MoDEST: Multimodal Image Description Evaluation with Semantic Scene Graph

Anonymous ACL submission

Abstract

Image description evaluation aims to automatically assess the semantic consistency between an image and its generated description. Existing metrics suffer from two major limitations: they rely on coarse-grained feature similarity without deep semantic understanding and fine-grained cross-modal alignment, and they act as black-box scoring systems lacking interpretable evidence. To address these issues, we propose MoDEST (Multimodal Image Description Evaluation with Semantic Scene Graph), a novel interpretable framework that integrates semantic scene graphs with multimodal large language models. By structuring image and text semantics into explicit object–attribute–relation triplets, MoDEST enables fine-grained, element-level graph comparison between reference and candidate content, while multimodal large language models jointly reason over visual evidence and reference semantics to produce consistency scores, factual reasoning, inconsistency analysis, and automatic correction suggestions. Extensive experiments on multiple benchmarks demonstrate that MoDEST significantly outperforms existing metrics and achieves substantially higher correlation with human judgments.

1 Introduction

Vision-language models such as Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023), and InstructBLIP (Dai et al., 2023) have driven rapid advances in image captioning. The task has diversified into paragraph captioning (Krause et al., 2017), dense captioning (Johnson et al., 2016), and controllable captioning, placing higher demands on evaluation metrics. Because captioning produces free-form natural language that must align with visual content (Figure 1), reliable automatic metrics are essential for measuring progress.

Existing metrics suffer from two fundamental limitations. Early methods such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) rely on lexical overlap, becoming unreliable as captions grow longer and more diverse (Vedantam et al., 2015). Image-text matching methods such as UMIC (Lee et al., 2021) and CLIPScore (Hessel et al., 2021) compute global feature similarity, yet they collapse images and sentences into single vectors, making them blind to fine-grained errors. For instance, "A man feeds a dog" and "A dog feeds a man" are nearly identical in embedding space despite opposite relations.

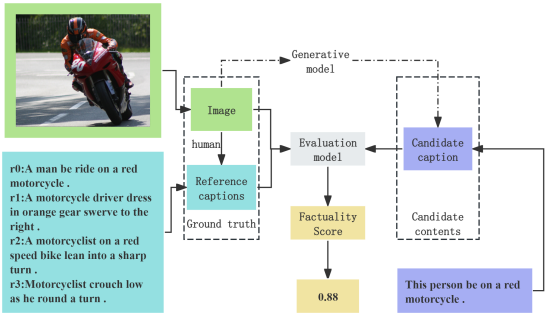

Figure 1: The image description evaluation task evaluates the consistency of candidate descriptions based on an image and a set of reference descriptions (collectively referred to as the ground truth).

Recent efforts have attempted to bridge this gap, but from opposing directions. Fine-grained methods such as InfoMetIC (Hu et al., 2023) and HiFi-Score (Yao et al., 2024) construct hierarchical parsing graphs or perform region-word alignment, yet they remain at the feature level, measuring similarity between distributed representations rather than explicitly verifying semantic facts. LLM-based evaluators such as CLAIR (Chan et al., 2023) and FLEUR (Lee et al., 2024) leverage strong semantic understanding and can generate explanations, but they operate as *black-box* systems: they produce holistic scores without structured verification, making it impossible to pinpoint which object is hallucinated or which relation is reversed.

More importantly, existing methods fail to jointly exploit two complementary sources of guidance: the visual facts present in the image, and the implicit salience rankings embedded in human-written references. When people describe an image, they do not list every visible detail; they selectively emphasize what they consider important. Reference captions therefore encode which objects, attributes, and relations humans prefer to mention. Image-only metrics can verify visual facts but lack access to these human priorities; text-only metrics can see these priorities but cannot verify them against the image. The fundamental limitation is thus not merely a lack of accuracy, but the absence of a structured, verifiable, and interpretable mechanism that can simultaneously ground descriptions in visual evidence and align them with human salience preferences.

As shown in Figure 2, vector-similarity-based methods may assign nearly identical scores to captions with fundamentally different relational correctness.

We propose MoDEST, an interpretable evaluation framework that integrates semantic scene graphs with multimodal large language models (MLLMs). Its central innovation is unifying image and text into explicit object–attribute–relation triplets, enabling fine-grained alignment at the semantic rather than embedding level. The framework operates through three stages: (1) **structured extraction**, parsing reference captions into scene graphs; (2) **multimodal fusion**, where MLLMs verify and supplement these graphs against visual evidence to build a joint "ground-truth" representation that preserves both image-supported facts and semantically preferred information reflected by references; and (3) **explicit comparison**, aligning candidate caption graphs against this joint graph for element-level consistency assessment. MLLMs serve as explainable evaluators that output scores alongside factual rationales, inconsistency analyses, and automatic corrections.

Our contributions are:

(1) A **unified structured representation** bridging images and text via scene graphs, moving beyond shallow embedding similarity;

(2) A **fine-grained graph comparison paradigm** enabling precise object–attribute–relation matching that existing metrics cannot perform;

(3) An **interpretable evaluation framework** powered by MLLMs that jointly reasons over visual evidence and reference semantics, producing scores, factual evidence, and corrective feedback.

Candidate captions	CLIPScore
A: A person feeds a strawberry to a turtle.	0.873
B: A turtle feeds a strawberry to a person.	0.861

Figure 2: Using CLIPScore as an illustrative example: although candidate A is factually accurate while B reverses the subject–object relationship, their proximity in vector space yields nearly identical scores.

