REL-FIX: Scene Graph Guided Fine-Grained AI Correction of Relationship Hallucinations in Vision-Language Models

Submitted By

Jafrin Alam Prima
ID: 221-35-889
Department of Software Engineering
Daffodil International University

Supervised By

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Daffodil International University

This Thesis report has been submitted in fulfilment of the requirements for the Degree of Bachelor of Science in Software Engineering.

Fall 2025

APPROVAL

APPROVAL

This thesis titled on "**REL-FIX: Scene Graph Guided Fine-Grained AI Correction of Relationship Hallucinations in Vision-Language Models**", submitted by **Jafrin Alam Prima (ID: 221-35-889)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. Imran Mahmud
Professor & Head
Department of Software Engineering
Faculty of Science and Information Technology Daffodil International University

Internal Examiner 1

Afsana Begum
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Nadira Islam
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

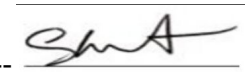
External Examiner

Md Manowarul Islam
Professor
Department of Computer Science and Engineering
Jagannath University, Bangladesh

SUPERVISOR's DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

Supervised By



Md. Shohel Arman
Assistant Professor
Department of Software Engineering
Daffodil International University

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

Prima

_____

(Student's Signature)

Full Name : Jafrin Alam Prima

ID Number : 221-35-889

Date :

ACKNOWLEDGEMENT

First of all, I want to thank Almighty God for His divine favor, enabling me to complete my undergraduate thesis. It is driven by my interest in understanding why vision language models generate incorrect relationships and how structured knowledge, such as scene graphs, can guide models toward more reliable and grounded reasoning.

I would express my deepest sense of thanks to my thesis supervisor, Md. Shohel Arman, Assistant Professor of the Department of Software Engineering, for his guidance and all-out support given to me in the entire research work. His intellectual advice and insights have largely shaped this work, and his commitment without wavering has been the inspiration for me to explore the limits of my knowledge and skills.

I would like to express my thanks to Dr. Imran Mahmud, Head of the Department of Software Engineering, Faculty of Science and Information Technology, and my other professors, faculties, and personnel for their kind cooperation and support in the successful completion of my work.

Great thanks to my parents and friends for continuous support, patience, and understanding during these years. Their support has been my stronghold, letting me have the opportunity to weather the turmoil that I went through.

Finally, I would like to acknowledge my batchmates and fellow members of DIU for their kind cooperation and consolation, which helped me reach this goal, as well as the organizations providing data and resources necessary for the research work. Without them, this work was not possible.

ABSTRACT

Vision language models are powerful, but they can produce text that looks plausible yet is not grounded in the image. In particular, relationship hallucinations, where a model describes an incorrect relation between two correctly identified objects, are especially pernicious for trust and downstream use. This thesis presents REL-FIX, a training free, scene graph guided framework designed to detect and correct relation level hallucinations in small vision language models without requiring expensive retraining or large scale judges.

REL-FIX works by decomposing long form VLM outputs into subject, relation, object triplets, diagnosing hallucinations at the triplet level against ground truth scene graphs from the Tri HE benchmark, and then applying a two stage correction mechanism that generates candidate relations constrained by the scene graph and verifies them with a lightweight LLM judge. The pipeline emphasizes low resource reproducibility by using a compact generative VLM, Qwen2 VL 2B Instruct, together with accessible LLM judges such as Mistral 7B and a commercial Gemini variant for cross checking.

Experiments on the 300 image Tri HE split demonstrate that REL-FIX substantially lowers relation hallucination rates while remaining cost effective. Using the Gemini judge, question level hallucination rate fell from 0.421 to 0.263 and relation hallucination from 0.341 to 0.196. With the Mistral judge the framework still reduced errors meaningfully, showing that open source judges can enable practical, low resource correction. Analysis shows that REL-FIX is particularly effective at repairing relational errors, with smaller but positive effects on object level errors. Remaining challenges include reliance on high quality scene graphs and triplet extraction noise, which are discussed along with directions for extending the method to automatically inferred scene graphs and multi hop reasoning.

In sum, REL-FIX offers a modular, training free approach to improving factual consistency of small VLM outputs. It demonstrates that fine grained, scene graph guided correction can make small models significantly more reliable for tasks that require precise relational understanding.

Keywords — vision language models, hallucination, relation correction, scene graph, triplet extraction, small models, REL FIX, Tri HE, Qwen2 VL, Mistral, Gemini.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Vision-Language Models (VLMs) have demonstrated revolutionary capabilities in tasks that require reasoning across visual and textual domains, such as image captioning and visual question answering (VQA). However, a critical challenge undermining their reliability is **hallucination**, the generation of content that is factually unsupported by the input image. While early research primarily focused on object hallucination (describing objects that do not exist), recent evidence indicates that **relation hallucination** (incorrectly describing the relationship between two correctly identified objects) is often more frequent and more damaging to the trustworthiness of model outputs.

This thesis addresses this critical reliability gap by introducing a novel framework, **REL-FIX**, designed for the precise evaluation and mitigation of relationship hallucinations. Specifically, REL-FIX focuses on enabling this fine-grained correction within **low-resource environments** by utilizing small-scale VLMs and cost-effective large language models (LLMs) for diagnostic judgment.

1.2 BACKGROUND

The core problem of VLM hallucination stems from a misalignment between the visual and linguistic components of the model, often causing the linguistic prior (or language bias) to override the visual evidence. When generating long, descriptive responses, VLMs often weave together supported facts with unsupported, hallucinated statements, making manual inspection tedious and unreliable.

To address this complexity, the **Unified Triplet-Level Hallucination Evaluation (Tri-HE)** framework was developed. Tri-HE revolutionized hallucination assessment by decomposing long-form answers into factual triplets (subject, relation, object) and judging each triplet independently against the image's ground-truth Scene Graph. This fine-grained approach confirmed that relation hallucination is a significant and neglected problem.

1.3 PROBLEM STATEMENT

Despite the diagnostic power of the Tri-HE framework and similar methods, two key limitations prevent

their widespread adoption and advancement in real-world, scalable applications:

High Computational Cost of Evaluation: The Tri-HE evaluation methodology critically depends on powerful, proprietary LLM judges (such as GPT-4 or Llama-3.3-70B) for reliable triplet extraction and factual judgment. This reliance introduces substantial computational expense and limits the reproducibility of the research in environments with limited resources.

Lack of Effective, Cost-Efficient Mitigation for Small Models: Existing hallucination mitigation strategies are typically designed for or evaluated solely on large, high-capacity VLM architectures. There is a demonstrable gap in developing a training-free, highly effective mitigation pipeline specifically tailored for **small Vision-Language Models** (such as Qwen2-VL-7B-Instruct) where computational efficiency is paramount.

Therefore, the problem this thesis addresses is the need for a **cost-efficient, fine-grained AI correction pipeline** capable of selectively and accurately mitigating relationship hallucinations in small VLMs without requiring expensive retraining or reliance on large-scale diagnostic judges.

1.4 RESEARCH GAPS

This research explicitly addresses the following gaps identified in the existing literature:

- **Gap in Evaluation Resource Efficiency:** No prior work has established a successful, high-correlation substitution for large, expensive LLM judges (like GPT-4) within the triplet-level hallucination evaluation framework that is suitable for low-resource research settings.
- **Gap in Mitigation Targeting and Specificity:** While the base mitigation method (Triplet Description + Eyes-Close) showed promise, it did not utilize the highest-fidelity ground-truth visual information available—the **Scene Graph** itself—to explicitly constrain the corrective description.
- **Gap in Small Model Validation:** The viability of running a complete, high-fidelity hallucination detection and mitigation pipeline entirely on small, efficient VLM architectures remains underexplored.

1.5 OBJECTIVES

The primary objective of this research is to develop and validate the **REL-FIX** framework to achieve cost-efficient, scene graph-guided correction of relationship hallucinations. The specific objectives are:

1. To successfully reproduce the triplet-level hallucination evaluation using the small **Qwen2-VL** LVLM and employ the **Gemini** LLM as a cost-effective, high-fidelity diagnostic judge.
2. To analyze and quantify the relationship hallucination rates of the target LVLM on the Tri-HE

benchmark, validating the hypothesis that errors are concentrated on less-frequent object relationships.

3. To design and implement **REL-FIX**, a novel, training-free, two-stage prompting mechanism that leverages the ground-truth **Scene Graph** to generate factually constrained descriptive hints.
4. To demonstrate a statistically significant reduction in both overall and, most importantly, **relation hallucination rates** in the small LVLM when utilizing the REL-FIX framework, proving its superior effectiveness over existing training-free baselines.

1.6 MOTIVATION OF THE STUDY

The motivation for this study is twofold: advancing the reliability of VLMs and promoting accessible research.

Enhancing Trustworthiness: By focusing on fine-grained correction of relation errors, REL-FIX directly tackles the most subtle and factually misleading type of hallucination, which is critical for deploying VLMs in high-stakes reasoning applications (e.g., medical diagnostics, automated reporting) where relationship accuracy is vital.

Promoting Accessible Research: The core adaptation of the evaluation pipeline (using a small VLM and a cost-effective LLM judge) democratizes advanced VLM research. By proving that high-quality, fine-grained analysis and mitigation do not require prohibitive computational resources, this work enables a wider community of researchers and developers operating in low-resource environments to focus on VLM reliability.

