FROM STRUCTURED FACTS TO LOGICAL CONFLICTS: A BENCHMARK AND FRAMEWORK FOR VISUAL-TEXT CONFLICT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large multimodal models are typically evaluated on image-text inputs that are semantically consistent, leaving their ability to handle visual-textual contradictions underexplored. We introduce LogicCon, a benchmark for visual-textual logical conflict reasoning that requires models to detect conflicts, localize the conflicting fact pair, and diagnose the conflict type. LogicCon constructs controllable contradiction samples by converting visually grounded question-answer pairs into consistent declarative statements, applying structured fact mutations, and verifying each mutation against symbolic visual evidence. We evaluate 12 representative open-source and closed-source multimodal models under a unified zero-shot protocol. Results show that current models can often detect whether a conflict exists, but perform substantially worse when required to recover the complete conflicting fact pair or reason over compositional samples. A large gap between the strongest models and human performance further indicates that visual-textual logical conflict reasoning remains a challenging and unresolved capability. Comparative experiments further reveal that hallucination mitigation methods mainly improve basic conflict detection but provide limited gains in recovering complete conflicting fact pairs, while general reasoning methods still struggle with precise visual evidence grounding. By explicitly parsing target-aware claims, verifying visual evidence, and aggregating conflicting facts, our method achieves more consistent improvements on higher-level logical conflict reasoning tasks. Code and benchmark artifacts are available at `https://anonymous.4open.science/r/LogicCon-ICLR2027/`.

1 INTRODUCTION

In recent years, multimodal large language models Guo et al. (2025); Shao et al. (2024); Sun et al. (2024); Team et al. (2025) have made significant progress in visual question answering, image-text understanding, cross-modal retrieval, and visual reasoning Team et al. (2025). Benefiting from large-scale image-text pretraining and strong language modeling capabilities, these models can already generate fluent and visually grounded responses in many open-world scenarios. However, such progress is largely built upon the assumption that the image and the text are generally semantically consistent. When the text contains descriptions that are inconsistent with visual facts in the image, whether these models can still reliably identify such inconsistencies, further localize where the conflict occurs, and explain why the conflict arises remains insufficiently studied Yin et al. (2024).

This problem is not merely a general image-text mismatch judgment, but a form of visual–textual logical conflict reasoning. Unlike shallow semantic inconsistency, logical conflicts are often implicit: most of the text may agree with the image, while only a key attribute, object category, relation, spatial position, or quantity is altered. In other cases, local facts may appear plausible in isolation, but their combination contradicts the actual visual evidence. For such samples, models cannot rely on overall semantic similarity; they must decompose the statement into verifiable factual units and compare each unit against the image. Therefore, visual–textual logical conflict reasoning evaluates whether a model can perform fine-grained factual verification and conflict localization under cross-modal conditions. Although existing multimodal benchmarks have advanced evaluation in image-text consistency, visual grounding, and hallucination mitigation, systematic study of visual–textual logical conflict reasoning remains limited. First, many datasets focus on overall consistency judgments or shallow

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

object/attribute perturbations, with limited coverage of conflicts that require relation binding, multi-step localization, nested reference resolution, or compositional reasoning. Second, most evaluations provide only sample-level labels and lack explicit conflict-point annotations, making it difficult to determine whether a model has identified the exact factual pair that contradicts the image. Third, existing benchmarks rarely analyze model behavior by conflict type and reasoning complexity. These limitations prevent current evaluations from fully revealing the capability boundaries of multimodal large language models in conflict reasoning scenarios. To address this problem, we propose a new benchmark for visual–textual logical conflict reasoning. Rather than relying on random textual perturbations, our benchmark constructs contradictions through a logic-driven process grounded in structured visual facts. We convert GQA question-answer pairs Hudson & Manning (2019) into image-consistent declarative statements, apply controlled mutations to target objects and their attribute, relation, spatial, or quantity facts, and verify each mutation against the scene graph through symbolic consistency checking. This produces genuine factual conflicts with fine-grained annotations, including target objects, textual claims, visual evidence, transformation rules, and conflict types. Based on these annotations, we define a multi-level evaluation covering conflict detection, conflict point analysis, conflict type diagnosis, and compositional conflict reasoning. We further introduce a plug-and-play reasoning framework for existing MLLMs. Instead of directly judging the whole image-text pair, the framework decomposes the text into target-aware factual claims, verifies each claim against visual evidence, and aggregates the verified results into conflict detection, conflict localization, and type diagnosis. This explicit verification process reduces reliance on language priors and overall semantic similarity, making the framework more effective for implicit and compositional visual–textual conflicts.

Overall, the main contributions of this paper are as follows: 1. We propose a new benchmark for visual-textual logical conflict reasoning. The benchmark systematically covers five types of conflicts, including attribute, object, relation, spatial, and quantity conflicts, and additionally annotates a compositional subset to evaluate model capabilities in complex conflict reasoning scenarios. 2. We design a logic-driven and verifiable data generation pipeline to construct high-quality samples with clear conflict sources and structured annotations. 3. We propose a plug-and-play conflict reasoning framework that enhances the conflict reasoning ability of multimodal large language models through factual parsing, visual verification, and conflict aggregation. 4. We systematically analyze the performance of existing multimodal large language models on the proposed benchmark, reveal their significant limitations in implicit conflicts, compositional conflicts, and conflict point analysis, and validate the effectiveness of the proposed method on multi-level conflict reasoning tasks.

2 Related Work

2.1 Large vision-language models.

Large language models (LLMs) Brown et al. (2020) have profoundly transformed the field of natural language processing and have, in many tasks, effectively replaced traditional approaches. Recent commercial models Achiam et al. (2023); Team et al. (2024) have achieved remarkable performance on text benchmarks, while open-weight models have rapidly caught up within a short time frame Lu et al. (2024); Grattafiori et al. (2024); Nath et al. (2026). Building on this progress, researchers further combined large language models with visual encoders, enabling them to process multimodal inputs such as images and text. Early large vision-language models (LVLMs) Li et al. (2022); Alayrac et al. (2022) laid the foundation for truly multimodal language models and helped drive the development of subsequent commercial systems Jaech et al. (2024); Team et al. (2024). A series of open-source LVLMs have also been developed on top of open-weight models Bai et al. (2023); Chen et al. (2024); Dai et al. (2023); Li et al. (2025); Liu et al. (2023). These models generally follow a similar architecture: a visual encoder first extracts latent representations from images, which are then fed into a large language model together with textual inputs to enable multimodal understanding and generation.

2.2 Inconsistency Detection in Vision-Language

Existing studies on multimodal inconsistency mostly focus on narrow or specific scenarios. CrossCheck-Bench Tian et al. (2026) focuses on diagnosing compositional failures in cross-modal conflict resolution. MMIR Yan et al. (2025) investigates multimodal inconsistency reasoning in

layout-rich contents such as webpages, posters, and slides. Although these works advance the study of cross-modal conflicts, they either focus on shallow controlled conflicts or target complex document scenarios, and thus do not systematically characterize fine-grained logical conflict recovery under unified natural-image inputs. Works such as cross Popordanoska et al. further examine modality gaps in specific attributes, such as color and shape, but do not extend to real-world conflicts that require multi-step localization and compositional reasoning. Meanwhile, benchmarks such as VisuLogic Xu et al. (2025b) focus on general visual logical reasoning, emphasizing abilities such as quantity, spatial, and attribute comparison, but they do not address explicit contradictions between visual facts and textual claims. It is also worth noting that VLM2-Bench Zhang et al. (2025) studies visual cue matching across images or videos, such as matching the same person or object, which is orthogonal to aligning, comparing, and diagnosing contradictory visual-textual evidence within a unified input. In contrast, our work focuses on visual-textual logical conflict reasoning: models are required not only to determine whether the image and text conflict, but also to recover the target object, the textual conflicting detail, and the corresponding visual evidence, and further diagnose the conflict type. This provides a more systematic characterization of the capability boundaries of multimodal models in structural alignment and logical comparison.

3 Task Definition

Given an image I and a textual statement T, we define visual-textual logical conflict evaluation as a multi-level diagnostic task. The model first needs to determine whether T is consistent with the visual facts in I: $y \in \{\text{consistent}, \text{conflicting}\}$. If $y = \text{conflicting}$, the model is required to identify the inconsistent fact pair between the visual fact and the textual claim. We do not require the model to determine which modality is incorrect. Instead, we evaluate whether the model can identify the same conflict fact as the standard record. In addition to conflict detection, we also evaluate the model's ability to localize the conflict point. Each conflicting sample contains a structured mutation record during dataset construction, which records the target object, textual claim, visual fact, transformation rule, and conflict type. Based on this record, we separately evaluate whether the model correctly identifies the target object, the contradictory detail in the text, and the corresponding visual evidence in the image. Only when all three components are correctly identified do we consider the model to have fully localized the conflict point. We divide conflict types into five categories: attribute conflict, object conflict, relation conflict, spatial conflict, and counting conflict. The conflict type is obtained by mapping the semantic intent of the source question, rather than being directly determined by the surface form of the generated text. This category is used during dataset construction to constrain the generation direction of candidate contradictory statements. For samples that require multi-step relations or nested references to locate the target object, we additionally annotate them as compositional. On this subset, we report the accuracy of conflict detection and conflict point analysis. Overall, our benchmark contains three evaluation levels: (1) **conflict detection**, which determines whether the text is inconsistent with the image; (2) **conflict point analysis**, which identifies the target object, textual claim, and visual evidence; and (3) **conflict type diagnosis**, which determines the category of visual-factual inconsistency. In addition, based on the compositional annotation, we further analyze how model performance changes on complex reasoning samples.

4 LogicCon: Benchmark for Visual-Textual Contradiction Reasoning

4.1 Source Data and Task Taxonomy

We construct our vision-language logical contradiction benchmark based on GQA, which provides real images, question-answer pairs, scene graphs, and functional programs. Formally, each source sample is represented as $S = (I, Q, A, G, P)$, where I is the image, Q and A are the original question and answer, G is the scene graph, and P is the functional program. The scene graph specifies objects, attributes, and relations in the image, while the functional program describes the reasoning path needed to answer the question. Together, they allow us to align natural-language QA pairs with explicit visual facts and provide traceable evidence for contradiction generation and validation. We reformulate VQA into an image-text consistency judgment task. Given a source sample, we first convert the question-answer pair into an image-consistent declarative statement. We then use this statement as the linguistic

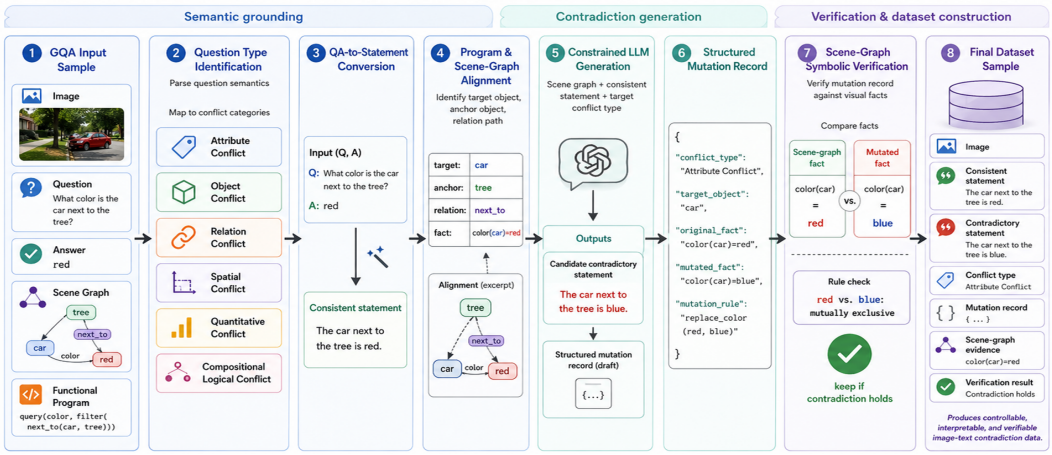

Figure 1: Overview of the LogicCon dataset construction pipeline. Starting from a sample containing an image, question, answer, scene graph, and functional program, we first identify the semantic intent of the original question, map it to a target conflict category, convert the question–answer pair into a statement consistent with the image, and align it with symbolic scene-graph evidence. We then use a constrained large language model to generate a contradictory statement together with a structured mutation record. The resulting mutation is further verified through symbolic checking on the scene graph to ensure that the original fact is visually supported and that the mutated fact forms a valid contradiction. The final dataset consists of image–text pairs annotated with conflict labels, conflict types, mutation records, visual evidence, and verification results.

backbone for generating candidate texts that contradict the image. During generation, we align the target object, anchor object, relational path, and answer fact with the scene graph and functional program, so that each contradiction is grounded in verifiable visual evidence. To support controllable generation and fine-grained evaluation, we define five conflict types according to the visual fact that is intentionally modified and verified: $\mathcal{C} = \{\text{Attribute}, \text{Object}, \text{Relation}, \text{Spatial}, \text{Quantitative}\}$. Specifically, attribute conflicts modify visual properties such as color, size, material, or state; object conflicts modify the target category or identity; relation conflicts modify non-spatial interactions such as holding, wearing, or riding; spatial conflicts modify spatial relations such as left/right, above/below, or on/under; and quantitative conflicts modify numbers, existence, or comparative relations. Importantly, the conflict type is determined by the verified mutation record rather than by the surface form of the question or answer. In addition, we label a sample as compositional if identifying the contradicted fact requires multi-step relations, nested references, or compositional conditions; otherwise, it is labeled as atomic. This label is used only for additional analysis.