2 Related Work

Text-only metrics. These metrics evaluate candidate captions using only textual information. BLEU-4 (Papineni et al., 2002) measures 4-gram precision; ROUGE-L (Lin, 2004) calculates recall based on the longest common subsequence; METEOR (Banerjee and Lavie, 2005) extends precision and recall with synonym matching and WordNet; CIDEr (Vedantam et al., 2015) computes tf-idf weighted n-gram cosine similarity emphasizing consensus across multiple references. SPICE (Anderson et al., 2016) compares semantic propositions through scene-graph matching. All these methods rely only on synonym-level matching and cannot verify deeper semantics against visual evidence. Recently, CLAIR (Chan et al., 2023) prompts an LLM with reference and candidate pairs to produce a score and justification, but as it never sees the image, it cannot verify details absent from the references.

Image-only metrics. These metrics assess candidate descriptions by directly comparing the source image with the generated text, ignoring reference captions. UMIC fine-tunes the UNITER (Chen et al., 2020) model with contrastive learning, enabling reference-free evaluation. CLIP-S leverages the vision-text alignment of CLIP (Radford et al., 2021) and evaluates quality by computing cosine similarity between the image and the generated description. InfoMetIC introduces a fine-

grained token-level scoring mechanism that can identify erroneous words and unmentioned image regions. HiFi-Score constructs a hierarchical semantic graph through multimodal fusion, yet its comparisons remain similarity-driven rather than explicit semantic verification. FLEUR (Lee et al., 2024) feeds the image into an MLLM during inference, but treats the whole image and text holistically, struggling to ground judgments on fine-grained semantic elements.

Text+Image metrics. These metrics exploit both reference captions and visual information. REO (Jiang et al., 2019a) employs learning-to-rank to assess description quality. TIGEr (Jiang et al., 2019b) uses a text-image grounding model to measure the match between description and image content as well as the similarity between machine-generated and human-generated descriptions. ViLBERTScore (Lee et al., 2020) extracts vision-aware text embeddings with ViLBERT (Lu et al., 2019) and considers image context when computing cosine similarity. FAIEr (Wang et al., 2021) bridges image and text modalities with scene graphs and fuses scene-graph embeddings via an attention module; however, its scene graphs serve mainly as feature carriers rather than explicit semantic reasoning structures. In short, no existing method simultaneously combines explicit semantic verification, fine-grained visual-semantic parsing, and human-preferred reference guidance while producing interpretable element-level explanations. We present MoDEST to close this gap.

3 Method

3.1 Problem Definition

Evaluating the consistency between images and text requires a structured form that can uniformly represent the semantic information of both modalities. We adopt scene graphs as an intermediate representation. Formally, a scene graph is defined as a triplet set $\mathcal{G} = (\mathcal{E}, \mathcal{A}, \mathcal{R})$, where $\mathcal{E}$ denotes the set of entities (objects), $\mathcal{A}$ the set of attributes, and $\mathcal{R}$ the set of relations.

Given an image $\mathcal{I} = \{E_k\}_{1 \le k \le n}$ and a corresponding descriptive text $\mathcal{T} = \{E_j\}_{1 \le j \le m}$, the entity mapping from image to text is formulated as:

$$f\big(\{E_k\}_{1 \le k \le n}\big) \to \{E_j\}_{1 \le j \le m}. \qquad (1)$$

To construct structured scene graphs, information extractors transform the semantics of each modality into unified structured representations:

$$\mathrm{LLM}(\mathcal{I}) \to \mathcal{G}_I, \quad \mathrm{LLM}(\mathcal{T}) \to \mathcal{G}_T, \qquad (2)$$

where $\mathrm{LLM}(\cdot)$ converts raw modal data into structured scene graphs.

$\mathcal{G}_I$ and $\mathcal{G}_T$ are represented in subject-centric triplet form (Figure 3). In $\mathcal{G}_I$, E_n is the n-th entity, A_{ni} its i-th attribute, and R_{nj} the relation between the n-th and j-th entities. An attribute triplet takes the form $\{\text{entity}, \text{attribute key}, \text{attribute value}\}$ (e.g., $\{\text{dog}, \text{color}, \text{yellow}\}$); a relation triplet takes the form $\{\text{entity}_A, \text{relation}, \text{entity}_B\}$ (e.g., $\{\text{dog}, \text{near}, \text{cat}\}$). Relations between entities may be mutual.

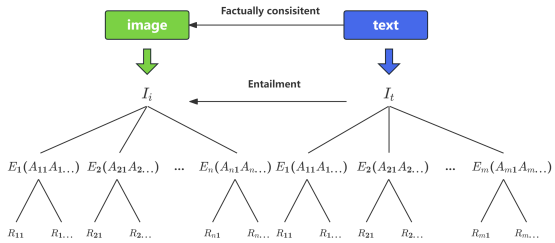

Figure 3: Structured representation of scene graphs (local tree expansion centred on a node). $\mathcal{G}_I$ and $\mathcal{G}_T$ denote the scene graphs of image and text, respectively; E_n represents an entity; A_{ni} denotes the i-th attribute associated with entity E_n, and R_{nj} denotes a directed relation from E_n to E_j.

3.2 MoDEST

To implement the above idea, we propose MoDEST, a structured evaluation framework for image caption assessment (Figure 4). It consists of three stages: information extraction, semantic fusion, and comparative evaluation.

