

MedF²Route: Evidence-Routed Graph Verification for Hallucination Detection in Clinical LVLMs

Xiao Song^{1*}, Ruiying Chen^{2*}, Haonan Qin¹, Jianyuan Huang¹,
Yuqi Fang^{1†}, Caifeng Shan^{1†}

¹Nanjing University, ²Wuhan University of Technology
{x.song.matt,ruiyingc007,haonanqin392}@gmail.com,
241880208@smail.nju.edu.cn, {yqfang,caifeng.shan}@nju.edu.cn

Abstract

Hallucination detection in clinical large vision-language models requires cross-verifying heterogeneous evidence, including patient text and images for faithfulness, and medical knowledge for factuality. However, existing pipelines either verify these dimensions in isolation or conflate them into a uniform process, potentially causing evidence misalignment that limits detection precision. Consequently, we posit that the essence of black-box hallucination detection is an evidence routing problem and propose **MedF²Route**, an Evidence-Routed Graph Verification Framework for both Faithfulness and Factuality hallucination detection. Specifically, it first transforms free-text clinical queries and model responses into structured graphs, partitioning the response graph into independent subgraphs to separate patient-specific grounding from knowledge-based diagnostic inference. An evidence-aware routing and verification mechanism then dynamically dispatches response triplets in these subgraphs to their dedicated heterogeneous evidence sources for targeted verification. Experiments on our MedF²Bench diagnostic benchmark show that our framework outperforms state-of-the-art baselines, notably improving detection precision. Overall, MedF²Route enables professional and precise detection of clinical hallucinations, underpinning the trustworthiness of clinical LVLMs. The code and data are available at <https://github.com/Agentic-CliniAI/MedF2Route>.

1 Introduction

Clinical large vision-language models (LVLMs) are increasingly deployed for auxiliary diagnostic reasoning, yet their reliability remains constrained by hallucinations. As illustrated in Figure 1, these hallucinations manifest along two dimensions (Huang

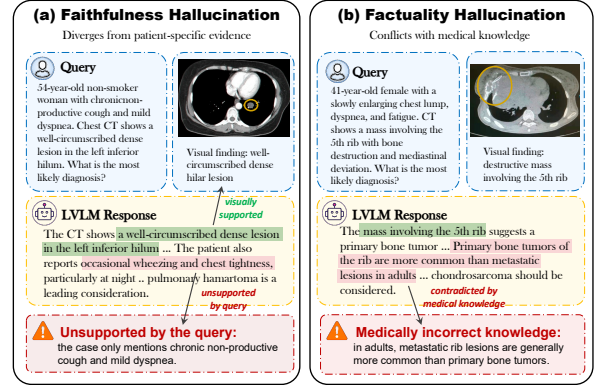


Figure 1: Two types of hallucinations in the context of heterogeneous multi-source evidence. (a) Faithfulness-oriented hallucination: the response diverges from patient text and medical images, while factuality hallucination denotes discrepancies with external clinical knowledge bases. In clinical practice, these two are often intertwined, as a clinically coherent diagnosis must remain grounded in patient-specific evidence while conforming to established medical knowledge. Therefore, developing a unified hallucination detector specifically designed for diagnostic scenarios is critical to ensuring the trustworthiness of clinical LVLMs.

et al., 2025) that require cross-verifying heterogeneous multi-source evidence: faithfulness hallucination refers to divergence from patient text and medical images, while factuality hallucination denotes discrepancies with external clinical knowledge bases. In clinical practice, these two are often intertwined, as a clinically coherent diagnosis must remain grounded in patient-specific evidence while conforming to established medical knowledge. Therefore, developing a unified hallucination detector specifically designed for diagnostic scenarios is critical to ensuring the trustworthiness of clinical LVLMs.

Existing detection methods typically employ a claim-level decomposition pipeline. Specifically, faithfulness-focused methods prioritize claim-level alignment with patient-specific evidence through natural language inference classifiers (Falke et al., 2019; Laban et al., 2022) or visual grounding metrics (Gunjal et al., 2024; Guan et al., 2024; Song et al., 2026). Meanwhile, factuality-oriented

*These authors contributed equally to this work.

†Corresponding authors.

frameworks decompose generated text into atomic claims and verify them against external medical knowledge through retrieval-augmented fact-checking (Min et al., 2023; Dhuliawala et al., 2024; Ceccarelli et al., 2025).

However, these approaches largely decompose responses into isolated claims, thereby fragmenting the contextual relationships essential for clinical reasoning, while predominantly evaluating hallucinations along a single dimension. To address these limitations, GraphEval (Sansford et al., 2024) attempts to preserve contextual relationships via graph structures, while UniHD (Chen et al., 2024) tries to jointly evaluate both dimensions in a uniform process. Nevertheless, GraphEval primarily focuses on unimodal text-grounded faithfulness, and UniHD aggregates heterogeneous evidence into an undifferentiated prompt, potentially risking evidence misalignment. Therefore, a unified detection architecture is required, one that preserves complex contextual relationships while explicitly navigating heterogeneous evidence sources.

To address these challenges, we posit that the essence of black-box hallucination detection is an evidence routing problem, and we propose **MedF²Route**, an Evidence-Routed Graph Verification Framework for both Faithfulness and Factuality hallucination detection. Specifically, it first transforms free-text clinical queries and model responses into structured graphs, partitioning the response graph into independent subgraphs to separate patient-specific grounding from knowledge-based diagnostic inference. An evidence-aware routing mechanism then dynamically dispatches response triplets in these subgraphs to their dedicated heterogeneous evidence sources for targeted verification. Subsequently, a dual-path verification pipeline assesses each routed triplet for faithfulness or factuality by aligning it with the routed evidence. Finally, a conservative aggregation rule maps fine-grained triplet-level verification signals to sentence-level outputs, ensuring that any unsupported clinical relation triggers a hallucination flag.

Experiments on the proposed MedF²Bench validate the effectiveness of MedF²Route across both faithfulness and factuality hallucination detection settings. Compared with baselines that often achieve high hallucination recall at the expense of precision, MedF²Route achieves notable improvements in detection precision and overall F1 score while maintaining a favorable precision-recall trade-off. Furthermore, its consistent per-

formance across diverse LVLM backbones underscores its model-agnostic reliability and strong potential for supporting downstream hallucination correction.

Our main contributions are as follows:

- We posit that the essence of black-box hallucination detection is an evidence routing issue and propose **MedF²Route**, an evidence-routed graph verification framework for faithfulness and factuality hallucination detection.
- We transform free-text clinical contexts into structured graphs and partition the response graph into independent subgraphs, which preserves complex contextual relationships while separating patient-specific grounding from knowledge-based diagnostic inference.
- We design a novel evidence-aware routing mechanism that dynamically dispatches response triplets to their dedicated heterogeneous evidence sources for targeted dual-path verification.
- We introduce **MedF²Bench**, a diagnostic benchmark specifically designed for evaluating hallucination detection methods. Extensive experiments on this benchmark demonstrate that MedF²Route significantly outperforms state-of-the-art baselines, notably improving detection precision and providing a reliable foundation for downstream correction.

2 Method

2.1 Task Formulation

Formally, given a multimodal query $X = \{x_t, x_v\}$ comprising case text x_t and medical image x_v , an LVLM generates a response sequence $Y = f_{\text{LVLM}}(x_t, x_v)$. The response is partitioned into T sentences $Y = \{s_1, \dots, s_T\}$. Clinical hallucinations in Y manifest along two dimensions: **Faithfulness-oriented hallucinations** (claims unsupported by or contradicted by patient-specific evidence in X), and **Factuality-oriented hallucinations** (claims inconsistent with established medical knowledge). Accordingly, we define a hallucination detector $\mathcal{D}$ that maps (X, Y) to a sentence-level binary prediction vector:

$$\hat{\mathbf{z}} = \mathcal{D}(X, Y), \quad \hat{\mathbf{z}} = \{\hat{z}_1, \dots, \hat{z}_T\}, \quad (1)$$

