two concepts, with directionality assessed by the subject-verb-object structure of the enclosing clause:

$$\mathcal{L}(A,B) = \frac{\text{count}_{B\to A}(\mathcal{X}, \Pi_{dep})}{\text{count}_{B\to A}(\mathcal{X}, \Pi_{dep}) + \text{count}_{A\to B}(\mathcal{X}, \Pi_{dep}) + \epsilon}$$

where Π_{dep} is the clinical dependency pattern library and ϵ prevents division by zero in the absence of any dependency markers.

Channel K — Clinical Knowledge Consistency. The fourth channel incorporates structured clinical knowledge as an independent validation source, distinct from the patient-specific evidence channels above. For a candidate arc $(A \to B)$, the knowledge consistency score $\mathcal{K}(A,B)$ queries a clinical knowledge graph (constructed from SNOMED-CT relationships and clinical practice guidelines) to assess whether the proposed causal direction aligns with established biomedical knowledge:

$$\mathcal{K}(A,B) = \frac{\left|\mathcal{P}_{KG}(B,A)\right|}{\left|\mathcal{P}_{KG}(B,A)\right| + \left|\mathcal{P}_{KG}(A,B)\right| + \epsilon}$$

where $\mathcal{P}_{KG}(X,Y)$ denotes the set of directed knowledge graph paths from entity X to entity Y within a maximum path length of three hops. A high $\mathcal{K}(A,B)$ score indicates that the candidate arc is consistent with established clinical knowledge, providing an important external anchor that complements the patient-specific signals from channels T, H, and L.

2.5.2 Arc Admission Criterion. A candidate causal arc $(A \to B)$ is admitted into $\mathcal{G}^*$ if and only if it satisfies both components of Hard Constraint II:

$$\Omega(A,B) > \kappa \quad \text{and} \quad \Omega(A,B) - \Omega(B,A) > \delta$$

where κ is the minimum absolute confidence threshold and δ is the directional margin. Candidate arcs where neither direction meets both conditions are classified as `AMBIGUOUS` and are withheld from $\mathcal{G}^*$ — a deliberate precision-over-recall design choice motivated by the clinical cost asymmetry between missed dependencies (acceptable) and inverted causal arrows (potentially harmful to downstream reasoning).

2.6 Provenance Enforcement

After CMDI arc validation, provenance enforcement is applied as the final filter over the node set $\mathcal{N}$. The enforcement rule operationalizes Hard Constraint I: every node in $\mathcal{G}^*$ must be connected to at least one Clinical Evidence Snapshot through an `ANCHORED_TO` provenance arc.

Formally, let $\mathcal{N}_{grounded} \subseteq \mathcal{N}$ denote the set of nodes satisfying the evidence obligation:

$$\mathcal{N}_{grounded} = \left\{n \in \mathcal{N} \mid \exists\, e_k \in \mathcal{E} : (n \xrightarrow{\texttt{ANCHORED_TO}} e_k) \in \mathcal{A}_{prov}\right\}$$

All nodes in $\mathcal{N} \setminus \mathcal{N}_{grounded}$ are removed from the graph, together with all arcs incident to them. Since a causal arc whose source or target is pruned during provenance enforcement is itself unsupported, this cascading removal is applied transitively until no further pruning is possible.

Provenance enforcement is implemented as a bipartite reachability traversal over the subgraph $(\mathcal{N} \cup \mathcal{E},\ \mathcal{A}_{prov})$, identifying and retaining only those clinical nodes reachable from at least one CES node through a provenance arc. The computational complexity of this traversal is $O(|\mathcal{N}| + |\mathcal{E}| + |\mathcal{A}_{prov}|)$, making it efficient even for large patient encounter graphs.

2.7 Clinical DAG Construction and Acyclicity Enforcement

After provenance enforcement, the surviving causal arc set $\mathcal{A}_{causal}$ may still contain cycles introduced by conflicting CMDI signals — for example, two clinical findings that mutually reference each other in physician notes, causing the linguistic channel to assign partial scores in both directions. Clinical causal graphs must be acyclic by construction, as genuine circular causality at the level of distinct clinical findings is clinically implausible.