3.2.1 Information Extraction

Information extraction converts images and texts into structured semantic facts. Textual information is represented as subject-centric facts, where each entity is associated with attributes and directed relations, forming explicit semantic triplets. A large language model (LLM) extracts these facts and outputs structured JSON representations (Appendix A.1). Candidate captions are transformed into a **candidate scene graph** $\mathcal{G}_C$. For multiple reference captions $\{r_1, r_2, \ldots\}$, extracted semantic facts are merged to construct a comprehensive **reference scene graph** $\mathcal{G}_R$.

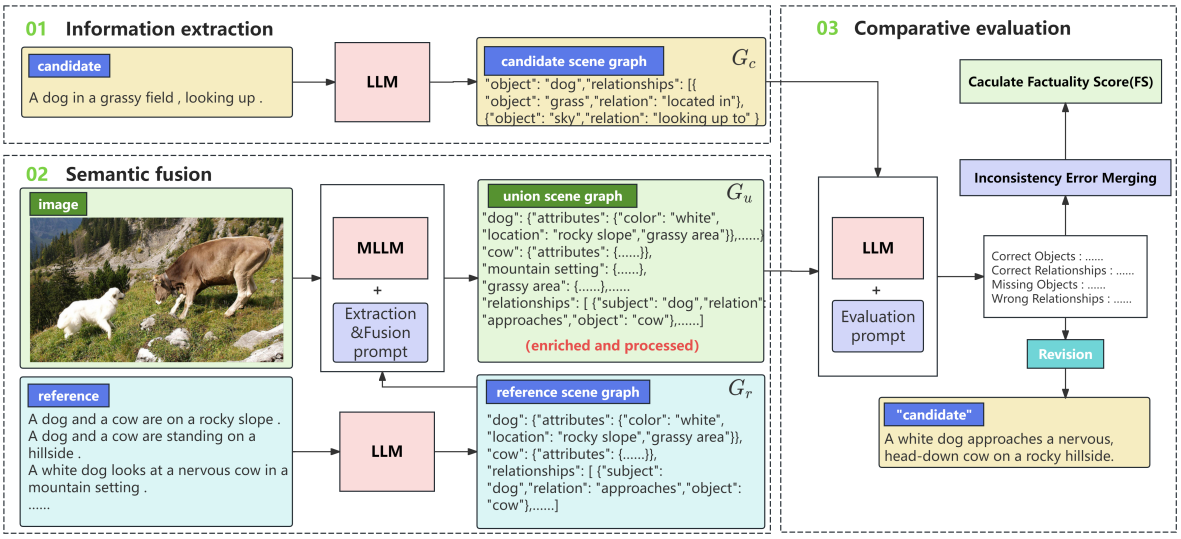

Figure 4: MoDEST consists of three stages. **(1) Information extraction.** A text-only LLM extracts facts from both the reference and the candidate caption, yielding a reference scene graph and a candidate scene graph. **(2) Semantic fusion.** An MLLM parses the original image and fuses its content with the reference graph to produce a union scene graph $\mathcal{G}_U$ (reference semantic representation). **(3) Comparative evaluation.** A comparative LLM matches the union scene graph against the candidate scene graph, returns correction hints and outputs the final assessment score.

3.2.2 Semantic Fusion

Although reference captions provide important semantic cues, they rarely cover all observable image content due to differences in descriptive habits and selective attention. Relying solely on $\mathcal{G}_R$ may incorrectly penalize candidate captions containing valid but unmentioned information.

We introduce a multimodal large language model (MLLM) to extract visual semantic facts from the original image and fuse them with $\mathcal{G}_R$, yielding a **union scene graph** $\mathcal{G}_U$. The fusion process optimizes semantic consistency through three mechanisms:

- **Retention with enhancement:** Semantic facts supported by both references and images are preserved and assigned higher importance weights.
- **Visual supplementation:** Additional visually grounded entities, attributes, and relations absent from references are incorporated into $\mathcal{G}_U$.
- **Inconsistency filtering:** Reference facts lacking identifiable visual counterparts are suppressed to improve semantic fidelity.

The resulting union scene graph maintains the same structure as $\mathcal{G}_R$ while providing a more complete and visually verified semantic representation. Consequently, $\mathcal{G}_U$ adaptively balances *fidelity* and *adequacy*, serving as the semantic basis for subsequent evaluation.

3.2.3 Comparative Evaluation

Given the union scene graph $\mathcal{G}_U$ and candidate scene graph $\mathcal{G}_C$, factual consistency is evaluated by comparing entities, attributes, and relations at the semantic level (Figure 5).

An LLM acts as a semantic comparator, determining whether each candidate element is supported by $\mathcal{G}_U$. Element-wise matching produces consistency judgments and identifies semantically correct content, inconsistencies, and hallucinated information.

To distinguish semantic information of different importance, we introduce a hierarchical weighting mechanism based on three source tags assigned during semantic fusion:

- **Tag 1:** Elements jointly supported by both reference captions and visual evidence;
- **Tag 2:** Elements visually grounded but absent from the references (visual supplementation);
- **Tag 0:** Elements in the candidate that find no support in $\mathcal{G}_U$ (hallucinations or factual errors).