4.2 Question-to-Statement Conversion and Visual Fact Alignment

Given a source sample $S = (I, Q, A, G, P)$, we first convert the original question-answer pair (Q, A) into a natural-language declarative statement that is consistent with the image. The goal of this step is to preserve the linguistic structure of the original question while transforming the VQA format into a declarative statement required for image-text consistency judgment. Formally, we denote the converted statement as T^+, where T^+ represents a text description filled with the original answer A and consistent with the image facts. This conversion is not a simple concatenation of the answer after the question. Different question types require different statement-conversion templates. Attribute questions are typically converted into attribute statements, object questions into relation-object statements, yes/no questions into affirmative or negative factual statements, and counting questions into counting statements. However, relying only on the surface text of the question may lead to ambiguous target objects. The object explicitly mentioned in the original question is not always the object being answered or modified. For example, in the question "What is next to the tree?", the tree serves as a locating anchor, while the answer object is the true target entity. To avoid such target mismatch, we further use the functional program P and the scene graph G to align the statement with visual facts. Specifically, according to the reasoning path described by P, we identify the target object, anchor objects, relational path, and answer fact in G, and construct a grounded semantic frame: $F = (o_t, O_a, R, f_a)$, where o_t denotes the target object, O_a denotes the set of anchor objects used to locate the target, R denotes the relational path between the target

object and the anchor objects, and f_a denotes the answer fact. For attribute questions, f_a can be instantiated as $\text{color}(o_t) = \text{red}$; for relation questions, f_a can be instantiated as $\text{holding}(o_t, o_a)$; for counting questions, f_a can be instantiated as the number of objects satisfying a given condition. This alignment process binds textual spans in the statement to specific visual entities and facts in the scene graph. For example, if T^+ is "The car next to the tree is red," the corresponding grounded frame can be represented as $o_t = \text{car}$, $O_a = \{\text{tree}\}$, $R = \text{next_to}(\text{car}, \text{tree})$, and $f_a = \text{color}(\text{car}) = \text{red}$. Subsequent contradiction generation is allowed to operate only on the explicitly aligned visual facts in this grounded frame, ensuring that the generated text remains semantically related to the original image and avoids modifying objects or attributes unrelated to the question target. In practice, we filter out samples that cannot be reliably converted into declarative statements or visually aligned. These include cases where the target object cannot be uniquely determined, the answer fact cannot be verified in the scene graph, or the statement is inconsistent with the reasoning target of the functional program. Finally, T^+ and its corresponding grounded semantic frame F are used together as the input to the next stage of constrained contradiction generation.

4.3 Constrained Contradictory Text Generation

After obtaining the consistent statement T^+ and its corresponding grounded semantic frame $F = (o_t, O_a, R, f_a)$, we generate candidate texts that contradict the image facts. Unlike directly applying surface-level word replacement to the statement, we adopt a constrained large language model generation strategy: the model is required not only to generate a contradictory statement, but also to output a structured mutation record that specifies which visual fact is modified, how it is modified, and which conflict type the modification corresponds to. Specifically, given the consistent statement T^+, scene-graph evidence, the target conflict type c, and the grounded semantic frame F, we construct a constrained prompt for the large language model to generate a candidate contradictory statement T^-. To reduce the noise introduced by open-ended generation, we explicitly restrict the modification scope in the prompt. For attribute conflicts, the model is allowed to change only the attribute value of the target object while keeping the object, anchor, and relational path unchanged. For object conflicts, the model is allowed to replace only the answer object category while preserving the original relational context. For relation conflicts, the model is allowed to modify or negate only the target relation. For spatial conflicts, the model is allowed to replace the spatial relation only with its predefined inverse relation. For quantitative conflicts, the model is allowed to change only the number, existence, or comparative relation. In this way, the generation process is constrained to the semantic slot corresponding to the target conflict type, rather than allowing the model to freely choose arbitrary conflict points. In addition to the candidate contradictory text T^-, the model is required to output a structured mutation record: $M = (c, o_t, f_{\text{orig}}, f_{\text{mut}}, r)$, where c denotes the conflict type, o_t denotes the modified target object, f_{orig} denotes the original visual fact, f_{mut} denotes the mutated textual fact, and r denotes the specific mutation rule. For example, for an attribute conflict, the mutation record can be written as $M = (\text{Attribute}, \text{car}, \text{color}(\text{car}) = \text{red}, \text{color}(\text{car}) = \text{blue}, \text{replace_color})$. This record provides a verifiable symbolic intermediate representation for the natural-language generation result. In other words, we do not directly trust the contradictory text generated by the model. Instead, we require the model to explicitly describe its semantic transformation process. In the next stage, the mutation record is compared with scene-graph facts through symbolic validation to determine whether the candidate text indeed forms a valid image-text contradiction. By binding candidate text generation with the structured mutation record, we convert open-ended language generation into a verifiable symbolic transformation. The mutation record provides a checkable intermediate representation for subsequent automatic filtering, evidence annotation, and fine-grained evaluation.

4.4 Symbolic Validation and Data Annotation

After obtaining the candidate contradictory statement T^- and its corresponding mutation record M, we further perform symbolic validation to ensure that the generated text indeed forms a logical conflict with the image facts. The core idea of this step is not to directly rely on the large language model to judge whether a contradiction exists, but to compare its structured transformation record with the visual facts in the scene graph. A candidate sample is retained only when the original fact is supported by the scene graph and the mutated fact satisfies the predefined contradiction constraints with respect to the original fact. Formally, given a mutation record $M = (c, o_t, f_{\text{orig}}, f_{\text{mut}}, r)$, we first check whether the original fact f_{orig} exists in the scene graph G, or whether it can be derived from the objects, attributes, and relations in G. Then, according to the conflict type c, we select the

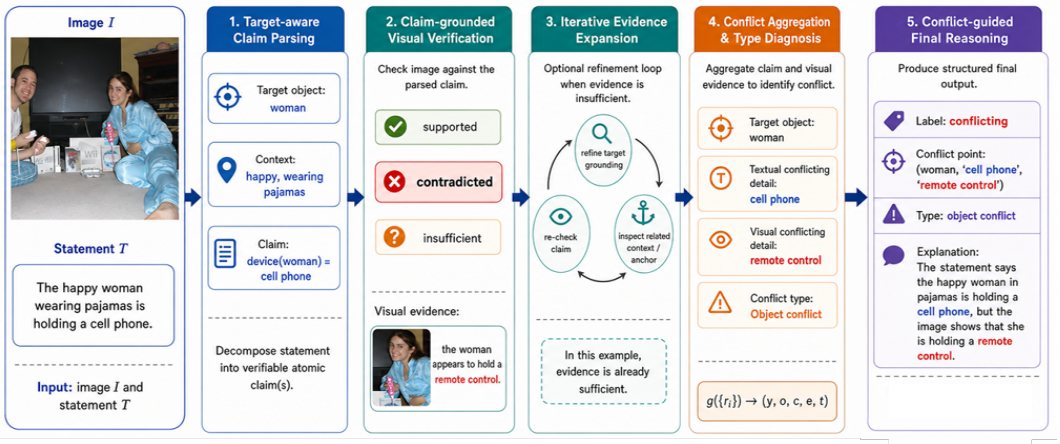

Figure 2: Overview of our reasoning framework. Given an image and a textual statement, the framework first parses the statement into target-aware atomic claims, then verifies each claim against visual evidence, optionally refines the evidence when needed, and finally aggregates the verified claim-level results to produce conflict detection, conflict point analysis, and conflict type diagnosis.

corresponding symbolic validation rule to determine whether f_{mut} is mutually exclusive with f_{orig}. More details are provided in Appendix K. In addition to mutation-record-level symbolic validation, we also perform a lightweight text consistency check to reduce inconsistencies between natural-language generation and the structured record. Specifically, we check whether the candidate text T^- preserves the target object and necessary anchor context, whether it expresses the mutated fact specified in the mutation record, and whether it introduces uncertain expressions or additional objects. If the target object in the candidate text is inconsistent with the target object in the mutation record, or if the generated text does not reflect the mutated fact, the sample is discarded. After the above validation, we retain only candidates that pass both the symbolic conflict check and the text-record consistency check. Any candidate that fails either step is discarded before constructing the final dataset. Finally, each data sample contains the image, original question, answer, consistent statement, contradictory statement, conflict type, reasoning complexity, target object, anchor object, original fact, mutated fact, mutation rule, and scene-graph evidence. By combining large language model generation with symbolic validation, our data construction pipeline preserves linguistic diversity while improving the reliability and interpretability of sample labels.

Human audit. To further assess the reliability of the automatically constructed samples, we conduct a stratified human audit of 500 conflicting examples, covering all five conflict types and both atomic and compositional cases. Overall, 94.6% of the audited samples satisfy all seven validity criteria simultaneously, corresponding to a post-stratified dataset-level validity estimate of 95.3%. Detailed protocols and analyses are provided in Appendix D.

5 METHOD

Existing multimodal large language models have demonstrated strong capabilities in general visual question answering and image-text understanding, but still struggle with visual–textual logical conflict reasoning. For implicit conflicts, such as local detail substitution, mismatched relational binding, reversed spatial relations, and quantity perturbation, models often make judgments based on overall semantic similarity rather than decomposing the statement into verifiable factual units and checking them against the image. As a result, they may overlook local but critical contradictions, or perceive an inconsistency in intermediate reasoning while still producing a final answer biased by language priors and surface fluency. To address this issue, we propose a plug-and-play visual–textual conflict reasoning framework that improves cross-modal conflict checking through explicit claim parsing, visual verification, and conflict aggregation. Given an image I and a textual statement T, our goal is to output a structured conflict reasoning result: $\mathcal{O} = (y, \hat{o}, \hat{c}, \hat{e}, \hat{t})$, where: $y \in \{\text{consistent}, \text{conflicting}\}$ indicates whether there is a conflict between the image and the text; $\hat{o}$ denotes the target object; $\hat{c}$ denotes the conflicting textual claim; $\hat{e}$ denotes the corresponding conflicting visual evidence in the image; $\hat{t}$ denotes the conflict type, which belongs to one of five categories: attribute conflict, object conflict, relation conflict, spatial conflict, and quantity conflict. To obtain the above output, we design a four-module reasoning framework. **Target-aware Claim Parsing** decomposes the

text into verifiable atomic claims with target objects and contextual constraints. **Claim-grounded Visual Verification** checks each claim against visual evidence around the target object. **Conflict Aggregation and Type Diagnosis** aggregates local verification results into conflict detection, conflict localization, and type diagnosis. **Conflict-guided Final Reasoning** produces the final response from these structured intermediate results. Scene graphs, functional programs, and mutation records are used only for dataset construction and symbolic verification. At test time, the framework requires only an image and a textual statement, making it a general inference-time plug-in for existing MLLMs.

