K-GIP: Diagnosing Logical Fractures in Large Vision-Language Models via Verification Scene Graphs and Sequential Pruning

Yujun Hu[1] **Xiaoyu Zhou**[2] **Changbo Wang**[3] **Gaoqi He**[1]*
[1]School of Computer Science and Technology, East China Normal University
[2]School of Information Science and Engineering, East China University of Science and Technology
[3]School of Data Science and Engineering, East China Normal University
gqhe@cs.ecnu.edu.cn

Abstract

Diagnosing fine-grained hallucinations in *Large Vision-Language Models* (LVLMs) can greatly advance their reliable deployment in real-world applications. Nevertheless, current benchmarks predominantly employ flat metrics that treat errors in isolation, leaving a gap in evaluating the complex causal dependencies between visual perception and textual reasoning. Motivated by this, we introduce the **K**nowledge-**G**uided **I**n-Context **P**robing (K-GIP) framework to fill this gap. Specifically, K-GIP constructs a high-fidelity dual-perception ground truth to transform abstract priors into multi-granularity queries. Furthermore, we propose a *Verification Scene Graph* metric equipped with a *Sequential Logic Pruning* protocol, which explicitly models existence-attribute dependencies to strictly penalize logical fractures. We conduct comprehensive evaluations of mainstream LVLMs across three datasets using K-GIP. The experimental results highlight that our methodology successfully isolates deep reasoning failures from simple perceptual misses. We hope K-GIP can serve as a valuable and rigorous standard to assess logical robustness in multimodal systems.

1 Introduction

Recent advancements in Large Vision-Language Models (LVLMs) have revolutionized multimodal reasoning, with models such as GPT-4 (Achiam et al., 2023), InstructBLIP (Dai et al., 2023), and the Qwen-VL series (Bai et al., 2025b,a) demonstrating remarkable capabilities through large-scale pre-training (Zhang et al., 2024). Despite their undeniable proficiency in generating fluent descriptions, these models remain highly susceptible to Object Hallucination in rigorous Visual Question Answering tasks. Existing benchmarks such as POPE (Li et al., 2023) and HallusionBench (Guan et al., 2024) establish existence baselines but predominantly treat hallucinations as isolated errors. This flat evaluation paradigm neglects causal dependencies between visual perception and textual reasoning, failing to distinguish simple perceptual misses from severe logical inconsistencies.

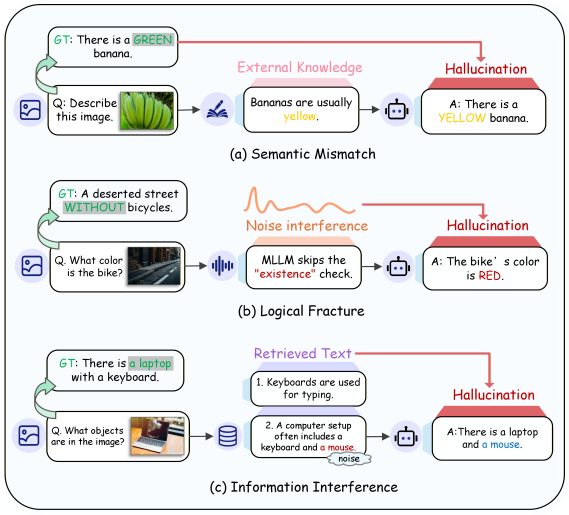


Figure 1: Illustration of three hallucination types triggered by priors and interference: Semantic Mismatch (priors overriding visual facts), Logical Fracture (attributes of absent objects), and Information Interference (co-occurring semantic noise).

Current metrics struggle when models over-rely on internal parametric knowledge. Although LLMs possess vast commonsense (Gui et al., 2022a; Lin and Byrne, 2022a), these priors often trigger insidious hallucinations evading standard benchmarks as models revert to language probabilities rather than grounding generated answers in concrete visual evidence. As illustrated in Figure 1, we categorize these failures into three distinct types.

Semantic Mismatch occurs when generic priors supersede specific visual instances. As shown in Figure 1 (a), pre-trained knowledge encapsulates generic facts (e.g., *Bananas are usually yellow.*),

*Corresponding author: gqhe@cs.ecnu.edu.cn.

Findings of the Association for Computational Linguistics: ACL 2026, pages 10222–10236
July 2-7, 2026

contrasting with atypical instances like an *unripe green banana*. Existing metrics often overlook this Modality Bias, where models prioritize textual priors over pixel-level evidence (Chen et al., 2024c).

Logical Fracture emerges from absent existential verification. Visual reasoning requires verifying object existence before discussing attributes. However, as shown in Figure 1 (b), leading questions often mislead models into fabricating descriptions for absent objects. This intensifies when language priors disrupt topological dependencies (Chen et al., 2024a; Yan et al., 2025). Neglecting these dependencies causes fine-grained hallucinations that accuracy metrics miss.

Information Interference results from strong semantic co-occurrence priors. As shown in Figure 1 (c), deeply ingrained associations often introduce relevant but visually absent objects, like hallucinating a *mouse* with a standalone *keyboard*. Since LLMs struggle to suppress learned redundancies (Liu et al., 2024c), this noise induces hallucinations based on textual probability rather than concrete visual presence within the scene.

We introduce the Knowledge-Guided In-Context Probing framework K-GIP to diagnose failures by dynamically verifying visual alignment and logical consistency. The pipeline first establishes a Dual-Perception Knowledge Alignment paradigm synergizing micro-entity detection with macro-context parsing to construct high-fidelity Ground Truth. We subsequently employ Multi-Granularity Probing Generation to transform priors into coupled existence-attribute queries. Finally we implement a verification scene graph metric utilizing sequential logic pruning to penalize logical fractures and ensure scores reflect valid reasoning chains.

The contributions are summarized as follows:

(1) Established a Dual-Perception Ground Truth mechanism. Integrating micro-entity detection with macro-context parsing anchors abstract commonsense to visual instances resolving ambiguities regarding general priors versus specific scene facts.

(2) Proposed a Dependency-Aware Scoring Protocol. Utilizing the Verification Scene Graph to model logical dependencies our sequential pruning metric dynamically penalizes invalid reasoning branches precisely diagnosing hallucinations driven by false premises.

(3) Comprehensive Diagnostic Evaluation. Deploying K-GIP on mainstream LVLMs reveals significant performance stratification and identifies cognitive bottlenecks including Logical Fractures in long-context generation previously masked by flat metrics.

2 Related Work

Hallucination Mechanisms and Mitigation in LVLMs. Despite the success of advanced models, hallucinations persist in fine-grained visual reasoning. Dai et al. (Dai et al., 2023) attribute this to low-quality data forcing reliance on language priors while Zhu et al. (Zhu et al., 2023) note this dominance intensifies during long-text generation. Current mitigation strategies like Lure (Zhou et al., 2023) and Woodpecker (Yin et al., 2024) primarily focus on post-hoc intervention via rewriting or detector guidance. In contrast, K-GIP functions as a fine-grained diagnostic framework that evaluates underlying logical robustness by verifying the strict consistency between foundational existence premises and subsequent attribute reasoning.