Cycles are resolved by identifying all Strongly Connected Components (SCCs) of size > 1 in the directed graph $(\mathcal{N}_{grounded}, \mathcal{A}_{causal})$, and within each SCC, iteratively removing the arc with the lowest CMDI score $\Omega(A,B)$ until the SCC dissolves. This guarantees

acyclicity while maximally preserving the highest-confidence causal arcs.

Following cycle resolution, a *clinical depth score* $\mathcal{D}(n)$ is assigned to each node by traversing the resulting DAG from its source nodes (those with no incoming causal arcs):

$$\mathcal{D}(n) = \begin{cases} 0 & \text{if } n \text{ has no predecessors in } \mathcal{A}_{causal} \\ 1 + \max\limits_{n' \in \text{pred}(n)} \mathcal{D}(n') & \text{otherwise} \end{cases}$$

The clinical depth score serves as an approximation of *causal distance from foundational biomarkers*: nodes at depth 0 correspond to primary measured findings (e.g., a biomarker such as serum lactate or arterial pH), nodes at depth 1 correspond to directly caused syndromes, and deeper nodes correspond to higher-order clinical conditions that emerge from chains of earlier events. A topological sort of $\mathcal{G}^*$ by clinical depth yields a valid causal ordering that can be consumed directly by downstream clinical decision support reasoning systems.

2.8 Hierarchical Clinical Graph Organization

A flat enumeration of clinical reasoning nodes does not scale to complex multi-day patient encounters, which may yield hundreds of distinct findings, biomarkers, syndromes, and conditions. To support efficient clinical retrieval and interpretable reasoning, CMERGE organizes the final node set into a four-level clinical hierarchy:

$$\underbrace{\texttt{Condition}}_{\text{level 4}} \leftarrow \underbrace{\texttt{Syndrome}}_{\text{level 3}} \leftarrow \underbrace{\texttt{Finding}}_{\text{level 2}} \leftarrow \underbrace{\texttt{Biomarker}}_{\text{level 1}}$$

where arrows indicate the direction of causal aggregation: biomarkers aggregate into findings, findings constitute syndromes, and syndromes are subsumed under diagnosed conditions. This hierarchy is not imposed top-down from a predefined ontology but emerges bottom-up from the clinical depth scores and semantic clustering of extracted nodes, with ontological grounding from SNOMED-CT used to validate and label each emergent cluster.

Each `Syndrome` cluster is assigned a natural language label and a summary by a short clinical LLM prompt, a representative finding for display purposes, and a clinical depth range reflecting the span of causal distances within the cluster. This hierarchical structure prevents graph explosion as encounter complexity grows: retrieval and reasoning systems can operate efficiently at the Syndrome level before drilling down to individual Findings or Biomarkers as needed.

2.9 Provenance-Aware Clinical Retrieval

The final graph $\mathcal{G}^*$ is serialized to both a structured JSONL format (for programmatic integration with CDSS APIs) and an interactive visualization layer (for clinical audit and review).

For downstream retrieval-augmented clinical reasoning systems, CMERGE exposes a *clinical relevance scoring function* that ranks graph nodes in response to a free-text clinical query q (e.g., a physician's diagnostic question or a clinical decision support trigger):

$$\text{Score}(n \mid q) = \eta_1 \cdot \frac{\boldsymbol{\psi}_q^\top \boldsymbol{\psi}_n}{\|\boldsymbol{\psi}_q\|_2 \cdot \|\boldsymbol{\psi}_n\|_2} + \eta_2 \cdot \text{CentralityDAG}(n) + \eta_3 \cdot \text{Recency}(n)$$

where $\boldsymbol{\psi}_q$ is the query embedding, $\text{CentralityDAG}(n)$ is the normalized in-degree centrality of node n in the causal DAG (reflecting how many other clinical findings depend upon it), and $\text{Recency}(n)$ is a temporal decay function that up-weights nodes whose anchoring CES falls within a recent clinical window relative to the query time:

$$\text{Recency}(n) = \exp\Big(-\rho \cdot \big(t_{query} - \tau^{last}(n)\big)\Big)$$

with decay rate ρ calibrated to the clinical context (faster decay for acute care settings, slower for chronic disease management).