5.1 Target-aware Claim Parsing and Visual Verification

For an input textual statement T, we do not directly judge whether it is consistent with the image. Instead, we first decompose it into a set of atomic factual units that can be verified against the image: $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$. Each factual unit is represented as $c_i = (o_i, p_i, a_i, m_i)$, where o_i denotes the target object, p_i denotes the predicate type, a_i denotes the attribute value, relational object, or numerical value to be verified, and m_i denotes the contextual constraints, such as anchor objects, modifiers, or nested reference information. After obtaining $\mathcal{C}$, we perform target-aware visual verification for each factual unit to determine whether the image supports the fact, contradicts it, or provides insufficient evidence. Specifically, the model outputs $r_i = (o_i, c_i, e_i, v_i)$, where e_i denotes the candidate visual evidence in the image, and $v_i \in \{\text{supported}, \text{contradicted}, \text{insufficient}\}$ indicates whether the image supports the factual unit, contradicts it, or does not provide sufficient evidence for judgment. Unlike methods that mainly focus on object existence verification, our framework uniformly handles five types of visual conflicts at the factual-unit level, including attribute, object, relation, spatial, and quantity conflicts. For simple samples, a single round of verification is usually sufficient to obtain a stable judgment. However, in the compositional subset, the target object often depends on multi-step relations or nested references for reliable localization. Therefore, a single verification step may only produce an **insufficient** result. To address this issue, we further introduce an iterative evidence supplementation mechanism. When the initial verification result is insufficient, the model further refines the localization of the target object, supplements visual evidence related to contextual anchors, decomposes the compositional fact into finer-grained sub-facts, and re-performs visual verification in a more focused manner. Let the verification result at the t-th iteration be $r_i^{(t)} = (o_i, c_i, e_i^{(t)}, v_i^{(t)})$. When the maximum number of iterations is reached or a stable judgment is obtained, we denote the final result as $\hat{r}_i = (o_i, c_i, \hat{e}_i, \hat{v}_i)$. This mechanism treats **insufficient evidence** as an intermediate state rather than a failure state, thereby improving the stability of conflict point analysis for complex conflict samples.

5.2 Conflict Aggregation and Final Reasoning

After obtaining the final verification results for all factual units, $\hat{\mathcal{R}} = \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n\}$, we generate the sample-level output through an aggregation function: $(y, \hat{o}, \hat{c}, \hat{e}, \hat{t}) = g(\hat{\mathcal{R}})$, where the outputs correspond to conflict detection, conflict point analysis, and conflict type diagnosis, respectively. Specifically, if any key factual unit is judged as **contradicted**, the sample is classified as **conflicting**. If all key factual units are judged as **supported**, the sample is classified as **consistent**. If multiple key factual units remain in the **insufficient** state, the framework outputs a more conservative judgment. For conflicting samples, we further recover the conflicting fact pair from $\hat{\mathcal{R}}$, including the target object $\hat{o}$, the conflicting textual claim $\hat{c}$, and the corresponding visual evidence $\hat{e}$ in the image. Based on this recovered conflict point, we further determine which type of visual factual inconsistency the conflict belongs to. Specifically, the conflict type is jointly determined by the predicate structure of the textual factual unit, the semantic type of the visual evidence, and their matching relationship. The final predicted conflict type is given by $\hat{t} \in \{\text{attribute}, \text{object}, \text{relation}, \text{spatial}, \text{quantity}\}$. For samples that require multi-step relational reasoning or nested reference resolution to localize the target object, we preserve the compositional flag and conduct separate analysis on this subset in our experiments. In the current implementation, $g(\cdot)$ is instantiated by calling the base model with a structured prompt. However, from the perspective of the method definition, it is an independent sample-level conflict diagnosis module rather than merely a prompt template. Finally, we feed the aggregated structured diagnostic results back to the base multimodal model to generate the final response. Compared with directly making a one-step judgment over (I, T), this process is grounded in explicit cross-modal factual verification results, and therefore can more stably perform conflict detection, conflict point analysis, and conflict type diagnosis.

Models	Avg.	Detection		Conflict Point Analysis				Conflict Type Diagnosis					Compositional		
		All	Conf	TO	TC	VE	CP	Attr	Obj	Rel	Spa	Qty	Comp-Det	Comp-CP	Comp-Type
Qwen3-VL-8B-Instruct	52.8	66.4	61.2	58.1	50.6	43.8	31.5	60.2	57.8	46.3	44.1	48.7	52.6	24.8	39.5
Qwen3.5-4B	51.9	65.1	59.7	56.4	49.1	42.3	29.8	58.9	56.2	45.0	43.4	47.1	51.0	23.6	38.2
MiMo-VL-7B-RL	56.3	69.2	64.4	61.5	54.2	46.8	35.7	63.8	61.1	49.7	47.9	51.2	56.4	28.9	42.7
InternVL3.5-14B	60.1	72.5	68.0	65.3	58.6	50.9	39.6	67.5	64.8	53.2	50.8	55.4	59.8	32.5	46.9
Qwen3.5-27B	61.4	73.8	69.1	66.8	60.4	52.5	41.2	68.9	66.2	54.4	51.9	56.7	61.2	34.1	48.5
Gemma 4-31B-IT	63.0	75.2	70.6	68.2	61.8	54.6	43.0	70.1	67.8	56.0	53.1	58.4	62.9	36.0	50.2
Qwen3.6-27B	62.6	74.7	70.0	67.9	61.1	54.0	42.3	69.8	67.1	55.6	52.8	57.9	62.1	35.4	49.6
Qwen3.8-27B	66.8	78.6	74.1	71.5	65.0	57.8	47.5	73.6	71.8	59.9	57.6	61.3	66.2	40.5	54.7
GPT-5.6 Luna	68.9	80.6	76.2	73.8	67.4	60.5	50.4	75.6	73.9	62.1	59.7	63.6	68.5	43.7	57.2
Grok 4.3 non-reasoning	70.1	81.7	77.5	75.0	68.8	61.7	51.9	77.0	75.1	63.2	60.5	64.8	69.8	45.1	58.7
GPT-5.4	72.1	83.4	79.3	76.8	71.2	64.1	54.6	79.2	77.8	66.0	63.1	67.4	72.0	48.0	61.5
Claude Opus 4.7	71.6	82.9	78.7	76.1	70.4	63.3	53.8	78.6	77.0	65.1	62.4	66.7	71.4	47.2	60.8
Human	92.3	95.8	94.1	93.6	90.8	88.2	82.7	94.2	93.0	89.8	88.4	90.7	90.1	78.6	85.9

Table 1: Main results of the LogicCon benchmark. All and Conf refer to overall detection and conflicting-sample detection accuracy. TO, TC, VE, and CP denote target object, textual conflicting detail, visual evidence, and full conflict point accuracy. Attr, Obj, Rel, Spa, and Qty denote attribute, object, relation, spatial, and quantity conflicts. Comp-Det, Comp-CP, and Comp-Type denote detection, conflict point analysis, and type diagnosis accuracy on the compositional subset.

6 Experiments

6.1 Experimental Setup

We evaluate 12 multimodal large models on our proposed visual-textual logical conflict reasoning benchmark. All models are evaluated under a unified zero-shot prompting protocol. Given only the input image and textual statement, each model is required to complete three levels of tasks: conflict detection, conflict point analysis, and conflict type diagnosis. We additionally report results on the compositional subset. Specifically, we evaluate detection accuracy, target object identification, textual conflicting detail identification, visual evidence identification, full conflict point analysis accuracy, and conflict type diagnosis accuracy. Closed-source models are accessed through their official APIs, while open-source models are evaluated using their official implementations on NVIDIA A100 GPUs. We adopt a hybrid scoring strategy: deterministic judgment and multiple-choice questions are evaluated by exact matching, while open-ended responses are semantically judged by GPT-4o. Human baseline performance is collected from five expert annotators following the same question-answering protocol. Details of the evaluation metrics are provided in Appendix G

6.2 Main Results

As shown in Table 1, existing multimodal large models still exhibit substantial limitations on LogicCon. Closed-source models generally outperform open-source models, yet a considerable gap to human performance remains. Across task levels, performance is substantially higher on conflict detection than on conflict point analysis and conflict type diagnosis. While models can often recognize that an inconsistency exists, their accuracy drops markedly when they must jointly identify the target object, the conflicting textual detail, and the corresponding visual evidence. This gap is particularly pronounced for the full conflict point (CP) metric, which requires all three components to be correct simultaneously. The difficulty also varies systematically across conflict types and reasoning complexity. Attribute and object conflicts are relatively easier, whereas relation, spatial, and quantitative conflicts are more challenging, suggesting that models are more reliable at recognizing direct visual properties than at reasoning over inter-object relations, spatial configurations, and constrained quantities. Performance further decreases on the compositional subset, especially for conflict point recovery. When target identification requires multi-step relations, nested references, or contextual anchors, models are more likely to fail to recover the complete conflicting fact pair. Finally, model scaling provides uneven gains. Stronger models generally achieve higher conflict detection accuracy, but the improvements are smaller for conflict point recovery, type diagnosis, and compositional reasoning. Moreover, the advantage of closed-source models becomes more apparent on these higher-level metrics. These results suggest that improving basic inconsistency recognition alone is insufficient; visual-textual logical conflict reasoning additionally requires fine-grained visual grounding, fact-level verification, and reliable conflict aggregation. We further evaluate our method on the MMIR dataset Yan et al. (2025) to examine its generalization beyond LogicCon. Detailed results and analysis are provided in Appendix E.

Backbone	Method	Avg.	Detection	CP	Type	Comp-CP
InternVL3.5-8B	LURE Zhou et al. (2023)	45.0	60.1	24.0	43.8	18.1
	VCD Leng et al. (2024)	47.2	62.8	25.7	45.6	19.8
	LLaVA-CoT Xu et al. (2025a)	51.5	66.1	31.8	50.8	25.9
	Chain-of-Action Pan et al. (2024)	50.2	64.9	29.6	49.1	23.7
	LogicCheckGPT Wu et al. (2024)	52.6	67.3	33.4	52.3	27.6
	Ours	**56.5**	**70.0**	**37.4**	**55.6**	**31.0**
GPT-4o-mini	LLaVA-CoT Xu et al. (2025a)	60.8	73.5	40.8	57.0	33.6
	Chain-of-Action Pan et al. (2024)	59.5	72.1	38.9	55.2	31.5
	LogicCheckGPT Wu et al. (2024)	61.9	74.4	42.6	58.4	35.4
	Ours	**65.8**	**77.0**	**46.5**	**61.7**	**39.4**
Qwen3.5-9B	LURE Zhou et al. (2023)	59.9	72.9	38.2	58.1	31.7
	VCD Leng et al. (2024)	61.8	75.0	40.4	59.8	33.9
	LLaVA-CoT Xu et al. (2025a)	66.4	78.6	46.7	65.2	40.8
	Chain-of-Action Pan et al. (2024)	65.2	77.4	44.9	63.7	38.9
	LogicCheckGPT Wu et al. (2024)	67.5	79.2	48.3	66.5	42.6
	Ours	**71.2**	**81.7**	**52.1**	**69.5**	**46.8**
Qwen3.5-27B	LURE Zhou et al. (2023)	63.1	75.1	42.6	60.9	35.2
	VCD Leng et al. (2024)	-	77.0	44.5	62.3	37.1
	LLaVA-CoT Xu et al. (2025a)	68.9	80.3	50.2	67.5	43.5
	Chain-of-Action Pan et al. (2024)	67.8	79.1	48.4	65.8	41.6
	LogicCheckGPT Wu et al. (2024)	70.1	81.2	52.1	69.0	45.7
	Ours	**73.6**	**83.4**	**55.5**	**72.0**	**49.6**

Table 2: Main comparison of plug-and-play reasoning methods on representative backbones for visual-text logical conflict reasoning.