Retrieval-Augmented Generation and Knowledge Conflicts. Retrieval-Augmented Generation addresses parametric limitations through Wikipedia paragraphs (Lin and Byrne, 2022b), multimodal graphs (Gui et al., 2022b), or in-context image-text pairs (Yang et al., 2023). However, Chen et al. (Chen et al., 2024b) demonstrate that models often prioritize conflicting external text over visual evidence, particularly in fine-grained OCR tasks (Singh et al., 2019; Biten et al., 2022). Our framework constructs dynamic visual verification queries to quantify this phenomenon. These queries act as probing signals to determine whether models successfully anchor retrieved citations in definitive visual facts or succumb to textual interference.

Scene Graph Generation and Structured Visual Reasoning. Scene graphs bridge pixel perception and symbolic reasoning (Wang et al., 2023a; Hildebrandt et al., 2020) enhancing VQA interpretability (Mishra et al., 2024). While TSG Bench (Ang et al., 2023) identifies LLM structural deficits visual models (Li et al., 2022; Yang et al., 2022) maintain authenticity despite noise. Distinct from input enhancement methods (Wang et al., 2023b) we leverage graph structure for evaluation topology. Our Verification Scene Graph metric enforces strict dependency scoring to penalize logical fractures ensuring scores reflect valid reasoning chains.

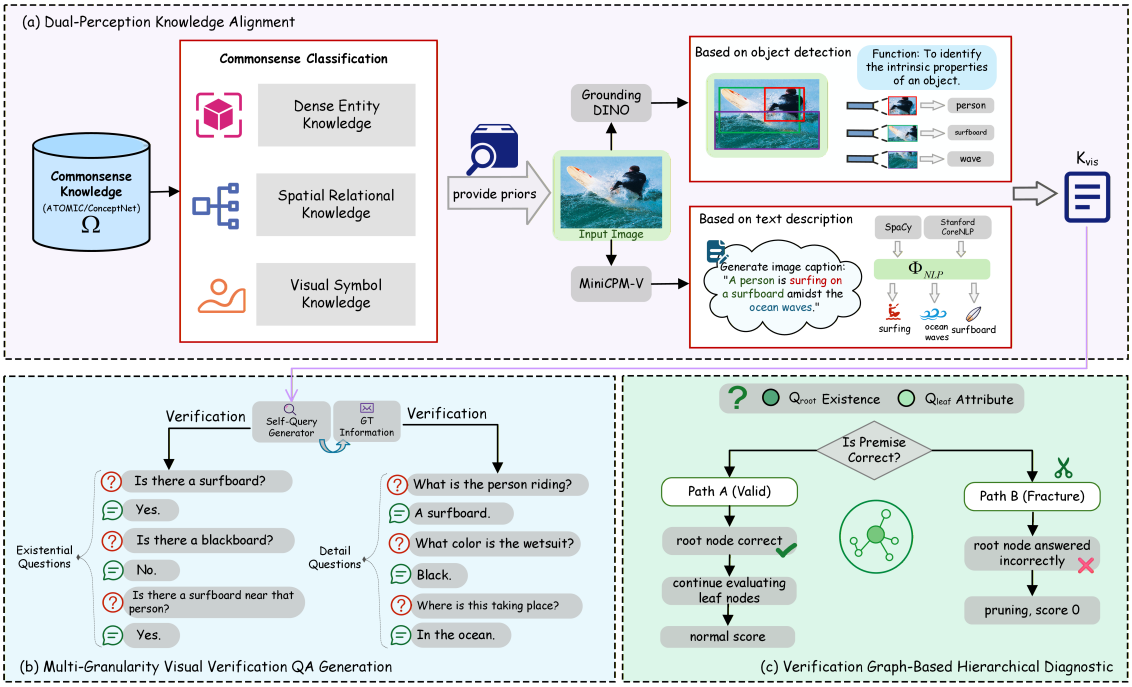


Figure 2: Overview of K-GIP. K-GIP integrates MiniCPM-V dual-perception with hierarchical queries for grounded benchmarking. The verification graph enforces dependency by pruning invalid branches to penalize hallucinations.

3 Methodology

In this section, we introduce the Knowledge-Guided In-Context Prompting framework, called K-GIP. Unlike traditional inference interventions, K-GIP is designed as a fine-grained diagnostic evaluation framework. To effectively address Factual Hallucinations in contemporary Multimodal Large Language Models like LLaVA-1.6 (Liu et al., 2024b) and InternVL2 (Chen et al., 2024d) on fine-grained tasks, we propose this novel, non-parametric, and automated benchmarking pipeline.

As shown in Figure 2, K-GIP operates in three stages: (1) Dual-perception visual retrieval to establish a ground-truth knowledge background; (2) Multi-granularity QA generation transforming static facts into probing queries; (3) Verification Scene Graph modeling to perform hierarchical diagnostic evaluation via graph topology, quantifying logical fractures in LVLM inference.

3.1 Dual-Perception Knowledge Alignment

To strictly ground evaluation in visual facts, as shown in Figure 3, we categorize the prior knowledge base Ω into three Visual Verification Dimensions to construct the ground truth:

- **Dense Entity Knowledge.** Targeting scenes with minute or densely distributed objects, this dimension benchmarks fine-grained recall. It strictly challenges models to distinguish between the valid presence and hallucination of dense entities within highly cluttered and inherently complex visual environments.
- **Spatial Relational Knowledge.** Tailored for structured scenes involving highly complex interactions, this dimension establishes the definitive ground truth for spatial topology and action logic. It necessitates parsing compositional logic within the visual scene graph to strictly verify structural consistency.
- **Visual Symbol Knowledge.** Focusing on embedded text and semantic symbols, this dimension validates scene text recognition and pragmatic interpretation. It specifically evaluates the capability to read and comprehend symbolic information beyond mere object recognition within complex visual scenes.

To extract image-specific facts from Ω, we propose a dual-perception mechanism synergizing micro-entity detection and macro-context parsing.

Micro-Entity Perception. To establish foundational visual facts, we employ Grounding DINO (Liu et al., 2024d) to retrieve core entities. Formally, given image I, the detector yields a set of ground-truth entities $\mathcal{E}_{det}$:

$$\mathcal{E}_{det} = \{(o_i, b_i, s_i)\}_{i=1}^{N} = DINO(I) \quad (1)$$

where o_i denotes the open-vocabulary phrase label,

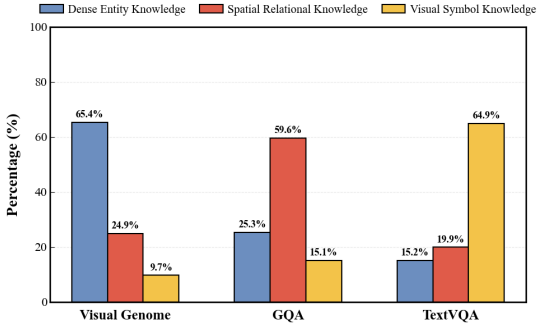


Figure 3: Statistical distribution of visual verification dimensions across three core evaluation datasets.

and N represents the number of detections. We enforce a strict threshold $s_i > \tau$ to prioritize precision, thereby establishing conservative hard anchors for dense entity knowledge to strictly ground probes in high-confidence evidence. A comprehensive sensitivity analysis regarding the impact of perception thresholds is provided in Appendix C.1.