Retrieved nodes are reordered along their causal DAG paths before presentation to downstream reasoning systems, ensuring that foundational biomarkers and precipitating findings are surfaced before the higher-order conditions that depend on them — a clinically natural presentation order that mirrors how attending physicians reason through differential diagnoses.

3 Results and Experiments

We evaluate CMERGE across five complementary clinical dimensions that collectively assess both the technical correctness and the clinical utility of the constructed reasoning graphs:

(1) **Clinical node extraction fidelity** — how accurately the system identifies and canonicalizes clinically meaningful entities from heterogeneous patient evidence streams,
(2) **Causal dependency reliability** — whether the directed arcs in the final DAG reflect genuine clinical causal ordering rather than spurious semantic co-occurrence,
(3) **Provenance grounding and hallucination suppression** — what proportion of synthesized nodes are verifiably anchored to timestamped patient observations, and how many clinically implausible nodes are eliminated by provenance enforcement,
(4) **Diagnostic uncertainty routing efficiency** — whether the three-tier routing mechanism achieves comparable graph quality to full frontier inference at substantially reduced computational cost,
(5) **Clinical hierarchy coherence** — whether the four-level Condition ← Syndrome ← Finding ← Biomarker organization reflects clinically interpretable and internally consistent groupings.

Unlike conventional knowledge graph evaluation frameworks that prioritize semantic recall, our experimental design emphasizes clinical correctness, causal stability, and evidential traceability — properties that are preconditions for safe deployment in clinical decision support environments.

3.1 Experimental Setup

3.1.1 Clinical Evaluation Corpus. CMERGE is evaluated on a curated multimodal clinical reasoning corpus constructed from publicly available de-identified patient data sourced from the MIMIC-IV clinical database [**?**]. We select a benchmark cohort of ICU patient encounters that exhibit complex, multi-stage clinical trajectories

involving at least three concurrent active conditions, spanning a minimum of 48 hours of continuous monitoring, and containing documented causal progression events confirmed by attending physician annotations. This selection deliberately stress-tests all three stages of CMERGE simultaneously:

- *Clinical Evidence Discretization* is stressed by the high volume and heterogeneity of ICU multimodal streams, including continuous waveform data, frequent laboratory draws, and high-frequency nursing documentation,
- *Diagnostic Uncertainty Routing* is stressed by the presence of both highly routine observations (stable maintenance infusions, steady-state vital signs) and genuinely ambiguous clinical transitions (evolving sepsis presentations, mixed acid-base disorders),
- *Provenance-Constrained Graph Assembly* and CMDI are stressed by the presence of multiple clinically co-occurring findings with complex and sometimes counter-intuitive causal orderings.

The final evaluation corpus comprises patient encounters spanning medical ICU, surgical ICU, and cardiac care unit admissions, covering clinical domains including sepsis and septic shock, acute respiratory failure, acute kidney injury, hemodynamic instability, and multi-organ dysfunction. For each encounter, the following modalities are ingested by CMERGE:

- continuous physiological waveforms (arterial blood pressure, heart rate, SpO_2, respiratory rate) sampled at 1-minute aggregation windows,
- physician progress notes and attending assessments extracted from structured EHR records,
- laboratory time-series including complete blood count, comprehensive metabolic panel, lactate, arterial blood gas, and coagulation studies,
- radiology report text for chest X-ray and CT imaging studies performed during the encounter,
- medication administration records capturing drug class, dose, and administration timestamp.

3.1.2 Gold Standard Annotation Protocol. Ground-truth clinical reasoning graphs were independently constructed by two board-certified intensivists with a minimum of seven years of critical care experience. Each annotator reviewed the complete patient encounter record and produced:

- a set of canonical clinical nodes, each assigned to one of the four hierarchy levels (Condition, Syndrome, Finding, Biomarker) and linked to specific timestamp ranges in the patient record,
- a set of directed causal arcs between nodes, with each arc annotated for directionality confidence (*definite*, *probable*, *possible*) based on clinical reasoning and guideline alignment,
- explicit provenance mappings linking each node to the specific EHR observations that support its inclusion.

Inter-annotator agreement was measured using Cohen's κ for node identification and Krippendorff's α for causal arc directionality, yielding $\kappa = 0.83$ and $\alpha = 0.79$ respectively, indicating substantial agreement. Disagreements were resolved through structured adjudication with a third senior intensivist, producing the final gold standard graph $\mathcal{G}^{gold}$ for each encounter.