6.3 Comparison with Existing Models

Table 2 compares different plug-and-play methods on four representative backbones. Overall, all methods improve the base models to some extent, but the gains vary noticeably across metrics and models. Our method achieves the best results on all backbones, showing that it can consistently enhance visual-textual logical conflict reasoning across models of different scales. In terms of metrics, the gains on Detection are generally smaller than those on CP and Comp-CP. This suggests that identifying whether a conflict exists is relatively easier, while recovering the complete conflicting fact pair is more challenging. Methods such as LURE and VCD mainly focus on hallucination mitigation or visual grounding, and therefore provide certain improvements on detection. However, their gains on conflict point analysis and compositional samples remain limited, indicating that simply reducing visual hallucinations is insufficient for resolving fine-grained logical conflicts. In contrast, methods with more structured reasoning processes, such as LLaVA-CoT and Chain-of-Action, perform better on CP, Type, and Comp-CP. This shows that high-level conflict reasoning requires not only detecting inconsistency, but also localizing the target object, identifying the textual conflicting detail, grounding the corresponding visual evidence, and integrating them into a coherent conflict diagnosis. Therefore, compared with generic hallucination correction, multi-step reasoning and verification mechanisms are better suited to this task. Among all baselines, LogicCheckGPT achieves the strongest overall performance, suggesting that explicit logical checking is effective for this task. Nevertheless, our method consistently outperforms LogicCheckGPT across all backbones, with clearer advantages on CP and Comp-CP. This indicates that our method not only improves conflict awareness, but also more effectively supports target-aware fact parsing, visual evidence verification, and conflict aggregation. In addition, improvements are usually larger on weaker backbones and smaller on stronger ones. Even so, our method still brings clear gains in conflict point recovery and compositional reasoning on Qwen3.5-27B, suggesting that stronger backbones alone cannot fully resolve the bottleneck of visual-textual logical conflict reasoning. More detailed results are provided in Appendix C.

7 Conclusion

We presented LogicCon, a benchmark for visual-textual logical conflict reasoning in multimodal large models. LogicCon moves beyond standard vision-language evaluation by requiring conflict detection, conflict point localization, and conflict type diagnosis under controlled and verifiable image-text contradictions. Experiments show that current models can often detect whether a conflict exists, but still struggle to recover the full conflicting fact pair, especially on compositional samples. Comparative results further suggest that generic enhancement strategies are insufficient for this setting, while conflict-oriented reasoning and verification are more effective. We hope LogicCon will support future research on more reliable and logically grounded multimodal.

AI USE STATEMENT

Generative AI tools were used in this work in several roles. First, they were used to assist with Chinese–English translation and language polishing during manuscript preparation. Second, a large language model was used as part of the LogicCon data construction pipeline to generate candidate contradictory statements and their associated structured mutation records under constrained prompts. These generated candidates were not accepted directly. They were subjected to scene-graph-based symbolic validation and text–record consistency checks, and the resulting dataset was further evaluated through a stratified human audit, as described in the paper. Third, GPT-4o was used as an automatic semantic judge for open-ended model responses. Deterministic labels and multiple-choice outputs were evaluated using exact matching. The evaluation criteria and reference annotations were defined independently of the judge model. Finally, the multimodal language models reported in our experiments were used as subjects of evaluation and, where applicable, as backbone models within the proposed inference framework. Their use in benchmark experiments should therefore be distinguished from the use of generative AI as a writing assistant. All AI-assisted manuscript text was reviewed and revised by the authors. Reported experimental results were obtained through the evaluation procedures described in the paper and were not manually generated or altered by generative AI. The authors reviewed the resulting analyses and take full responsibility for the final manuscript, experimental methodology, dataset, reported results, and scientific conclusions.

ETHICS STATEMENT

LogicCon is constructed from existing vision–language data rather than from newly collected personal data. The benchmark builds on images and structured annotations from GQA and introduces derived textual statements, conflict annotations, and evaluation records. Any public release of LogicCon will follow the licenses and terms of use of the underlying data sources. The benchmark is intended for research on the reliability, grounding, and logical reasoning capabilities of multimodal models. Although the source images may contain people, LogicCon does not introduce identity annotations or perform biometric identification, and the benchmark is not designed for high-stakes decision-making about individuals. Because LogicCon is derived from an existing vision–language dataset, it may inherit biases and coverage limitations present in the source data, including biases in visual concepts, object distributions, geographic or cultural representation, and scene annotations. Benchmark performance should therefore not be interpreted as evidence of model reliability across all real-world populations, domains, or safety-critical applications. Our data construction pipeline uses automated language-model generation, which may introduce incorrect or ambiguous examples. To mitigate this risk, generated samples undergo structured symbolic validation and consistency filtering, and we additionally conduct a stratified human audit of the final dataset. We report the remaining limitations and error patterns rather than assuming that automatic verification guarantees perfect annotation quality. The human audit evaluates the validity of benchmark examples and does not use annotator demographic characteristics as part of the analysis. We will release the benchmark and associated code with documentation describing the data sources, construction procedure, validation process, and known limitations. The authors take responsibility for the responsible use and release of the resulting research artifacts.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we release the LOGICCON benchmark, evaluation pipeline, and implementation of our reasoning framework through an anonymous project page at `ANONYMOUS-LINK`. The release contains the dataset annotations and official train/validation/test splits, source image identifiers, preprocessing scripts, evaluation and scoring code, prompting templates, inference configurations, and the code used to reproduce the main benchmark and method-comparison experiments. For reproducibility of model evaluation, we document the model versions, inference settings, input/output formats, and metric computation procedures used throughout the experiments. The released evaluation scripts implement the conflict detection, target-object, textual-conflict, visual-evidence, conflict-point, conflict-type, and compositional metrics reported in the paper. LogicCon is derived from GQA, and the original image files are therefore not redistributed as part of our release. Each LogicCon sample retains its source image identifier, and we provide scripts for resolving the corresponding images

from a locally downloaded copy of the official GQA image distribution. All materials are provided anonymously during the double-blind review period. A permanent public release will be provided after the review process.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, pp. 24185–24198, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Drew A Hudson and Christopher D Manning. 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*, pp. 6700–6709, 2019.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Mriganka Nath, Anurag Das, Jiahao Xie, and Bernt Schiele. Clipttt: Clip-guided test-time training helps lvlms see better. *arXiv preprint arXiv:2603.26486*, 2026.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. Chain-of-action: Faithful and multimodal question answering through large language models. *arXiv preprint arXiv:2403.17359*, 2024.

Teodora Popordanoska, Jiameng Li, and Matthew B Blaschko. Crosscheck: A vision-language conflict detection benchmark.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. Visual agents as fast and slow thinkers. *arXiv preprint arXiv:2408.08862*, 2024.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Baoliang Tian, Yuxuan Si, Jilong Wang, LingYao Li, Zhongyuan Bao, Zineng Zhou, Tao Wang, Sixu Li, Ziyao Xu, Mingze Wang, et al. Crosscheck-bench: Diagnosing compositional failures in multimodal conflict resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25887–25895, 2026.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6944–6962, 2024.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025a.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025b.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18829–18845, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R Fung. Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7510–7545, 2025.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

A LIMITATIONS

Although LogicCon provides a fine-grained benchmark for visual–textual logical conflict reasoning, several aspects remain for future exploration. First, our construction pipeline relies on structured visual annotations to generate and verify conflicts. This design improves controllability and annotation quality, but it also means that the current benchmark is mainly built from data sources where such structured information is available. Extending the construction process to more diverse visual domains without explicit scene-graph annotations would further improve dataset coverage. Second, LogicCon focuses on five common types of visual–textual conflicts: attribute, object, relation, spatial, and quantitative conflicts. These categories cover many important forms of factual inconsistency, but they may not include all possible real-world multimodal conflicts, such as temporal changes, causal inconsistencies, or conflicts involving external knowledge. Future work can extend the taxonomy to broader conflict scenarios. Third, although our plug-and-play reasoning framework improves conflict localization and compositional reasoning, it still depends on the underlying multimodal model's visual perception and grounding ability. When the base model fails to recognize the relevant object, read visual text, or localize the correct evidence region, the framework may also produce incomplete or incorrect verification results. Improving low-level perception and evidence localization remains an important direction. Finally, our evaluation uses structured annotations and GPT-based semantic judgment for open-ended conflict point analysis. While this allows flexible evaluation of natural-language responses, automatic judging may still introduce minor inconsistencies. Future work may further combine automatic evaluation with larger-scale human validation to improve evaluation reliability.

B ABLATION STUDY

Table 3 reports the ablation results on Qwen3.5-9B and Qwen3.5-27B. Overall, each component contributes to the final performance, and the full framework achieves the best results across both backbones. Target-aware claim parsing brings the first consistent improvement by decomposing the input statement into verifiable factual units and explicitly identifying the target object and textual conflicting detail. This mainly improves TO and TC, indicating that making the textual conflict structure explicit is helpful for locating what should be checked.

The largest improvement comes from claim-grounded visual verification, especially on VE, CP, and Comp-CP. This shows that the main bottleneck of visual-textual logical conflict reasoning is not simply detecting whether an inconsistency exists, but grounding the corresponding visual evidence and comparing it with the textual claim. After adding visual verification, the model becomes better at recovering the complete conflicting fact pair, rather than only producing a sample-level conflict judgment.

Conflict aggregation further improves conflict type diagnosis and compositional performance. This suggests that local verification results still need to be organized into a coherent sample-level decision; otherwise, models may detect partial inconsistencies but fail to produce a stable conflict label or type. Finally, conflict-guided final reasoning provides smaller but consistent gains by reducing answer drift and making the final output follow the structured intermediate results. The improvements are larger on Qwen3.5-9B, while Qwen3.5-27B still benefits clearly on CP and Comp-CP, indicating that stronger backbones do not fully solve fine-grained conflict localization and compositional reasoning.

C COMPARISON WITH EXISTING MODELS

Table 4 provides a fine-grained comparison of different plug-and-play methods across four representative backbones. Overall, our method consistently achieves the best performance on all backbones and across most evaluation dimensions, including overall average score, conflict detection, conflict point analysis, conflict type diagnosis, and compositional reasoning. This demonstrates that the proposed framework is not tied to a specific backbone, but can stably improve visual–textual logical conflict reasoning for models with different capacities. A closer look at the detection results shows that most plug-and-play methods improve the base models on All and Conf accuracy. Methods such as LURE and VCD mainly improve this level, indicating that hallucination mitigation and visual grounding can help models better recognize whether an inconsistency exists. However, their gains

Backbone	Ablation Setting	Avg.	Detection		Conflict Point Analysis				Conflict Type Diagnosis					Compositional		
			All	Conf	TO	TC	VE	CP	Attr	Obj	Rel	Spa	Qty	Comp-Det	Comp-CP	Comp-Type
Qwen3.5-9B	Base MLLM	63.4	75.2	83.5	67.8	61.2	53.0	40.6	69.1	66.8	58.4	55.6	56.3	65.1	34.5	50.2
	+ Claim Parsing	66.0	77.3	85.5	72.4	66.3	55.8	44.2	71.2	69.0	60.7	58.0	58.4	67.2	38.0	53.7
	+ Visual Verification	68.9	79.6	87.8	75.8	69.5	61.5	49.1	74.8	72.5	63.9	61.4	60.5	70.5	43.2	57.8
	+ Conflict Aggregation	70.2	80.8	89.0	77.1	71.0	63.7	50.8	76.5	74.3	65.8	63.1	61.4	72.3	45.4	60.0
	Full Framework	**71.2**	**81.7**	**90.1**	**78.2**	**72.0**	**65.1**	**52.1**	**78.0**	**75.8**	**67.1**	**64.4**	**62.2**	**74.0**	**46.8**	**61.3**
Qwen3.5-27B	Base MLLM	66.8	78.4	86.6	72.1	65.7	58.0	45.2	72.6	70.1	61.4	58.8	56.2	67.8	38.3	54.0
	+ Claim Parsing	68.8	79.8	88.0	75.8	69.8	60.3	48.2	74.5	72.0	63.6	61.0	58.2	69.8	41.5	56.8
	+ Visual Verification	71.2	81.6	89.4	78.3	72.2	65.4	53.0	78.2	75.7	67.1	64.4	61.1	72.8	46.7	61.0
	+ Conflict Aggregation	72.5	82.5	90.2	79.4	73.5	66.9	54.4	79.6	77.1	68.7	66.0	62.4	74.3	48.3	62.8
	Full Framework	**73.6**	**83.4**	**91.0**	**80.5**	**74.4**	**68.2**	**55.5**	**81.0**	**78.5**	**70.0**	**67.4**	**63.1**	**76.0**	**49.6**	**64.2**