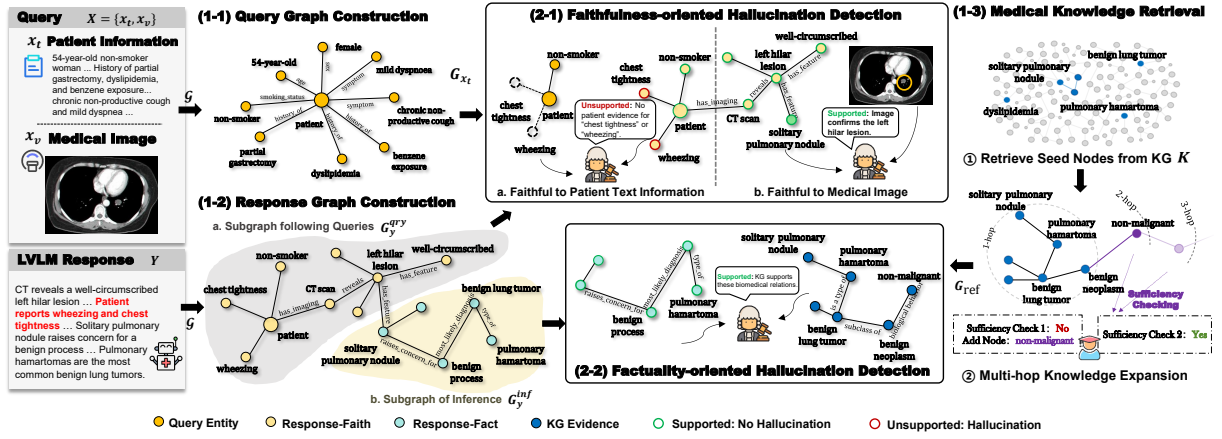


Figure 2: Overview of the proposed MedF²Route framework. Given the multimodal query X and the LVLm response Y as inputs, it first transforms the free-text context into a structured query graph G_{x_t} and a response graph $G_y = G_y^{\text{qry}}, G_y^{\text{inf}}$, where G_y^{qry} and G_y^{inf} are the query-following and inference subgraphs, respectively. Next, it retrieves relevant medical knowledge via multi-hop expansion based on sufficiency verification. Then, dual-path verification is applied to identify faithfulness- and factuality-oriented hallucinations, respectively.

where $\hat{z}_t = 1$ indicates a hallucination in s_t and $\hat{z}_t = 0$ otherwise. The core challenge is to preserve the contextual relationships inherent in diagnostic reasoning, which are often fragmented by atomic claim-level verification, while validating both dimensions in the context of heterogeneous multi-source evidence concurrently in a unified framework.

2.2 Framework Overview

To tackle the aforementioned challenges, in this paper, we theoretically posit that the essence of black-box hallucination detection is an evidence routing problem and instantiate $\mathcal{D}$ with **MedF²Route**, an evidence-routed graph verification framework for jointly detecting faithfulness- and factuality-oriented hallucinations in clinical LVLms. As illustrated in Figure 2, MedF²Route contains four stages. First, it transforms the case text x_t and the LVLm response Y into a structured query graph G_{x_t} and response graph G_y (Section 2.3). Second, it routes each triplet in G_y to the appropriate verification dimension and evidence source, including the query graph, medical image, or medical knowledge (Section 2.4). Third, dual-path verification checks each routed triplet for faithfulness or factuality and assigns a triplet-level hallucination label (Section 2.5). Finally, triplet-level predictions are conservatively aggregated into sentence-level outputs $\hat{z}$, where a sentence is labeled as hallucinated if any of its constituent triplets is judged hallucinated (Section 2.6). The following subsections detail each stage.

2.3 Clinical Graph Construction

To preserve contextual relationships that are often fragmented by prior claim-level methods, MedF²Route represents free-text clinical content as structured relational graphs. Specifically, the constructor $\mathcal{G}$ operates as an instruction-guided clinical relation extractor. $\mathcal{G}$ is implemented using the same backbone as the verifier. Given the case text x_t or response Y , $\mathcal{G}$ identifies clinically salient concepts C and connects them with concise relation labels to form directed clinical triplets. C encompasses symptoms, anatomical structures, imaging findings, diagnoses, procedures, etc. The extractor explicitly retains negation, certainty, laterality, and severity attributes when they alter clinical meaning, ensuring that relational semantics are preserved during structuring.

Formally, each clinical triplet is represented as $r = \langle c_h, \ell, c_t \rangle$, where c_h and c_t denote the head and tail clinical concepts, and ℓ denotes their relation. The extracted concepts constitute the node set C , and the clinical triplets constitute the directed edge set E . As illustrated in Figure 2 (1-1) and (1-2), the case text and response are mapped to:

$$\begin{aligned} G_{x_t} &= (C_{x_t}, E_{x_t}) = \mathcal{G}(x_t), \\ G_y &= (C_y, E_y) = \mathcal{G}(Y), \end{aligned} \quad (2)$$

where G_{x_t} and G_y represent the query and response clinical graphs, respectively.

2.4 Evidence-Aware Routing

Since each response triplet depends on distinct reference sources (query text, image, or knowledge

base) and requires dimension-specific verification, precise routing is essential to allocate it to the corresponding validation pathway and evidence source. Therefore, we propose the evidence-aware routing mechanism consisting of two sequential steps. Formally, we denote the evidence-aware routing function as ρ :

$$\{G_y^{\text{qry},\text{txt}}, G_y^{\text{qry},\text{img}}, G_y^{\text{inf},\text{kg}}\} = \rho(G_y). \quad (3)$$

Here, $G_y^{\text{qry},\text{txt}}$, $G_y^{\text{qry},\text{img}}$, and $G_y^{\text{inf},\text{kg}}$ denote response subgraphs routed to text-grounded faithfulness verification, image-grounded faithfulness verification, and knowledge-grounded factuality verification, respectively.

Specifically, ρ is implemented through two sequential routing steps. First, we separate G_y into two semantically distinct subgraphs: a query-following subgraph G_y^{qry} and an inference subgraph G_y^{inf} . Triplets in G_y^{qry} directly reference or summarize patient-specific evidence in X , and are routed to faithfulness verification. In contrast, triplets in G_y^{inf} introduce diagnostic deductions, causal mechanisms, differential diagnoses, or prognostic assessments beyond the explicit query content, and are routed to factuality verification. This subgraph partition determines the verification dimension: query-following triplets are checked for faithfulness, while inference triplets are checked for factuality.

Second, evidence-source routing determines the corresponding evidence source for verification. For triplets in G_y^{qry} , those describing directly observable imaging attributes are routed to the medical image x_v , forming $G_y^{\text{qry},\text{img}}$, while the remaining faithfulness triplets are routed to the query graph G_{x_t} , forming $G_y^{\text{qry},\text{txt}}$. For triplets in G_y^{inf} , the evidence source is set to the reference knowledge subgraph G_{ref} , forming $G_y^{\text{inf},\text{kg}}$. Specifically, image-routed triplets in $G_y^{\text{qry},\text{img}}$ describe directly observable imaging attributes, such as lesion morphology, anatomical location, laterality, or density. Through these two steps, each response triplet is assigned both a verification dimension and a concrete evidence source. This constrained routing space reduces cross-dimensional signal interference and clarifies the evidence scope for each response triplet, providing structured evidence alignment for the subsequent verification stage.

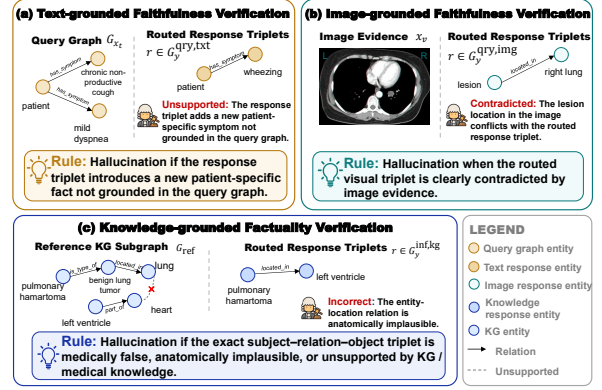


Figure 3: Illustration of dual-path verification with evidence-specific branches in MedF2Route.

2.5 Dual-Path Verification

Following routing, MedF2Route applies evidence-specific verifiers to each routed response triplet, which assess whether the triplet is supported by its allocated evidence source, producing a triplet-level hallucination prediction $\hat{y}_r \in \{0, 1\}$, where $\hat{y}_r = 1$ denotes that the triplet is unsupported by or contradicted by its routed evidence. This unified verification process is written as:

$$\hat{y}_r = \mathcal{H}(r, E_r), \quad E_r \in \{G_{x_t}, x_v, G_{\text{ref}}\}. \quad (4)$$

where E_r denotes the routed evidence source for response triplet r , and $\mathcal{H}$ is instantiated as a prompt-based VLM backbone, where the prompt is adapted according to the routing result. Figure 3 illustrates the evidence-specific branches under the dual-path verification design.

Faithfulness-oriented Hallucination Detection.