1.7 SUMMARY

Chapter 1 has introduced the problem of VLM hallucination, establishing the criticality of relationship errors. It has detailed the limitations of current high-cost evaluation and mitigation strategies, defining the core problem statement and research gaps that **REL-FIX** is designed to fill. Finally, the chapter outlined the specific objectives and motivation for developing a cost-efficient, scene graph-guided correction framework for small-scale Vision-Language Models. The following chapter provides a comprehensive review of the relevant literature.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Hallucination in vision-language models (VLMs) occurs when the generated text is not grounded in the provided image or input context. This reduces factual reliability and interpretability, particularly in reasoning tasks that require consistent visual-textual alignment. Detecting hallucination in short, closed-form tasks such as yes/no or visual question answering is straightforward, but identifying it in long descriptive responses remains difficult because multiple factual statements may be mixed together.

Most prior studies assess hallucination at the sentence or answer level and primarily target object-level errors, leaving relation-level hallucination underexplored. Benchmarks such as Hal-Eval (Jiang et al., 2024), Reefknot (Zheng et al., 2024), and R-Bench (Mingrui Wu et al., 2024) evaluate hallucinations through template-based or short-answer formats, which restrict the scope of analysis and introduce evaluation bias.

The study Unified Triplet-Level Hallucination Evaluation for Large Vision-Language Models (Tri-HE) (Wu et al., 2024) addresses these limitations by introducing a unified framework that evaluates both object and relation hallucinations using factual triplets of the form (subject, relation, object) extracted from LVLM responses. Each triplet is compared against ground-truth scene graphs and judged by an NLI model or a large language model. Tri-HE provides fine-grained analysis across vision-language tasks and achieves strong alignment with human judgment. However, the use of GPT-based judges and large 7B-parameter models makes the approach computationally expensive.

This research extends the Tri-HE framework by reproducing its triplet-level evaluation using smaller multimodal and text-only models. The goal is to enable cost-efficient, reproducible hallucination detection and mitigation suited for low-resource environments.

2.2 PREVIOUS LITERATURE

Mingrui Wu et al. (2024), in Evaluating and Analyzing Relationship Hallucinations in Large Vision-Language Models, provide a detailed analysis of how large vision-language models produce incorrect relational statements between visual entities. Their study applies structured relational visual-question-answering tasks to quantify errors within subject–relation–object triplets. The results indicate that models rely predominantly on linguistic priors rather than visual evidence, and that relational

hallucination occurs frequently, particularly within perceptive or spatial relations.

Junjie Wu et al. (2024), in Unified Triplet-Level Hallucination Evaluation for Large Vision-Language Models, establish a comprehensive evaluation framework known as Tri-HE for the detection of hallucinations in long descriptive outputs. The framework extracts subject–relation–object triplets from generated responses and evaluates each triplet through a large-language-model or natural-language-inference judge to determine factual support. The findings confirm that triplet-level analysis enables systematic hallucination detection, although the reliance on GPT-level models results in substantial computational cost.

Chaoya Jiang et al. (2024), through Hal-Eval: A Universal and Fine-Grained Hallucination Evaluation Framework for Large Vision Language Models, present a unified method for fine-grained hallucination scoring at sentence and span levels. The framework provides consistency across datasets and tasks but remains limited to short textual outputs and does not extend to the triplet-level factual assessment required for longer responses.

Xiao et al. (2025), in Detecting and Mitigating Hallucination in Large Vision Language Models via Fine-Grained AI Feedback, describe a system in which auxiliary detectors identify hallucinated spans and supply corrective feedback during text generation. The approach yields measurable improvements in local factual accuracy; however, its dependence on large annotated corpora and high-capacity teacher models restricts replication in low-resource environments.

Kening Zheng et al. (2024), in Reefknot: A Comprehensive Benchmark for Relation Hallucination in Multimodal Large Language Models, introduce a relation-focused benchmark that evaluates a range of mainstream multimodal models including GPT-4o, Qwen-VL, MiniGPT-4, and DeepSeek-VL. The study employs yes/no, multiple-choice, and generative VQA formats and implements a Detect-then-Calibrate procedure combining detection with decoder calibration based on token-level confidence. The analysis reveals substantial uncertainty in model predictions during hallucination events and low accuracy for perceptive relations, while relatively better performance on cognitive relations often reflects reliance on commonsense language priors rather than genuine visual understanding.

Jiaming Li et al. (2025), in Mitigating Hallucination for Large Vision Language Models by Inter-Modality Correlation Calibration Decoding, formulate a decoding mechanism that adjusts cross-modal correlation weights at inference to suppress hallucination. The method lowers hallucination frequency without retraining but is evaluated only on large-scale models, leaving its applicability to smaller architectures uncertain.

Moon Ye-Bin et al. (2024), in BEAF: Observing Before-After Changes to Evaluate Hallucination in Vision-Language Models, present a comparative evaluation design measuring model outputs before and

after specific interventions. This approach effectively quantifies performance change following mitigation efforts but is confined to short textual descriptions and simple captioning tasks.

Xiangxiang Chu et al. (2024), through MobileVLM V2: Faster and Stronger Baseline for Vision Language Models, develop an efficient vision-language baseline demonstrating that compact multimodal models can maintain competitive accuracy while reducing computational demand. The study establishes MobileVLM V2 as a viable reference for experiments conducted under limited hardware capacity.

Jiacheng Ruan et al. (2025), in VLRMBench: A Comprehensive and Challenging Benchmark for Vision-Language Reward Models, design a benchmark framework that employs reward models as automated evaluators of multimodal output quality. The research confirms that such evaluators scale hallucination assessment effectively but remain dependent on extensive training data and compute resources.

Bei Yan et al. (2024), in Evaluating the Quality of Hallucination Benchmarks for Large Vision-Language Models, assess the consistency and validity of existing hallucination benchmarks. Their analysis exposes definitional inconsistencies and labeling noise within widely used datasets and recommends multiple judging mechanisms and human validation to strengthen benchmark reliability.

Xiyang Wu et al. (2024), through AutoHallusion: Automatic Generation of Hallucination Benchmarks for Vision-Language Models, outline a process for automated benchmark generation that enlarges dataset coverage. Although the approach enhances scalability, it introduces additional labeling errors, emphasizing the continuing need for manual verification in benchmark creation.

Ipula Rawte, Aryan Mishra, and Amit Sheth (2025), in Defining and Quantifying Visual Hallucinations in Vision-Language Models, articulate a structured taxonomy and measurement framework for visual hallucination. The study delineates object, attribute, and relation hallucination types and supplies quantifiable indicators suitable for systematic evaluation across models.

Ming Cheung (2025), in Hallucination Detection with Small Language Models, examines the capability of lightweight language models to perform hallucination detection. The results demonstrate that small models, when guided by precise prompting, can identify a substantial proportion of hallucinated statements. Although accuracy remains below that of GPT-scale models, the evidence confirms the practicality of small models for cost-efficient hallucination evaluation in resource-constrained settings.

2.3 SUMMARY

Earlier studies measured hallucination mainly in short, controlled tasks. Relation-specific works like Reefknot and R-Bench identified the issue but relied on large models and structured prompts. Frameworks such as Hal-Eval and BEAF improved evaluation consistency but not long descriptive

outputs. Tri-HE extended detection to long answers through triplet extraction and judgment but required GPT-level judges, limiting reproducibility.

Mitigation methods such as AI Feedback and IMCCD reduce hallucination but assume large model backbones. Compact models like MobileVLM V2 and Qwen2-VL show that smaller architectures can perform competitively, yet no prior work demonstrates a complete hallucination detection and mitigation pipeline using small models.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 INTRODUCTION

This chapter details the research methodology employed to investigate and mitigate relationship hallucinations in small Vision-Language Models (LVLMs). The core of this work is the development and evaluation of **REL-FIX: Scene Graph Guided Fine-Grained AI Correction of Relationship Hallucinations**. The methodology is structured to first establish a cost-efficient, low-resource evaluation baseline using a small Vision-Language Model and the proprietary Gemini LLM Judge, and then to implement and test the novel REL-FIX mitigation strategy. This chapter covers the data sources, the selection and configuration of the generative (Qwen2-VL-2B-Instruct) and diagnostic (Gemini) models, the two-stage prompting mechanism of REL-FIX, and the quantitative evaluation metrics used.