Table 3: Ablation study of our plug-and-play visual-textual conflict reasoning framework. We report results on Qwen3.5-9B and Qwen3.5-27B by progressively adding target-aware claim parsing, claim-grounded visual verification, conflict aggregation, and final reasoning.

become much smaller in the conflict point analysis block, especially on VE and CP. Since CP is counted as correct only when the target object, textual conflicting detail, and visual evidence are all correctly identified, this gap suggests that improving inconsistency awareness alone is insufficient for recovering the complete conflicting fact pair. Reasoning-oriented methods, such as LLaVA-CoT and Chain-of-Action, bring larger gains than hallucination-mitigation methods on CP and conflict type diagnosis. This indicates that explicit reasoning is helpful for connecting local observations with a final conflict decision. However, these methods still rely heavily on the model's internal reasoning trajectory and do not explicitly enforce target-aware visual evidence verification. As a result, they may identify the general conflict direction but still fail to precisely ground the visual evidence or recover all components of the conflict point. LogicCheckGPT is the strongest baseline among the compared methods. Its advantage over LURE, VCD, and general CoT-style methods shows that explicit logical checking is better aligned with our benchmark than generic hallucination correction. Nevertheless, our method consistently outperforms LogicCheckGPT on all four backbones. The improvement is especially clear on VE, CP, and Comp-CP, suggesting that logical self-consistency alone is not enough for fine-grained visual–textual conflict reasoning. Compared with LogicCheckGPT, our framework explicitly parses target-aware claims, verifies the corresponding visual evidence, and aggregates verified facts into conflict localization and type diagnosis, which leads to more reliable conflict point recovery. The conflict type diagnosis block further reveals different levels of difficulty across conflict categories. Attribute and object conflicts are generally easier for all methods, because they often depend on direct visual recognition. In contrast, relation, spatial, and quantity conflicts are more challenging, as they require models to compare objects, reason about spatial layouts, or verify numerical constraints. Our method obtains larger gains on these harder categories, indicating that structured claim parsing and evidence verification are particularly useful when the conflict cannot be resolved by simple object or attribute recognition. The compositional block shows the largest performance gap between our method and prior plug-and-play baselines. While most methods improve Comp-Det to some extent, their gains on Comp-CP and Comp-Type remain limited. This suggests that detecting a conflict in compositional samples is easier than locating the full conflicting fact pair or diagnosing its type. Our method achieves the strongest Comp-CP and Comp-Type results because it explicitly decomposes complex claims, verifies contextual anchors, and aggregates local evidence before producing the final diagnosis. This is especially important when the target object must be located through multi-step relations, nested references, or compositional constraints. The improvements are also more pronounced on weaker backbones such as InternVL3.5-8B and GPT-4o-mini, where the base models have limited fine-grained grounding and reasoning ability. On stronger backbones such as Qwen3.5-9B and Qwen3.5-27B, the gains are smaller but remain consistent, especially on CP and compositional metrics. This trend suggests that stronger backbones already provide better basic perception and conflict detection, but still lack a reliable mechanism for structured conflict localization and visual evidence verification. Therefore, the bottleneck of LogicCon cannot be fully addressed by backbone scaling alone. Overall, the fine-grained comparison supports three conclusions. First, hallucination mitigation methods mainly improve basic conflict detection, but are insufficient for complete conflict point recovery. Second, general reasoning methods improve high-level reasoning to some extent, but still lack precise target-aware evidence grounding. Third, our framework is more effective because it directly matches the structure of the task: it decomposes textual claims, verifies visual evidence, and aggregates local contradictions into conflict

labels, conflict points, and conflict types. These results explain why our method brings the largest gains on the most challenging metrics, especially CP, VE, Comp-CP, and Comp-Type.

Backbone	Method	Avg.	Detection		Conflict Point Analysis				Conflict Type Diagnosis					Compositional		
			All	Conf	TO	TC	VE	CP	Attr	Obj	Rel	Spa	Qty	Comp-Det	Comp-CP	Comp-Type
InternVL3.5-8B	LURE Zhou et al. (2023)	45.0	60.1	69.0	46.5	39.2	32.8	24.0	51.6	49.3	39.2	36.8	42.1	44.7	18.1	31.8
	VCD Leng et al. (2024)	47.2	62.8	71.6	49.0	41.8	34.7	25.7	53.1	51.0	40.9	38.6	44.4	47.1	19.8	33.6
	LLaVA-CoT	51.5	66.1	75.3	55.7	48.4	40.6	31.8	59.0	56.3	45.7	43.2	49.8	52.6	25.9	39.7
	Chain-of-Action	50.2	64.9	74.1	53.0	46.1	38.4	29.6	56.8	54.6	43.8	41.5	48.8	50.9	23.7	37.8
	LogicCheckGPT	52.6	67.3	76.2	57.2	50.0	42.1	33.4	60.7	58.2	47.2	44.6	50.8	54.1	27.6	41.4
	Ours	**56.5**	**70.0**	**79.0**	**61.6**	**54.8**	**47.3**	**37.4**	**64.8**	**62.1**	**51.3**	**48.9**	**55.0**	**58.7**	**31.0**	**45.6**
GPT-4o-mini	LLaVA-CoT	60.8	73.5	82.4	65.8	59.6	52.1	40.8	65.9	63.5	54.4	52.2	58.9	62.7	33.6	48.7
	Chain-of-Action	59.5	72.1	81.0	63.7	57.4	50.0	38.9	64.0	61.8	52.7	50.5	57.0	60.8	31.5	46.9
	LogicCheckGPT	61.9	74.4	83.2	67.2	60.9	54.0	42.6	67.2	64.8	55.9	53.6	60.5	64.2	35.4	50.5
	Ours	**65.8**	**77.0**	**85.6**	**71.4**	**65.2**	**58.9**	**46.5**	**70.5**	**68.4**	**59.2**	**56.8**	**63.6**	**68.3**	**39.4**	**54.7**
Qwen3.5-9B	LURE Zhou et al. (2023)	59.9	72.9	82.0	64.2	57.3	50.6	38.2	65.6	63.1	54.8	52.0	55.1	62.5	31.7	47.3
	VCD Leng et al. (2024)	61.8	75.0	83.7	66.5	59.6	52.7	40.4	67.5	65.2	56.5	53.9	55.9	64.4	33.9	49.0
	LLaVA-CoT	66.4	78.6	87.0	72.5	66.3	58.8	46.7	73.0	70.7	62.3	59.8	60.4	69.0	40.8	55.8
	Chain-of-Action	65.2	77.4	86.0	70.4	64.2	56.5	44.9	71.4	69.0	60.6	58.1	59.4	67.2	38.9	53.9
	LogicCheckGPT	67.5	79.2	87.8	74.0	67.6	60.2	48.3	74.6	72.3	63.9	61.2	60.5	70.1	42.6	57.0
	Ours	**71.2**	**81.7**	**90.1**	**78.2**	**72.0**	**65.1**	**52.1**	**78.0**	**75.8**	**67.1**	**64.4**	**62.2**	**74.0**	**46.8**	**61.3**
Qwen3.5-27B	LURE Zhou et al. (2023)	63.1	75.1	84.0	68.4	61.5	54.1	42.6	69.8	67.2	58.3	55.7	53.5	65.0	35.2	51.6
	VCD Leng et al. (2024)	64.8	77.0	85.6	70.6	63.9	56.3	44.5	71.6	69.1	60.0	57.4	53.4	66.8	37.1	53.0
	LLaVA-CoT	68.9	80.3	88.6	75.4	69.0	62.1	50.2	76.2	73.8	64.9	62.3	60.3	71.2	43.5	58.4
	Chain-of-Action	67.8	79.1	87.4	73.5	67.2	60.0	48.4	74.4	72.1	63.2	60.6	58.8	69.5	41.6	56.7
	LogicCheckGPT	70.1	81.2	89.3	77.0	70.9	64.0	52.1	77.6	75.1	66.5	63.8	62.0	72.8	45.7	60.0
	Ours	**73.6**	**83.4**	**91.0**	**80.5**	**74.4**	**68.2**	**55.5**	**81.0**	**78.5**	**70.0**	**67.4**	**63.1**	**76.0**	**49.6**	**64.2**

Table 4: Fine-grained comparison of plug-and-play reasoning methods on representative backbones. We report the overall average score, detection accuracy, conflict point analysis, conflict type diagnosis, and compositional performance.

D HUMAN VERIFICATION

To assess the reliability of the automatically generated conflict samples, we conduct a stratified human audit on 500 samples from the final LogicCon dataset. We stratify the data along two dimensions: five conflict types (Attribute, Object, Relation, Spatial, and Quantitative) and two reasoning-complexity levels (Atomic and Compositional), resulting in ten strata. We randomly sample 50 examples from each stratum, yielding 500 audited samples in total. This balanced design provides sufficient coverage of all conflict categories while ensuring that compositional cases, which typically involve multi-step relations and more complex reference resolution, are evaluated separately from atomic cases.

Each sample is independently evaluated by three annotators. The annotators are provided with the image, the original consistent statement, the mutated statement, and the associated structured annotations. They assess seven aspects of sample validity: (1) whether the original statement is supported by the image; (2) whether the mutated statement forms a genuine conflict with the visual content; (3) whether the target-object annotation is correct; (4) whether the textual conflicting fact is correctly identified; (5) whether the annotated visual evidence correctly supports the conflict; (6) whether the conflict-type label is correct; and (7) whether the sample contains no ambiguity or unintended secondary conflict. Majority voting across the three annotators is used to determine the final label for each criterion.

Overall, 473 out of 500 samples satisfy all seven criteria simultaneously, corresponding to a fully valid rate of 94.6%. The associated 95% Wilson confidence interval is 92.3–96.3%. As shown in Table 6, individual audit dimensions achieve pass rates between 95.8% and 99.0%. In particular, conflict-type annotations achieve the highest reliability at 99.0%, while the "no ambiguity or unintended conflict" criterion is relatively more challenging, with a pass rate of 95.8%. Inter-annotator agreement is consistently strong, with Krippendorff's α ranging from 0.76 to 0.93 across individual dimensions and a macro-averaged α of 0.84.

Because the audit intentionally uses an equal number of samples across strata rather than following the empirical distribution of LogicCon, we additionally post-stratify the audit results according to the actual distribution of the full dataset. This yields an estimated dataset-level validity of 95.3%, further indicating that the automatically generated samples are reliable at the dataset level.

Conflict Type	Atomic	Compositional	Total
Attribute	50	50	100
Object	50	50	100
Relation	50	50	100
Spatial	50	50	100
Quantitative	50	50	100
Total	**250**	**250**	**500**

Table 5: Stratified sampling protocol for the human verification study.