Macro-Context Perception. To address the lack of dynamic interaction descriptions in micro-detection, we capture higher-order spatial relational and visual symbol knowledge for complex reasoning probes. We employ MiniCPM-V 2.6 (Yao et al., 2024) using a *Descriptive Prompting* strategy to synthesize a high-fidelity panoramic description $C = \text{MiniCPM}_\theta(I, \text{Prompt}_{description})$. To extract the semantic structure from this unstructured sequence, we utilize a parsing function Φ_{NLP}, integrating SpaCy (Danenas and Skersys, 2022) and Stanford CoreNLP (Kovriguina et al., 2017), to map C into a structured ground-truth query set:

$$\mathcal{Q}_{sem} = \Phi_{NLP}(C) = \{\mathcal{Q}_{rel}\} \cup \{\mathcal{Q}_{sym}\} \quad (2)$$

where $\mathcal{Q}_{rel}$ denotes the relation predicate set, derived by traversing syntax trees to extract core verbs and prepositions (e.g., *stands next to*) for indexing topological logic; conversely, $\mathcal{Q}_{sym}$ represents the symbolic context set, constructed by identifying nouns or modifiers (e.g., *stop sign*) to retrieve pragmatic priors. These parsed semantic units serve as the standard $\mathcal{K}_{vis}$ to generate probes evaluating alignment with visual reality.

3.2 Multi-Granularity Visual Verification QA Generation

Subsequently, we execute multi-granularity visual verification QA generation to transform the extracted ground truth into actionable visual queries, serving as atomic nodes for our evaluation graph. To rigorously assess the model, we utilize a Self-Query Generator to synthesize two categories of probing questions derived from $\mathcal{K}_{vis}$. The specific prompt templates and the adversarial query generation framework are illustrated in Appendix B.

Type I: Existence Anchor Questions (The Premises). These questions aim to check whether the model can perceive the premise objects, serving as the foundation of logical reasoning. The generation logic targets the dense entity knowledge, creating questions such as "*Is there a* $\{obj\}$*?*" or "*Is a* $\{obj\}$ *visible?*". To evaluate hallucination resistance, we introduce *negative questions*. These are created by replacing real objects in the original QA pairs with non-existent objects selected from a pre-constructed set. The expected answer is *None* or *No*, forcing the model to distinguish between hallucinated and real dense entities.

Type II: Attribute and Logic Detail Questions (The Reasoning). These questions aim to assess finer reasoning capabilities. We diversify the questions to cover the remaining dimensions:

- For **Spatial Relational Knowledge**, questions focus on topology, such as "*Where is the* $\{obj\}$ *relative to the* $\{obj2\}$*?*".
- For **Visual Symbol Knowledge**, questions focus on pragmatics, such as "*What is written on the sign?*".

Unlike traditional VQA evaluation which treats all questions equally, our approach distinguishes between existence (Type I) and detail (Type II) to enable a structured diagnostic.

3.3 Verification Graph-Based Hierarchical Diagnostic

Motivation. Conventional flat metrics such as accuracy frequently fail to characterize logical inconsistencies where a model might hallucinate attributes despite correctly identifying the absence of an object. To rigorously penalize such logical fractures, we restructure discrete QA pairs into a verification scene graph, denoted as VSG, and implement a dependency-aware scoring protocol.

3.3.1 Structure: Verification Scene Graph

To model the causal dependency between visual perception and reasoning, we organize probing questions into a hierarchical topology consisting of two distinct levels. Root Nodes function as

the premise by anchoring fundamental visual facts through Type I Existence Anchor Questions. Leaf Nodes represent dependent reasoning tasks corresponding to Type II Attribute and Logic Questions. We establish a directed edge $v_{root} \rightarrow v_{leaf}$ to signify an existential dependency, implying that the validity of an attribute evaluation is strictly contingent upon the correctness of the premise.

3.3.2 Dependency-Aware Pruning Protocol

Departing from traditional methodologies that treat evaluation questions independently, we propose a Sequential Pruning Protocol to compute the final reliability score. For any given root-leaf pair, the diagnostic mechanism operates through a conditional evaluation path described below.

Step 1: Premise Verification. Initially, the model addresses the root question q_{root}. We define a validity indicator V to determine whether the prediction aligns with the ground truth:

$$V = \mathbb{I}(A_{root} \equiv \text{Ground Truth}) \tag{3}$$

where $\mathbb{I}$ denotes the indicator function.

Step 2: Conditional Scoring. The subsequent assessment of the leaf question q_{leaf} is conditioned on the state of V.

1. Valid Inference ($V = 1$). Under the condition that the premise is correctly identified, the system evaluates q_{leaf}, assigning a score based on the reasoning accuracy of the model.

2. Logical Fracture ($V = 0$). Conversely, if the premise is incorrect due to hallucination or omission, the evaluation of the leaf node is pruned. Consequently, the score for q_{leaf} is automatically penalized to zero.

Significance. This protocol mitigates the inflation of benchmark scores caused by unfounded conjectures. Through the enforcement of this hierarchical constraint, our $\mathcal{C}_{verified}$ metric ensures that high scores reflect robust reasoning chains rather than isolated correct responses.

4 Experiments

In this section, we conduct a comprehensive evaluation on three widely adopted benchmark datasets to validate the effectiveness of the proposed framework. We first introduce the experimental setup, followed by a quantitative analysis using K-GIP as a fine-grained evaluation tool to assess the visual perception and hallucination resistance of mainstream Multimodal Large Language Models.

4.1 Setup

Datasets. We construct the benchmark using three representative datasets to assess reasoning across distinct granularities. GQA (Hudson and Manning, 2019) challenges compositional logic with 22 million questions over 113K images, requiring multi-step spatial and semantic inference. Visual Genome (Krishna et al., 2017) stress-tests fine-grained recall, comprising 108K images densely annotated with 3.8 million objects and 2.3 million relationships. TextVQA (Singh et al., 2019) evaluates pragmatic understanding via 45,336 questions, compelling models to reason based on OCR-extracted scene text rather than pure object recognition.

Models. We select representative mainstream LVLMs for evaluation: mPLUG-Owl (Ye et al., 2023), MultiModal-GPT (Gong et al., 2023), InstructBLIP (Dai et al., 2023), LLaVA-1.5-7B/13B (Liu et al., 2024a), GPT-4V (Zhou et al., 2024), Qwen2.5-VL (Bai et al., 2025b), and Qwen3-VL (Bai et al., 2025a). Detailed implementation configurations and inference settings for the evaluated models are provided in Appendix A.

Metrics. We employ Accuracy, Precision, Recall, and F1-score as primary evaluation metrics. Specifically, precision on caption-sourced questions serves as a key indicator for hallucination resistance, while recall on image-sourced questions reflects visual perception limits.