The final factuality score is computed as:

$$\mathrm{FS} = \frac{\sum_{x \in \mathcal{G}_C} w(x) \cdot 1_{x \in \mathrm{match}(x)}}{\sum_{x \in \mathcal{G}_C} w(x)} \times 100\%, \quad (3)$$

where $w(x)$ denotes the importance weight assigned to semantic element x according to its source tag, and $1_{x \in \text{match}(x)}$ indicates whether the element is semantically consistent with $\mathcal{G}_U$.

This mechanism penalizes hallucinated or incorrect information while emphasizing key semantics, enabling fine-grained and balanced evaluation.

Interpretable Evaluation. Unlike conventional metrics that only output similarity scores, MoDEST provides structured semantic alignment results, including correctly matched information as well as inconsistencies and hallucinations. These alignment results can be further transformed into natural-language explanations by the LLM, making scoring decisions interpretable and traceable. Moreover, inconsistent semantic elements can be used to generate correction suggestions for candidate captions based on $\mathcal{G}_U$, forming an integrated framework of *evaluation–diagnosis–correction*.

4 Experiments

4.1 Evaluation Datasets

Flickr8k-Expert (Hodosh et al., 2013) contains 5,644 image–machine-generated-caption pairs scored by three expert annotators on a scale of 1–4.

Flickr8k-CF (Hodosh et al., 2013) comprises 47,830 image–caption pairs judged "yes" or "no" by at least three annotators; the final score is the proportion of "yes" votes.

Composite (Aditya et al., 2018) contains 3,995 images from MSCOCO (Lin et al., 2014), Flickr30K (Young et al., 2014) and Flickr8K. For each image there are two machine-generated captions and one human-written caption; every pair is rated from 1 (irrelevant) to 5 (fully relevant).

Pascal-50S (Vedantam et al., 2015) contains 4,000 triplets, each consisting of one image and two captions. Annotators decide which caption is better. The dataset is split into four subsets: "HC" (two human-written correct captions), "HI" (two human captions, one incorrect), "HM" (one human and one machine caption), and "MM" (two machine-generated captions).

4.2 Implementation Details

Information extraction. We design a structured-data template as the expected output of extraction. Experiments use Qwen3-32B to extract semantic information from text, deployed locally via an API endpoint. Extracted entities, attributes and relations

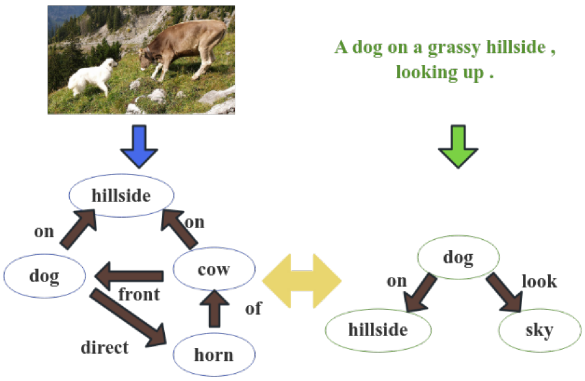

Figure 5: Since the grounded scene graph G_U and the candidate scene graph G_C share an identical structure, the evaluation process can be reduced to an element-wise comparison of triplet semantic information.

are combined into a reference scene graph and a candidate scene graph stored as JSON files.

Semantic fusion. Qwen3-VL-instruct reads the original image and the reference scene graph, fuses their information, and generates the union scene graph in the same template.

Comparative assessment. Qwen3-32B-thinking receives the union and candidate scene graphs, detects missing or inconsistent parts, and produces the final score. We prompt it to revise the candidate text by resolving identified factual inconsistencies.

Hierarchical weighting. During semantic fusion, each element in $\mathcal{G}_U$ is tagged by its provenance: Tag 1 for facts jointly supported by references and images, Tag 2 for facts introduced solely through visual supplementation, and Tag 0 for candidate elements that fail to match $\mathcal{G}_U$. We set the corresponding weights empirically as $w_1 = 0.15$, $w_2 = 0.10$, and $w_0 = 0.10$.

4.3 Evaluation Metrics

We follow previous works (Hessel et al., 2021; Vedantam et al., 2015; Kasai et al., 2021). Kendall-τ_c correlation is used on Flickr8k-Expert, Kendall-τ_b on Flickr8k-CF, Kendall-τ_c on Composite, and classification accuracy on Pascal-50S.

4.4 Comparison with State of the Art

Table 1 reports results on four benchmarks. On F-Ex and F-CF, MoDEST achieves the highest Kendall correlation among all compared methods. The improvement on F-Ex is particularly pronounced, indicating that explicitly verifying entities, attributes, and subject–action–object relations