Human Audit Dimension	Majority-Vote Pass	Pass Rate	95% Wilson CI	Krippendorff's α
Original statement supported by the image	492/500	98.4%	96.9–99.2	0.82
Mutated statement forms a genuine conflict	487/500	97.4%	95.6–98.5	0.86
Target-object annotation is correct	490/500	98.0%	96.4–98.9	0.88
Textual-conflict annotation is correct	491/500	98.2%	96.6–99.1	0.89
Visual-evidence annotation is correct	483/500	96.6%	94.6–97.9	0.79
Conflict-type annotation is correct	495/500	99.0%	97.7–99.6	0.93
No ambiguity or unintended conflict	479/500	95.8%	93.7–97.2	0.76
All seven criteria satisfied	**473/500**	**94.6%**	**92.3–96.3**	**0.84**

Table 6: Human verification results across seven dataset-quality dimensions.

We further analyze validity with respect to reasoning complexity and conflict type. As reported in Table 7, atomic samples achieve a fully valid rate of 96.4% (241/250), whereas compositional samples achieve 92.8% (232/250). The performance gap is consistent with the greater difficulty of resolving multi-step references and relational dependencies in compositional examples. Attribute conflicts achieve the highest overall validity rate of 97.0%, followed by Object conflicts at 96.0%. Spatial conflicts are the most challenging category, with an overall validity rate of 92.0%. In particular, compositional spatial conflicts exhibit the lowest validity rate among all strata at 90.0%, primarily due to ambiguity in visual reference resolution under complex anchor relations.

To better understand the residual generation errors, we manually categorize the 27 samples that fail at least one of the seven validity criteria. Table 8 summarizes the resulting error taxonomy. The most frequent error source is ambiguous or insufficient visual evidence, accounting for 33.3% of invalid samples, followed by unintended secondary conflicts introduced by text mutation (25.9%) and non-unique target or anchor references (18.5%). By contrast, incorrect conflict-type annotations account for only one case (3.7%), suggesting that most remaining failures arise from visual grounding and reference resolution rather than from systematic errors in the symbolic conflict-generation procedure.

Overall, the human verification demonstrates that the automatic generation and symbolic verification pipeline produces highly reliable conflict samples. The remaining errors are primarily concentrated in ambiguous visual evidence and complex reference resolution, especially for compositional spatial conflicts, rather than in systematic errors in conflict-type assignment or structured annotations. We emphasize that this dataset-quality audit is distinct from the human-performance evaluation: the former evaluates the validity and reliability of the generated LogicCon samples, whereas the latter measures human performance on the LogicCon reasoning task. “‘

E Evaluation on MMIR

Adaptation to MMIR. MMIR differs from LogicCon in that its inconsistencies mainly appear in layout-rich multimodal artifacts such as webpages, office documents, and posters, rather than in natural images with object-centric scene structure. Therefore, when transferring our framework to MMIR, we keep the same reasoning principle but adapt the evidence verification stage to the document-style setting. Specifically, target-aware claim parsing decomposes the input question or statement into verifiable claims involving key entities, values, regions, or layout elements. The claim-grounded visual verification module is then extended from object-level visual evidence to more general multimodal evidence, including OCR text regions, titles, tables, charts, icons, logos, spatial layout, and cross-region correspondences. For each claim, the model is prompted to identify the

Conflict Type	Atomic	Compositional	Overall
Attribute	98.0% (49/50)	96.0% (48/50)	97.0%
Object	98.0% (49/50)	94.0% (47/50)	96.0%
Relation	96.0% (48/50)	92.0% (46/50)	94.0%
Spatial	94.0% (47/50)	90.0% (45/50)	92.0%
Quantitative	96.0% (48/50)	92.0% (46/50)	94.0%
Overall	**96.4% (241/250)**	**92.8% (232/250)**	**94.6% (473/500)**

Table 7: Fully valid rates across conflict types and reasoning-complexity levels.

Main Error Source	Count	Percentage
Ambiguous or insufficient visual evidence	9	33.3%
Mutated text introduces an unintended secondary conflict	7	25.9%
Target or anchor reference is not unique	5	18.5%
Original statement is insufficiently supported by the image	3	11.1%
Minor inconsistency between natural-language text and mutation record	2	7.4%
Incorrect conflict-type annotation	1	3.7%
Total	**27**	**100.0%**

Table 8: Error taxonomy for the 27 samples that fail at least one human-audit criterion.

relevant evidence region, compare the claimed content with the observed evidence, and determine whether the evidence supports or contradicts the claim. Finally, the conflict aggregation module combines evidence from multiple regions to produce the final open-ended answer or multiple-choice prediction. This adaptation allows our framework to evaluate whether explicit claim decomposition and evidence verification can generalize beyond natural-image conflicts to layout-rich multimodal inconsistency reasoning, while still requiring only the original image and textual input at test time.

Model	Method	Open-ended				Multiple-choice			
		Web	Office	Poster	Overall	Web	Office	Poster	Overall
Qwen2.5-VL-72B	Base	18.33	40.80	14.78	27.24	33.33	44.39	34.50	38.10
	Ours	**25.60**	**48.70**	**22.90**	**34.00**	**39.80**	**50.60**	**41.90**	**44.10**
Llama-3.2-90B-Vision-Instruct	Base	7.08	23.76	7.04	14.04	20.62	23.31	29.57	22.94
	Ours	**11.50**	**29.80**	**12.20**	**19.10**	**25.80**	**28.60**	**34.70**	**28.50**
Qwen2.5-VL-7B	Base	8.54	29.14	11.97	17.60	14.37	33.18	16.90	22.56
	Ours	**13.90**	**35.80**	**18.70**	**23.40**	**20.50**	**39.20**	**24.10**	**29.40**
LLaVA-NeXT-7B	Base	10.20	21.97	7.04	14.70	11.45	25.33	5.63	16.47
	Ours	**14.00**	**26.50**	**10.80**	**18.60**	**15.20**	**29.90**	**9.80**	**20.80**
InternVL2.5-8B	Base	7.70	24.21	4.92	14.23	9.37	23.54	11.97	15.63
	Ours	**11.60**	**29.20**	**8.90**	**18.40**	**14.20**	**28.10**	**16.40**	**20.20**
Phi-3.5-Vision-4B	Base	6.87	24.43	7.04	14.23	1.66	8.52	0.00	4.30
	Ours	**8.90**	**27.20**	**9.10**	**16.00**	**3.50**	**11.00**	**2.00**	**6.10**

Table 9: Transfer evaluation on MMIR. We compare the base model with our claim-grounded evidence verification framework under open-ended and multiple-choice settings.

Analysis. The results show that our framework consistently improves all evaluated open-source models on MMIR under both open-ended and multiple-choice settings. This suggests that the proposed claim decomposition and evidence verification strategy is not limited to LogicCon-style natural-image conflicts, but can also benefit layout-rich multimodal inconsistency reasoning. Compared with the base models, the gains are generally larger in the multiple-choice setting, where the structured verification results can more directly guide the model to select the most evidence-consistent option. In the open-ended setting, the improvement is also clear, but the final answer quality is more affected by the model's generation ability and its capacity to express the verified evidence accurately.

Across different domains, the improvement is more evident on Web and Poster samples than on some Office samples. This is likely because webpages and posters often contain localized textual or visual elements, such as titles, dates, prices, icons, and event information, which can be naturally decomposed into claim-evidence pairs. Office documents, in contrast, may involve denser text, tables, or more implicit cross-region dependencies, making evidence localization and comparison more challenging. This indicates that claim-grounded verification is helpful, but its effectiveness still depends on the backbone model's ability to read OCR content, understand layout structure, and identify the relevant evidence region.

The results also show different gains across backbone models. Qwen2.5-VL-72B benefits the most from our framework, which is consistent with its stronger OCR, document understanding, and layout perception ability. Qwen2.5-VL-7B and Llama-3.2-90B-Vision-Instruct obtain moderate improvements, showing that the framework can also help models with weaker or different perception capacities. For smaller models such as Phi-3.5-Vision-4B, the improvement is relatively limited, especially in the multiple-choice setting. This suggests that our framework can guide reasoning, but cannot fully compensate for weak low-level perception or poor OCR ability.

Overall, the MMIR transfer results support the generality of our framework. By adapting visual verification into claim-grounded evidence verification, the same reasoning principle can be applied to layout-rich multimodal artifacts. However, the gains also reveal an important limitation: the framework relies on the backbone model's ability to perceive and localize textual, visual, and layout evidence. Therefore, the proposed method is most effective when the base model already has reasonable OCR and document-level visual understanding, and it further improves performance by organizing these perceptual signals into explicit inconsistency reasoning.

F Additional Dataset Statistics

In this appendix, we provide additional statistics of LogicCon to better characterize the dataset scale, conflict distribution, compositional complexity, annotation structure, and construction quality control.

F.1 Dataset Scale and Label Distribution

Table 10 reports the overall split statistics of LogicCon. The benchmark contains 20,000 image-text pairs from 9,793 unique images, including 6,000 consistent pairs and 14,000 conflicting pairs. We include more conflicting samples because LogicCon is designed primarily for diagnostic visual–textual conflict reasoning rather than balanced image-text matching. This design provides sufficient coverage for conflict point analysis, conflict type diagnosis, and compositional reasoning, while still retaining enough consistent samples for conflict detection. Since the label distribution is intentionally imbalanced, we report both overall detection accuracy and conflicting-sample detection accuracy in the main experiments to avoid misleading conclusions from label priors.

Split	Images	Consistent	Conflicting	Total
Train	7,284	4,200	9,800	14,000
Val	1,246	900	2,100	3,000
Test	1,263	900	2,100	3,000
Total	9,793	6,000	14,000	20,000

Table 10: Dataset split statistics of LogicCon. The benchmark contains more conflicting than consistent samples, reflecting its focus on fine-grained visual–textual conflict reasoning rather than balanced binary matching.

F.2 Conflict Type Distribution

Table 11 shows the distribution of the five conflict categories: attribute, object, relation, spatial, and quantitative conflicts. The distribution is naturally imbalanced but controlled. Attribute and object conflicts are slightly more frequent because they are more directly supported by scene-graph annotations and are common in visual question-answering samples. Relation, spatial, and quantitative

conflicts are also well represented, since they are important for evaluating whether models can reason beyond direct object recognition. This distribution enables fine-grained analysis of model behavior across different forms of visual–textual inconsistency instead of relying only on an overall detection score.

Conflict Type	Train	Val	Test	Total
Attribute	2,240	480	470	3,190
Object	2,030	435	445	2,910
Relation	1,925	405	420	2,750
Spatial	1,735	370	365	2,470
Quantitative	1,870	410	400	2,680
Total	9,800	2,100	2,100	14,000

Table 11: Distribution of conflict types in LogicCon. The conflicting samples cover attribute, object, relation, spatial, and quantitative conflicts, with a naturally imbalanced but controlled distribution.

F.3 Atomic and Compositional Conflicts

Table 12 summarizes the proportion of atomic and compositional conflicts. A sample is labeled as compositional if identifying the target fact requires multi-step relations, nested references, or compositional constraints; otherwise, it is labeled as atomic. In total, 4,645 out of 14,000 conflicting samples are compositional, accounting for 33.2% of all conflict samples. This provides a sufficiently large subset for evaluating complex reasoning ability. The compositional subset is important because many visual–textual conflicts cannot be verified by checking a single object or attribute. Instead, models may need to first resolve the target object through relational anchors or contextual descriptions before determining whether the textual claim contradicts the image.

Split	Atomic	Compositional	Total Conflicts	Comp. Ratio
Train	6,548	3,252	9,800	33.2%
Val	1,392	708	2,100	33.7%
Test	1,415	685	2,100	32.6%
Total	9,355	4,645	14,000	33.2%

Table 12: Statistics of atomic and compositional conflicts. A sample is labeled as compositional if identifying the target fact requires multi-step relations, nested references, or compositional constraints.

F.4 Conflict Type and Compositionality

Table 13 further breaks down the conflict types under atomic and compositional settings. Attribute and object conflicts contain a larger proportion of atomic samples, since the target fact can often be checked through direct visual recognition. In contrast, relation and spatial conflicts contain a higher proportion of compositional samples because they often require resolving anchor objects and inter-object dependencies before the conflicting fact can be verified. Quantitative conflicts also include a considerable number of compositional cases, especially when counting or comparison depends on a constrained set of objects. This analysis shows that compositionality is not restricted to one conflict category, but appears across multiple types with different difficulty levels.