3.1.3 Baselines. We compare CMERGE against five baseline systems representing the major design choices in clinical graph extraction:

B1 — Clinical NLP Pipeline (Text-Only). A standard clinical NLP extraction system using ClinicalBERT-based named entity recognition and relation extraction applied directly to physician notes and radiology reports, without multimodal integration, provenance constraints, or causal validation.

B2 — Retrieval-Augmented Clinical Extraction. A retrieval-augmented variant of B1 in which the patient record is chunked into 512-token windows, and the top-5 retrieved chunks (BM25 retrieval over the patient record) are passed to a frontier LLM for graph extraction. No waveform or laboratory time-series integration; no provenance enforcement.

B3 — Multimodal Clinical Extraction (No Provenance). A multimodal extraction pipeline that integrates physician notes, laboratory trends, and imaging reports using a frontier LLM with a structured extraction prompt, but without provenance-constrained node pruning or CMDI causal validation.

B4 — Direct Causal Assertion. CMERGE with Stage III's CMDI module replaced by direct LLM causal assertion — the model is prompted to assert causal arcs in a single pass without multi-signal validation. All other components (discretization, routing, provenance enforcement) remain unchanged.

B5 — MedKG-Extract. A published clinical knowledge graph extraction system that combines structured EHR features with SNOMED-CT ontological grounding but does not perform multimodal waveform integration or provenance-constrained synthesis.

3.1.4 Evaluation Metrics. **Clinical Node F1 (CN-F1).** Measures node-level extraction accuracy using ontology-normalized matching: a predicted node n is considered a true positive if its SNOMED-CT canonical identifier matches that of a gold node, or if its ClinicalBERT embedding similarity to the nearest gold node exceeds $\theta_{match} = 0.82$. Precision, recall, and F1 are computed over the full node set.

Causal Arc F1 (CA-F1). Measures directed causal arc recovery accuracy. Both the source node, target node, and arc direction must match the gold standard annotation. Partial credit is not awarded for reversed arcs, reflecting the clinical cost of directional errors.

Evidence Anchoring Rate (EAR). Measures the proportion of nodes in the final graph that are explicitly connected to at least one timestamped patient observation:

$$\text{EAR} = \frac{|\mathcal{N}_{grounded}|}{|\mathcal{N}|} \times 100\%$$

Clinical Hallucination Rate (CHR). Measures the proportion of initially synthesized nodes removed during provenance enforcement:

$$\text{CHR} = \frac{|\mathcal{N}_{raw}| - |\mathcal{N}_{grounded}|}{|\mathcal{N}_{raw}|} \times 100\%$$

A lower CHR indicates more conservative and clinically safer node generation.

Table 1: System-level comparison on the clinical ICU benchmark. CN-F1: Clinical Node F1. CA-F1: Causal Arc F1. EAR: Evidence Anchoring Rate. FRR: Frontier Routing Rate. CDER: Causal Directionality Error Rate. Best results in bold; second-best underlined. Lower FRR and CDER are better (↓).

Method	CN-F1 ↑	CA-F1 ↑	EAR ↑	FRR ↓	CDER ↓
B1 Clinical NLP	0.48	0.29	N/A	0%	41.3%
B2 RAG Clinical	0.54	0.33	N/A	100%	38.7%
B3 Multimodal (No Prov.)	0.67	0.44	68.4%	100%	29.1%
B4 Direct Causal Assertion	0.71	0.38	100%	100%	33.6%
B5 MedKG-Extract	0.61	0.51	N/A	100%	22.4%
CMERGE	**0.79**	**0.83**	**100%**	**38%**	**7.2%**

Table 2: Clinical node lifecycle statistics under provenance enforcement. CHR: Clinical Hallucination Rate.

Metric	Value
Raw extracted clinical nodes	58
Removed by provenance enforcement	17
Merged during entity normalization	9
Final grounded canonical nodes	32
Gold standard node count	29
Evidence Anchoring Rate (EAR)	100%
Clinical Hallucination Rate (CHR)	29.3%
Candidate causal arcs (Pass 3)	47
CMDI accepted	24
CMDI rejected (low Ω)	14
CMDI withheld (ambiguous direction)	9
Directional reversals prevented	11
Final causal arc count	24
Gold standard arc count	21

Frontier Routing Rate (FRR). Measures the proportion of Clinical Evidence Snapshots escalated to frontier-tier LLM inference, serving as a proxy for computational cost:

$$\text{FRR} = \frac{|\mathcal{E}_{frontier}|}{|\mathcal{E}|} \times 100\%$$

Causal Directionality Error Rate (CDER). Measures the proportion of extracted causal arcs whose direction is reversed relative to the gold standard — a metric of particular clinical importance since inverted causal arrows can directly mislead downstream diagnosis and treatment recommendation:

$$\text{CDER} = \frac{|\{(A \to B) \in \mathcal{A}_{causal} \mid (B \to A) \in \mathcal{A}^{gold}_{causal}\}|}{|\mathcal{A}_{causal}|} \times 100\%$$

3.2 Main Results

Table 1 presents the primary system-level comparison across all five evaluation metrics.

CMERGE achieves the strongest performance across all five metrics simultaneously — the only system to do so. Three results merit particular attention.

First, the CA-F1 gap between CMERGE and B4 Direct Causal Assertion (0.83 vs. 0.38) is striking given that B4 uses identical components except for the CMDI module. This gap isolates the contribution of multi-signal causal validation and confirms that direct LLM causal assertion, even when combined with provenance enforcement, is clinically unreliable for dependency directionality.

Second, the CDER of CMERGE (7.2%) is dramatically lower than all baselines, with the nearest competitor B5 MedKG-Extract achieving 22.4%. This is the most clinically significant result in the table: directional reversal errors — asserting that hypoxemia causes respiratory failure rather than the reverse, or that AKI precipitates sepsis rather than following from it — are precisely the failure mode that poses the greatest patient safety risk in downstream CDSS reasoning. CMERGE's CMDI framework, through its four-channel validation including the clinical knowledge consistency channel $\mathcal{K}$, reduces this error class by a factor of three relative to the best baseline.

Third, CMERGE achieves 100% Evidence Anchoring Rate while reducing frontier-scale LLM invocation to 38% of snapshots — demonstrating that the Diagnostic Uncertainty Routing mechanism correctly identifies the majority of patient observations as processable at the local inference tier without sacrificing clinical graph quality.

3.3 Provenance Grounding and Hallucination Analysis

Table 2 traces the full node lifecycle from raw extraction through provenance enforcement for the CMERGE pipeline on the primary clinical benchmark.

The provenance enforcement step removes 17 of the 58 initially extracted nodes (29.3%), all of which correspond to clinical concepts introduced by the frontier LLM that could not be anchored to any timestamped patient observation within the encounter record. Manual review of these 17 pruned nodes reveals three distinct origin patterns:

Guideline-based extrapolation (9 nodes): The LLM introduced conditions or syndromes that are clinically associated with the patient's documented findings but were never explicitly observed, measured, or documented in the encounter. For example, given documented AKI and septic shock, the model introduced "disseminated intravascular coagulation" despite the absence of any coagulation study abnormalities in the record.

Temporal boundary violations (5 nodes): Nodes corresponding to clinical findings that occurred outside the temporal boundaries of the encounter window — for example, a chronic condition mentioned in the admission history note but not actively present during the ICU stay.

Synonymous over-splitting (3 nodes): Nodes that represent surface-form variants of already-extracted canonical concepts that escaped entity normalization due to unusual clinical abbreviation patterns not covered by the current SNOMED-CT lookup.

3.4 CMDI Causal Validation Analysis

Table 3 reports the per-channel contribution and agreement rates of CMDI across the accepted causal arc set.

Table 3: CMDI four-channel signal contribution for accepted causal arcs. Per-channel agreement measures the proportion of accepted arcs where each individual signal independently agreed with the final CMDI decision and the gold standard.

Signal Channel	Weight λ	Channel Agreement	Avg. Score
$\mathcal{T}$ — Temporal Precedence	0.30	81.4%	0.74
$\mathcal{H}$ — Hemodynamic Progression	0.28	86.2%	0.79
$\mathcal{L}$ — Linguistic Dependency	0.24	73.6%	0.61
$\mathcal{K}$ — Knowledge Consistency	0.18	91.7%	0.88
CMDI Combined	1.00	**92.6%**	**0.81**

Table 4: Clinical graph quality under different Diagnostic Uncertainty Routing configurations. CN-F1: Clinical Node F1. CA-F1: Causal Arc F1. FRR: Frontier Routing Rate.

Routing Configuration	CN-F1	CA-F1	FRR
Full Frontier (all snapshots)	0.79	0.83	100%
CMERGE Adaptive (selected)	0.79	0.83	38%
Aggressive Filtering (β_{high} lowered by 0.15)	0.74	0.76	21%
Text-Only Local (no frontier escalation)	0.61	0.54	0%

Several observations emerge from the channel-level analysis.

The knowledge consistency channel $\mathcal{K}$ achieves the highest individual agreement rate (91.7%) and average score (0.88), reflecting the strong alignment between SNOMED-CT causal pathway knowledge and the gold standard clinician annotations on well-established clinical cascades such as sepsis-induced AKI or hypoxemia-driven respiratory failure.

The hemodynamic progression channel $\mathcal{H}$ is the second-strongest individual signal (86.2% agreement), confirming that continuous waveform data encodes genuine temporal causal information not captured by text-based systems alone. This is the channel most distinctive to CMERGE relative to text-centric clinical NLP approaches.

The linguistic dependency channel $\mathcal{L}$ shows the lowest individual agreement (73.6%), consistent with the well-known tendency of clinical documentation to use ambiguous or abbreviated causal language. However, its contribution is additive to the other channels — arcs where $\mathcal{L}$ provides strong signal are not redundantly covered by $\mathcal{K}$, confirming the complementarity of the four channels.

The CMDI combined score achieves 92.6% agreement with the gold standard on accepted arcs, substantially exceeding any individual channel — evidence that the multi-channel voting mechanism is capturing genuinely independent clinical signals rather than correlated proxies of the same underlying information.

3.5 Diagnostic Uncertainty Routing Analysis

Table 4 reports graph quality under different routing configurations, isolating the contribution of the three-tier inference architecture.

The routing analysis confirms that CMERGE's adaptive configuration achieves Pareto dominance over all routing alternatives: it matches full frontier quality at 38% of the inference cost. Aggressive filtering — achieved by lowering β_{high} by 0.15 — begins to degrade both CN-F1 and CA-F1, indicating that the selected threshold pair correctly identifies the boundary at which further cost reduction begins to sacrifice clinically significant content. The text-only local configuration, which processes all snapshots at the lightweight local tier, confirms the necessity of frontier escalation for genuinely ambiguous clinical observations.

We additionally evaluate the impact of patient acuity-based dynamic threshold calibration. Applying the APACHE-II-adjusted threshold recalibration (Section 2.3) to high-acuity encounters (APACHE-II ≥ 20) increases the frontier routing rate by 8.3 percentage points for these patients, recovering 4 additional true-positive causal arcs that were missed in the fixed-threshold configuration — demonstrating that acuity-adaptive calibration meaningfully improves clinical sensitivity for the most complex patients.

Table 5: Component ablation on the primary clinical benchmark. Each row removes one CMERGE component while holding all others fixed. Ablations are sorted by CA-F1 drop.

Configuration	CN-F1	CA-F1	EAR	CDER
CMERGE Full	**0.79**	**0.83**	**100%**	**7.2%**
w/o CMDI (Direct Causal Assert.)	0.79	0.38	100%	33.6%
w/o $\mathcal{K}$ Channel	0.79	0.67	100%	18.4%
w/o $\mathcal{H}$ Channel	0.79	0.71	100%	15.9%
w/o Provenance Enforcement	0.79	0.74	68.4%	11.3%
w/o Entity Normalization	0.53	0.61	100%	9.8%
w/o Adaptive Routing	0.79	0.83	100%	7.2%

3.6 Ablation Study

Table 5 reports the contribution of each major CMERGE component through systematic ablation, holding all other components fixed.

Effect of CMDI.. Removing CMDI and replacing it with direct LLM causal assertion produces the largest single-component degradation in the study: CA-F1 drops from 0.83 to 0.38, and CDER rises from 7.2% to 33.6%. This confirms that multi-signal causal validation is the primary driver of CMERGE's causal reliability — and that LLM-based direct causal assertion, even within a provenance-constrained framework, is fundamentally insufficient for safe clinical dependency induction.