For faithfulness-oriented response triplets, this pathway determines whether each triplet is supported by patient-specific evidence. Depending on the routed evidence source, faithfulness verification is performed either against the query graph G_{x_t} or against the medical image x_v . Formally, the faithfulness verifier is instantiated as:

$$\hat{y}_r = \begin{cases} \mathcal{H}(r, G_{x_t}), & r \in G_y^{\text{qry},\text{txt}}, \\ \mathcal{H}(r, x_v), & r \in G_y^{\text{qry},\text{img}}. \end{cases} \quad (5)$$

First, for $r \in G_y^{\text{qry},\text{txt}}$, text-grounded faithfulness verification performs subgraph matching against G_{x_t} . A response triplet $r = \langle c_h, \ell, c_t \rangle$ is considered supported if its relation is explicitly present in or semantically entailed by the query graph through prompted semantic verification, permitting minor lexical variations in concept names

or relation labels while flagging unsupported or contradictory assertions. Second, for $r \in G_y^{\text{qry, img}}$, image-grounded faithfulness verification performs visual consistency checking against the medical image x_v . A response triplet $r = \langle c_h, \ell, c_t \rangle$ is examined by comparing its visually assessable finding or attribute with x_v , focusing on directly observable aspects such as presence, anatomical location, laterality, morphology, contour, density, signal, enhancement, size, and spatial extent when applicable. The triplet is flagged as hallucinated when the image does not support the described visual finding or attribute, or provides a clear visual contradiction. In practice, this comparison is implemented by presenting the triplet and the corresponding medical image x_v to a vision-language verifier, which returns a binary hallucination label based on visual support or contradiction.

Factuality-oriented Hallucination Detection. In contrast to faithfulness-oriented hallucination detection, which is grounded in patient-specific evidence, factuality-oriented hallucination detection must validate whether the triplets are consistent with established medical knowledge. Therefore, MedF²Route first constructs a compact reference subgraph G_{ref} from a local medical knowledge graph K . The resulting G_{ref} serves as structured knowledge evidence for factuality verification. Formally, the factuality verifier is instantiated as:

$$\hat{y}_r = \mathcal{H}(r, G_{\text{ref}}), \quad r \in G_y^{\text{inf, kg}}. \quad (6)$$

Specifically, the construction process builds G_{ref} in three stages. First, the construction process identifies medically salient concepts from the query and response graphs as seed nodes. Second, the construction process retrieves one-hop neighbors of the seed nodes from K and filters out nodes whose embedding similarity to the target context falls below a threshold δ , where the target context is represented by all clinical triplets extracted from the query and response graphs. Third, the construction process performs coverage-guided multi-hop expansion by iteratively incorporating frontier nodes that either help connect the head and tail concepts of factuality triplets $r = \langle c_h, \ell, c_t \rangle$, or introduce KG paths for judging whether the relation ℓ between c_h and c_t is medically supported or contradicted. Expansion terminates when no remaining frontier node can add a missing head or tail concept, a bridge concept, or a relevant relation path for any factuality triplet, or when the frontier is exhausted or the

predefined depth budget is reached.

Consequently, based on G_{ref} , the knowledge-grounded factuality verifier produces a triplet-level label $\hat{y}_r$ for each triplet $r \in G_y^{\text{inf, kg}}$. Here, $\hat{y}_r = 1$ indicates that the biomedical relation expressed by r is judged unsupported by or inconsistent with the medical evidence provided by the reference knowledge subgraph. This knowledge-grounded factuality verification complements faithfulness-oriented detection by verifying medical correctness rather than direct grounding in patient-specific evidence.

2.6 Granularity Alignment

The final stage maps fine-grained triplet-level verification signals to sentence-level granularity for a fair comparison with other detection methods. Since each response triplet is explicitly anchored to its source sentence during graph construction, triplet-level predictions can be aggregated without introducing sentence-assignment ambiguity. For each response sentence s_t , MedF²Route computes the sentence-level hallucination label via conservative max-pooling:

$$\hat{z}_t = \max_{r \in R(s_t)} \hat{y}_r, \quad (7)$$

where $R(s_t)$ denotes the set of response triplets extracted from sentence s_t . A diagnostic statement is classified as hallucinated if any constituent clinical relation is unsupported by or contradicted by patient-specific evidence, or inconsistent with established medical knowledge.

3 Experiments

3.1 Benchmark Construction

To evaluate clinical hallucination detection methods under the diagnostic scenario, we construct a sentence-level benchmark, named **MedF²Bench**, covering both faithfulness- and factuality-oriented hallucinations. The benchmark is built from Multi-CaRe (Nievas Offidani et al., 2025), an open-source multimodal clinical case dataset derived from open-access case reports. Each case contains clinical text and associated medical images, which we reorganize into diagnosis-oriented question-answer instances. The benchmark spans five imaging modalities, including CT, ECG, MRI, ultrasound, and X-ray, supporting rigorous multimodal assessment.

Because identifying fine-grained hallucination signals in naturally generated clinical LVLm responses requires substantial annotation effort, and

manually annotated hallucination datasets may quickly become outdated as model capabilities evolve, we follow established controlled and traceable construction practices adopted in FactCC, HaluEval, and MedHallu (Kryscinski et al., 2020; Li et al., 2023; Pandit et al., 2025) and construct an evaluation set by introducing targeted sentence-level errors into diagnostic responses. Specifically, a model-assisted pipeline rewrites a predefined target sentence according to a predefined perturbation rule to introduce a controlled hallucination. The target sentence, binary label, and hallucination type are therefore fixed by the perturbation protocol and its construction record. The introduced errors correspond to two clinically important dimensions: patient-specific faithfulness errors, where responses add claims unsupported by the case text or medical image, and medical factuality errors, where responses contain incorrect clinical relations, explanations, or diagnostic implications. Each generated sample is further filtered for image-text consistency by excluding cases with missing images, incomplete image information, or mismatched image-text evidence. The final benchmark contains 500 cases, with detailed statistics, construction procedures, and annotation protocols provided in Appendix A.

3.2 Experimental Setup

Baselines. On MedF²Bench, we compare MedF²Route with representative hallucination detection methods operating with different verification granularities. We organize the baselines into three groups: sentence-level detection methods, claim-level verification methods, and graph-structured verification methods. Sentence-level methods include VLM-as-Judge and SelfCheckGPT (Manakul et al., 2023). Claim-level methods include FActScore (Min et al., 2023) and UNIH (Chen et al., 2024), which decompose responses into finer-grained claims or atomic facts before verification. For graph-structured verification, we include FactSelfCheck (Sawczyn et al., 2026), which represents responses with graph-structured intermediate representations for hallucination detection.

Implementation Details. All methods are evaluated on the same MedF²Bench instances and response sentences, at sentence-level granularity. For deterministic judgment and verification modules, the decoding temperature is set to 0.0,

while sampling-based baselines follow their original sampling settings. For the knowledge-grounded factuality verification in MedF²Route, we use LMKG (Yang et al., 2024) as the local medical knowledge graph because of its broad coverage of medical entities and relations. Detailed KG preprocessing and retrieval parameters are provided in Appendix B.2. For baselines that produce claim-, fact-, or graph-level predictions, we map their outputs back to sentence-level labels using the same conservative aggregation rule, ensuring that all methods are compared at the same granularity.

Evaluation Metrics. The main comparison is conducted under the sentence-level binary hallucination detection setting. We report hallucination-class precision, recall, and F1-score, together with overall accuracy and macro-F1 over hallucination and non-hallucination classes. In the main results, cases are grouped by the target hallucination type, i.e., faithfulness-oriented or factuality-oriented cases, and metrics are first computed within each case and then averaged across cases within each group, which we refer to as the case-averaged protocol.

For ablation studies, we additionally report routed triplet-level metrics. Joint macro-F1 is computed over all routed response triplets, while Fact.-branch and Faith.-branch macro-F1 are computed over response triplets routed to factuality and faithfulness verification, respectively. The sentence-level score in the ablation table is computed over all cases. This case-averaged protocol prevents cases with more sentences or response triplets from dominating the results. We additionally report instance-averaged results, obtained by pooling all sentences within each subset, in Appendix C.3. Since hallucination detection is class-imbalanced, macro-F1 is used as the primary overall metric.

3.3 Main Results

Table 1 compares MedF²Route with SOTA hallucination detection baselines on MedF²Bench. Considering that different baselines operate at different verification granularities, we unify their outputs at the sentence level for fair comparison. Experiments are conducted across three VLM backbones: GPT-5.4-nano, InternVL-3.5, and Gemini-3.1-flash-lite.