F.5 Annotation Statistics

Table 14 reports the fine-grained annotation statistics of LogicCon. Each conflicting sample is paired with a structured mutation record, including the target object, textual conflicting detail, visual evidence, transformation rule, conflict type, and compositional label. These annotations support evaluation beyond sample-level consistency judgment. In particular, target-object, textual-conflict, and visual-evidence annotations allow us to evaluate whether a model truly identifies the conflicting fact pair, rather than simply predicting that a conflict exists. The average statement length is 12.18 words, and the average conflicting-detail length is 2.41 words, indicating that the statements are

Conflict Type	Atomic	Compositional	Total
Attribute	2,386	804	3,190
Object	2,132	778	2,910
Relation	1,683	1,067	2,750
Spatial	1,395	1,075	2,470
Quantitative	1,759	921	2,680
Total	9,355	4,645	14,000

Table 13: Conflict type distribution under atomic and compositional settings. Relation and spatial conflicts contain a higher proportion of compositional samples because they often require resolving anchors and inter-object dependencies.

concise while still containing localized factual changes. On average, each conflict involves 1.84 visual evidence facts, 1.27 anchor objects, and 2.46 objects, reflecting the need for evidence grounding and relational reasoning in many samples.

Statistic	Value
Total image-text pairs	20,000
Unique images	9,793
Consistent pairs	6,000
Conflicting pairs	14,000
Conflict categories	5
Annotation coverage for conflicting samples	100.0%
Mutation record coverage	100.0%
Target-object annotation coverage	100.0%
Textual-conflict annotation coverage	100.0%
Visual-evidence annotation coverage	100.0%
Conflict-type annotation coverage	100.0%
Compositional-label coverage	100.0%
Avg. statement length	12.18 words
Avg. conflicting-detail length	2.41 words
Avg. visual evidence facts per conflict	1.84
Avg. anchor objects per conflict	1.27
Avg. objects involved per conflict	2.46

Table 14: Annotation statistics of LogicCon. We report dataset scale, annotation coverage over conflicting samples, and average textual/visual evidence statistics.

F.6 Verification and Filtering Statistics

Table 15 reports the filtering process used during dataset construction. Starting from 56,420 initial QA candidates, we first retain samples that can be converted into image-consistent declarative statements. We then align the statements with scene-graph facts and functional-program traces to identify the target object, queried semantic slot, and supporting visual evidence. Candidate contradictory statements are generated only after a valid mutation target is identified. Finally, symbolic verification checks whether the original fact is supported by the scene graph and whether the mutated fact forms a valid contradiction under type-specific rules. The final benchmark retains 14,000 verified conflicting samples, corresponding to 24.8% of the initial candidates. This selective filtering process helps ensure that retained samples correspond to genuine factual conflicts rather than surface-level linguistic perturbations or ambiguous image-text pairs.

Overall, these statistics show that LogicCon is designed for both broad model comparison and fine-grained diagnostic analysis. The dataset contains enough consistent samples for conflict detection, while assigning more capacity to conflicting samples to support conflict point analysis, conflict type diagnosis, and compositional reasoning. The five conflict categories enable type-specific evaluation, and the compositional subset provides an additional dimension for analyzing multi-step reasoning difficulty. More importantly, the structured mutation records and symbolic verification process allow

Stage	Samples	Retained Ratio
Initial QA candidates	56,420	100.0%
QA-to-statement conversion passed	49,386	87.5%
Scene-graph alignment passed	38,712	68.6%
Valid mutation target identified	29,458	52.2%
Contradictory generation succeeded	21,736	38.5%
Symbolic verification passed	16,284	28.9%
Final balanced conflict subset	14,000	24.8%

Table 15: Filtering statistics of the LogicCon construction pipeline. We retain only samples whose original fact is supported by the scene graph and whose mutated fact forms a valid contradiction under type-specific symbolic verification.

us to evaluate conflict point recovery, visual evidence grounding, and conflict type diagnosis in a transparent and reproducible manner.

G Evaluation Metrics

We provide formal definitions of the evaluation metrics used in LogicCon. For each sample, let the standard mutation record be $M = (o^*, f^*_{\text{text}}, f^*_{\text{vis}}, c^*)$, where o^* denotes the target object, f^*_{text} denotes the textual conflicting detail corresponding to the mutated fact, f^*_{vis} denotes the original visual fact supported by the image, and c^* denotes the ground-truth conflict type. Given a model prediction, we extract the predicted target object $\hat{o}$, textual conflicting detail $\hat{f}_{\text{text}}$, visual evidence $\hat{f}_{\text{vis}}$, and conflict type $\hat{c}$.

Conflict Detection. Conflict detection evaluates whether the model correctly predicts the sample-level consistency label: $y \in \{\text{consistent}, \text{conflicting}\}$. We report two detection metrics. **All** denotes detection accuracy over all samples, including both consistent and conflicting samples. **Conf** denotes detection accuracy restricted to conflicting samples only. The latter is used to measure whether a model can recognize actual conflicts without being affected by the ratio of consistent and conflicting samples.

Target Object Accuracy (TO). TO measures whether the model correctly identifies the target object involved in the conflict. A prediction is counted as correct if the predicted target object $\hat{o}$ refers to the same object as o^* in the mutation record. We allow minor surface-form variation, such as "man" vs. "person", when the reference is semantically equivalent and points to the same visual entity.

Textual Conflicting Detail Accuracy (TC). TC measures whether the model correctly identifies the conflicting detail in the textual statement. A prediction is counted as correct if $\hat{f}_{\text{text}}$ expresses the same mutated fact as f^*_{text}. For example, if the text incorrectly claims that an object is "red" while the mutation record marks the mutated attribute as "red", the TC prediction is correct even if the model phrases it as "the text says the object has a red color." Thus, TC is evaluated based on semantic equivalence rather than exact string matching.

Visual Evidence Accuracy (VE). VE measures whether the model correctly identifies the visual evidence that contradicts the textual claim. A prediction is counted as correct if $\hat{f}_{\text{vis}}$ expresses the original visual fact f^*_{vis} supported by the image. For example, if the image shows that the target object is blue while the text claims it is red, the VE prediction is correct only when the model recovers the visual fact that the object is blue. This metric evaluates whether the model grounds the conflict in the correct visual evidence rather than merely detecting that a conflict exists.

Conflict Point Accuracy (CP). CP measures whether the model fully localizes the conflicting fact pair. A prediction is counted as correct only when TO, TC, and VE are all correct: $\text{CP} = [\text{TO} = 1 \wedge \text{TC} = 1 \wedge \text{VE} = 1]$. Therefore, CP is a stricter metric than each individual component, since it requires the model to identify the target object, the textual conflicting detail, and the corresponding visual evidence simultaneously.

Conflict Type Diagnosis. Conflict type diagnosis evaluates whether the predicted conflict type matches the type specified in the mutation record: $\hat{c} = c^*$. The conflict type belongs to one of

five categories: $c^* \in \{\text{attribute}, \text{object}, \text{relation}, \text{spatial}, \text{quantity}\}$. We report both the overall type diagnosis accuracy and type-specific accuracies for attribute, object, relation, spatial, and quantity conflicts.

Compositional Evaluation. In addition to evaluating all conflicting samples, we separately report performance on the compositional subset. A sample is labeled as compositional if locating the target fact requires multi-step relations, nested references, or contextual constraints. On this subset, we report **Comp-Det**, **Comp-CP**, and **Comp-Type**, corresponding to conflict detection accuracy, conflict point accuracy, and conflict type diagnosis accuracy on compositional samples, respectively.

Open-ended Answer Judging. For open-ended model responses, we evaluate TO, TC, VE, and conflict type based on semantic equivalence to the mutation record. Deterministic labels and multiple-choice outputs are evaluated by exact matching, while natural-language conflict point analyses are judged according to whether they express the same target object, textual conflicting detail, and visual evidence as the standard record. This allows models to use different surface expressions while still being evaluated against the same underlying conflict facts.

H Implementation Details of the Inference Framework

Our plug-and-play reasoning framework is implemented as a prompt-based inference procedure. At test time, the framework only takes an image I and a textual statement T as input. Scene graphs, functional programs, and mutation records are used only during dataset construction and evaluation, and are not provided to the model during inference. Unless otherwise specified, all modules are executed by the same backbone MLLM without introducing additional trainable models, external object detectors, or extra structured annotations.

Module execution. The framework consists of four sequential modules. First, target-aware claim parsing decomposes the textual statement into atomic factual claims. Each claim is required to contain a target object, a predicate, a claimed detail, and optional contextual constraints. Second, claim-grounded visual verification verifies each parsed claim against the image and produces a textual evidence description together with a verification label. The verification label is selected from supported, contradicted, and insufficient. Third, conflict aggregation and type diagnosis aggregates claim-level verification results into a sample-level conflict judgment, conflict point, and conflict type. Finally, conflict-guided final reasoning generates the final answer based on the structured intermediate results.

Evidence format. We use textual evidence descriptions as the default evidence format. Specifically, the model is asked to describe the relevant visual evidence that supports or contradicts each claim, such as the object being checked, its attribute, its relation to another object, its spatial position, or the number of relevant instances. We do not require bounding boxes or cropped regions, because many closed-source MLLMs do not expose reliable grounding coordinates, and requiring box-level outputs would make the framework less generally applicable. This design also keeps the same inference interface for both open-source and closed-source models.

Iterative evidence supplementation. For simple samples, one round of claim verification is usually sufficient. However, compositional samples may require resolving anchors, nested references, or multi-step relations before the target fact can be verified. Therefore, when the initial verification result is insufficient, we perform iterative evidence supplementation. In each iteration, the model is prompted to refine the target object, identify missing contextual anchors, decompose the claim into finer sub-claims when necessary, and verify the claim again using more focused evidence.

We set the maximum number of verification iterations to $K = 3$ in all experiments. The iteration stops early if the model outputs either supported or contradicted with explicit visual evidence. If the model still cannot obtain sufficient evidence after K iterations, the final verification label remains insufficient. We choose $K = 3$ as a practical trade-off between reasoning depth and inference cost: it is sufficient for most compositional cases involving one or two contextual anchors, while avoiding excessive prompt calls and unstable long reasoning chains.

Conflict aggregation. After claim-level verification, conflict aggregation is performed over the structured verification results. If at least one target-relevant claim is verified as contradicted, the sample is predicted as conflicting. If all verifiable claims are supported, the sample is predicted as

consistent. If the results contain only insufficient evidence, the final reasoning module produces a conservative judgment based on the available evidence. For conflicting samples, the conflict point is constructed from the contradicted claim by reporting the target object, the textual conflicting detail, and the corresponding visual evidence. The conflict type is diagnosed according to the factual slot that causes the contradiction, including attribute, object, relation, spatial, and quantitative conflicts.

Prompt budget and fairness. For fair comparison, all plug-and-play methods are evaluated under the same image-text input setting. Our method uses the same maximum iteration number across all backbone models. For baseline methods that require multi-step reasoning, we keep a comparable prompt budget whenever possible. This ensures that the observed improvements are mainly due to the target-aware claim parsing, visual evidence verification, and conflict aggregation design, rather than simply using substantially more inference steps.

I Example of Framework Reasoning Process

To better illustrate how our plug-and-play reasoning framework works during inference, we provide a qualitative example in this section. Given an image and a textual statement, the model is required to detect whether a visual–textual conflict exists, identify the conflicting fact pair, and diagnose the conflict type. Different from direct prompting, our framework explicitly decomposes the statement into target-aware claims, verifies each claim against visual evidence, aggregates local verification results, and then produces the final answer.



Figure 3: A woman is sitting on a sofa and holding a remote

Input Example

Image: A woman is sitting on a sofa and holding a remote control in her hand.
Textual statement: "The woman is holding a cell phone."