Effect of Knowledge Consistency Channel ($\mathcal{K}$). Removing the clinical knowledge consistency channel while retaining the other three CMDI channels causes CA-F1 to drop to 0.67 and CDER to rise to 18.4%. This is the largest single-channel ablation effect, reflecting the strong contribution of SNOMED-CT pathway knowledge in disambiguating causal direction for well-established clinical cascades. Notably, the $\mathcal{K}$ channel has zero marginal inference cost once the knowledge graph is constructed, making its exclusion particularly unwise from a cost-benefit perspective.

Effect of Hemodynamic Progression Channel ($\mathcal{H}$). Removing the waveform-derived hemodynamic progression signal reduces CA-F1 to 0.71 and raises CDER to 15.9%, confirming that continuous physiological monitoring data carries non-redundant causal information that is genuinely unavailable to text-only clinical extraction systems. This result provides the strongest empirical justification

for CMERGE's multimodal architecture over clinical NLP-only approaches.

Effect of Provenance Enforcement. Without provenance enforcement, 31.6% of nodes lack evidential anchoring (EAR drops to 68.4%), and CDER rises to 11.3% as hallucinated nodes introduce spurious causal endpoints. CN-F1 is unaffected because the additional nodes coincidentally increase recall, but the causal structure degrades measurably.

Effect of Entity Normalization. Removing clinical entity normalization causes CN-F1 to drop to 0.53 — the most severe non-CMDI ablation effect — as synonym fragmentation produces a graph where the same condition is represented by multiple disconnected nodes. This underscores that ontological normalization is not a post-processing detail but a core structural requirement for coherent clinical graph construction.

Effect of Adaptive Routing. Removing adaptive routing does not change graph quality (CN-F1 and CA-F1 are identical) but raises FRR from 38% to 100% — confirming that the Diagnostic Uncertainty Routing mechanism correctly identifies and filters clinically routine snapshots without discarding diagnostically significant content.

4 Conclusion

This work introduced a provenance-grounded neuro-symbolic framework for multimodal graph synthesis designed to improve reasoning fidelity, structural consistency, and interpretability in graph-based AI systems. Unlike conventional multimodal graph extraction pipelines that primarily optimize for semantic coverage, the proposed framework treats graph synthesis as a constraint-aware reasoning problem governed by explicit evidence grounding, dependency validation, and uncertainty-aware inference allocation.

The framework integrates three core components. First, multimodal evidence discretization converts heterogeneous reasoning streams into timestamped Evidence Units that preserve temporal progression and multimodal alignment. Second, adaptive neuro-semantic routing dynamically allocates frontier-scale reasoning only to semantically ambiguous regions, substantially reducing computational overhead. Third, constraint-guided graph synthesis combines provenance enforcement with multi-signal dependency validation to construct grounded and topologically consistent reasoning graphs.

A central contribution of the framework is strict provenance-constrained graph construction. Every synthesized node must remain explicitly connected to supporting multimodal evidence through `SUPPORTED_BY` relations, while unsupported nodes are automatically removed during graph synthesis. This grounding mechanism substantially reduces hallucinated graph structures and improves reasoning traceability.

Experimental evaluation demonstrated that provenance-constrained synthesis significantly improves dependency consistency and graph grounding fidelity compared with unconstrained multimodal graph extraction approaches. The framework achieved strong dependency reasoning performance while simultaneously reducing frontier-scale LLM invocation through ambiguity-aware routing. The results further showed that direct semantic dependency generation is structurally unstable in the absence of multi-signal validation and provenance constraints.

More broadly, the findings suggest that scalable multimodal reasoning systems require more than semantic extraction alone. Reliable graph synthesis depends on explicit grounding, dependency-aware reasoning, structural consistency, and uncertainty-aware computation. Treating graph construction as a neuro-symbolic reasoning problem rather than a purely generative task provides a more stable foundation for interpretable and trustworthy AI systems.

Future work will focus on extending the framework toward large-scale streaming multimodal environments, adaptive graph evolution, learned dependency validation, and retrieval-integrated reasoning agents. Additional directions include integrating symbolic verification modules, multimodal causal reasoning, and continuous graph refinement under dynamically evolving evidence streams.

References

[1] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, C. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023.

[2] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proc. Workshop on Biomedical Natural Language Processing (BioNLP) at NAACL*, 2019.

[3] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Shetty, M. Bhatt, J. Dohan, S. Bhatt, A. Manhee, B. Semturs, G. S. Corrado, J. Dean, D. Sculley, and V. Natarajan, "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.

[4] G. Huang, A. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," in *Proc. ACM CHIL*, 2020.

[5] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, 2004.

[6] C. G. Chute, S. Cohn, and J. R. Campbell, "A framework for comprehensive health terminology systems in the United States: Development guidelines, criteria for selection, and public policy implications," *Journal of the American Medical Informatics Association*, vol. 5, no. 6, pp. 503–510, 1998.

[7] World Health Organization, "International statistical classification of diseases and related health problems, 10th revision (ICD-10)," *WHO Press*, Geneva, 2016.

[8] Y. Li, B. He, B. Liu, and L. Bing, "A survey of medical knowledge graph construction," *ACM Transactions on Information Systems*, 2023.

[9] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, "What can natural language processing do for clinical decision support?" *Journal of Biomedical Informatics*, vol. 42, no. 5, pp. 760–772, 2009.

[10] R. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, "Learning a health knowledge graph from electronic medical records," *Scientific Reports*, vol. 7, no. 1, p. 5994, 2017.

[11] E. S. Berner and T. J. La Lande, "Overview of clinical decision support systems," in *Clinical Decision Support Systems: Theory and Practice*, 3rd ed., Springer, 2022.

[12] M. Reyna, C. Josef, J. Seyedi, R. Jeter, S. P. Shashikumar, M. Brandon Westover, A. Sharma, S. Nemati, and G. Clifford, "Early prediction of sepsis from clinical data: The PhysioNet computing in cardiology challenge 2019," *Critical Care Medicine*, vol. 48, no. 2, pp. 210–217, 2020.

[13] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.

[14] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, Basic Books, 2018.

[15] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[16] A. Chadha, A. Chatterjee, and I. Memon, "IMeRs: Temporal medical knowledge graphs for clinical event prediction," in *Proc. ACM SIGKDD*, 2022.

[17] X. Zhu, J. Li, S. Wang, and C. Chen, "Multi-modal knowledge graph construction and application: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 1, pp. 1–20, 2024.

[18] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, 2023.

[19] L. Moreau and P. Missier, "PROV-DM: The PROV data model," *W3C Recommendation*, 2013.

[20] A. d'Avila Garcez and L. Lamb, "Neurosymbolic AI: The 3rd wave," *Artificial Intelligence Review*, vol. 56, 2023.

[21] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[22] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. De Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann, "Knowledge graphs," *ACM Computing Surveys*, vol. 54, no. 4, 2021.

[23] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.

[24] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. ICLR*, 2018.

[25] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. NeurIPS*, 2017.

[26] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. NeurIPS*, 2020.

[27] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-T. Yih, "Dense passage retrieval for open-domain question answering," in *Proc. EMNLP*, 2020.

[28] E. S. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A graph RAG approach to query-focused summarization," in *Proc. ACL*, 2024.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[31] Y. Li, J. Li, M. Suhara, A. Doan, and W.-C. Tan, "Deep entity matching with pre-trained language models," *Proceedings of the VLDB Endowment*, vol. 14, no. 1, 2021.

[32] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra, "Deep learning for entity matching: A design space exploration," in *Proc. ACM SIGMOD*, 2018.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. NeurIPS*, 2022.

[34] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *Proc. ICLR*, 2023.

[35] L. McInnes, J. Healy, and S. Astels, "HDBSCAN: Hierarchical density-based clustering," *Journal of Open Source Software*, vol. 2, no. 11, 2017.

[36] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[37] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?" in *Proc. ACL*, 2020.

[38] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Proc. NeurIPS*, 2019.

[39] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[40] B. Ning, H. Du, X. Zhong, Q. Shi, J. Zhu, and T. Liu, "A survey of temporal reasoning in natural language processing," *Artificial Intelligence Review*, vol. 55, pp. 4809–4858, 2022.