Overall, MedF²Route achieves consistently strong performance across both hallucination types on all three VLM backbones, demonstrating its effectiveness. Specifically, on faithfulness halluci-

Case Type	Granularity	Methods	GPT					InternVL					Gemini				
			Hall.			Overall		Hall.			Overall		Hall.			Overall	
			P	R	F1	Acc.	Mac.F1	P	R	F1	Acc.	Mac.F1	P	R	F1	Acc.	Mac.F1
Faithfulness	Sentence-level	VLM-as-Judge	29.79	61.33	<u>37.76</u>	<u>64.01</u>	<u>53.67</u>	27.44	36.00	29.85	<u>78.44</u>	<u>57.74</u>	31.13	76.89	<u>42.30</u>	57.27	<u>51.44</u>
		SelfCheckGPT	20.24	48.00	26.41	53.87	43.13	19.65	<u>79.82</u>	30.76	28.12	24.51	21.13	92.44	34.03	25.81	22.50
	Claim-level	FActScore	20.47	70.22	30.97	38.60	34.60	17.00	18.57	16.17	72.84	49.73	19.49	44.44	25.81	<u>58.47</u>	47.84
		UNIHD	24.83	<u>91.96</u>	37.74	36.60	32.21	16.40	49.78	23.26	54.42	39.30	23.80	76.96	35.51	45.39	40.40
	Graph-structured	FactSelfCheck	21.02	99.11	34.63	21.99	18.57	20.71	98.22	<u>34.15</u>	21.96	18.20	20.27	<u>89.33</u>	32.88	25.10	22.40
		MedF ² Route	36.68	66.67	45.03	67.44	59.64	44.85	56.00	48.33	79.73	67.20	46.20	73.78	54.61	76.46	68.58
Factuality	Sentence-level	VLM-as-Judge	<u>42.96</u>	84.73	<u>54.11</u>	<u>68.13</u>	<u>62.76</u>	32.98	41.45	35.30	78.75	60.53	43.04	91.64	55.53	<u>64.27</u>	<u>61.12</u>
		SelfCheckGPT	42.94	86.18	53.55	63.36	59.19	25.97	<u>98.18</u>	39.89	32.63	29.43	25.38	100.00	39.56	30.58	27.57
	Claim-level	FActScore	25.46	80.36	37.36	42.87	39.12	<u>31.92</u>	36.29	30.78	73.58	<u>56.95</u>	26.20	57.80	33.99	59.86	51.95
		UNIHD	29.73	<u>96.70</u>	43.59	41.41	38.24	27.25	61.82	<u>35.43</u>	60.89	50.01	33.09	84.27	45.02	51.99	48.67
	Graph-structured	FactSelfCheck	21.63	99.27	35.42	22.40	18.80	21.52	100.00	35.37	21.76	18.01	22.22	<u>97.45</u>	36.04	25.64	22.57
		MedF ² Route	48.40	93.45	60.84	70.36	67.32	27.00	36.73	30.02	<u>73.87</u>	56.24	<u>42.94</u>	71.64	<u>51.62</u>	74.45	66.05

Table 1: Main case-averaged sentence-level results on MedF²Bench. Cases are grouped by target hallucination type, with metrics first computed within each case and then averaged across cases in each group. The best and second-best results are bolded and underlined, respectively.

nation detection, MedF²Route achieves the highest sentence-level Macro-F1 across all three backbones, outperforming the strongest baseline by 5.97, 9.46, and 17.14 percentage points on GPT, InternVL, and Gemini, respectively. For factuality hallucination detection, MedF²Route achieves the best Macro-F1 on GPT and Gemini, with gains of 4.56 and 4.93 percentage points over the strongest baseline, respectively. The less consistent performance on factuality detection, particularly observed on InternVL, can be attributed to two primary factors. First, it is partly constrained by backbone-specific verification limitations. Second, a cross-granularity comparison across the three evaluation series reveals that decomposing responses into finer units (e.g., claims or triplets) inherently degrades factuality detection performance. This degradation occurs might because finer fragmentation isolates statements from their broader context, increasing the difficulty of standalone factual verification.

In contrast, the gains are particularly consistent in faithfulness-oriented cases, where MedF²Route achieves higher hallucination precision and macro-F1 across all backbones. Additionally, it can be observed that several baselines obtain the highest recall at the expense of precision, which means that they frequently over-flag correct responses as hallucinations. Such false positives are particularly problematic in downstream workflows when users attempt to correct hallucinations reported by the detection methods. In contrast, our method maintains relatively high precision among all comparisons, indicating that our method can better identify

hallucinations while reducing false alarms, which may better support downstream hallucination correction.

We attribute these performance gains primarily to our design for preserving contextual relationships. Existing sentence-level or claim-level methods often verify responses in a fragmented manner, making it difficult to preserve clinical contextual relationships. In contrast, MedF²Route transforms responses into structured clinical graphs and routes their triplets to patient-specific evidence or medical knowledge. Furthermore, by preserving the evidence structure behind visual findings, anatomical relations, and diagnostic implications, this structured verification leads to more robust clinical hallucination detection across various VLM backbones.

3.4 Ablation Study

The ablation study systematically evaluates the contribution of each core component to MedF²Route’s clinical hallucination detection capability. Table 2 reports the results. To ensure a fair comparison, all variants are evaluated using the GPT-5.4-nano backbone. We report sentence-level macro-F1 for the final prediction and triplet-level macro-F1 for the intermediate verification process, computed over routed response triplets. Specifically, we construct ablation variants by individually bypassing each key module while preserving its intermediate outputs to maintain the original context for downstream modules. This context-preserving design isolates the direct contribution of the disabled module without disrupting the data flow of the re-

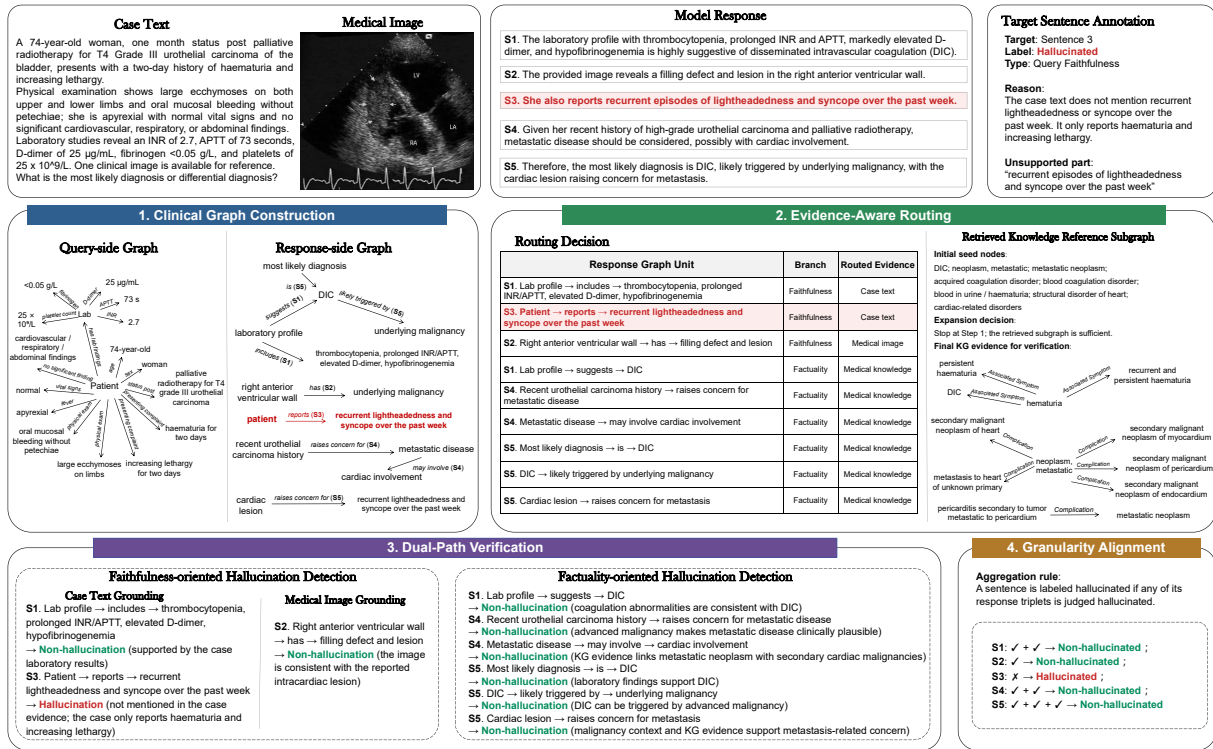


Figure 4: Qualitative case study of MedF²Route. The figure illustrates the four stages of our method: Clinical Graph Construction builds query- and response-side clinical graphs from the case and the LVLM response; Evidence-Aware Routing allocates each response triplet to patient-specific evidence or external medical knowledge; Dual-Path Verification checks faithfulness and factuality against the routed evidence; and Granularity Alignment identifies S3 as hallucinated because the claims of “recurrent lightheadedness and syncope” are unsupported by the case text.

Variants	Sent. Mac.F1	Triplet-level Mac.F1		
		Joint	Fact.	Faith.
MedF ² Route	63.86	67.12	66.93	41.38
w/o Evidence-aware Routing	51.91	60.27	58.07	44.83
w/o KG Retrieval	60.60	65.28	64.24	41.38
w/o Multi-hop KG expansion	63.18	66.56	65.86	41.38
w/o Visual Evidence Selection	60.95	63.69	66.93	20.02

Table 2: Case-averaged ablation results of MedF²Route on MedF²Bench, including sentence-level and triplet-level macro-F1. Gray values indicate unaffected results.

maining pipeline. For instance, when ablating the *evidence-aware routing*, the routing mechanism itself is disabled, but its pre-computed dimension-specific assignments (e.g., identifying fact-oriented triplets) are still utilized by the downstream knowledge retrieval. Following this principle, we ablate the *KG evidence retrieval* and *multi-hop KG expansion* to assess their marginal contributions to factual enrichment, and remove the *visual evidence selection* module to evaluate the importance of precise visual grounding.