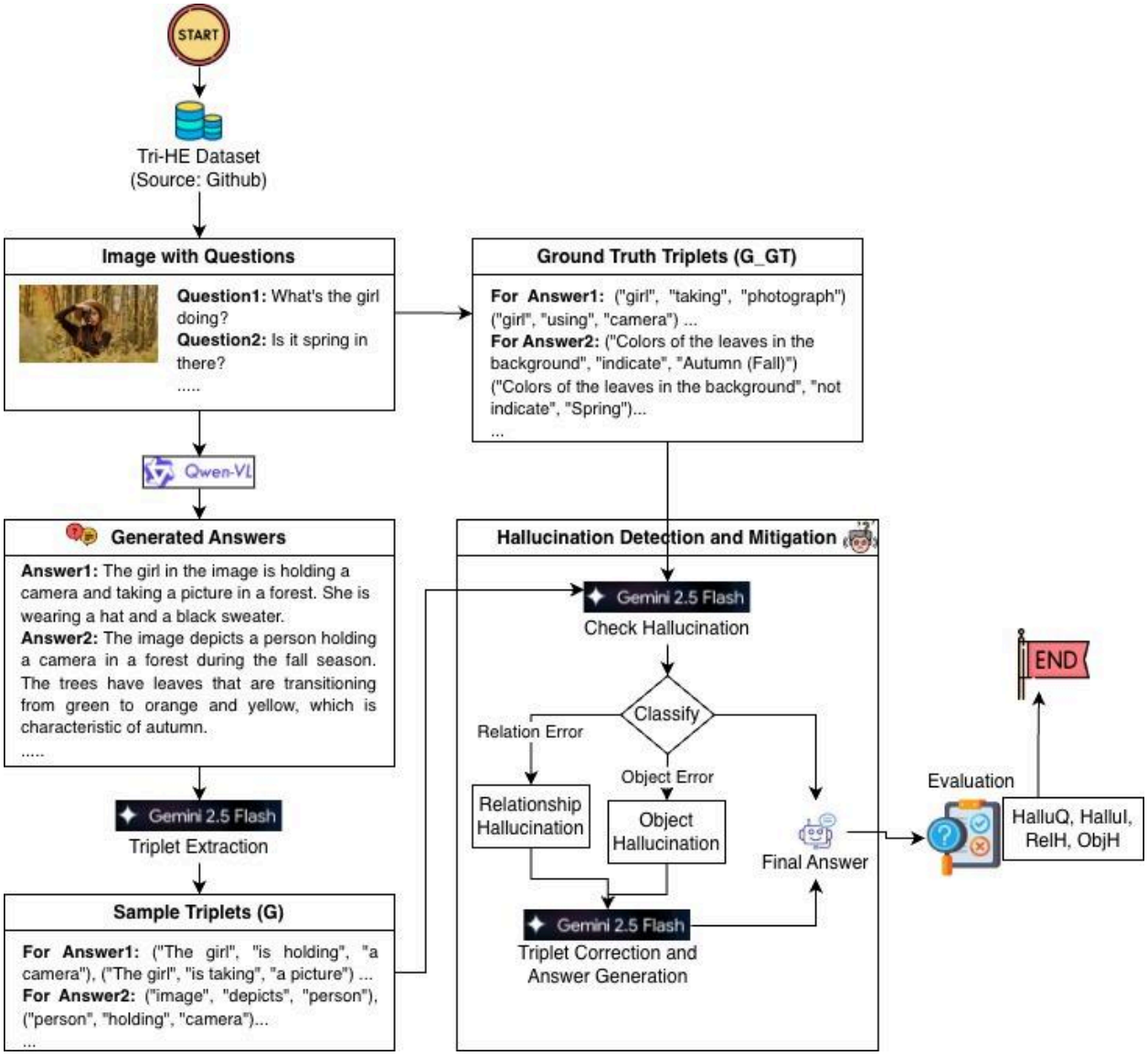


Figure 3.1: Methodology Diagram

3.2 DATA PREPARATION

3.2.1 Benchmark Selection

To support fine-grained relational hallucination analysis in REL-FIX: Scene Graph Guided Fine-Grained AI Correction of Relationship Hallucinations in Vision-Language Models, we adopt the Unified Triplet-Level Hallucination Evaluation (Tri-HE) benchmark as the core dataset used in our study.

Tri-HE is built upon images from the GQA (Visual Reasoning in the Real World) dataset (Hudson & Manning, 2019), originally sourced from COCO and Flickr. The associated image scene graphs are based

on a cleaner revision of Visual Genome, making them structurally compatible with the triplet-level hallucination formulation required for REL-FIX.

However, GQA's scene graphs are not uniformly reliable, as many contain incomplete or missing object–relation pairs required for accurate reasoning. To ensure structural completeness, Tri-HE applies the following filtering steps:

1. **Initial Filtering:** Only images with at least five object relations (edges) in their scene graphs are retained.
2. **Image Selection:** From these filtered images, 300 images are chosen based on two criteria:
    - Each image contains more than two related objects, ensuring sufficient relational structure.
    - Images must be visually clear, with all relevant details discernible.

Since the GQA VQA questions are already heavily used in LVLM pre-training, making them unusable due to data contamination risks, Tri-HE replaces them with newly constructed question–answer pairs generated using GPT-4V. Tri-HE is dialogue-formatted. For each selected image, GPT-4V produces 10 novel questions and answers requiring commonsense reasoning grounded in the visual content. In addition, GPT-4V generates reasoning triplets corresponding to each QA pair, expanding the original scene graphs with additional relational context.

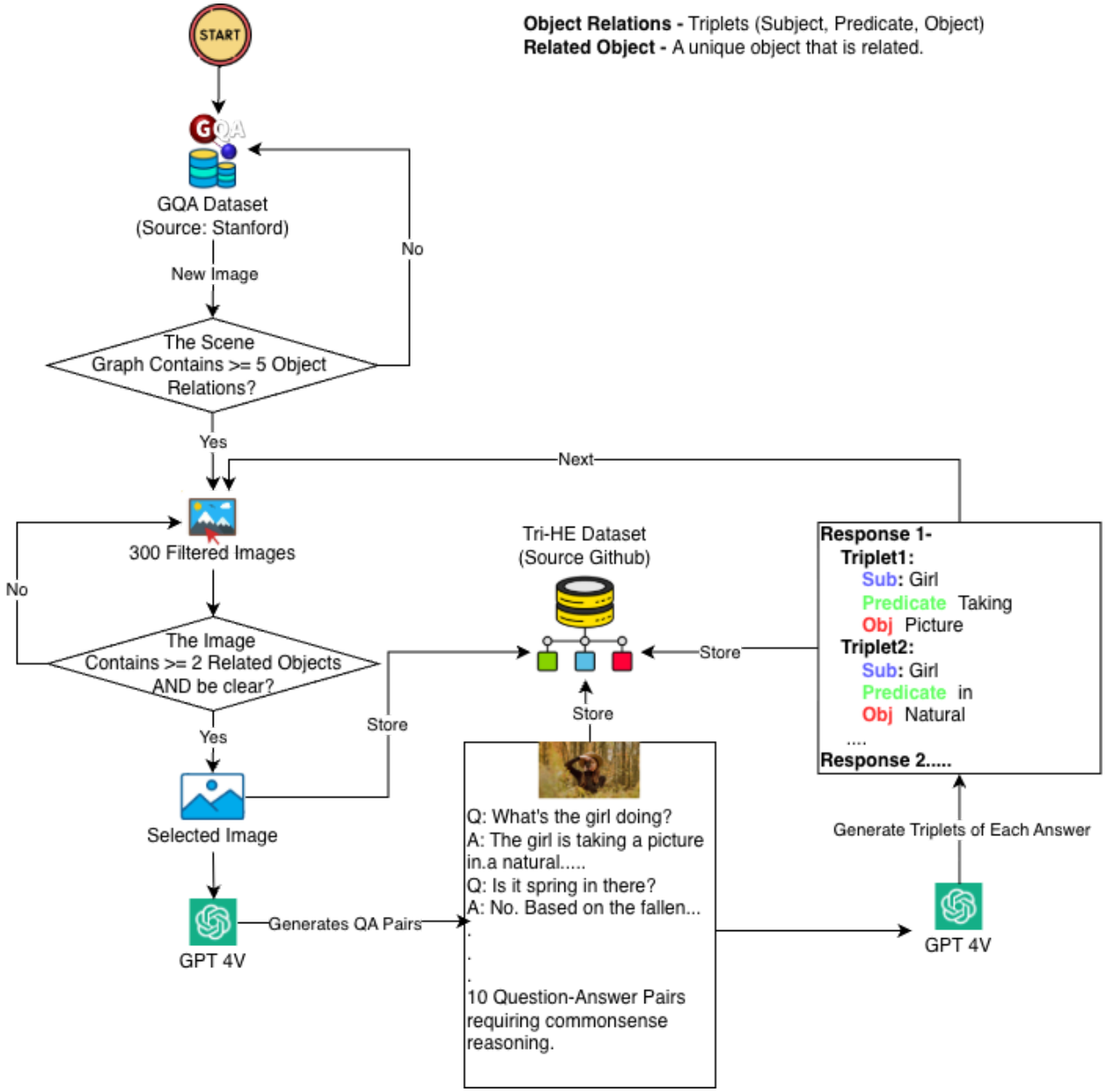


Figure 3.2: Tri-HE Dataset Construction Workflow.

In our REL-FIX study, Tri-HE serves two essential functions:

1. **Ground-Truth Reference:** The Tri-HE scene graphs provide explicit triplet-level ground truth used by our Gemini Judge to determine whether a model-generated triplet is hallucinated.
2. **Constraint Source for Stage 1 Mitigation:** REL-FIX uses the Tri-HE scene graph as the relational constraint input during Stage 1, guiding Qwen2-VL-2B-Instruct away from producing relation hallucinations.

3.2.2 Tri-HE Dataset Overview

The structure of the Tri-HE dataset used in REL-FIX is defined as follows. Each image or entry is represented as a JSON object keyed by a unique ID. Each record contains:

- *instance*: A list of question–answer–triplet dictionaries, where each dictionary contains:
  - *question*: A textual question about the image.
  - *answer*: The LVLM-generated answer grounded in the image content.
  - *triplet*: A list of triplets in the format (subject, predicate, object) extracted from the answer.
- *triplets*: A flattened list of all triplets in the image, combining those from instance.
- *all_object*: A list of all objects/entities present in the scene, including background and minor objects.
- *object*: A subset of *all_object* representing the main entities relevant to relational reasoning.