4.2 Data Processing and Verification

We utilized the K-GIP self-query generator to construct an automated benchmark containing Image-Sourced and Caption-Sourced questions. To validate pipeline reliability and mitigate oracle fallacy, we conducted a human verification study using stratified random sampling to select 1000 QA pairs across visual verification dimensions. Three linguistically trained annotators performed a double-blind evaluation based on *Syntactic Validity* for grammatical coherence and *Visual Consistency* for strict alignment with source visual facts. More details can be found in Appendix D.

As illustrated in Figure 4 the framework achieves an overall visual consistency rate of 92.1% with the dense entity knowledge dimension yielding a peak accuracy of 96.2% confirming object anchor-

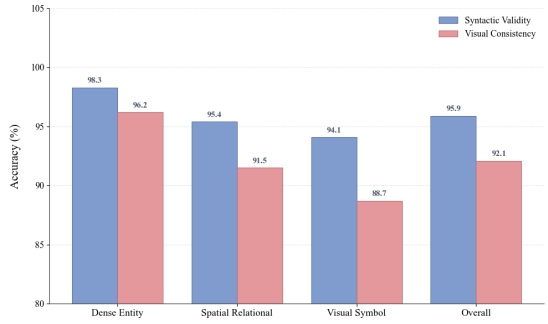


Figure 4: Human verification results of K-GIP generated QA pairs across three dimensions.

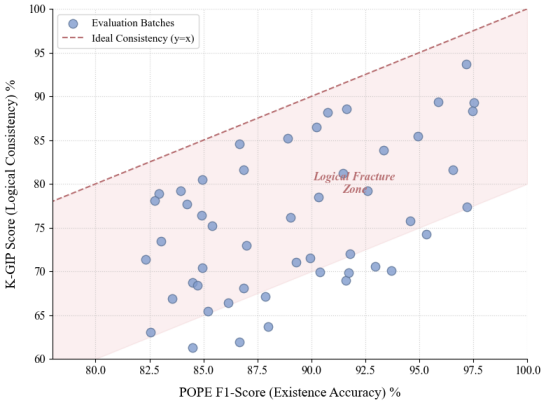


Figure 5: **Comparison of POPE (x-axis) and K-GIP (y-axis).** The shaded *Logical Fracture Zone* denotes models exhibiting high existence accuracy but poor attribute consistency, resulting in lower K-GIP scores.

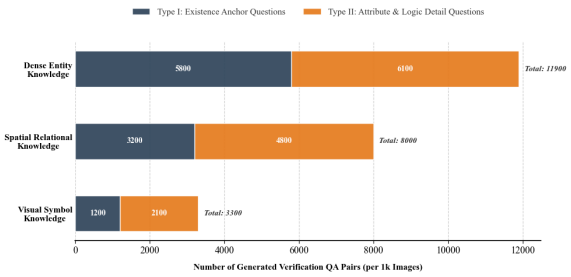


Figure 6: Distribution of generated verification QA pairs across the three Visual Verification Dimensions (Sampled 1k images each).

ing robustness. Even within the challenging visual symbol dimension the framework maintains 88.7% reliability. These results demonstrate that the automatically constructed benchmark exhibits minimal noise and constitutes a reliable gold standard for diagnosing LVLM hallucinations.

4.3 Overall Performance Analysis

Performance Stratification and Logical Fractures. Table 1 shows Qwen3-VL-8B (Bai et al., 2025a) outperforms baselines like Qwen2.5-VL-7B (Bai et al., 2025b), confirming the capability-robustness correlation. It surpasses GPT-4V in Precision and F1, highlighting superior hallucination suppression. However, a critical gap persists: most models maintain acceptable Image-Sourced Recall yet suffer Caption-Sourced Precision drops. This exposes a "logical fracture" where LVLMs hallucinate from textual priors rather than visual existence. Even advanced models show precision declines on Visual Genome captured by K-GIP. Detecting fine-grained errors missed by flat metrics emphasizes our structured verification's necessity.

Comparative Analysis with State-of-the-Art Metrics. We validated K-GIP superiority in diagnosing fine-grained hallucinations through comparative analysis against the prevalent POPE benchmark. Since POPE primarily evaluates binary object existence it often masks logical deficiencies in attribute binding and relational reasoning. Figure 5 visualizes the score distribution of randomly sampled batches exhibiting a distinct lower-triangular distribution where numerous samples achieve high POPE scores $> 90\%$ yet suffer substantial K-GIP performance drops. This score gap quantitatively reveals the Logical Fracture phenomenon where models correctly anchor objects but fail to maintain consistency in subsequent attribute reasoning. These findings demonstrate that K-GIP captures deep cognitive discrepancies overlooked by traditional flat metrics providing a stricter standard for assessing true LVLM robustness.

Ablation Study: Effectiveness of Sequential Logic Pruning. We validated the *Sequential Logic Pruning* protocol through an ablation study. We established a baseline w/o Pruning that removes dependency constraints to independently calculate leaf node accuracy, contrasting it with standard K-GIP. As shown in Table 2, eliminating pruning causes significant performance inflation. While mPLUG-Owl exhibits the largest discrepancy of 12.4%, Qwen3-VL-8B achieves a minimal gap of 0.7%, demonstrating superior logical consistency. This contrast confirms that traditional metrics remain susceptible to language priors, whereas K-GIP effectively filters out hallucinatory responses based on erroneous premises. Additional quantitative analysis demonstrating the visual gap caused by logical fractures is detailed in Appendix C.2.

Data Sampling and Multi-Dimensional Ques-

Table 1: Fine-grained evaluation results of mainstream LVLMs using the K-GIP framework. **Image-Sourced** questions assess visual fact perception (Recall focus), while **Caption-Sourced** questions assess hallucination resistance (Precision focus). **Bold** indicates the best performance.