The results demonstrate that each component contributes to the overall performance. First, by-

passing the *evidence-aware routing* module leads to the most substantial degradation in sentence-level F1 (-11.95) and factuality (-8.86). Since the downstream modules in this variant still receive the routed evidence assignments, this performance drop directly indicates that the dynamic coordination and context-aware processing provided by the routing mechanism are essential for accurate detection. Notably, this variant yields a marginal increase in faithfulness (+3.45). This suggests that in an unguided setting, the model more easily matches direct evidence for faithfulness-oriented triplets, but suffers from diminished sensitivity to complex factual knowledge, explaining the severe drop in factuality. Second, ablating the *KG evidence retrieval* and *multi-hop KG expansion* modules reduces factuality by 2.69 and 1.07, respectively, confirming their direct contribution to factual enrichment through external medical knowledge. Third, disabling the *visual evidence selection* module leads to a catastrophic drop in faithfulness (-21.36), proving it as the exclusive driver for precise visual grounding. Ultimately, these ablation results reveal a clear functional decoupling that KG modules bolster fac-

tuality, visual selection safeguards faithfulness, and the routing mechanism dynamically harmonizes them to optimize joint clinical reliability.

3.5 Qualitative Analysis

Figure 4 shows how MedF²Route provides interpretable hallucination detection through evidence-routed graph verification. Given a clinical response, the framework first extracts response triplets from each sentence and routes them to the corresponding evidence sources, including case text, medical image, and external medical knowledge. In this case, the laboratory findings, image-related description, and malignancy-related diagnostic implications are verified by their routed evidence sources. In contrast, *S3* introduces “*recurrent lightheadedness and syncope*”, which are not supported by the case text. MedF²Route therefore marks the corresponding response triplet as hallucinated and assigns a hallucination label to *S3* through granularity alignment. This example shows that our framework can localize unsupported statements while preserving clinically valid reasoning in other sentences.

4 Conclusion

In this paper, we propose **MedF²Route**, an evidence-routed graph verification framework for joint faithfulness- and factuality-oriented hallucination detection in LVLM-assisted diagnosis. By transforming free-text clinical queries and LVLM responses into structured graphs and routing response triplets to corresponding evidence sources through a dual-path verification pipeline, our framework successfully detects both hallucination dimensions in a unified manner and preserves the contextual relationships essential for accurate diagnostic reasoning. Experiments on our MedF²Bench benchmark demonstrate that MedF²Route outperforms SOTA baselines, achieving notable improvements in detection precision and overall F1 score while maintaining a favorable precision-recall trade-off. Its consistent performance across diverse VLM backbones indicates its model-agnostic robustness and strong potential for supporting downstream hallucination correction.

Limitations

To provide a further reference for future research, we investigate the limitations of the proposed approach. First, this work primarily focuses on the diagnostic scenario. While it represents one of

the most critical clinical applications, we acknowledge that other medical workflows equally warrant specialized investigation. These tasks often exhibit distinct evaluation priorities: for instance, radiology report generation predominantly hinges on faithfulness to image evidence, whereas medication recommendation requires consideration of complex drug-drug interactions that dictate pharmacological safety. Second, while our framework successfully unifies dual-dimension detection and preserves contextual relationships, the precision of fine-grained triplet validation remains partially constrained by the underlying backbones. Consequently, hallucination assessment inherently scales with the backbones’ intrinsic clinical reasoning capabilities, motivating future exploration with more advanced or domain-specialized backbones. Third, the reliability of factuality evaluation is intrinsically tied to the coverage, accuracy, and timeliness of the source knowledge base. This underscores the need for more comprehensive, rigorously curated, and dynamically updated clinical knowledge graphs to support robust, long-term hallucination mitigation.

Acknowledgments

This work was supported by the New Generation Artificial Intelligence–National Science and Technology Major Project (2025ZD0123502).

References

- Cristian Ceccarelli, Alessandro Raganato, and Marco Viviani. 2025. [Knowledge-grounded detection of factual hallucinations in large language models](#). In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 216–224. CEUR Workshop Proceedings.
- Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. [Unified hallucination detection for multimodal large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252, Bangkok, Thailand. Association for Computational Linguistics.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. [Chain-of-verification reduces hallucination in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

- Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. [Ranking generated summaries by correctness: An interesting but challenging application for natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. [Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. [Detecting and preventing hallucinations in large vision language models](#). In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18135–18143.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Transactions on Information Systems*, 43(2):1–55.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. [What disease does this patient have? a large-scale open domain question answering dataset from medical exams](#). *Applied Sciences*, 11(14):6421.
- Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. [Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval](#). *Bioinformatics*, 39(11):btad651.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. [Evaluating the factual consistency of abstractive text summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. [Summac: Re-visiting nli-based models for inconsistency detection in summarization](#). *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. [HaluEval: A large-scale hallucination evaluation benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. [SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Mauro Nieves Offidani, Facundo Roffet, María Carolina González Galtier, Miguel Massiris, and Claudio Delrieux. 2025. [An open-source clinical case dataset for medical image classification and multimodal ai applications](#). *Data*, 10(8):123.
- Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. 2025. [MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2858–2873, Suzhou, China. Association for Computational Linguistics.
- François Remy, Kris Demuyne, and Thomas Demeester. 2024. [BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights](#). *Journal of the American Medical Informatics Association*, 31(9):1844–1855.
- Alexey Romanov and Chaitanya Shivade. 2018. [Lessons from natural language inference in the clinical domain](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.
- Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. 2024. [GraphEval: A knowledge-graph based llm hallucination evaluation framework](#). In *Joint Proceedings of the Workshop on Knowledge-infused Learning and the workshop on Deep Learning and Large Language Models for Knowledge Graphs*, pages 20 – 31, Barcelona, Spain. CEUR Workshop Proceedings.
- Albert Sawczyn, Jakub Binkowski, Denis Janiak, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. 2026.

FactSelfCheck: Fact-level black-box hallucination detection for LLMs. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5603–5621, Rabat, Morocco. Association for Computational Linguistics.

Xiao Song, Haonan Qin, Zhaoxu Zhang, Jiong Zhang, Yuqi Fang, and Caifeng Shan. 2026. **Detecting clinical hallucinations in lvlms via counterfactual visual grounding uncertainty.** *Preprint*, arXiv:2606.28520.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. **Benchmarking retrieval-augmented generation for medicine.** In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Peiru Yang, Hongjun Wang, Yingzhuo Huang, Shuai Yang, Ya Zhang, Liang Huang, Yuesong Zhang, Guoxin Wang, Shizhong Yang, Liang He, and Yongfeng Huang. 2024. **Lmkg: A large-scale and multi-source medical knowledge graph for intelligent medicine applications.** *Knowledge-Based Systems*, 284:111323.

Appendix

A Benchmark Details

This section provides supplementary details of MedF²Bench. As described in Section 3.1, MedF²Bench is constructed from Multi-CaRe (Nievas Offidani et al., 2025), an open-access multimodal clinical case dataset derived from PubMed Central case reports. The final benchmark contains 500 diagnosis-oriented multimodal clinical cases, 585 medical images, and 2,383 labeled response sentences. As shown in Table 3, it covers five imaging modalities: CT, ECG, MRI, ultrasound, and X-ray.

Modality	Cases	Faith.	Fact.	Sent.	Avg. Sent.	Hall. Rate
CT	132	49	83	629	4.77	20.99%
ECG	65	28	37	311	4.78	20.90%
MRI	100	69	31	482	4.82	20.75%
Ultrasound	121	65	56	575	4.75	21.04%
X-ray	82	14	68	386	4.71	21.24%
Total	500	225	275	2383	4.77	20.98%

Table 3: Statistics of MedF²Bench. Faith. and Fact. denote faithfulness- and factuality-oriented cases, respectively.

A.1 Benchmark Construction

Source processing and query construction. Since case reports are typically written after diagnosis and treatment, we reorganize each selected

case into a pre-diagnostic setting before constructing queries. Specifically, each case is divided into three evidence fields: (i) *visible clinical evidence*, including demographics, relevant history, presenting symptoms, physical examination findings, and other information reasonably available before diagnosis; (ii) *withheld diagnosis-related evidence*, including definitive diagnosis, pathology, treatment, follow-up, outcome, and retrospective explanations, which is excluded from the query to avoid diagnosis leakage; and (iii) *image-grounded observable evidence*, containing diagnosis-neutral visual findings tied to the associated image, which is reserved for multimodal grounding and hallucination evaluation. Based on the visible clinical evidence, we construct a concise pre-diagnostic consultation query that asks for the most likely diagnosis or differential diagnosis, while only indicating that clinical image evidence is available for reference.

Data cleaning and annotation. We filter out cases with missing images, incomplete image information, mismatched image-text pairs, or image-grounded evidence that cannot be reasonably verified.

For each query, we construct a clinically realistic diagnostic response with controlled sentence-level perturbation. Each response is mostly grounded in the textual query and image-grounded observable evidence, with one primary target hallucinated sentence. We consider two target perturbation types: *query-faithfulness* errors, which introduce patient-specific information unsupported by the query or image evidence, and *knowledge-factuality* errors, which introduce medically incorrect or implausible claims within the case context. The target sentence, binary hallucination label, and hallucination type are determined by the predefined perturbation protocol and its construction record. The target type is used only for benchmark organization and grouped analysis; during inference, MedF²Route routes each response triplet according to its semantic role and required evidence source, without relying on the case-level target type.

Each response sentence is recorded with its sentence index, text, binary hallucination label, hallucination type, and a minimal correction when applicable. Sentences without hallucination are labeled as supported with a null type. In addition to sentence-level labels, we retain triplet-level records, including query-side triplets, response-side triplets, routed verification branches, evidence sources, and

triplet-level outcomes. These fine-grained records are used for component-level ablation and qualitative analysis, while the primary benchmark target remains sentence-level hallucination detection.

A.2 Controlled Construction Quality Audit

To assess the reliability of the controlled construction process, we conduct a stratified quality audit of 50 cases, corresponding to 10% of MedF²Bench and balanced across the two hallucination types and five imaging modalities. One author experienced in clinical NLP and familiar with the benchmark protocol assesses whether the intended hallucination is successfully introduced, whether its type is correctly assigned, and whether any unintended additional error is introduced. The results are shown in Table 4.

Audit criterion	Overall	Faith.	Fact.
Intended hallucination introduced	49/50	25/25	24/25
Hallucination type correctly assigned	50/50	25/25	25/25
No unintended additional error	49/50	25/25	24/25

Table 4: Quality audit of the controlled construction process on 50 stratified cases.

Overall, the intended hallucination is successfully introduced in 98% of the audited cases, the hallucination type is correctly assigned in all cases, and 98% contain no unintended additional error. These results support the quality of the controlled construction process within the audited subset.

A.3 Validation on Naturally Generated Responses

To complement the controlled benchmark, we further evaluate MedF²Route on naturally generated responses from the same 50 cases. The diagnostic responses are generated directly from the case text and images without targeted error injection or manual rewriting, such that the resulting hallucinations are not specified in advance. The same author subsequently labels each response sentence, assessing faithfulness against patient-specific textual and visual evidence and factuality against verifiable medical knowledge. MedF²Route and VLM-as-Judge use the same GPT-5.4-nano detection backbone.

Method	Overall	Faithfulness	Factuality
VLM-as-Judge	34.5	31.2	35.2
MedF ² Route	47.1	45.6	47.4
Δ	+12.6	+14.4	+12.2

Table 5: Macro-F1 on naturally generated responses from 50 cases under the GPT-5.4-nano configuration.

As shown in Table 5, MedF²Route improves Overall, Faithfulness, and Factuality Macro-F1 by 12.6, 14.4, and 12.2 percentage points over VLM-as-Judge, respectively. These results provide complementary evidence that the advantage of MedF²Route extends beyond the predefined perturbations used to construct MedF²Bench.

B Implementation Details

B.1 Baseline Implementation Details

For all baselines, we follow their original settings whenever applicable and make only necessary adaptations to fit the multimodal clinical setting and the sentence-level evaluation protocol of MedF²Bench. All methods are evaluated on the same cases and response sentences. For baselines that output claim-, fact-, or graph-structured predictions, we map them to sentence-level labels using a shared conservative aggregation rule: a sentence is labeled as hallucinated if any associated fine-grained prediction indicates a hallucination; otherwise, it is labeled as supported. If no valid fine-grained item is extracted from a sentence, the sentence is treated as supported.

VLM-as-Judge. We implement VLM-as-Judge as a direct sentence-level multimodal evaluator. For each case, the judge is given the pre-diagnostic clinical query, all available clinical image evidence, and one response sentence at a time. It is prompted to return one of two JSON labels, supported or hallucinated, with deterministic decoding at temperature 0.0. The normalized label is directly used as the sentence-level prediction.

SelfCheckGPT. SelfCheckGPT (Manakul et al., 2023) detects hallucinations by measuring consistency between the original response and sampled responses. Following its original setting, we generate four sampled responses for each case using the same clinical query and image evidence, with temperature set to 1.0. Each original response sentence is compared against the sampled responses,

and the resulting sentence-level score is converted into the unified granularity.

FactScore. FactScore (Min et al., 2023) decomposes generated text into atomic facts and verifies each fact against external evidence. Since MedF²Bench targets clinical diagnosis rather than open-domain biographies, we keep its decomposition-and-verification pipeline but replace the evidence source with medical reference documents. Specifically, we retrieve evidence from the MedRAG (Xiong et al., 2024) Textbooks (Jin et al., 2021) corpus using MedCPT (Jin et al., 2023). Fact-level predictions are mapped to sentence-level labels using the shared aggregation rule.

UNIHD. UNIHD (Chen et al., 2024) is a tool-augmented hallucination detection framework that extracts verifiable claims, formulates tool-specific questions, collects evidence, and verifies claim-level hallucinations. We adapt it to our clinical setting by providing the clinical query, image evidence, and response sentence as multimodal context. For fact-oriented evidence collection, we replace the original commercial web search backend with DDGS¹ for reproducibility, while keeping the remaining pipeline unchanged. The resulting claim-level predictions are aggregated to sentence-level labels.

FactSelfCheck. FactSelfCheck (Sawczyn et al., 2026) checks whether facts in the original response are consistent with multiple sampled responses. We adopt its frequency-based KG variant, which represents response facts as subject–relation–object triples and performs graph-structured consistency checking against KGs constructed from four sampled responses generated under the same setting as SelfCheckGPT. The graph-fact-level outputs are mapped to sentence-level labels using the shared aggregation rule.

B.2 Medical Knowledge Graph Processing

We use LMKG (Yang et al., 2024) as the local medical knowledge graph for knowledge-grounded factuality verification. LMKG is a large-scale and multi-source medical knowledge graph designed for intelligent medicine applications, providing structured medical entities and relations for external knowledge verification.

Since the clinical queries, model responses, and extracted clinical triplets in MedF²Bench are rep-

resented in English, we preprocess LMKG into a consistent English entity space. Specifically, we retain entities that either contain an explicit English name or whose canonical name is already written in English. Among the 403,784 entities in the original graph, 327,707 entities are retained, covering 81.16% of the entity set. We further keep relation triples whose head and tail entities are both included in the filtered entity set, ensuring that entity retrieval and relation lookup are conducted within the same processed graph space.

For entity retrieval, we encode the retained English entity names with BioLORD-2023 (Remy et al., 2024), a biomedical representation model for clinical concepts and sentences. The resulting 768-dimensional normalized embeddings are indexed with FAISS using inner-product similarity. For each medically salient concept extracted from the query graph and response graph, we retrieve the top-3 candidate KG entities and retain candidates with similarity above 0.75 as seed entities. During semantic pruning, we use threshold $\delta = 0.7$, where a KG node is retained if it is sufficiently similar to any concept from the query graph or response graph. For sufficiency-guided expansion, the maximum depth budget is set to $T = 5$, and one high-relevance frontier node is selected at each expansion step.

To assess the practical coverage of LMKG on MedF²Bench, we additionally measure entity-level matching under the GPT-5.4-nano configuration. Approximately 70% of the clinical entities extracted from MedF²Bench can be matched to LMKG. We therefore treat LMKG as a structured external evidence source rather than an exhaustive knowledge base, and the absence of a matched entity or relation is not by itself considered evidence of hallucination.

C Additional Experiments and Analyses

C.1 Triplet Extraction and Evidence Routing

Response-triplet extraction quality. To evaluate the reliability of the graph construction stage, we stratify 50 cases across the two hallucination types and five imaging modalities and manually assess all 560 response triplets extracted under the GPT-5.4-nano configuration. One author familiar with the benchmark protocol evaluates whether each triplet preserves the entities, central relation, and clinical meaning of its source sentence, together with clinically important attributes such as

¹<https://github.com/deedy5/ddgs>

negation, laterality, temporality, uncertainty, and severity. The results are shown in Table 6.

Triplet quality	Count	Percentage
Fully correct	518	92.50%
Partially correct	4	0.70%
Incorrect	38	6.80%
Total	560	100.00%

Table 6: Intrinsic evaluation of response-triplet extraction on 50 stratified cases.

Overall, 518 of 560 response triplets (92.50%) are fully correct, indicating that the graph construction stage preserves the source clinical information in most evaluated cases.

Evidence routing evaluation. To evaluate routing independently of extraction quality, we annotate reference routes for the 518 fully correct triplets. Each triplet is assigned to patient text, medical image, or external medical knowledge according to the evidence required by its central relation. Table 7 reports the per-class results.

Evidence source	P	R	F1	Support
Medical image	99.0	88.8	93.6	107
Patient text	88.0	95.4	91.6	108
Medical knowledge	98.0	98.7	98.4	303
Macro average	95.0	94.3	94.5	518

Table 7: Evidence-routing performance on 518 fully correct response triplets.

Routing achieves 95.9% overall accuracy and 94.5 macro-F1. As shown in Table 8, most remaining errors occur between patient-text and medical-image evidence.

Gold / Pred.	Image	Text	Knowledge
Image	95	11	1
Text	0	103	5
Knowledge	1	3	299

Table 8: Confusion matrix of evidence routing. Rows denote reference routes and columns denote predicted routes.

C.2 Robustness and Statistical Significance

Stability across random seeds. The original experiments use seed 123. We additionally evaluate MedF²Route and VLM-as-Judge on the same fixed subset of 100 cases using seeds 456 and 789, while

keeping the inputs, prompts, model parameters, and evaluation procedure unchanged. The results are shown in Table 9.

Method	123	456	789	Mean $\pm$ SD
MedF ² Route	67.79	70.66	67.29	68.58 $\pm$ 1.82
VLM-as-Judge	61.44	59.39	59.52	60.12 $\pm$ 1.15
Δ	+6.35	+11.27	+7.77	+8.46 $\pm$ 2.53

Table 9: Macro-F1 across three random seeds on the same fixed subset of 100 cases.

MedF²Route outperforms VLM-as-Judge under all three seeds, with improvements ranging from 6.35 to 11.27 percentage points.

Paired-bootstrap comparison. We further conduct a case-level paired bootstrap comparison over all 500 cases, resampling cases with replacement to preserve dependencies among sentences from the same case. The results are summarized in Table 10.

Subset	N	Ours	Judge	Δ	95% CI	p
All	500	63.86	58.67	+5.19	[2.88, 7.50]	< 0.0001
Faith.	225	59.64	53.67	+5.97	[2.24, 9.68]	0.0018
Fact.	275	67.32	62.76	+4.56	[1.64, 7.48]	0.0023

Table 10: Case-level paired-bootstrap comparison between MedF²Route and VLM-as-Judge.

For the full benchmark and both hallucination subsets, the 95% confidence intervals of the Macro-F1 differences exclude zero, with all p -values below 0.01.

C.3 Medical NLI Baseline and Aggregation Analysis

Medical NLI baseline. We additionally compare MedF²Route with PubMedBERT-MedNLI, a PubMedBERT model (Gu et al., 2021) fine-tuned on MedNLI (Romanov and Shivade, 2018). Both methods use the same triplet extraction, evidence routing, and sentence-level aggregation, with only the verification module changed. For each text-verifiable triplet, its textualized form is used as the hypothesis, while the routed patient text or medical-knowledge evidence is used as the premise. Because PubMedBERT-MedNLI does not support image inputs, both methods are evaluated on the same 1,767 text-verifiable sentences from all 500 cases.

Method	Hall. P	Hall. R	Hall. F1	Acc.	Macro-F1
PubMedBERT	34.71	73.80	45.01	54.01	49.38
MedF ² Route	46.99	75.40	55.39	70.70	66.16

Table 11: Comparison with PubMedBERT fine-tuned on MedNLI on 1,767 text-verifiable sentences.

As shown in Table 11, MedF²Route improves hallucination F1 and macro-F1 by 10.38 and 16.78 percentage points, respectively, on the text-verifiable subset.

Case-averaged and instance-averaged evaluation. The main experiments compute Macro-F1 within each case and then average across cases. We additionally report instance-averaged results by pooling all sentences within each target subset. Table 12 compares the two protocols.

Target	Case-averaged		Instance-averaged	
	Ours / Judge	Δ	Ours / Judge	Δ
Faith.	59.64 / 53.67	+5.97	61.37 / 57.64	+3.73
Fact.	67.32 / 62.76	+4.56	67.35 / 64.65	+2.70

Table 12: Case-averaged and instance-averaged Macro-F1. Each comparison reports MedF²Route / VLM-as-Judge.

MedF²Route achieves higher Macro-F1 for both hallucination types under both aggregation protocols, indicating that its advantage does not depend on case averaging.

C.4 InternVL Factuality Error Analysis

To better understand the factuality regression under the InternVL configuration, we analyze the 1,296 sentences from the 275 factuality-oriented cases. Here, factuality-oriented cases contain a targeted factuality perturbation, while evaluation is performed over all response sentences. We focus on the 196 sentences correctly classified by VLM-as-Judge but misclassified by MedF²Route and categorize their primary error sources.

Primary error source	Count	Percentage
Backbone capability limitation	137	69.90%
KG coverage gap	28	14.29%
Insufficient retrieved evidence	18	9.18%
Extraction or routing error	11	5.61%
Ambiguous or other	2	1.02%
Total	196	100.00%

Table 13: Error analysis of the 196 InternVL factuality disagreement cases.

As shown in Table 13, 137 of the 196 errors (69.90%) arise from the InternVL-based verifier incorrectly rejecting supported triplets. In comparison, KG coverage gaps and insufficient retrieved evidence jointly account for 46 cases (23.47%). This analysis indicates that the observed InternVL regression is driven primarily by backbone-specific verification limitations, while knowledge coverage and retrieval are secondary sources of error.

In addition, as analyzed in the Section.3.3, a cross-granularity comparison across the three evaluation series in Table.1 reveals that decomposing responses into finer units (e.g., claims or triplets) inherently degrades factuality detection performance. This degradation occurs because finer fragmentation isolates statements from their broader context, increasing the difficulty of standalone factual verification. This observation provides a vital insight for future research: rather than blindly pursuing finer granularity, it is imperative to dynamically align the splitting granularity with the contextual reasoning capacity of the detection algorithm.

D Cost Analysis

We further report the inference cost of different methods under the GPT-5.4-nano backbone in Table 14. For each method, we report macro-F1, the average number of API calls, and average token usage over the evaluated set. MedF²Route introduces additional inference overhead compared with direct sentence-level judgment, mainly due to graph construction, evidence routing, and fine-grained verification. Nevertheless, the cost remains within a practical range while providing more structured and interpretable hallucination detection.

Method	Macro F1	Avg. Calls	Avg. Tokens
VLM Judge	0.5867	4.76	3100.24
SelfCheckGPT	0.5196	23.06	6808.72
FActScore	0.3709	23.80	34203.65
UNIHD	0.3552	6.00	18424.79
FactSelfCheck	0.1869	14.77	20634.39
MedF²Route	0.6386	12.92	16735.96

Table 14: Cost analysis of different methods. We report macro-F1, average API calls, and average token usage over the evaluated set.

E Prompt Templates

The prompt templates for query graph construction and response graph construction are shown in Fig-

ures 5 and 6, respectively. The prompt templates for response triplet routing and image evidence routing are shown in Figures 7 and 8, respectively. The prompt template for sufficiency-guided knowledge subgraph expansion is shown in Figure 9. The prompt templates for image-grounded faithfulness verification, text-grounded faithfulness verification, and knowledge-grounded factuality verification are shown in Figures 10, 11, and 12, respectively.

Prompt template for query graph construction

Extract clinically meaningful triples from the following clinical case input.

Clinical case input:

{{case_text}}

Task:

Build the query-side clinical graph by extracting structured triples from the input case. The triples should capture patient-specific evidence explicitly stated in the case and preserve the context needed for later evidence routing and verification.

Rules:

Extract only information explicitly stated in the input. Do not add external medical knowledge, diagnostic assumptions, or unstated facts. Each triple should describe one clinically meaningful relation using concise subject, relation, and object fields. Preserve clinically important modifiers, including negation, severity, duration, certainty, location, laterality, and imaging descriptions when they affect the meaning. Extract image findings only when they are described in the input text, and do not treat the mere availability of an image as a clinical fact. If the input contains a diagnostic question, include the question target without predicting the answer.

Output format:

```
[  
  ["subject", "relation", "object"],  
  ["subject", "relation", "object"]  
]
```

Output raw valid JSON only. Do not wrap the output in markdown code fences.

Figure 5: Prompt template for constructing the query-side clinical graph from the clinical case input.

Prompt template for response graph construction

Extract clinically meaningful triples from the following clinical response.

The response has been split into numbered sentences:

{{numbered_sentences}}

Task:

Build the response-side clinical graph by extracting structured triples from the response. The triples should capture the main patient-specific statements, imaging descriptions, diagnostic conclusions, differential diagnoses, and clinical reasoning expressed in the response. Each triple must be allocated to the sentence in which the corresponding claim is mainly stated.

Rules:

Extract only information explicitly stated in the response. Do not add external medical knowledge, diagnostic assumptions, or unstated facts. Each triple should describe one clinically meaningful relation using concise subject, relation, and object fields. Preserve clinically important modifiers, including certainty, likelihood, negation, location, laterality, severity, duration, and imaging descriptions when they affect the meaning. When the response explicitly attributes a finding or claim to an image or clinical test, preserve that evidence source in the triple. Resolve simple local references when necessary, such as “this lesion”, “this finding”, “this mass”, or “it”. Keep each triple clinically meaningful and avoid unnecessary fragmentation of a single claim.