Type	Metric	F-Ex(τ_c)↑	F-CF(τ_b)↑	Com(τ_c)↑	Pascal-50S Acc. ↑				
					HC	HI	HM	MM	Mean
Text-only Reference	BLEU-4	30.8	16.9	30.6	52.5	90.4	63.0	42.3	55.8
	ROUGE-L	32.3	19.9	32.4	55.0	95.3	93.1	58.7	75.5
	METEOR	41.8	22.2	38.9	59.0	97.7	93.9	62.0	78.2
	CIDEr	43.9	24.6	37.7	53.7	98.1	90.8	63.7	76.6
	SPICE	44.9	24.4	40.3	56.9	96.3	87.1	66.4	76.7
	BERT-S	39.2	22.8	30.1	54.4	96.1	94.3	56.4	75.3
	BERT-S++	46.7	–	44.9	65.4	98.1	96.4	60.3	80.1
	CLAIR	61.6	–	60.4	52.4	99.5	89.8	73.0	78.7
Image-grounded (No Text Ref)	UMIC	46.8	–	56.1	66.1	99.8	**98.1**	76.2	85.1
	FAIEr-r	50.1	32.4	50.5	–	–	–	–	–
	CLIP-S	51.5	34.4	53.8	60.4	99.4	97.8	77.1	83.7
	CLIP-S_{tune}	54.3	36.6	57.3	61.0	99.5	95.9	**82.0**	84.6
	InfoCLIP	32.6	23.5	15.3	37.3	87.3	58.9	72.9	64.1
	InfoCLIP$_{tune}$	37.7	27.7	24.6	37.3	92.5	62.7	74.7	66.8
	InfoMetIC	54.2	36.3	59.2	69.0	99.8	94.0	78.3	85.3
	InfoMetIC+	55.5	36.6	**59.3**	**69.9**	99.7	96.8	79.6	**86.5**
	FLEUR	53.0	38.6	63.5	61.3	99.7	97.6	74.2	83.2
Hybrid (Image + Text Ref)	TIGEr	49.3	–	45.4	56.0	99.8	92.8	74.2	80.7
	ViLBERTScore-F	50.1	–	52.4	49.9	99.6	93.1	75.8	79.6
	FAIEr-4	52.6	35.4	57.7	59.7	**99.9**	92.7	73.4	81.4
	RefCLIP-S	53.0	36.4	55.4	57.9	99.5	96.1	80.8	83.6
	MoDEST	**64.1**	**40.8**	58.1	60.8	99.8	**98.1**	72.2	82.2

Table 1: Comparison of automatic image caption evaluation metrics under different evaluation paradigms. **F-Ex**, **F-CF**, and **Com** denote Flickr8k-Expert, Flickr8k-CF, and Composite, respectively (Section 4.1). Kendall's τ_c is used on F-Ex and Com, Kendall's τ_b on F-CF, and classification accuracy on Pascal-50S. All scores are reported as percentages; higher values indicate better agreement with human judgments.

is effective for identifying subtle factual errors that global feature similarity misses.

On Composite, MoDEST performs competitively with CLIP-S_{tune} and other strong image-grounded baselines. This suggests that when evaluation focuses primarily on overall content relevance, feature-based similarity remains effective, reducing the relative advantage of explicit relational verification.

Beyond quantitative performance, MoDEST differs fundamentally from CLIP-S_{tune} in its ability to provide semantically grounded and actionable evaluation. While CLIP-S_{tune} outputs a single scalar similarity score, MoDEST performs element-wise verification over entities, attributes, and relations extracted from the candidate caption and compares them against the union scene graph. This enables MoDEST to distinguish captions that receive similar similarity scores but differ in factual correctness. Figure 6 presents a qualitative comparison: CLIP-S_{tune} assigns similarly high scores to captions with incorrect relations, hallucinated attributes, or missing salient entities, whereas MoDEST explicitly localizes these semantic inconsistencies through structured verification and provides targeted corrections.

Comparison with LLM-based evaluators. CLAIR and FLEUR represent the prevailing end-to-end paradigm. CLAIR, being text-only, lacks visual grounding and achieves lower Kendall correlation on Flickr8k-Expert (61.6 vs. 64.1). FLEUR incorporates visual input but processes image and text holistically; it surpasses MoDEST on Composite (63.5 vs. 58.1), where overall semantic relevance dominates, yet trails on Flickr8k-Expert (53.0 vs. 64.1) and Flickr8k-CF (38.6 vs. 40.8). These results suggest that end-to-end semantic matching excels at coarse-grained relevance assessment but struggles with fine-grained factual verification—exactly the scenario where MoDEST's explicit triplet-level comparison shines.

More critically, both CLAIR and FLEUR operate as black-box generators: their scores and explanations stem from unconstrained language generation, making it impossible to deterministically trace which specific object, attribute, or relation caused a score reduction. In contrast, MoDEST constrains the LLM/MLLM to structured graph operations (extraction, fusion, and element-wise matching), yielding not only scores but also verifiable alignment records. This structural constraint simultane-

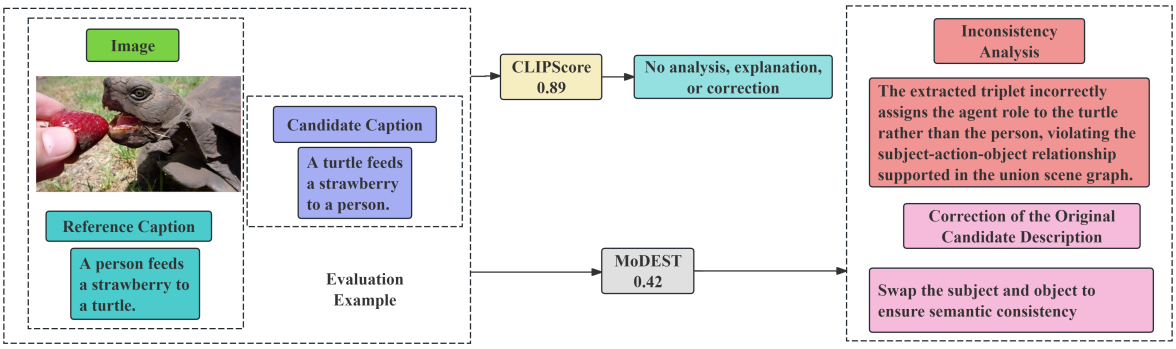

Figure 6: While CLIPScore, based on feature-vector similarity, yields a high score of 0.89, it provides no scoring rationale. In contrast, MoDEST avoids vague similarity computation and instead relies on structured scene graph parsing, assigning a low score of 0.42 to the candidate description, explicitly identifying the cause of the inconsistency, and providing a correction for the erroneous description.