Dataset	Model	Image-Sourced Questions				Caption-Sourced Questions			
		Acc	Prec	Recall	F1	Acc	Prec	Recall	F1
GQA	mPLUG-Owl (Ye et al., 2023)	42.1	40.5	41.2	41.5	31.4	30.2	31.5	30.8
	MultiModal-GPT (Gong et al., 2023)	53.5	52.8	51.9	52.3	42.5	41.7	42.1	41.9
	InstructBLIP (Dai et al., 2023)	65.8	64.2	65.1	64.8	50.2	49.5	50.1	49.8
	LLaVA-1.5-7B (Liu et al., 2024a)	70.5	69.8	70.2	70.1	55.4	54.8	55.2	55.0
	LLaVA-1.5-13B (Liu et al., 2024a)	75.4	74.1	75.0	74.8	60.1	59.5	60.2	59.9
	GPT-4V (Zhou et al., 2024)	87.2	81.4	86.3	85.5	70.3	71.5	75.8	69.3
	Qwen2.5-VL-7B (Bai et al., 2025b)	88.5	82.8	87.5	86.8	72.5	73.1	77.2	71.5
	Qwen3-VL-8B (Bai et al., 2025a)	**90.1**	**84.5**	**89.2**	**88.6**	**74.8**	**75.6**	**79.5**	**73.9**
Visual Genome	mPLUG-Owl (Ye et al., 2023)	41.8	40.2	40.5	41.1	35.6	34.2	33.8	34.5
	MultiModal-GPT (Gong et al., 2023)	50.2	48.7	48.1	48.5	40.1	39.5	38.8	39.2
	InstructBLIP (Dai et al., 2023)	60.5	59.1	58.5	59.2	48.5	47.8	47.1	47.5
	LLaVA-1.5-7B (Liu et al., 2024a)	68.7	67.5	66.8	67.2	56.4	55.8	55.1	55.5
	LLaVA-1.5-13B (Liu et al., 2024a)	72.5	71.2	70.5	71.1	62.8	61.5	60.8	61.2
	GPT-4V (Zhou et al., 2024)	87.0	81.2	86.0	85.3	84.2	78.9	84.1	82.2
	Qwen2.5-VL-7B (Bai et al., 2025b)	88.4	82.6	87.2	86.5	85.6	80.2	85.5	83.5
	Qwen3-VL-8B (Bai et al., 2025a)	**89.8**	**84.2**	**88.8**	**87.9**	**86.9**	**81.8**	**86.8**	**84.9**
TextVQA	mPLUG-Owl (Ye et al., 2023)	38.5	37.2	36.5	37.1	30.2	29.5	28.8	29.2
	MultiModal-GPT (Gong et al., 2023)	45.8	44.5	43.2	44.1	36.5	35.8	35.1	35.5
	InstructBLIP (Dai et al., 2023)	52.6	51.5	50.8	51.2	42.8	41.9	41.2	41.6
	LLaVA-1.5-7B (Liu et al., 2024a)	62.5	61.2	60.5	61.1	50.5	49.8	49.1	49.5
	LLaVA-1.5-13B (Liu et al., 2024a)	68.2	67.5	66.8	67.1	58.6	57.5	56.8	57.2
	GPT-4V (Zhou et al., 2024)	78.5	77.2	76.5	77.1	72.8	71.5	70.8	71.2
	Qwen2.5-VL-7B (Bai et al., 2025b)	81.2	80.5	79.8	80.1	75.5	74.2	73.5	74.1
	Qwen3-VL-8B (Bai et al., 2025a)	**83.5**	**82.8**	**81.9**	**82.5**	**77.2**	**76.5**	**75.8**	**76.1**

Table 2: **Ablation on Sequential Logic Pruning.** "w/o Pruning" calculates reasoning accuracy independently (flat metric), while K-GIP enforces strict dependency penalties. The Gap (Δ) quantifies lucky guesses derived from language priors.

Model	w/o Pruning	K-GIP	Gap (Δ)
mPLUG-Owl	55.4	43.0	-12.4
MiniGPT-4	56.8	46.2	-10.6
InstructBLIP	68.5	62.1	-6.4
LLaVA-1.5-13B	70.2	65.3	-4.9
GPT-4V	80.1	78.5	-1.6
Qwen2.5-VL-7B	81.5	80.3	-1.2
Qwen3-VL-8B	**83.5**	**82.8**	**-0.7**

tion Distribution. To ensure a comprehensively balanced and representative benchmark, we employed a rigorous stratified random sampling strategy across the three predefined Visual Verification Dimensions, constructing a Core Evaluation Subset of 3,000 images carefully curated from Visual Genome, GQA, and TextVQA. As illustrated in Figure 6, the resulting distribution of generated QA pairs precisely maps to the distinct cognitive hierarchies inherent within each dataset. Specifically, the Dense Entity Knowledge dimension yields the highest volume with nearly 50% Type I Existence Anchors, aligning seamlessly with the fundamental requirement for accurate micro-entity recall. Conversely, the Spatial Relational and Visual Symbol dimensions are predominantly characterized by Type II Attribute and Logic Details questions at 60% and 64% respectively. This distribution inherently prioritizes complex topological parsing and pragmatic symbolic decoding over simple object counting. Ultimately, these statistics robustly validate that K-GIP adaptively imposes differentiable, fine-grained verification constraints, which encompass a spectrum ranging from basic existence checks to intricate logic detailing, and are custom-tailored to the specific cognitive demands of each visual dimension.

4.4 Fine-Grained Results

Hierarchical Cognitive Evaluation. We comprehensively assess micro-entity perception, topological logic, and symbolic semantics across three diverse benchmarks: Visual Genome (Krishna et al., 2017), GQA (Hudson and Manning, 2019), and TextVQA (Singh et al., 2019). This structured

Table 3: Fine-grained evaluation of LVLMs across three visual verification dimensions on TextVQA dataset. **Bold** indicates best performance.

Model	Dense Entity				Spatial Relational				Visual Symbol			
	Acc.	P.	R.	F1	Acc.	P.	R.	F1	Acc.	P.	R.	F1
mPLUG-Owl (Ye et al., 2023)	51.6	54.1	51.5	42.5	16.0	39.2	42.3	21.0	17.1	30.8	39.6	21.8
MultiModal-GPT (Gong et al., 2023)	57.3	75.7	47.3	48.0	20.6	55.7	53.5	27.4	22.7	56.5	55.8	32.7
InstructBLIP (Dai et al., 2023)	70.8	80.6	70.2	68.3	34.3	56.4	59.7	33.7	42.1	58.3	66.6	48.1
LLaVA-1.5-7B (Liu et al., 2024a)	79.2	82.4	77.5	78.4	27.9	55.6	56.7	27.8	47.9	59.1	69.7	44.7
LLaVA-1.5-13B (Liu et al., 2024a)	84.6	87.7	81.4	84.2	61.0	62.2	76.2	55.6	57.5	61.0	75.7	52.1
GPT-4V (Zhou et al., 2024)	90.8	87.7	89.8	88.6	83.6	77.7	85.2	79.8	66.2	61.2	73.2	58.3
Qwen2.5-VL-7B (Bai et al., 2025b)	91.5	88.5	90.3	89.4	84.8	79.2	86.5	81.5	70.8	66.1	76.9	63.8
Qwen3-VL-8B (Bai et al., 2025a)	**92.4**	**89.6**	**91.2**	**90.4**	**86.2**	**81.5**	**87.8**	**83.2**	**72.5**	**68.4**	**78.5**	**66.1**

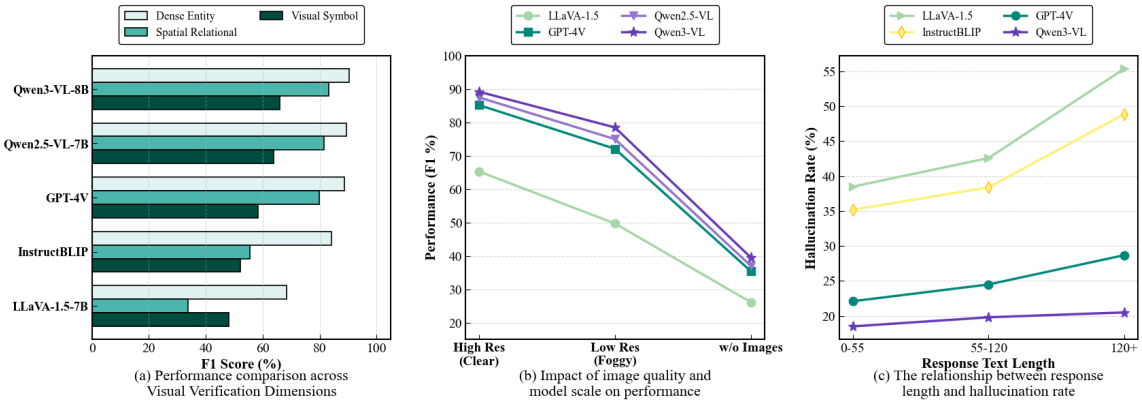


Figure 7: Analysis of the underlying causes of fine-grained hallucinations in LVLMs.

approach successfully validates K-GIP's Dual-Perception Mechanism, revealing distinct performance stratifications between foundational visual perception and complex high-order reasoning.