Output format:

```
[  
  ["subject", "relation", "object", sentence_id],  
  ["subject", "relation", "object", sentence_id]  
]
```

Output raw valid JSON only. Do not wrap the output in markdown code fences.

Figure 6: Prompt template for constructing the response-side clinical graph from numbered response sentences.

Prompt template for response triplet routing

You are a medical response-triplet router.

Query graph triples:

{{query_triples}}

Response triples with sentence ids:

{{response_triples}}

Task:

Classify each response triple as either query or knowledge. Use the query graph as case context to decide whether each response triple mainly follows patient-specific evidence in the current case or expresses knowledge-dependent clinical reasoning. A query triple describes a patient-specific or case-specific fact, while a knowledge triple expresses a diagnostic conclusion, differential diagnosis, etiological explanation, clinical interpretation, or other knowledge-dependent reasoning claim.

Rules:

Use query for current-case findings and patient-specific evidence, including symptoms, history, examination findings, laboratory results, procedures, and directly described imaging findings or attributes. Use knowledge when the triple mainly links the clinical presentation to a diagnosis, disease, etiology, differential diagnosis, or medical interpretation. Do not judge whether the triple is correct or hallucinated at this stage. Return one label for every response triple.

Output format:

```
[  
[sentence_id, "subject", "relation", "object", "query"],  
[sentence_id, "subject", "relation", "object", "knowledge"]  
]
```

Output raw valid JSON only. Do not wrap the output in markdown code fences.

Figure 7: Prompt template for partitioning response triplets into query-following and inference subgraphs.

Prompt template for image evidence routing

You are a medical image evidence router.

Candidate response triples from the query-side branch:

{{candidate_triples}}

Task:

Select the response triples that should be verified against the medical image. A triple should be selected only when it describes a direct imaging or visual observation in the current case.

Rules:

Select triples that explicitly describe findings shown, revealed, demonstrated, or seen on a medical image or imaging examination, such as CT, MRI, X-ray, ultrasound, radiograph, scan, or ECG. Direct visual observations may include visible lesions, masses, anatomical location, laterality, contour, border, density, signal, enhancement, calcification, size, extent, adjacency, or spatial relationships. Do not select symptoms, history, physical examination findings, laboratory results, treatments, prognosis, diagnoses, differential diagnoses, etiologies, mechanisms, or medical reasoning claims unless the response explicitly states that they are visually shown by the image or imaging examination. This is only an evidence-routing step; do not judge whether the triple is true or hallucinated.

Output format:

```
{
  "selected_indices": [0, 2],
  "reason": "short reason"
}
```

Output raw valid JSON only. Indices are 0-based over the candidate response triple list. Do not wrap the output in markdown code fences.

Figure 8: Prompt template for routing query-side response triples to the image-grounded verification branch.

Prompt template for sufficiency-guided knowledge subgraph expansion

You are a medical knowledge subgraph expansion assistant.

Knowledge-related response triples to verify:

{{response_triples}}

Current selected nodes:

{{selected_nodes}}

Current selected subgraph triples:

{{selected_subgraph_triples}}

Frontier candidate nodes:

{{candidate_nodes}}

Task:

Decide whether the current selected subgraph is sufficient for knowledge-grounded factuality verification, or whether one additional frontier node should be added. The goal is to improve relational coverage for the knowledge-related response triples while keeping the reference subgraph compact.

Rules:

Stop expansion when the current subgraph already provides sufficient evidence for judging the relevant medical relations, or when adding another frontier node is unlikely to improve verification. Select one frontier node only if it is likely to provide useful supporting evidence, contradicting evidence, or a bridging relation for diagnoses, etiologies, mechanisms, differential diagnoses, or clinical interpretations. Do not select a node only because it is broadly related to the topic. Do not select isolated visual details, anatomical terms, laterality terms, devices, or report containers unless they are directly needed for a knowledge claim. Do not judge whether the response triples are hallucinated at this stage.

Output format:

```
{
  "decision": "stop|expand",
  "node_id": "selected node id or null",
  "reason": "short reason"
}
```

Output raw valid JSON only. If the decision is stop, set node_id to null. Do not wrap the output in markdown code fences.

Figure 9: Prompt template for sufficiency-guided multi-hop expansion of the medical knowledge reference subgraph.

Prompt template for image-grounded faithfulness verification

You are a medical image visual-support judge.

Input:

The original clinical image and multiple target triples with indices.

Task:

For each target triple, decide whether the core visual claim is supported or contradicted by the medical image.

Labels:

non_hallucination: the claim is visually supported by or consistent with the image, or the image quality and coverage do not allow a reliable contradiction judgment.

hallucination: the image clearly contradicts the claim, shows the opposite finding, or makes the claimed visual finding or attribute clearly absent.

Rules:

Use only the original image and judge the core visual claim rather than exact wording. Do not use clinical history, diagnosis priors, or outside medical knowledge. For visual attributes such as location, laterality, contour, density, signal, enhancement, size, or extent, allocate hallucination only when the attribute is clearly inconsistent with the image. If the image is insufficient for a reliable judgment, do not infer unsupported conclusions. Do not infer diagnosis, cause, treatment, prognosis, or clinical meaning. Return one result for every input triple.

Output format:

```
{
  "results": [
    {
      "index": 0,
      "label": "non_hallucination|hallucination",
      "visual_evidence": "short visible evidence or uncertainty",
      "reason": "short reason"
    }
  ]
}
```

Output raw valid JSON only. Do not wrap the output in markdown code fences.

Figure 10: Prompt template for image-grounded faithfulness verification of image-routed response triplets.

Prompt template for text-grounded faithfulness verification

You are a medical query-faithfulness checker.

Query-side graph triples:

{{query_triples}}

Response triples with sentence ids:

{{response_triples}}

Task:

Identify response triples that introduce patient-specific or case-specific facts not grounded in the query-side graph. Judge only case grounding, and do not evaluate whether the medical reasoning, diagnosis, or medical knowledge is correct.

Decision criteria:

A response triple is supported if its core case fact is directly stated, clearly entailed, faithfully reformulated, or extracted as a faithful subfact from the query-side graph. A response triple is unsupported only if it adds new patient-specific or case-specific information beyond the query-side graph.

Rules:

Treat faithful paraphrases, summaries, abbreviation changes, relation-name changes, and simple temporal or numeric reformulations as supported when the core clinical fact is unchanged. Do not use external medical knowledge to infer missing case facts. Do not flag diagnosis, differential diagnosis, or general medical reasoning unless it introduces a new patient-specific fact. Mark a triple as unsupported only when it clearly adds new current-case information, such as a new finding, location, size, severity, extent, invasion, compression, obstruction, or patient-specific relation or mechanism. Return every unsupported response triple separately with its sentence id and a short reason. If no unsupported triple is found, return an empty list.

Output format:

```
{
  "unsupported_query_faithfulness": [
    {
      "sent_id": 1,
      "triple": ["subject", "relation", "object"],
      "reason": "short reason"
    }
  ]
}
```

Output raw valid JSON only. Do not wrap the output in markdown code fences.

Figure 11: Prompt template for text-grounded faithfulness verification of response triplets routed to the query graph.

Prompt template for knowledge-grounded factuality verification

You are a medical knowledge factuality checker.

Query graph triples:

{{query_triples}}

Reference subgraph:

{{reference_subgraph}}

Response graph triples:

{{response_triples}}

Current triples to verify:

{{current_triples}}

Task:

For each current triple, judge only its medical knowledge factuality. Focus on whether the medical relation, diagnosis, cause, interpretation, or factual statement is medically correct in the given context. Do not judge whether a patient-specific fact is newly introduced or absent from the query.

Labels:

non_hallucination: the medical claim is supported by the reference subgraph or established medical knowledge, and is medically reasonable in the given case context.

hallucination: the medical claim is medically false, contradicted by the reference subgraph or established medical knowledge, or not medically justified in the given case context.

Rules:

Use the reference subgraph first when it is relevant, and treat it as supportive only when its relation or path matches the medical relation being checked. Do not regard related entities alone as sufficient support for a specific diagnosis, cause, mechanism, likelihood, or interpretation. If the reference subgraph is incomplete or irrelevant, use established medical knowledge together with the case context. Do not label a claim as non_hallucination only because it is loosely possible. Judge each triple independently and provide a brief reason.

Output format:

```
{
  "results": [
    {
      "triple": ["subject", "relation", "object"],
      "label": "non_hallucination|hallucination",
      "reason": "short reason"
    }
  ]
}
```

Output raw valid JSON only. Do not wrap the output in markdown code fences.

Figure 12: Prompt template for knowledge-grounded factuality verification of response triplets routed to the reference knowledge subgraph.