ously improves interpretability and reduces output variance, as evidenced by the low standard deviations reported in Table 5 (Std < 0.18 across all subsets). Thus, MoDEST does not merely apply LLMs to evaluation; it harnesses them within a structured framework that combines the semantic richness of large models with the precision and auditability of explicit graph comparison.

Performance on Pascal-50S. Pascal-50S measures pairwise preference rather than absolute quality. MoDEST achieves near-saturated performance on **HI** (99.8%) and **HM** (98.1%) because these subsets contain at least one factually incorrect caption, and our explicit triplet verification reliably detects hallucinated objects, reversed relations, or missing attributes.

The weaker results on **HC** (60.8%) and **MM** (72.2%) reveal a fundamental limitation. In **HC**, both captions are factually valid, so MoDEST assigns high scores to both but lacks granularity to discriminate based on stylistic nuance or descriptive richness. In **MM**, both candidates contain partial errors; MoDEST evaluates each against the absolute standard of $\mathcal{G}_U$, penalizing every deviation, which yields moderate scores for both and hinders reliable relative ranking. By contrast, embedding-based metrics such as CLIP-S_{tune} (82.0% on MM) excel by measuring vector-space proximity, naturally supporting relative comparison even when both candidates are imperfect. InfoMetIC+ achieves the highest mean accuracy (86.5%) by combining fine-grained token-level scoring with learned ranking, explicitly optimizing for the pairwise paradigm.

These results do not indicate a failure of factual verification; rather, they highlight a *paradigm mismatch*. MoDEST is designed for absolute consistency assessment ("Is this caption faithful to the image and references?"), whereas Pascal-50S rewards relative preference modeling ("Which of these two imperfect captions is preferable?"). When at least one candidate is clearly wrong, absolute verification suffices; when both are partially correct or both are human-quality, relative discrimination requires additional signals—such as salience ranking, stylistic preference, or learned pairwise margins—that lie outside MoDEST's current scope. Additional analysis of model scaling, cross-model generalization, and scoring stability is provided in Appendix B.

4.5 Ablation Study

We conduct an ablation study on Flickr8k-Expert (F-Ex), Flickr8k-CF (F-CF), and Composite (Com) to analyze the impact of reference modalities and architectural components (Table 2).

Effect of reference modality. Using only the image as reference (r1) consistently outperforms using only text references (r3): 52.8 vs. 48.5 on F-Ex, 36.4 vs. 28.3 on F-CF, and 49.6 vs. 45.2 on Com. The gap is particularly pronounced on F-CF (8.1 points), indicating that visual grounding is crucial for verifying fine-grained factual consistency, whereas text-only supervision suffers from insufficient coverage and potential hallucinations.

Effect of information extraction. Adding the information extraction module to image-only input (r2) yields steady gains on all metrics (F-Ex: +2.5, F-CF: +2.3, Com: +2.2), showing that struc-

ID	Reference Modal		Architecture		F-Ex(τ_c)↑	F-CF(τ_b)↑	Com(τ_c)↑
	Image	Text	Info. Ex.	Se. Fusion			
r1	✓				52.8	36.4	49.6
r2	✓		✓		55.3	38.7	51.8
r3		✓			48.5	28.3	45.2
r4		✓	✓		51.2	26.5	47.3
r5	✓	✓			56.2	37.8	52.4
r6	✓	✓	✓		58.6	39.2	54.5
r7	✓	✓	✓	✓	64.1	40.8	58.1

Table 2: Ablation study of reference modalities and architectural components. **Info. Ex.** denotes the information extraction module, and **Se. Fusion** denotes the semantic fusion module between image and text. Results are reported on Flickr8k-Expert (F-Ex), Flickr8k-CF (F-CF), and Composite (Com).

tured parsing helps capture relational errors. However, when only text references are available, enabling extraction (r4) improves F-Ex (+2.7) and Com (+2.1) yet degrades F-CF (-1.8). This reveals a dual effect: structured representation enhances semantic clarity, but without visual constraint it may amplify textual biases, compromising factual verification.

Effect of multimodal inputs and semantic fusion. Simply combining both modalities without extraction or fusion (r5) already surpasses the best single-modal configuration (56.2 vs. 52.8 on F-Ex), confirming that visual facts and human preferences are complementary. Adding information extraction (r6) yields further gains (F-Ex: +2.4, F-CF: +1.4, Com: +2.1), yet the full model with semantic fusion (r7) achieves the largest leap, especially on F-Ex ($58.6 \rightarrow 64.1$, +5.5). This demonstrates that semantic fusion is essential for cross-modal alignment: it corrects textual inconsistencies against visual evidence and supplements visually grounded facts missing from the references. Consequently, we treat **r7** as the full **MoDEST** model.

5 Conclusion

We presented **MoDEST**, a semantic-level evaluation framework for image captioning that goes beyond feature similarity by explicitly verifying entities, attributes, and subject–action–object relations between a candidate caption and a reference union scene graph. By combining text-derived reference guidance with image-grounded semantic fusion, MoDEST evaluates captions from both *fidelity* (penalising hallucinated facts) and *adequacy* (rewarding coverage of salient visual content) perspectives. Empirically, MoDEST achieves strong alignment with human judgments across multiple benchmarks, and the ablation results confirm the importance of hybrid references together with information extraction and semantic fusion. Beyond metric scores, MoDEST provides interpretable, fine-grained diagnostic evidence for factual mismatches and uses these explanations to guide targeted caption revisions, making the evaluation process transparent and actionable.

Limitations

MoDEST relies on LLMs and MLLMs for semantic extraction, fusion, and comparison. While fixed prompts and deterministic decoding improve stability, future work could explore model-agnostic extraction strategies or lightweight verification modules for greater robustness, and optimize the multi-stage pipeline whose computational cost exceeds that of feature-based metrics. Although MoDEST generates structured explanations, inconsistency analyses, and correction suggestions, no human evaluations have been conducted to assess their quality or utility. This work demonstrates the mechanism for interpretability—element-wise graph matching and natural-language rationales—but leaves open whether these outputs effectively help users identify errors, trust the metric, or improve their captions. Future studies should therefore systematically evaluate explanation correctness, completeness, and actionability against human judgments, and compare them with free-form justifications from end-to-end LLM evaluators. Finally, extending MoDEST beyond reference-caption approximations of human preference to broader domains, languages, and task settings remains an important research direction.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermüller. 2018. Image

understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding*, 173:33–45.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Richard Ring, Eliza Rutherford, Serkan Cabi, Tao Han, Zheng Gong, Sina Samangooei, Matheus Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, pages 382–398.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

David M. Chan, Suzanne Petryk, Joseph E. Gonzalez, Trevor Darrell, and John Canny. 2023. Clair: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*, pages 104–120.

Wenhai Dai, Junnan Li, Dongxu Li, Anthony Tiong, Jie Zhao, Wenhai Wang, Baishan Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Anwen Hu, Shizhe Chen, Lei Zhang, and Qingzhao Jin. 2023. Infometic: An informative metric for reference-free image caption evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3171–3185.

Ming Jiang, Junjie Hu, Qiuyuan Huang, Lei Zhang, Jana Diesner, and Jianfeng Gao. 2019a. Reorelevance, extraness, omission: A fine-grained evaluation for image captioning. *arXiv preprint*, arXiv:1909.02217.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019b. Tiger: Text-to-image grounding for image caption evaluation. *arXiv preprint*, arXiv:1909.02050.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2021. Transparent human evaluation for image captioning. *arXiv preprint*, arXiv:2111.08940.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hwanhee Lee, Seong Yoon, Franck Dernoncourt, Trong Bui, and Kyunghyun Jung. 2021. Umic: An unreferenced metric for image captioning via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 220–226.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Vilbertscore: Evaluating image captions using vision-and-language bert. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3732–3746.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. 2021. Faier: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14050–14059.

Ziwei Yao, Ruiping Wang, and Xilin Chen. 2024. Hifi-score: Fine-grained image description evaluation with hierarchical parsing graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

A Prompt Templates

A.1 Information Extraction Prompt

We use a frozen text-only large language model to convert each caption into a structured scene graph representation. Given a caption, the model is required to perform deterministic semantic decomposition into a subject-centric graph.

Input. A single image caption.

Instruction Specification. The model is explicitly instructed to perform three atomic operations:

(i) **Entity extraction:** identify all explicit noun phrases referring to objects, people, or scenes;

(ii) **Attribute grounding:** assign visually observable attributes (e.g., color, size, state) to each entity;

(iii) **Relation extraction:** extract directed subject–predicate–object triples for all explicit interactions or actions.

All outputs must follow a strict JSON schema to ensure structural consistency across samples.

Output Constraint. The output must be a subject-centric scene graph in JSON format with: `{entities, attributes, relations}`.

Example. Input: "A man is feeding a dog in the park."

Output: entities: {man, dog, park}; relations: man $\rightarrow$ feeding $\rightarrow$ dog; attributes: (implicit spatial context: park).

A.2 Semantic Fusion Prompt

We introduce a multimodal large language model (MLLM) to align textual semantic facts with visual evidence. This step reduces annotation bias in reference captions and improves factual completeness.

Input. An image and a reference scene graph extracted from captions.

Instruction Specification. The MLLM performs three constrained operations:

(i) **Verification:** validate each entity, attribute, and relation against visual evidence;

(ii) **Preservation:** retain all visually supported facts from the reference graph;

(iii) **Completion:** add missing but visually observable elements not present in the reference.

Output Constraint. The output preserves the same JSON schema as the input reference graph to ensure structural alignment.

Role Clarification. This step acts as a *weakly supervised grounding refinement module*, bridging human annotation bias and visual completeness.

A.3 Comparative Evaluation Prompt

A multimodal large language model is used as a structured semantic comparator.

Input. A union scene graph $\mathcal{G}_U$ and a candidate scene graph $\mathcal{G}_C$.

Instruction Specification. The model performs element-wise alignment:

(i) match entities between $\mathcal{G}_U$ and $\mathcal{G}_C$;

(ii) verify attribute consistency;

(iii) check relation directionality correctness;

(iv) identify missing or hallucinated elements.

Each decision must be accompanied by a short justification.

Aggregation Rule. Final score is computed by aggregating valid matches over total candidate elements.

B Additional Analysis of Foundation Models

We further investigate the behavior of MoDEST from three perspectives: (i) model scale and reasoning mode, (ii) cross-model generalization, and (iii) stability of scoring outputs. All experiments are conducted on Pascal-50S.

B.1 Effects of Model Scale and Reasoning Mode

We first study how model capacity and reasoning strategy affect evaluation performance within the Qwen3 family.

Table 3 reports results across model scales from 0.6B to 32B under both standard and thinking modes. A clear pattern of **non-linear performance scaling** emerges.

At small scales (0.6B–4B), performance remains close to random level (around 50%), indicating that lightweight models lack sufficient semantic grounding and structured reasoning ability to perform reliable scene-graph-based evaluation. This suggests that MoDEST requires a minimum representational capacity before meaningful semantic alignment can emerge.

A sharp performance jump occurs at the 14B scale, where Mean accuracy increases from 52.50% to 76.70%. This transition indicates a **capability emergence threshold**, beyond which the model acquires sufficient semantic abstraction ability to reliably perform entity-level alignment, relation verification, and structured reasoning over scene graphs.

Another important observation is that improvements are particularly pronounced on HI and HM subsets, which require fine-grained error detection and human-vs-machine discrimination. This suggests that MoDEST places stronger demands on relational reasoning than on shallow lexical understanding.

Finally, reasoning mode further enhances performance. As shown in Table 3, switching from standard to thinking mode improves Mean performance from 80.35% to 82.20% for the 32B model. This gain is especially evident in the HM subset, where explicit chain-of-thought reasoning helps the model better detect subtle inconsistencies between candidate captions and structured semantic graphs.

However, this improvement comes at a significant computational cost: inference time nearly doubles. This highlights a fundamental trade-off between accuracy and efficiency in structured multimodal evaluation systems.

B.2 Cross-Model Generalization Across LLM Families

We next evaluate whether MoDEST depends on a specific backbone model by comparing multiple representative LLMs, including GPT-4o, GLM-Z1, DeepSeek-R1, and Llama-3.3.

Table 4 shows that performance is remarkably consistent across all models, with Mean scores tightly clustered between 81.88% and 82.20%. This narrow variance strongly suggests that MoDEST is **model-agnostic**.

More importantly, the consistency is not only reflected in overall scores but also across individual subsets. For example, all models achieve near-saturated performance on the HI subset (above 99%), while maintaining similarly strong results on HM and MM subsets.

This convergence suggests that MoDEST does not depend on model-specific priors, but instead induces a shared evaluation behavior through its structured scene-graph representation. In other words, the architecture itself constrains different LLMs to perform similar semantic reasoning operations.

B.3 Stability and Reproducibility of Scoring Results

Finally, we evaluate the stability of MoDEST by running Qwen3-32B-Thinking for 10 independent trials under identical conditions.

As shown in Table 5, MoDEST exhibits extremely low variance across all subsets, with standard deviation consistently below 0.18. This indicates that the evaluation process is highly stable and nearly deterministic under structured constraints.

Table 3: Performance (Kendall correlation, %) and inference time (s) of Qwen3 models on Pascal-50S.

Model	Params	Mode	HC	HI	HM	MM	Mean	Time
Qwen3-0.6B	0.6B	Std	48.52	51.37	50.86	49.25	50.00	8.50
Qwen3-1.7B	1.7B	Std	49.18	52.64	51.43	48.92	50.54	12.30
Qwen3-4B	4B	Std	50.75	54.28	53.62	51.35	52.50	18.70
Qwen3-14B	14B	Std	56.35	95.42	89.76	65.28	76.70	28.40
Qwen3-32B	32B	Std	58.94	98.15	94.83	69.47	80.35	35.60
Qwen3-32B	32B	Thinking	60.80	99.80	98.10	72.20	82.20	72.40

Table 4: Performance comparison of different LLMs on Pascal-50S (Kendall correlation, %).

Model	Params	Mode	HC	HI	HM	MM	Mean
Qwen3-32B-Thinking	32B	Thinking	60.80	99.80	98.10	72.20	82.20
GPT-4o	Closed	Std	60.65	99.72	97.95	72.05	82.08
GLM-Z1-32B	32B	Thinking	60.58	99.68	97.88	71.92	82.02
DeepSeek-R1-Distill-Qwen-32B	32B	Thinking	60.45	99.55	97.75	71.78	81.88
Llama-3.3-70B	70B	Std	60.72	99.75	98.02	72.12	82.15

Table 5: Stability statistics across 10 runs using Qwen3-32B-Thinking (Kendall correlation, %).

Subset	Mean	Std	Max–Min
HC	60.80	0.12	0.32
HI	99.80	0.05	0.14
HM	98.10	0.08	0.22
MM	72.20	0.18	0.48
Mean	82.20	0.10	0.29

The stability can be attributed to the decomposition of the evaluation task into explicit semantic matching operations. Instead of free-form generation, MoDEST performs constrained graph-level comparison, which significantly reduces stochastic variability. This design is particularly effective for the HI subset, where decision boundaries are clear and objective.

Even in the MM subset, which is inherently more ambiguous due to partial correctness of both candidates, the variance remains low, confirming the robustness of the framework in practical evaluation scenarios.