Entity Perception and Model Stratification. As demonstrated in Table 3, advanced models such as Qwen3-VL-8B (Bai et al., 2025a) and GPT-4V (Zhou et al., 2024) maintain superior perceptual robustness. While the majority of evaluated models achieve consistently high Recall within the Dense Entity Knowledge dimension, earlier architectures like mPLUG-Owl (Ye et al., 2023) exhibit markedly lower Precision. This critical discrepancy highlights their inherent susceptibility to hallucinating minute targets, ultimately failing to accurately distinguish concrete visual existence from their deeply ingrained internal generative priors.

Higher-Order Reasoning Bottlenecks. Rigorous evaluations on the Spatial Relational and Visual Symbol dimensions expose significant cognitive deficiencies across all tested model architectures. The declining F1 scores observed in LLaVA-1.5 (Liu et al., 2024a) and InstructBLIP (Dai et al., 2023) imply severe logical fractures, indicating reasoning breakdowns despite the successful grounding of valid existence premises. Notably, while Qwen3-VL-8B achieves a state-of-the-art F1 score of 66.1% in the challenging Visual Symbol dimension, deep pragmatic understanding remains a pervasive challenge for most contemporary LVLMs.

5 Diagnostic Analysis

To systematically investigate the underlying etiology of hallucinations, we analyzed the impacts of knowledge distribution, image interference, and generation length, as illustrated in Figure 7.

Specifically, Figure 7 (a) reveals a distinct performance bottleneck where models excel in Dense Entity tasks but falter significantly in complex Visual Symbol processing. While GPT-4V (Zhou et al., 2024) demonstrates considerable strength, Qwen3-VL-8B (Bai et al., 2025a) establishes a new state-of-the-art, effectively bridging the cognitive gap between basic visual perception and advanced OCR capabilities. Furthermore, Figure 7

(b) demonstrates that F1 scores decay linearly with declining image quality; however, Qwen3-VL exhibits superior noise robustness compared to GPT-4V, maintaining performance consistency even in challenging low-resolution settings. Finally, Figure 7 (c) depicts the correlations between response length and error rates. Where models like LLaVA (Liu et al., 2024a) suffer from compounding logical drift, Qwen3-VL and GPT-4V maintain remarkably flat curves. This resilience confirms that structured sequential pruning effectively blocks erroneous premises, thereby preventing severe downstream error propagation in extended textual outputs.

6 Conclusion

We introduce K-GIP, a fine-grained diagnostic framework for detecting hallucination and logical inconsistencies in Large Vision-Language Models. We employ a novel Verification Scene Graph with sequential dependency pruning to rigorously evaluate the causal link between visual perception and textual reasoning. Our findings demonstrate that current models, despite high recognition rates, suffer from Logical Fractures by relying on priors rather than visual evidence. Acknowledging upstream detector constraints, we hope K-GIP serves as a rigorous standard for assessing logical robustness and guiding the future development of more reliable zero-hallucination multimodal systems.

Limitations

The diagnostic capabilities of K-GIP framework, while robust, remain subject to specific constraints inherent to the automated model-based construction pipeline. A primary limitation stems from the intrinsic dependence on frozen upstream perception models, specifically Grounding DINO and MiniCPM-V. Consequently, the generated ground truth focuses predominantly on distinct, salient foreground entities and their attributes, while potentially overlooking minute or heavily occluded background objects. This deliberate design choice results from our strict precision-first confidence threshold strategy, intended to prioritize precision over recall and ensure that all existence anchors constitute definitive visual facts. Additionally, the current question generation process utilizes rigid structured templates to strictly maintain the logical dependencies required for the sequential pruning protocol. While this mechanism guarantees evaluation rigor, it inevitably limits the linguistic diversity compared to naturally occurring open-ended conversational data. Future iterations will aim to enhance syntactic variability through advanced controlled paraphrasing mechanisms while preserving the necessary hierarchical logical structure.

Ethical Considerations

To ensure benchmark reliability, we conducted a human verification study with three expert NLP graduate students from our affiliated institution. We strictly followed fair labor practices by compensating participants at $15.00 USD per hour, a rate exceeding the local minimum wage. Electronic informed consent was obtained before the study began. The protocol was exempt from full Institutional Review Board (IRB) review as it involved non-invasive annotation of publicly available datasets like Visual Genome, GQA, and TextVQA without collecting any sensitive or personally identifiable information. We further verified that these datasets are free of offensive content and used consistent with their intended research purposes. Regarding broader impacts, this work focuses on diagnosing hallucinations in Large Vision-Language Models to improve system safety and trustworthiness. We foresee no negative societal consequences or dual-use risks.

Acknowledgments

This article is partially supported by Natural Science Foundation of China under Grants 62472178, Fundamental Research Funds for the Central Universities, the Key Technology Research and Development Program Project of the Shanghai Science and Technology Commission under Grants 25511107200, the Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems (No.MAIS2024111).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yihao Ang, Qiang Huang, Yifan Bao, Anthony KH Tung, and Zhiyong Huang. 2023. Tsgbench: Time series generation benchmark. *arXiv preprint arXiv:2309.03755*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R Manmatha. 2022. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16548–16558.

X Chen, R A Chi, X Wang, and 1 others. 2024a. Premise order matters in reasoning with large language models. *arXiv preprint arXiv:2402.08939*.

Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. 2024b. Premise order matters in reasoning with large language models. *arXiv preprint arXiv:2402.08939*.

Zhe Chen, Weiyun Wang, Y Cao, and 1 others. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267.

Paulius Danenas and Tomas Skersys. 2022. Exploring natural language processing in model-to-model transformations. *IEEE Access*, 10:116942–116958.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Tianrui Guan, Fuxiao Liu, Xudi Wu, Ruiqian Xian, Zongxia Li, Xiaoyu Liu, Xiyang Wang, Lijie Chen, Furong Furrer, Yanjie Ren, and 1 others. 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*, pages 14375–14385.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022a. Kat: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022b. Kat: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 956–968.

Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. 2020. Scene graph reasoning for visual question answering. *arXiv preprint arXiv:2007.01072*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Liubov Kovriguina, Ivan Shilin, Alexander Shipilo, and Alina Putintseva. 2017. Russian tagging and dependency parsing models for stanford corenlp natural language toolkit. In *International Conference on Knowledge Engineering and the Semantic Web*, pages 101–111. Springer.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Rongjie Li, Songyang Zhang, and Xuming He. 2022. Sgtr: End-to-end scene graph generation with transformer. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19486–19496.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Weizhe Lin and Bill Byrne. 2022a. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Weizhe Lin and Bill Byrne. 2022b. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llavanext: Improved reasoning, ocr, and world knowledge.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024c. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024d. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer.

Aakansha Mishra, Miriyala Srinivas Soumitri, and Vikram N Rajendiran. 2024. Learning representations from explainable and connectionist approaches for visual question answering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6420–6424. IEEE.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, and 1 others. 2023a. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. 2023b. Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21582–21592.

B Yan, Z Zhang, L Jing, and 1 others. 2025. Fiha: Automated fine-grained hallucinations evaluations in large vision language models with davidson scene graphs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12014–12026.

Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. 2022. Panoptic scene graph generation. In *European conference on computer vision*, pages 178–196. Springer.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, and 1 others. 2023. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11844–11857.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Jing Zhang, Jianbeom Huang, S Jin, and 1 others. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.

Yiliang Zhou, Hanley Ong, Patrick Kennedy, Carol C Wu, Jacob Kazam, Keith Hentel, Adam Flanders, George Shih, and Yifan Peng. 2024. Evaluating gpt-4v (gpt-4 with vision) on detection of radiologic findings on chest radiographs. *Radiology*, 311(2):e233270.

Yiyang Zhou, Chen Cui, Jihun Yoon, Linjun Zhang, Zhun Deng, A Tung, Chelsea andTung, and Liangke Hu. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

A Implementation Details

To ensure the full reproducibility of the K-GIP framework, this section details the parameter configurations for the dual-perception mechanism and the inference settings for the evaluated Multimodal Large Language Models (LVLMs).

A.1 Dual-Perception Module Configuration

Micro-Entity Perception. We utilize the pre-trained Grounding DINO (Swin-T backbone) (Liu et al., 2024d) as the core detector for extracting foundational visual facts. To strictly mitigate the risk of introducing noise into the ground truth, we adopt a *Precision-First* strategy. Specifically, we set a rigorous confidence threshold of $\tau = 0.35$. Detected objects with confidence scores below this value are filtered out. This conservative filtering ensures that the generated *Existence Anchor Questions* are grounded in definitive visual evidence, serving as a reliable lower bound for hallucination detection.

Macro-Context Perception. For the macro-context perception stage, we upgrade the captioning module to MiniCPM-V 2.6 (8B) (Bai et al., 2025a), a state-of-the-art edge-side LVLM. MiniCPM-V 2.6 is built upon the SigLip-400M visual encoder and the Qwen2-7B language model, exhibiting superior capabilities in Optical Character Recognition (OCR) and dense captioning compared to previous baselines like BLIP-2. To obtain high-fidelity panoramic descriptions, we employ a deterministic decoding strategy with `temperature=0` and a length penalty of 1.0. This configuration encourages the model to generate comprehensive and logically coherent descriptions, which are essentially required for constructing the *Visual Symbol Knowledge* and *Spatial Relational Knowledge* dimensions.

A.2 LVLM Inference Settings

For all evaluated LVLMs (e.g., Qwen-VL series, GPT-4V, LLaVA series), we conduct experiments in a zero-shot setting to assess their intrinsic capabilities without task-specific tuning. To eliminate randomness and ensure benchmark stability, we enforce greedy decoding by setting the sampling `temperature` to 0 and `top-p` to 1.0. The maximum generation length is restricted to 128 tokens, as verification tasks typically require concise responses.

B Prompt Templates

To facilitate future research, we provide the core prompt templates used in the K-GIP pipeline.

B.1 Adversarial Query Generation Framework

To facilitate reproducibility of the adversarial evaluation mechanism we present the exact prompt structure utilized in the Multi-Granularity Probing Generation stage as illustrated in Figure 9. This framework employs a rigorous system instruction to guide the language model in synthesizing high-quality visual verification queries based on the extracted dual-perception ground truth. The generation process strictly adheres to three distinct cognitive dimensions including Dense Entity Focus which targets minute object recall Spatial Logic Focus which evaluates topological reasoning capabilities and Visual Symbol Focus which assesses optical character recognition and pragmatic understanding.

The prompt design incorporates a specific color-coding strategy to distinguish between valid visual anchors and adversarial traps. The green text within the examples represents Positive Anchors derived directly from the verified ground truth entities and relations ensuring that existence queries are grounded in definitive visual evidence. Conversely the red text signifies Hallucination Traps which correspond to semantically plausible but visually absent objects designed to diagnose Logical Fractures. This structured approach forces the evaluated model to distinguish between actual visual perception and internal knowledge priors thereby providing a robust assessment of fine-grained hallucination resistance.

C Additional Quantitative Analysis

C.1 Sensitivity Analysis of Perception Thresholds

One potential limitation inherent in model-based ground truth construction lies in the dependency on the hyperparameters of the upstream detection module. To evaluate the stability of our proposed K-GIP framework, we performed a comprehensive sensitivity analysis by systematically adjusting the confidence threshold τ of the Grounding DINO detector from 0.25 to 0.45. This interval captures

Table 4: **Sensitivity Analysis of K-GIP Scores.** We report the Average F1 scores under different confidence thresholds τ for the Dense Entity Knowledge dimension. The strictly consistent ranking order across varying strictness levels demonstrates the robustness of the proposed metric.

Model	$\tau = 0.25$	$\tau = 0.35$	$\tau = 0.45$
mPLUG-Owl	42.1	43.0	43.5
LLaVA-1.5-13B	64.8	65.3	65.9
GPT-4V	77.9	78.5	79.1
Qwen2.5-VL-7B	79.8	80.3	80.6
Qwen3-VL-8B	**82.2**	**82.8**	**83.1**

a decisive trade-off region where lower threshold values prioritize high recall rates but introduce the risk of including noise or false positives into the ground truth. In contrast, higher threshold values enhance precision but increase the likelihood of missing subtle objects or false negatives.

The quantitative results displayed in Table 4 indicate that the absolute F1 scores exhibit minor fluctuations ranging between 0.5 percent and 0.9 percent as the filtering criteria become more stringent. Despite these numerical variations, the relative performance ranking among the evaluated models remains strictly consistent across all settings. Specifically, Qwen3-VL-8B maintains its position as the top-performing model followed by Qwen2.5-VL-7B, GPT-4V, LLaVA-1.5-13B, and mPLUG-Owl. This rank invariance confirms that K-GIP functions as a reliable diagnostic tool capable of reflecting the intrinsic reasoning capabilities of large vision-language models rather than being influenced by artifacts arising from the ground truth generation process.

C.2 Quantitative Impact of Logic Pruning

To intuitively demonstrate the regulatory effect of the proposed Sequential Logic Pruning protocol we present a comparative visualization across different model architectures in Figure 8. This analysis contrasts the performance using standard evaluation metrics against our dependency aware K-GIP scoring mechanism. Conventional evaluation paradigms treat attribute queries as isolated statistical events which frequently allows models to accrue points for correctly guessing properties such as color or location even when they fail to perceive the underlying object. In contrast our protocol imposes a strict topological constraint that nullifies these invalid reasoning branches.

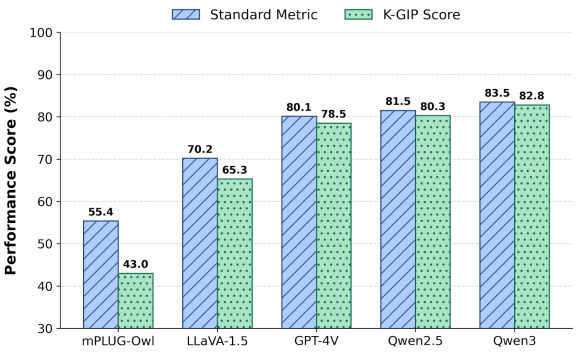


Figure 8: **Quantitative Analysis of Logical Fractures.** This figure contrasts the inflated performance under standard flat metrics versus the rigorous K-GIP scores after pruning invalid reasoning chains. The visual delta between the bars represents the Hallucination Gap where models fabricate attributes for non-existent objects.

The empirical results illustrated in Figure 8 reveal a ubiquitous performance regression across all architectures when these strict dependency constraints are applied. The vertical disparity between the standard metric and the K-GIP score quantitatively represents the Hallucination Gap where models successfully hallucinate attributes despite failing the existential premise. This phenomenon is notably pronounced in earlier architectures such as mPLUG-Owl and LLaVA-1.5 which exhibit severe logical fractures and a heavy reliance on language priors. Conversely state-of-the-art models including GPT-4V and Qwen3-VL demonstrate a significantly narrower margin. This stratification confirms that advanced models possess superior grounding capabilities whereas the pruning protocol effectively filters out unfounded conjectures to provide a rigorous assessment of true multimodal understanding.

D Human Verification Details

To ensure the reliability of the K-GIP benchmark and to validate the quality of the automatically generated Question-Answer (QA) pairs, we conducted a human verification study as mentioned in Section 4.2. This appendix provides detailed information regarding the annotation process, adhering to established responsible research guidelines.

D.1 Task Instructions

We provided the annotators with the original images and the corresponding generated QA pairs, instructing them to evaluate each pair based on a strictly double-blind protocol where they remained

unaware of which model or rule generated the specific question. The verification task was governed by two primary criteria. First, annotators assessed **Syntactic Validity** by checking whether the questions were grammatically correct, fluent, and unambiguous; any questions failing these standards were marked as Invalid. Second, they evaluated **Visual Consistency** to verify that the pipeline-generated ground-truth answers were strictly supported by the visual facts present in the image. This included confirming the presence or absence of objects for existence questions and validating specific visual details for attribute and relation queries. Furthermore, specific instructions were provided regarding adversarial negatives, where annotators were explicitly directed to mark an answer as correct if the system successfully identified a hallucinated object (such as a non-existent pen) with a negative response like "No" or "None". Ultimately, annotators labeled each QA pair as either Pass, meeting both criteria, or Fail.

D.2 Recruitment and Demographics

We recruited three annotators to perform the verification task through internal departmental announcements at the authors' affiliated institution. To ensure high-quality judgments, we required participants to have a background in linguistics or computational linguistics. Consequently, the selected annotators were graduate students specializing in Natural Language Processing (NLP). All participants demonstrated proficiency in English at a C1 level or higher, ensuring they could accurately judge the syntactic validity of the generated questions.

D.3 Compensation

We ensured that all annotators were fairly compensated for their time and effort. Annotators were paid at a rate of $15.00 USD per hour. The verification of 1,000 QA pairs required approximately 10 hours per annotator. This hourly rate exceeds the local minimum wage and was determined to be commensurate with the cognitive load required for the fine-grained verification task.

D.4 Data Consent and Ethics

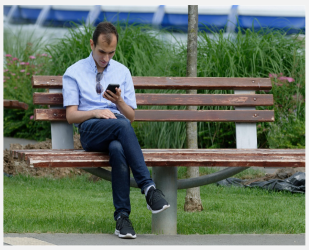
Prior to the commencement of the task, all participants were provided with a consent form explaining the purpose of the study, the nature of the task, and their right to withdraw at any time. Informed consent was obtained electronically before the task began. Regarding ethical approval, an official Institutional Review Board (IRB) approval was determined to be exempt. This decision relied on the fact that the study involved expert annotation of publicly available, non-sensitive datasets (such as Visual Genome, GQA, and TextVQA) and did not involve psychological experiments, intervention in the lives of the participants, or the collection of personally identifiable information (PII).

System Instruction:
You are an adversarial evaluator for Vision-Language Models. Your goal is to generate Visual Verification QA Pairs based on the provided Ground Truth (GT) information derived from the image.

Generation Rules:
1. Existence Anchors (Positive): For objects present in the GT, generate "Yes/No" questions confirming their presence.
2. Adversarial Negatives (Hallucination Traps): Select objects that are semantically similar to the GT objects but NOT present in the image. Generate questions asking about these absent objects.
3. Logical Dependency: If querying a non-existent object, ensure the expected answer is "No" or "None".
4. Format: Output the result as a structured JSON containing the Question, Answer, and Knowledge Type.

Example 1 (Dense Entity Focus):
[Ground Truth]: {'objects': ['Laptop', 'Mouse', 'Coffee Cup']}
[Generated Queries]:
(+) Positive Anchor: "Is there a mouse next to the laptop?" ——>Yes.
(-) Hallucination Trap: "Is there a pen lying on the desk?" ——> No. (Trap: Pens often co-occur with notebooks/laptops)

Example 2 (Spatial Logic Focus):
[Ground Truth]: {'objects': ['Man', 'Bench', 'Phone'], 'relation': ['Man holds Phone']}
[Generated Queries]:
(+) Positive Anchor: "Is the man sitting on a bench?" ——>Yes.
(-) Hallucination Trap: "Is the man holding a book?" ——> No. (Trap: 'Reading' often implies a book, testing fine-grained perception)



Example 3 (Visual Symbol Focus):
[Ground Truth]: {'objects': ['Sign'], 'text': ['STOP']}
[Generated Queries]:
(+) Positive Anchor: "Does the sign contain the text 'STOP'?" ——>Yes.
(-) Hallucination Trap: "Does the sign say 'DO NOT ENTER'?" ——> No. (Trap: Semantic similarity vs. Visual reality)

Figure 9: **The Adversarial Query Generation Prompt Framework.** We employ a structured instruction template to generate diverse visual verification queries targeting three cognitive dimensions: **Dense Entity Focus**, **Spatial Logic Focus**, and **Visual Symbol Focus**. In the examples, **Red text** indicates Hallucination Traps, representing semantically plausible but visually absent objects designed to diagnose Logical Fractures.