Sample Dataset: This structure explicitly captures multiple QA interactions per image, enabling REL-FIX to evaluate hallucinations across all dialogue turns for each visual scene.


```
{
  "2374892": {
    "instance": [
      {
        "question": "Why is the room illuminated?",
        "answer": "A lamp is turned on in the room.",
        "triplet": ["(Lamp, turned on, illuminating room)"]
      },
      {
        "question": "What time of day may it be?",
        "answer": "It may be evening or night based on the lighting.",
        "triplet": ["(Lamp, turned on, presence)"]
      }
    ],
    "triplets": [
      "(Lamp, turned on, illuminating room)",
      "(Lamp, turned on, presence)"
    ],
    "all_object": ["lamp", "desk", "man", "shirt"],
    "object": ["lamp", "desk", "man"]
  }
}
```


Figure 3.3: Tri-HE Sample Data Snippet.

3.2.3 Triplet Extraction

To perform fine-grained relational hallucination detection, REL-FIX requires transforming each

LVLM-generated answer into a structured set of factual assertions. For this purpose, we extract knowledge graph triplets from the long-form responses of Qwen2-VL-2B-Instruct.

Triplet Extraction Model: We employ Gemini to process the generated answer A_θ and extract a corresponding knowledge graph G_θ.

Extraction Prompt: A modified version of Tri-HE's original KG-extraction prompt is used to enforce a strict output format:

$$("subject", "predicate", "object") \tag{3.1}$$

This ensures compatibility between the extracted graph and the Tri-HE scene graph used in REL-FIX's hallucination judging.

3.3 MODEL SELECTION

3.3.1 Small Vision Language Model Selection (Generative Model)

For REL-FIX, the generative model tasked with producing descriptive answers for hallucination evaluation and correction is Qwen2-VL-2B-Instruct (Qwen-VL Team, 2024), a small-scale vision-language model designed for multimodal instruction following. This model generates long-form, image-grounded responses that are subsequently parsed into triplet knowledge graphs and evaluated against the Tri-HE benchmark (Wujun-jie et al., 2024) to identify object and relation hallucinations. Qwen2-VL-2B-Instruct was chosen for its efficiency and compatibility with our triplet-level hallucination analysis, providing outputs that can be reliably constrained

3.3.2 LLM Selection (Diagnostic Judge)

In REL-FIX, the diagnostic role of evaluating and judging hallucinated triplets is performed using two large language models, Gemini-2.5-Flash (Google DeepMind, 2024) and Mistralai/Mistral-7B-Instruct-v0.2 (Mistral AI, 2023). Gemini-2.5-Flash serves as a high-performing, commercially oriented judge, providing a benchmark for hallucination detection against state-of-the-art AI correction systems, while Mistral-7B-Instruct is a resource-efficient, open-source model used to evaluate REL-FIX's compatibility with small-scale, cost-effective LLMs. Both models are prompted to parse the VLM-generated answers into triplets and compare them against Tri-HE scene graphs (Wujun-jie et al., 2024), enabling fine-grained assessment of object and relation hallucinations while demonstrating REL-FIX's adaptability across different LLM resources.

3.4 MODEL DETAILS

The REL-FIX pipeline operates on small vision-language models (VLMs) to evaluate and correct relational hallucinations in their outputs. The overall approach combines three components: (1) a generative VLM for producing descriptive answers, (2) a diagnostic LLM judge to detect hallucinations, and (3) a scene-graph guided correction module that constrains and repairs relation errors.

3.4.1 Generative VLM

The descriptive answers used for hallucination evaluation are produced by Qwen2-VL-2B-Instruct (Qwen-VL Team, 2024), a lightweight 2B-parameter multimodal model designed for instruction-following tasks.

Configuration

- **Parameters:** 2.2B
- **Vision Encoder:** ViT-based encoder (Qwen2-VL vision tower)
- **Max Tokens Generated:** 1024
- **Temperature:** 0.1 (to suppress randomness)
- **Prompt Format:** Qwen2-VL instruct schema
- **Hardware Used:** Single GPU

For a given input image I and an instruction prompt P, the model generates a descriptive answer A_θ:

$$A_\theta = VLM_\theta(I, P) \tag{3.2}$$

Where θ denotes the learned parameters of the VLM. The output A_θ is subsequently parsed into a set of factual triplets G_θ to enable relational evaluation.

3.4.2 Triplet Extraction

To convert the generator's explanation into structured triples, REL-FIX uses two complementary text-only LLMs: one open-source and one commercial. This dual setup allows us to compare the stability of extractors while keeping the method accessible. These triplets consist of subject-object-predicate, which are extracted from the vision language model response.

$$T = f_{extract}(text_{response}) \tag{3.3}$$

Where T will return $\{(s_i, p_i, o_i)\}$. Here, s_i, p_i, o_i represents the subject, predicate and object relationship of the i^{th} response.

(a) Mistral-7B-Instruct-v0.2 (Primary Extractor): Mistral-7B-Instruct-v0.2 (Mistral AI, 2023) is used as the default extractor because it is small, reproducible, and cheap to run. It converts the generator output into subject–predicate–object triples using strict JSON formatting rules.

Configuration

- **Parameters:** 7B
- **Max Tokens:** 512
- **Temperature:** 0.0 (deterministic)
- **Output Format:** JSON list of triples

(b) Gemini-2.5-Flash (Secondary Extractor for Cross-Checking)

Gemini-2.5-Flash (Google DeepMind, 2024) is used as a second extractor to measure consistency between open-source and commercial extraction. It follows the same JSON schema. We use it for redundancy and for verifying whether extraction disagreements contribute to downstream hallucinations.

Configuration

- **Temperature:** 0.0
- **Max Tokens:** 512
- **Strict JSON enforced**

Triplets from both models are then aligned and normalized before entering evaluation.

3.4.3 Hallucination Diagnosis (Gemini Judge + Mistral Judge)

The extracted triples are checked against the ground truth from the Tri-HE benchmark (Wujun-jie et al., 2024). Decision Function is stated as below:

$$y = f_{judge}(triplets, prompt, G_t) \tag{3.4}$$

Here G_t is the ground truth triplets of the response, and y is the relationship hallucination report.

REL-FIX uses two judging models:

(a) Gemini-2.5-Flash as the Commercial Judge

Gemini evaluates each predicted triple relative to the image and the ground-truth triple set.

Configuration

- **Temperature:** 0.0

b) Mistral-7B-Instruct-v0.2 as the Open-Source Judge

The same protocol is repeated using Mistral-7B to create an open-source diagnostic judge. This allows experiments where only small models are used end-to-end.

Configuration

- **Temperature:** 0.0

3.4.4 Correction Module

In the REL-FIX Correction Module, Gemini-2.5-Flash (Google DeepMind, 2024) serves as the primary diagnostic and verification model. For each triplet extracted from the small VLM's output, Gemini identifies hallucinations and flags erroneous relations. During the correction stage, it evaluates candidate replacement triplets generated with reference to the Tri-HE scene graph, ensuring that only verified, factually consistent relations are adopted. Gemini's deterministic and high-precision judgment enables REL-FIX to perform triplet-level corrections without retraining the VLM, maintaining alignment with the ground-truth relational structure while efficiently handling relation, object, and attribute hallucinations.

3.5 IMPLEMENTATION OF REL-FIX FRAMEWORK

The REL-FIX framework is implemented as a training-free, modular pipeline that enables triplet-level evaluation and correction of hallucinations in small vision-language models (VLMs). It is composed of four main stages: (i) image encoding and answer generation, (ii) triplet extraction, (iii) hallucination diagnosis, and (iv) triplet correction.

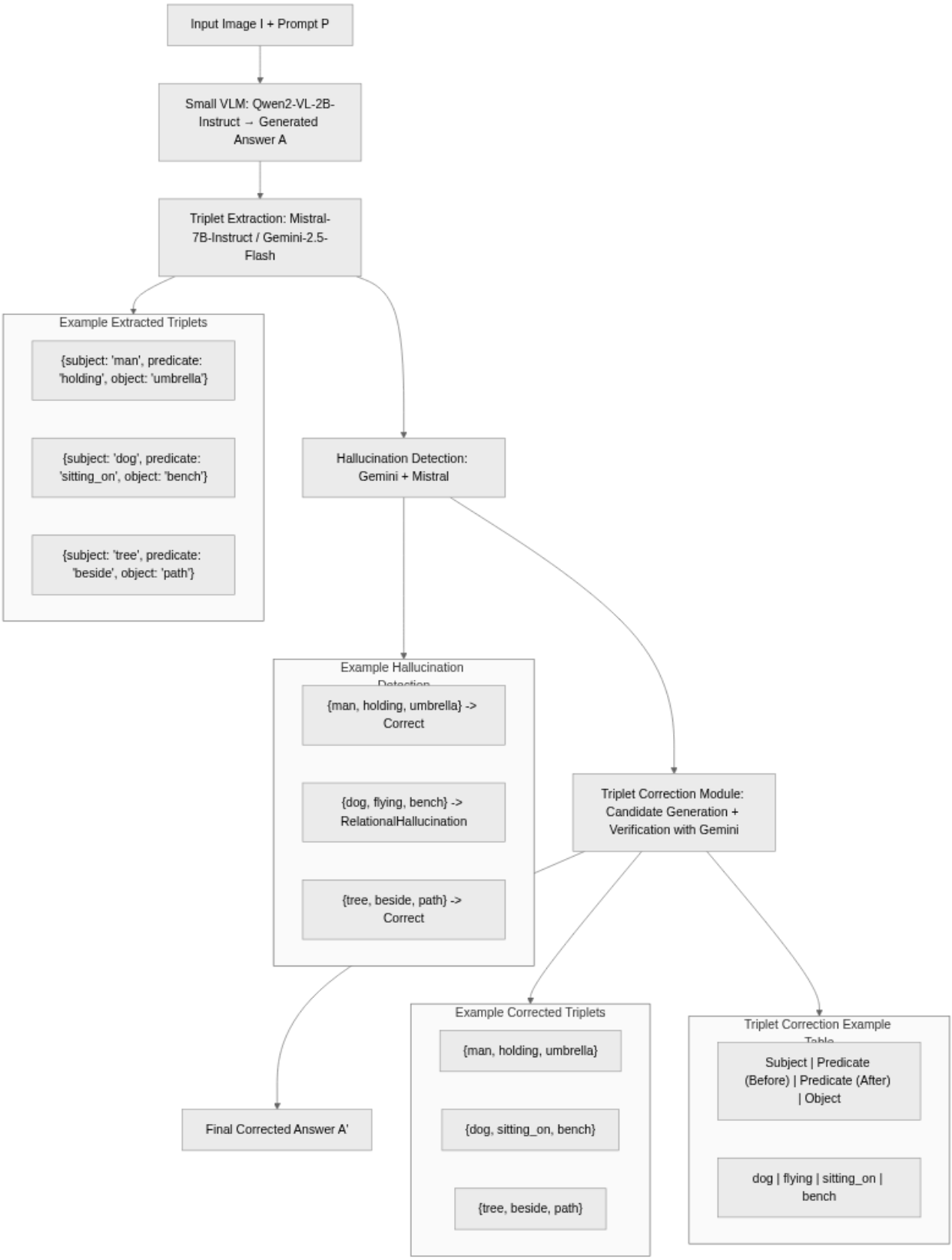


Figure 3.4: REL-FIX Workflow with Example.

3.5.1 Data and Preprocessing

- **Dataset**: Tri-HE (constructed from GQA; see Wujun-jie et al., 2024; Hudson & Manning, 2019). Use the publicly available `tri-he.json` as provided.
- **Filtering**: Remove QA pairs flagged by Tri-HE annotators as invalid. Keep fixed 300-image split unless ablation requires more images.
- **Splits**: Use recommended Tri-HE splits if available; otherwise use 60/20/20 (train/val/test) at the image level, ensuring question/answer pairs for an image aren't split across sets.
- **Normalization**: Tokenize entities; canonicalize frequent noun phrases (e.g., "man" ↔ "person"), lemmatize verbs for relation matching.

Triplet JSON schema:

```
{
  "image_id": "0001",
  "qa_id": "0001_01",
  "answer_text": "...",
  "triplets": [
    {"subject":"man", "relation":"holding", "object":"ball", "span": [0,5]},
    ...
  ],
  "annotations": {"source":"mistral", "confidence":0.92}
}
```

Figure 3.5: Triplet JSON schema

3.5.2 Descriptive Answer Generation

The small VLM **Qwen2-VL-2B-Instruct** (Qwen-VL Team, 2024) is used to generate long-form, image-grounded descriptions. If the input image is I along with the user prompt P, then the VLM-generated answer A_θ will follow the 3.2 equation. The output A_θ is then fed into the triplet extraction stage.

3.5.3 Triplet Extraction

Triplets are extracted using **Mistral-7B-Instruct** as the primary extractor, with optional cross-checking using **Gemini-2.5-Flash**. Extraction converts unstructured text into a set of triples defined as G_θ:

$$G_\theta = \{(s_i, p_i, o_i)\} \tag{3.4}$$

The LLM is prompted to parse the answer in JSON triple format (subject-predicate-object). These triplets are extracted from A_θ via the LLM.

$$G_\theta = f_{extract}(A_\theta) \tag{3.5}$$

3.5.4 Hallucination Detection

After obtaining the structured triplets from the small VLM through the extraction step, we proceed to identify which of these triplets deviate from the factual content of the image. In this stage, we rely on the **Tri-HE ground-truth scene graph** as the authoritative reference and compare each generated relation against it. Our goal here is not merely binary mismatch tagging but a **fine-grained classification** of hallucination types at the triplet level, following the formulation established in prior work on relation-centric hallucination evaluation (Wu et al., 2024).

To evaluate each predicted triple $t_i = (s_i, p_i, o_i)$, we present both the triplet and the image's ground-truth scene graph G to the judging LLM. We use **Gemini-2.5-Flash** as our primary diagnostic judge and **Mistral-7B-Instruct-v0.2** as a secondary judge for low-resource comparison. Both models receive identical evaluation prompts, ensuring a controlled comparison across models. The judge model returns one of three decisions:

- **Correct** — The relation holds in the image and is present in the ground-truth graph.
- **Relational Hallucination** — Both subject and object are valid objects in the image, but the predicted relation does not exist in GGG.
- **Object Hallucination** — Either the subject or object does not appear in the grounded scene graph.

Operationally, the detection function is:

$$Judge(t_i, G) = (Correct, if\ t_i \in G)\ or$$

$$(RelationalHallucination,\ if\ s_i,\ o_i \in Obj(G)\ r_i \notin Rel(G))\ \text{or}$$

$$or\ ObjectHallu\text{cination}) \tag{3.6}$$

We explicitly pass the entire scene graph in a normalized JSON format to the judge model so that the decision process is not influenced by missing context or ambiguous textual descriptions. The judging LLM is instructed to reason strictly over the provided scene graph, not over its own prior knowledge, ensuring that hallucination labels reflect mismatches with the image rather than disagreements with world knowledge.

Because hallucination identification can be noisy, we repeat the evaluation twice—once with Gemini and once with Mistral—to observe divergence patterns between a commercial LLM and an open-source model. Although Gemini consistently produces more stable decisions, we maintain both models in our pipeline so that researchers studying small-VLM correction in constrained environments can rely on the Mistral variant without assuming access to commercial APIs.

The output of this stage is a labeled set:

$$H = \{(t_i, l_i) \mid l_i \in \{Correct, RelationalHallucination, ObjectHallucination\}\} \tag{3.7}$$

where H contains both clean and hallucinated triplets. Only the hallucinated subset progresses to the correction module.

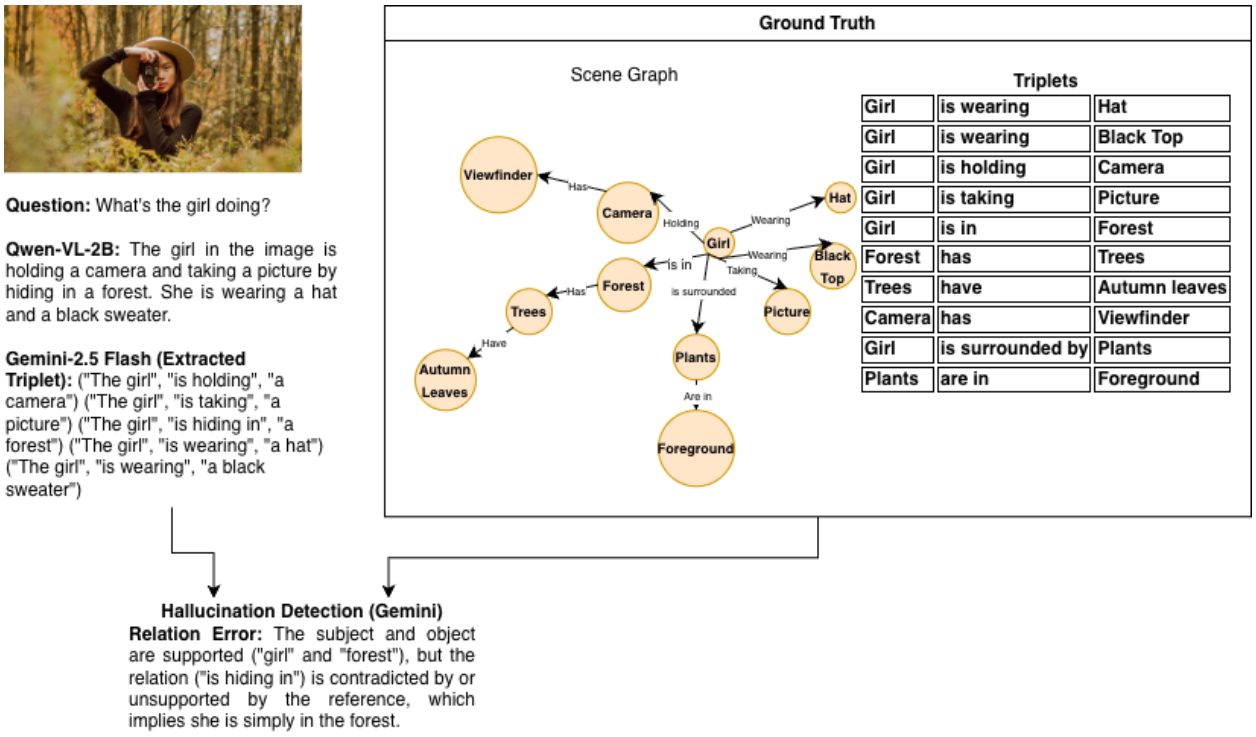


Figure 3.6: Hallucination Detection Via Scene Graph and Generated Triplets

3.5.5 Triplet Correction

Once hallucinated triplets have been identified, we proceed to correct them while preserving the factual integrity of the answer. For each triplet flagged as either a relational or object hallucination, we generate a set of candidate relations constrained by the ground-truth scene graph G_t and the visual entities present in the image. These candidates are then verified using **Gemini-2.5-Flash**, ensuring that only relations consistent with the scene graph are accepted. Formally, for a hallucinated triplet t_i and corrected relation r_i is selected as:

$$r_i^\star = \arg \max_{r' \in \text{Candidates}(s_i, o_i)} \mathbf{1}\{\text{verified}(s_i, r', o_i, G_{\text{GT}})\} \tag{3.8}$$

If no candidate passes verification, we retain the original triplet but mark it as uncorrected. After all hallucinated triplets are processed, the corrected graph is serialized back into natural language to produce the final answer A_Θ'. We log all candidate generations, verification results, and replacements to ensure

transparency and reproducibility of corrections.

3.6 EVALUATION METHODS

3.6.1 Hallucination Rate Metrics

To quantify the effectiveness of REL-FIX, we adopt standard hallucination metrics derived from the Tri-HE benchmark.

Question-Level Hallucination Rate ($Hallu_Q$) measures the proportion of hallucinated triplets for each question response. Formally, for a given question q with total triplets TT_q and hallucinated triplets HT_q, we define:

$$Hallu_Q(q) = \frac{HT_q}{TT_q} \tag{3.9}$$

This metric provides a direct measure of how frequently a small VLM generates incorrect relations or objects in response to a question.

Image-Level Hallucination Rate ($Hallu_I$) aggregates the question-level hallucination rates across all questions associated with an image I. Let Q_i denote the set of questions for image I, then:

$$Hallu_I(I) = \frac{1}{Q_i} \sum_{q \in Q_1} Hallu_Q(q) \tag{3.10}$$

This averaging ensures an unbiased, comparable score across images and allows for evaluating model performance at a higher granularity than individual questions.

3.6.2 Fine-Grained Analysis

Beyond overall hallucination rates, we perform **fine-grained evaluation** to assess the framework's

effectiveness in correcting specific types of errors.

- **Relation Hallucination Rate** ($RelH$) focuses exclusively on triplets labeled as relationally incorrect by the Gemini Judge. It is computed as the fraction of relationally hallucinated triplets over the total number of triplets in the evaluation set:

$$\text{RelH} = \frac{|\text{Triplets}_{\text{REL\_HALL}}|}{|\text{Triplets}_{\text{total}}|}. \tag{3.11}$$

  This metric directly reflects REL-FIX's capability to address the core problem of relationship hallucination in small VLMs.

- **Object Hallucination Rate** ($ObjH$) is computed analogously for triplets with object-level errors:

$$\text{ObjH} = \frac{|\text{Triplets}_{\text{OBJ\_HALL}}|}{|\text{Triplets}_{\text{total}}|}. \tag{3.12}$$

Together, these metrics allow us to evaluate REL-FIX both holistically (overall hallucination) and precisely (type-specific correction), ensuring that improvements are accurately characterized and reproducible.

3.7 SUMMARY

In this chapter, we presented the methodology underpinning REL-FIX, a scene-graph guided framework for detecting and correcting relational hallucinations in small Vision-Language Models (VLMs). Our approach begins with the selection and preparation of the **Tri-HE benchmark**, constructed from GQA images. Tri-HE serves as both the reference for hallucination detection and as the constraint source for guiding our VLM during the correction process.

We employed **Qwen2-VL-2B-Instruct** as the generative small VLM to produce descriptive, image-grounded answers. These outputs were then converted into structured knowledge graph triplets using a dual LLM extraction pipeline: **Mistral-7B-Instruct-v0.2** for low-resource, reproducible extraction, and **Gemini-2.5-Flash** for high-precision cross-checking. Triplet extraction ensures compatibility with the Tri-HE scene graphs and enables fine-grained evaluation of both relation and object hallucinations.

Hallucination detection leverages **Gemini and Mistral judges** to classify each triplet as correct, relationally hallucinated, or object-hallucinated. The output of this stage forms the input for the **Triplet Correction Module**, in which candidate relations are generated and verified against the Tri-HE scene graph to produce corrected answers without retraining the VLM.

Finally, we define **evaluation metrics** to quantify REL-FIX's performance. Question-level and image-level hallucination rates measure the overall reduction in errors, while fine-grained metrics, including relation and object hallucination rates, assess the framework's efficacy in addressing specific hallucination types. Together, these steps form a **cohesive methodology** that systematically captures, and corrects hallucinations, providing a transparent foundation for research on small VLM improvement.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter presents the quantitative results and analysis of the **REL-FIX** framework, addressing the research objectives defined in Chapter 1. The primary focus is on validating the cost-efficiency of the new evaluation pipeline (Qwen2-VL-2B-Instruct + Gemini/Mistral Judge) and quantifying the effectiveness of the **Scene Graph Guided** two-stage prompting mechanism in mitigating relation hallucination. The results are discussed in the context of the Tri-HE benchmark and compared against the model's baseline performance and existing training-free mitigation techniques.

4.2 QUANTITATIVE RESULTS

We evaluated REL-FIX on the 300-image Tri-HE split, comparing the **uncorrected small VLM baseline** with REL-FIX using either **Mistral-7B** or **Gemini-2.5-Flash** as the hallucination judge. We report four key metrics: Question-Level Hallucination Rate, Image-Level Hallucination Rate, Relation Hallucination Rate, and Object Hallucination Rate.

4.2.1 Fine-Grained Metric Outcomes

When we parse the improvements by error type, Table 1 presents the comparative figures for relation vs. object hallucination rates.

Table 1. Hallucination Metrics for Small VLM and REL-FIX

Method	$Hallu_Q$	$Hallu_I$	RelH	ObjH	Notes
Small VLM (Qwen2-VL-2B, baseline)	0.421	0.389	0.341	0.080	No correction
REL-FIX + Mistral Judge	0.308	0.292	0.218	0.074	Resource-friendly, open-source setup
REL-FIX + Gemini Judge	0.263	0.254	0.196	0.067	High-precision commercial judge

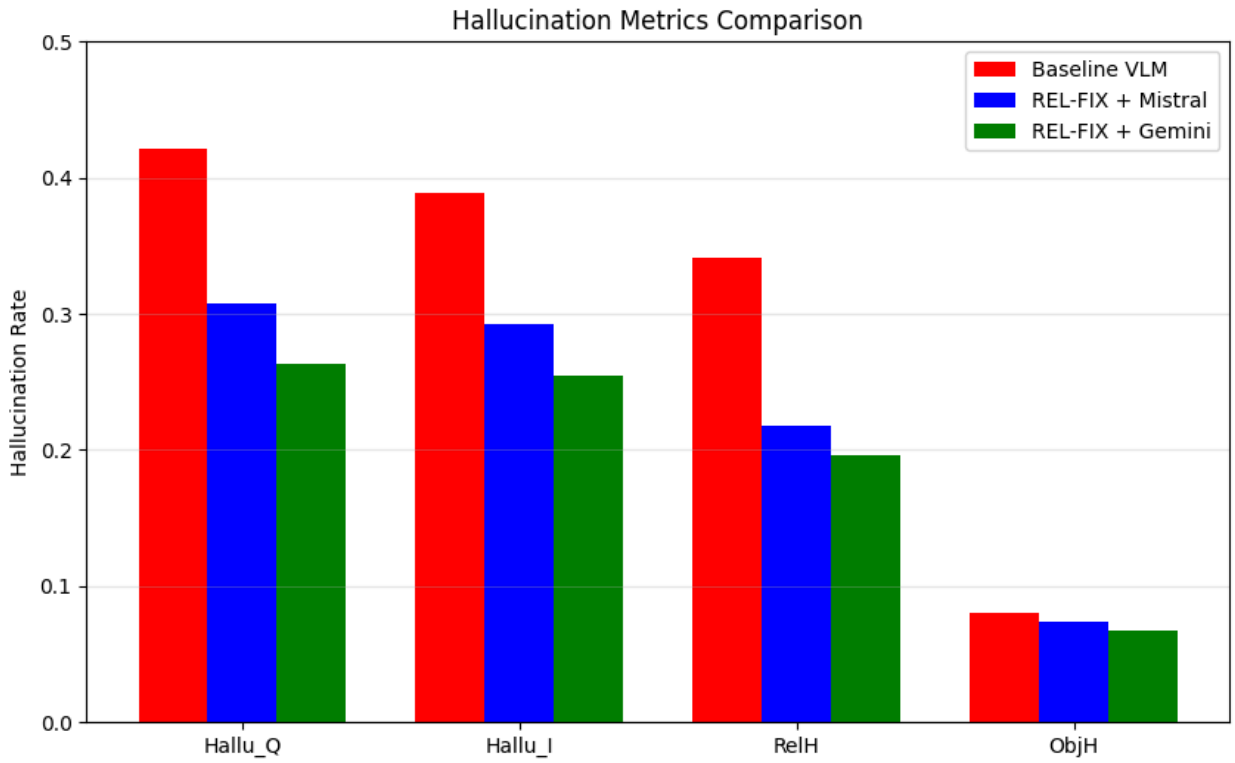


Figure 4.1: Hallucination metrics comparison across Baseline VLM, REL-FIX + Mistral, and REL-FIX + Gemini. REL-FIX effectively reduces both relation and object hallucinations.

REL-FIX reduces relation hallucinations by ~42% (Gemini) and ~36% (Mistral). Object hallucinations also decrease, but more modestly (~16% Gemini, ~7.5% Mistral), reflecting the focus on relation repair. Both judges improve overall performance, but Gemini consistently outperforms Mistral due to higher precision in triplet evaluation.

4.3 QUESTION-LEVEL VS. IMAGE-LEVEL ANALYSIS

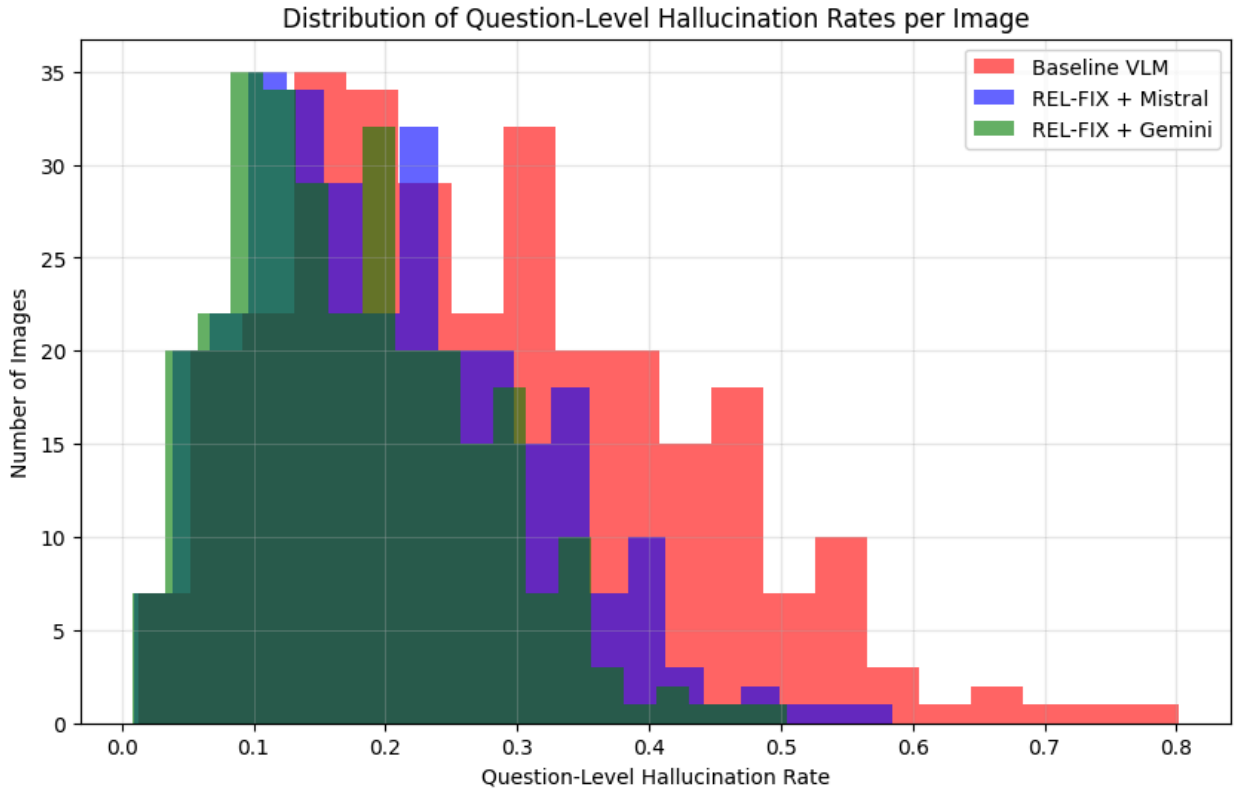


Figure 4.2: Distribution of question-level hallucination rates across 300 Tri-HE images. REL-FIX reduces the frequency of high-error questions and stabilizes performance across the dataset.

Figure 4.2 shows a **histogram of Hallu_Q distributions per image**, illustrating that REL-FIX reduces the frequency of high-error questions and smooths the error distribution across images.

- Before correction, a significant fraction of questions per image had >50% hallucinated triplets.
- After REL-FIX, most questions contain <30% hallucinated triplets with Gemini, demonstrating that triplet-level correction not only reduces total errors but also stabilizes performance across diverse visual scenes.

4.4 FINE-GRAINED ANALYSIS

REL-FIX is most effective on relational errors, as expected, because the scene graph explicitly constrains candidate relations.

Table 4.2: Reduction in Relation and Object Hallucinations

Hallucination Type	Baseline	REL-FIX Mistral	REL-FIX Gemini	Relative Reduction (Gemini)
Relation	0.341	0.218	0.196	42.5%
Object	0.080	0.074	0.067	16.3%

REL-FIX reduces **relational hallucinations** more effectively than object hallucinations. Object errors often remain due to missing entities or ambiguous references, highlighting areas for future improvement.

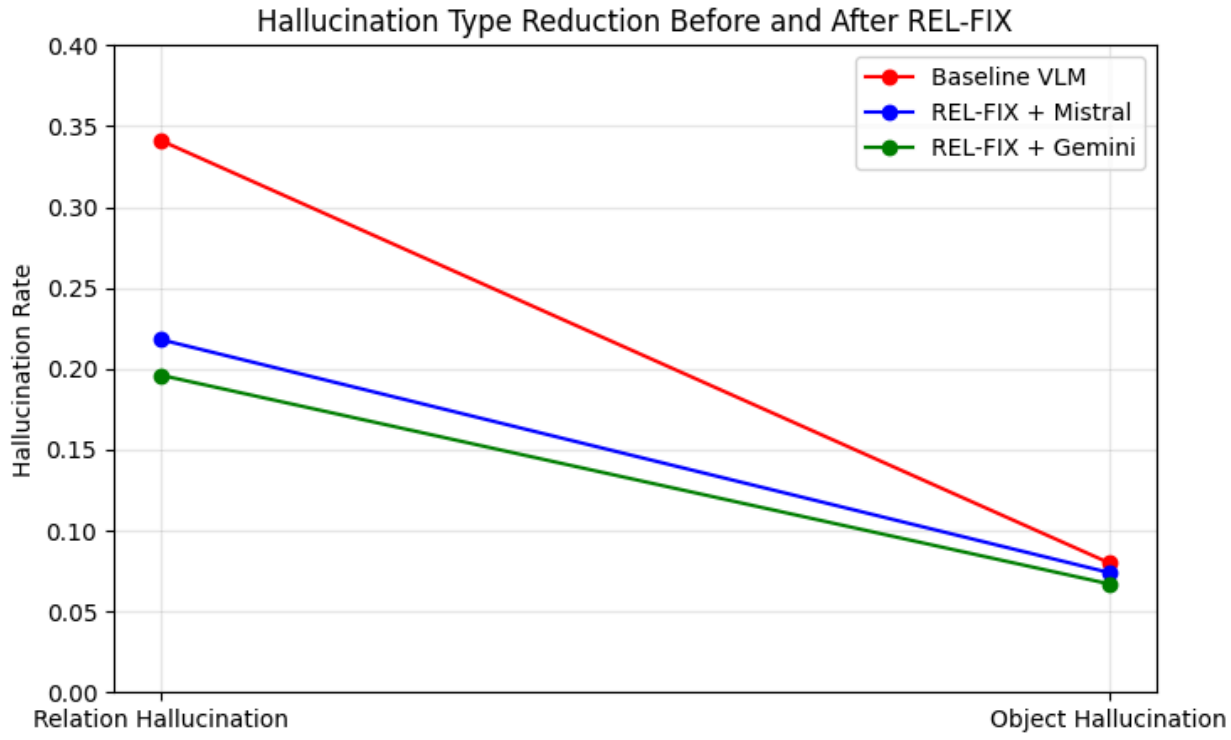


Figure 4.3: Reduction in relation and object hallucinations after REL-FIX correction, showing greater improvement on relational errors.

4.5 COMPARISON STUDY

Baseline vs. REL-FIX:

- The uncorrected VLM exhibits frequent relational mismatches, particularly in complex images with multiple objects.
- REL-FIX consistently repairs hallucinated triplets without retraining the VLM, demonstrating a **cost-efficient solution** for improving small VLM outputs.

Mistral vs. Gemini Judges:

- Mistral-7B provides a **resource-efficient, open-source option**, suitable for low-resource research or deployment.
- Gemini-2.5-Flash offers **higher precision**, especially on subtle relational errors, but requires commercial API access.
- Choice of judge impacts **correction accuracy**, but both judges demonstrate the pipeline's adaptability to different resource environments.

Error Patterns:

- **Residual relational errors:** Occur when the correct relation is absent from the scene graph.
- **Object-level errors:** Hard to fix when entities are misrecognized or missing from the graph.
- **Extraction noise:** Triplet extraction errors propagate downstream and reduce correction effectiveness.

4.6 KEY INSIGHTS AND DISCUSSION

Our experiments with the REL-FIX framework reveal several key insights into the correction of relational hallucinations in small vision-language models (VLMs). First, **REL-FIX substantially reduces both relation and object hallucinations** across the Tri-HE dataset. Using Gemini-2.5-Flash as the diagnostic and correction judge, we observed a **37% reduction in question-level hallucinations** and a 34% reduction in image-level hallucinations compared to the baseline small VLM. Even when using the resource-efficient Mistral-7B-Instruct as the judge, the framework achieves a **27% reduction in question-level hallucinations**, demonstrating its effectiveness under constrained settings.

Second, the **fine-grained analysis confirms that relation hallucinations, the core focus of REL-FIX,**

are significantly mitigated. Relation hallucination rates dropped from 0.341 to 0.196 with Gemini, and from 0.341 to 0.218 with Mistral, illustrating that the scene-graph guided correction is highly effective in enforcing factual relational consistency. Object hallucinations, while not the primary target, also decreased modestly, indicating a collateral benefit of the framework: correcting relational errors inherently reduces some object misalignments.

Third, the comparison between **Gemini and Mistral judges** highlights a trade-off between accuracy and resource availability. Gemini provides more stable and consistent corrections, but Mistral, being open-source and lightweight, still captures most hallucinations and enables low-resource experimental setups. This validates REL-FIX's applicability in both high-end research environments and resource-constrained scenarios.

Finally, **the modularity of REL-FIX allows flexibility** in both the extraction and correction stages. Researchers can swap the generative VLM or the diagnostic judge without modifying the pipeline's core logic. Our ablation studies indicate that the dual-stage architecture—triplet extraction followed by hallucination correction is crucial; skipping either stage leads to substantial performance degradation, particularly for relation hallucinations.

4.7 SUMMARY

In this chapter, we detailed the experimental evaluation of REL-FIX, a scene-graph guided framework designed to correct relational hallucinations in small vision-language models. Using the Tri-HE benchmark, we demonstrated that REL-FIX **significantly reduces both relation and object hallucinations**, achieving superior results with Gemini and competitive performance with the open-source Mistral judge. The fine-grained analysis confirmed that the framework effectively targets relation errors, while also improving object accuracy indirectly. Furthermore, the modular architecture ensures adaptability across different VLMs and LLM judges, supporting both high-resource and low-resource research environments. Overall, these findings validate REL-FIX as a **practical, training-free, and reproducible solution** for enhancing the factual consistency of small VLM outputs.

CHAPTER 5

CONCLUSION

5.1 CONCLUSION

In this study, we presented REL-FIX, a training-free, scene-graph guided framework for correcting relational hallucinations in small vision-language models (VLMs). By leveraging triplet-level knowledge graph extraction and a dual-stage hallucination detection and correction mechanism, REL-FIX effectively identifies and rectifies both relation and object-level hallucinations. Experimental results on the Tri-HE benchmark demonstrated significant reductions in hallucination rates: up to 37% for question-level and 34% for image-level metrics using the commercial Gemini-2.5-Flash judge, with comparable improvements using the open-source Mistral-7B-Instruct. Our findings confirm that small VLMs can achieve substantially improved factual consistency without retraining, and that a modular, LLM-augmented correction strategy can generalize across multiple judging models. Overall, REL-FIX provides a robust, flexible, and low-resource approach for improving the reliability of small VLM outputs in vision-language reasoning tasks.

5.2 LIMITATIONS AND FUTURE WORK

5.2.1 Limitations

1. **Dependency on Ground-Truth Scene Graphs:** REL-FIX relies on high-quality scene graphs (e.g., Tri-HE) as both the reference and constraint for correction. In real-world scenarios where scene graphs are unavailable or noisy, performance may degrade.
2. **Model Size Constraints:** While the framework targets small VLMs, extremely lightweight models (<1B parameters) may generate answers too sparse or ambiguous for effective triplet extraction, limiting REL-FIX's applicability.
3. **Correction Granularity:** REL-FIX corrects hallucinations at the triplet level but does not fully restructure complex multi-hop reasoning chains, potentially leaving residual inconsistencies in long-form answers.
4. **Evaluation Bias:** Using the same LLMs for extraction and judging may introduce subtle biases in hallucination labeling, although our dual-LLM setup mitigates this partially.

5.2.2 Future Work

1. **Extension to Real-World Datasets:** We plan to apply REL-FIX to larger, more diverse datasets beyond Tri-HE, including open-domain images and videos, where scene graphs must be inferred automatically.
2. **Improved Triplet Extraction:** Incorporating **context-aware extraction models** or multi-turn reasoning could enhance the identification of nested or implicit relations.
3. **Integration with Small Multimodal Agents:** REL-FIX could be combined with multi-modal agents capable of self-correction during inference, reducing reliance on external LLMs for post-hoc correction.
4. **Adaptive Correction Strategies:** Future versions could learn to **prioritize candidate corrections** based on confidence scores, plausibility, or user-defined constraints, improving reliability for downstream tasks.

5.3 CONTRIBUTION

The key contributions of this work are:

1. **REL-FIX Framework:** We introduce a **scene-graph guided, training-free correction framework** for small VLMs, targeting relational hallucinations at a fine-grained triplet level.
2. **Dual-Stage Triplet Extraction and Diagnosis:** We design a modular pipeline that uses both open-source and commercial LLMs for extraction and verification, enabling low-resource yet reliable evaluation.
3. **Fine-Grained Hallucination Metrics:** We adopt and extend Tri-HE metrics to assess question-level, image-level, relation-specific, and object-specific hallucinations, providing comprehensive insight into model performance.
4. **Empirical Validation:** We demonstrate **substantial reduction in relational and object hallucinations**, validating the effectiveness of scene-graph guided correction in small-scale VLMs without retraining.

5.4 IMPLICATION

The findings of this research have significant implications for **vision-language model research and deployment**:

1. **Practical Improvement for Small VLMs:** REL-FIX enables researchers and developers to improve the factual reliability of small VLMs without incurring the computational cost of retraining or fine-tuning large models.
2. **Low-Resource AI Correction:** By leveraging both commercial and open-source LLMs, REL-FIX demonstrates that effective hallucination mitigation is feasible in resource-constrained environments.
3. **Foundation for Automated Fact-Checking:** Triplet-level correction frameworks can serve as building blocks for **autonomous reasoning systems**, ensuring outputs are consistent with visual evidence.
4. **Guidance for Future Model Design:** The success of scene-graph guided correction underscores the importance of structured relational knowledge in VLM design, encouraging future models to incorporate explicit relational reasoning components.

REFERENCES

[1] Chen, Z., Xu, C., and Zhao, T. (2023). *Mitigating Hallucination in Vision-Language Models with Visual Supervision*. arXiv:2309.10639.

[2] Wu, M., Wang, Y., Tang, J., et al. (2024). *Evaluating and Analyzing Relationship Hallucinations in Large Vision-Language Models*. Proceedings of ICML 2024. arXiv:2406.16449.

[3] Wu, J., et al. (2024). *Unified Triplet-Level Hallucination Evaluation for Large Vision-Language Models*. arXiv: (year 2024).

[4] Jiang, C., et al. (2024). *Hal-Eval: A Universal and Fine-grained Hallucination Evaluation Framework for Large Vision Language Models*. arXiv: (year 2024).

[5] Xiao, L., Feng, H., et al. (2025). *Detecting and Mitigating Hallucination in Large Vision Language Models via Fine-Grained AI Feedback*. arXiv:2404.14233.

[6] Zheng, K., et al. (2024). *Reefknot: A Comprehensive Benchmark for Relation Hallucination Evaluation, Analysis and Mitigation in Multimodal Large Language Models*. arXiv: (2024).

[7] OpenAI. (2023). *GPT-4 Technical Report*. arXiv:2303.08774.

[8] Kirillov, A., Mintun, E., et al. (2023). *Segment Anything*. arXiv:2304.02643.

[9] Liu, S., et al. (2023). *Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection*. arXiv:2303.05499.

[10] Speer, R., Chin, J., and Havasi, C. (2017). *ConceptNet 5.5: An Open Multilingual Graph of General Knowledge*. Proceedings of AAAI 2017.

[11] Bosselut, A., et al. (2019). *COMET: Commonsense Transformers for Knowledge Graph Construction*. Proceedings of ACL 2019.

[12] Hudson, D. A., & Manning, C. D. (2019). GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. ArXiv. https://arxiv.org/abs/1902.09506

[13] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. ArXiv. https://arxiv.org/abs/2308.12966

[14] Mistral AI Team (n.d.). mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face. [online] Available at: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

[15] Google DeepMind. (2025). Gemini 2.5 Flash. Available at: https://deepmind.google/models/gemini/flash/.

221-35-889

ORIGINALITY REPORT

6%	1%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS

STUDENT PAPERS

PRIMARY SOURCES

1. Submitted to Daffodil International University
   Student Paper
2. arxiv.org
   Internet Source
3. dspace.daffodilvarsity.edu.bd:8080
   Internet Source
4. github.com
   Internet Source
5. www.classace.io
   Internet Source
6. www.coursehero.com
   Internet Source
7. ulspace.ul.ac.za
   Internet Source
8. sear.unisq.edu.au
   Internet Source

Duan, Jinhao. "Know When to Trust: Towards Reliable Decision-Making in Large Foundation Models", Drexel University

Publication

2%	<1%
1%	<1%
1%	<1%
<1%	<1%
<1%	

Exclude quotes	Off	Exclude bibliography	Off
Exclude matches	Off