Step 1: Target-aware Claim Parsing

The textual statement is decomposed into the following verifiable factual unit:

$$c_1 = (o_1, p_1, a_1, m_1)$$

Parsed claim:

- **Target object** o_1: the woman
- **Predicate** p_1: holding
- **Claimed object/detail** a_1: cell phone
- **Contextual constraint** m_1: object held in the woman's hand

The claim to be verified is therefore: the woman is holding a cell phone.

Step 2: Claim-grounded Visual Verification

The model checks the image region around the target object and the object held in her hand.
Candidate visual evidence:

- The woman is holding a small rectangular object.
- The object appears to be a remote control.
- The object does not visually match a cell phone.

Verification result:

$$r_1 = (o_1, c_1, e_1, v_1)$$

- **Target object**: the woman
- **Textual claim**: holding a cell phone
- **Visual evidence**: the woman is holding a remote control
- **Verification label**: contradicted

Step 3: Conflict Aggregation and Type Diagnosis

The local verification result indicates that the textual claim contradicts the visual evidence.
Conflict aggregation:

- The target object is correctly localized as the woman.
- The conflicting textual detail is "cell phone."
- The corresponding visual evidence is "remote control."
- The image supports that the woman is holding a remote control, not a cell phone.

Conflict type diagnosis: Object conflict.
This is an object conflict because the category of the held object in the text is inconsistent with the object category supported by the image.

Step 4: Conflict-guided Final Reasoning

Final prediction: conflicting.
Conflict point analysis: The text claims that the woman is holding a cell phone, but the image shows that the woman is holding a remote control.
Conflict type: object conflict.

J Challenging Qualitative Examples

We provide challenging qualitative examples to better illustrate the diagnostic value of LogicCon and the behavior of our reasoning framework. These cases are not simple object or color mismatches. Instead, they require resolving contextual anchors, distinguishing visually similar objects, verifying relations, or aggregating multiple pieces of visual evidence. In each example, a direct MLLM response may detect the conflict incorrectly or fail to recover the full conflict point, while our framework improves the result through target-aware claim parsing, claim-grounded visual verification, and conflict aggregation.

J.1 Attribute Conflict with Contextual Anchor

Textual statement. "The man holding the tennis racket is wearing a red shirt."

Visual evidence. The image contains multiple people. The man holding the tennis racket is wearing a blue shirt, while another nearby person is wearing red.



Figure 4: Challenging qualitative examples for visual–textual logical conflict reasoning.

Why this case is challenging. The conflict cannot be resolved by simply checking whether a red shirt exists in the image. The model must first identify the correct target object through the anchor relation "holding the tennis racket," and then verify the shirt color of that specific person.

Typical baseline failure. A direct MLLM may answer that the statement is consistent because it detects a red shirt somewhere in the image. In this case, the model recognizes the attribute but attaches it to the wrong person.

Our framework. The claim parser extracts the target as "the man holding the tennis racket" and the claimed detail as "red shirt." Visual verification then focuses on the anchored person and finds that his shirt is blue. The final output correctly predicts a conflict and localizes the conflict point as: textual claim = red shirt; visual evidence = blue shirt.

J.2 Object Conflict with Similar Handheld Objects

Textual statement. "The woman sitting on the sofa is holding a cell phone."

Visual evidence. The woman is holding a remote control. A cell phone-like object appears on the table nearby, but it is not in her hand.

Why this case is challenging. The incorrect object is visually similar to the true object, and another plausible distractor object appears in the scene. The model must verify not only object existence, but also the relation between the woman and the held object.

Typical baseline failure. A direct MLLM may predict the statement as consistent because it detects a phone-like object in the image or because a remote control and a phone share a similar rectangular shape.

Our framework. The framework decomposes the claim into target object = woman, predicate = holding, claimed object = cell phone. Visual verification checks the object in the woman's hand and identifies it as a remote control. The framework therefore predicts an object conflict, rather than merely confirming that a phone-like object exists somewhere in the image.

J.3 Relation Conflict with Entity Co-occurrence

Textual statement. "The man in the black jacket is riding the bicycle."

Visual evidence. The man in the black jacket is standing beside the bicycle and holding its handlebar, but he is not riding it.

Why this case is challenging. Both entities mentioned in the text are present, and the man is physically close to the bicycle. The conflict lies in the interaction predicate, not in object existence.

Typical baseline failure. A direct MLLM may mark the statement as consistent because it detects both the man and the bicycle and observes that they are spatially close. This reflects an entity-

co-occurrence bias: the model treats the presence of both entities as sufficient evidence for the relation.

Our framework. The framework separately verifies the predicate "riding." The visual evidence indicates "standing beside / holding the bicycle" rather than "riding." The aggregation module therefore identifies a relation conflict and reports the conflict point as: textual claim = riding the bicycle; visual evidence = standing beside the bicycle.

J.4 Spatial Conflict with Nested Reference

Textual statement. "The cup to the left of the plate is on the right side of the fork."

Visual evidence. The image contains two cups. The cup to the left of the plate is actually on the left side of the fork, while another cup on the right side of the fork is not the one referred to by the statement.

Why this case is challenging. The target object is defined by a nested reference: the model must first locate "the cup to the left of the plate," and only then verify its spatial relation to the fork. The existence of another cup creates a distractor.

Typical baseline failure. A direct MLLM may detect a cup on the right side of the fork and incorrectly conclude that the statement is consistent. This indicates failure in resolving the target object before checking the spatial relation.

Our framework. The claim parser separates the anchor constraint "cup to the left of the plate" from the spatial claim "on the right side of the fork." Visual verification first resolves the anchored cup and then compares its position relative to the fork. The framework correctly identifies a spatial conflict.

J.5 Quantity Conflict with Constrained Counting

Textual statement. "There are three people wearing hats near the bus."

Visual evidence. There are three people near the bus, but only two of them are wearing hats. A third person wearing a hat appears farther away from the bus.

Why this case is challenging. The model must count only people satisfying both conditions: wearing hats and being near the bus. Counting all people near the bus or all people wearing hats leads to the wrong answer.

Typical baseline failure. A direct MLLM may answer that the statement is consistent because it separately detects three people near the bus and also detects hats in the image. It fails to bind the counting condition to the correct subset.

Our framework. The framework decomposes the claim into a constrained quantity fact: target category = people, condition = wearing hats, spatial constraint = near the bus, claimed count = three. Visual verification enumerates the relevant instances and finds that only two people satisfy both conditions. The final output correctly identifies a quantity conflict.

J.6 Compositional Relation Conflict

Textual statement. "The child standing next to the woman in a yellow dress is holding a balloon."

Visual evidence. The child standing next to the woman in a yellow dress is holding a toy. Another child elsewhere in the image is holding a balloon.

Why this case is challenging. The target child is not directly specified by a simple category label. The model must first locate the woman in a yellow dress, then identify the child standing next to her, and finally verify what the child is holding.

Typical baseline failure. A direct MLLM may predict the statement as consistent because it detects a child holding a balloon somewhere in the image. The model fails to bind the balloon-holding relation to the correct child.

Our framework. The iterative evidence supplementation mechanism first resolves the anchor "woman in a yellow dress," then locates the child next to her, and finally verifies the held object. The framework finds that the target child is holding a toy rather than a balloon, leading to a correct relation conflict diagnosis.

K CONFLICT VALIDATION AND FILTERING

For example, for attribute conflicts, we require the original attribute value and the transformed attribute value to belong to the same semantic dimension and to be mutually exclusive. For instance, red and blue are both color attributes, but they cannot hold at the same time. For spatial conflicts, we require the transformed spatial relation to be the inverse of the original relation, such as left of and right of. For relation conflicts, we require the textual fact to explicitly negate or replace the target relation that exists in the scene graph. For counting conflicts, we check whether the transformed number or comparison result is inconsistent with the statistical result derived from the scene graph. For object conflicts, we require that the transformed object category cannot be supported by the scene graph under the same target and relational context.

L Prompt for Image-Text Conflict Detection

We use the following prompt to ask the model to determine whether a given textual statement conflicts with the image content. The model is required to first make a binary judgment, then explain the conflict point if a conflict exists, and finally assign the conflict to one of five predefined categories.

Image-Text Conflict Detection Prompt

You will be given an image and a textual statement. Please determine whether the textual statement conflicts with the image content.
Please note:
1. You do not need to decide whether the image is wrong or the text is wrong. You only need to determine whether there is an inconsistency between the visual facts and the textual claim.
2. If there is no conflict, answer "No conflict."
3. If there is a conflict, answer "Conflict" and explain the conflict point.
4. When explaining the conflict point, please include the following information as much as possible:

- the target object involved in the conflict;
- what the text claims;
- what the corresponding visual fact in the image is.

5. If there is a conflict, please select one conflict category from the following five types:

- **Attribute conflict:** The color, size, material, shape, state, or other attribute is inconsistent with the image.
- **Object conflict:** The category or identity of the target object is inconsistent with the image.
- **Relation conflict:** A non-spatial interaction relation is inconsistent with the image, such as holding, wearing, riding, eating, or looking at.
- **Spatial conflict:** A spatial relationship is inconsistent with the image, such as left/right, above/below, on/under, inside/outside, or in front of/behind.
- **Counting conflict:** The number, existence, or comparative relationship is inconsistent with the image.

6. If the text contains multi-step reference or a complex description, please first locate the final target object and then determine whether the description of this target object is consistent with the image.
7. The conflict category should be determined according to the actual contradictory visual fact, rather than whether the sentence itself is complex.
Textual statement:
{TEXT}
Please answer in the following format:
Judgment: Conflict / No conflict
Conflict point explanation: If there is a conflict, explain where the textual claim is inconsistent with the image. If there is no conflict, explain why the text is consistent with the image.
Conflict category: If there is a conflict, choose one from the following five categories: Attribute conflict / Object conflict / Relation conflict / Spatial conflict / Counting conflict. If there is no conflict, write "None."

M Judge Prompt for Conflict Analysis Evaluation

To evaluate whether the model correctly identifies the image-text conflict, we use an LLM-based judge. The judge is provided with the standard conflict record and the response generated by the evaluated model. It is asked to determine whether the model output identifies the same or semantically equivalent conflict point as the standard record.

Judge Prompt

You are an evaluation judge. Your task is to determine whether the evaluated model's analysis of the image-text conflict is correct.
You will be given a standard conflict record and the evaluated model's response. Based on the standard conflict record, judge whether the model identifies the same or semantically equivalent conflict point.
Standard conflict record:
Target object: {TARGET_OBJECT}
Text claim: {TEXT_CLAIM}
Visual evidence / image fact: {VISUAL_EVIDENCE}
Evaluated model response: {MODEL_OUTPUT}
Scoring rules:
1. **Target object:** If the model correctly points out the same or an equivalent target object as the standard record, the target is correct. If the model attributes the conflict to an incorrect object, an anchor object, or an irrelevant object, the target is incorrect.
2. **Text claim:** If the model correctly identifies the textual claim that is inconsistent with the image, the text claim is correct. Different wording is allowed as long as the meaning is semantically equivalent. If the model only vaguely states that "the text is wrong" without specifying the concrete textual claim, the text claim is incorrect.
3. **Visual evidence / image fact:** If the model correctly identifies the visual fact or evidence in the image that conflicts with the textual claim, the visual evidence is correct. Different wording is allowed as long as the meaning is semantically equivalent. If the model does not mention the corresponding visual fact, or describes an incorrect visual fact, the visual evidence is incorrect.
4. **Description direction:** Do not penalize the model for describing the conflict in a different direction. For example, "the text says the car is blue, but the image shows that the car is red" and "the image shows that the car is red, while the text says it is blue" should be considered equivalent.
5. **Overall judgment:** If the model correctly identifies the target object, the text claim, and the visual evidence, then the overall conflict analysis is correct. If any of these core components is incorrect, then the overall conflict analysis is incorrect. If the model response is too vague to determine whether it identifies the specific conflict point, then the overall conflict analysis is incorrect.
Please output in the following format:
Target correct: Yes / No
Text claim correct: Yes / No
Visual evidence correct: Yes / No
Overall conflict analysis correct: Yes / No
Brief reason: