arXiv:2605.04503v2 [cs.CV] 2 Nov 2026

DiffCap-Bench: A Comprehensive, Challenging, Robust Benchmark for Image Difference Captioning

Yuancheng Wei[♡*] **Haojie Zhang**[♡*] **Linli Yao**[◇] **Lei Li**[♠]
Jiali Chen[♡] **Tao Huang**[♣] **Yiting Lu**[‡]
Duojun Huang[‡] **Xin Li**[‡†] **Zhao Zhong**[‡]
[♡]South China University of Technology
[◇]Peking University
[♠]The University of Hong Kong
[♣]Tianjin University
[‡]Tencent Hunyuan

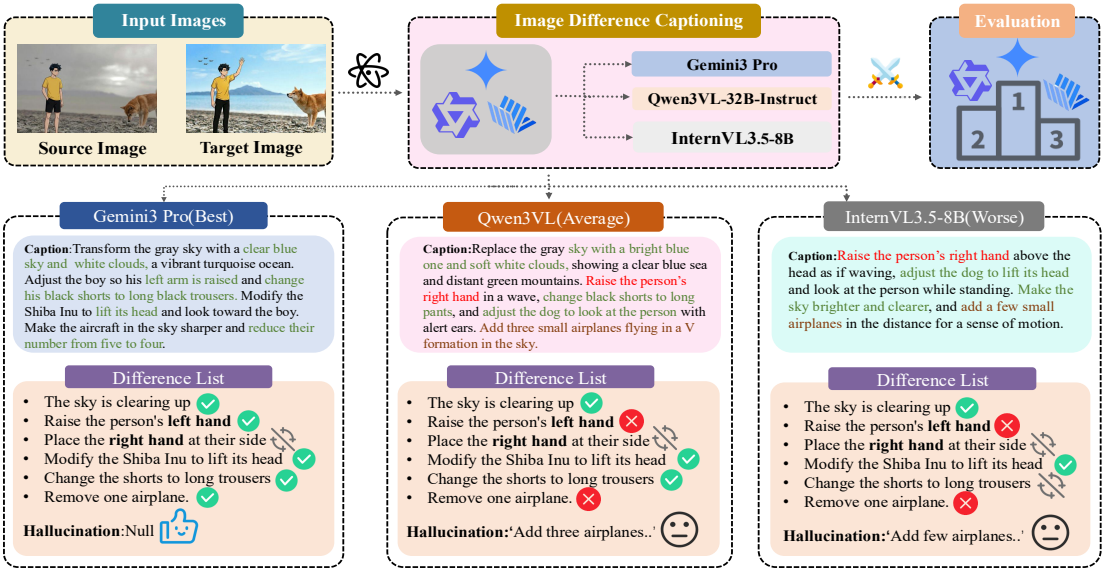


Figure 1: An example from DiffCap-Bench. Captions are checked against a human-validated Difference List: (1) ✓ correctly described; (2) × incorrectly described; (3) **Missed** omitted; **Hallucination** denotes fabricated changes. **Gemini 3 Pro** captures differences with few hallucinations, while open-source models show more errors.

Abstract

Image Difference Captioning (IDC) generates natural language descriptions that precisely identify differences between two images, serving as a key task for fine-grained change perception, cross-modal reasoning, and image editing data construction. However, existing benchmarks lack data diversity and compositional complexity, while their evaluation protocols rely on lexical-overlap metrics (e.g., BLEU, METEOR) that fail to capture semantic consistency or penalize hallucinations, preventing a comprehensive and robust evaluation of multimodal large language models (MLLMs) on IDC. To address these gaps, we introduce **DiffCap-Bench**, a comprehensive IDC benchmark covering ten difference categories. Furthermore, we propose an LLM-as-a-Judge evaluation protocol grounded in human-validated *Difference Lists*, enabling a robust assessment of models' ability to capture and describe visual changes. Through extensive evaluation of state-of-the-art MLLMs, we reveal significant gaps between proprietary and open-source models and the critical role of reasoning capability. Further analysis shows that our framework demonstrates stable cross-judge rankings, strong alignment with human expert judgments, and strong correlation with downstream image editing data construction quality. These findings establish DiffCap-Bench as both a reliable IDC evaluation framework and a practical predictor of downstream utility.

wyc528813339@gmail.com
https://github.com/wyclike/DiffCap-Bench
[*]Equal contribution.
[†]Corresponding author.

Introduction

Image Difference Captioning (IDC) provides precise natural language descriptions of visual discrepancies between image

pairs, serving as a foundational task for change detection and visual reasoning (Vinyals et al. 2015; Kim, Lee, and Park 2026). In particular, IDC also serves as a critical upstream component for image editing pipelines by identifying differences between two images to provide intermediate edit instructions whose accuracy directly determines the quality of the resulting image editing data (Hui et al. 2024; Ye et al. 2025b). With the rapid advancement of Multimodal Large Language Models (MLLMs) and their demonstrated strong performance on open-ended vision-language tasks such as image captioning (Xing et al. 2025; Lu et al. 2025), how well these models perform on IDC remains systematically unexplored. This gap stems primarily from the limitations of existing benchmarks that fall short on both the data side and the evaluation side, rendering them unable to provide a robust and reliable assessment of MLLMs' IDC capabilities.

On the data side, existing IDC benchmarks remain limited in two key aspects: they are neither comprehensive enough in difference coverage nor challenging enough in compositional complexity. Earlier datasets exhibit restricted diversity in different ways: Spot-the-Diff (Jhamtani and Berg-Kirkpatrick 2018) is mainly built from surveillance-style videos with constrained scenarios, while CLEVR-Change (Park, Darrell, and Rohrbach 2019) focuses on synthetic settings; both typically contain few differences per pair. More recent datasets, including Image-Edit-Request (Tan et al. 2019), OneDiff (Hu et al. 2024), DiffTell (Di et al. 2025), and OmniDiff (Liu et al. 2025), improve realism, yet still often exhibit limited compositionality, incomplete coverage, or annotation noise. As a result, they underrepresent the breadth and difficulty of IDC in modern image editing contexts, where multiple coupled differences can co-occur across objects, attributes, text, viewpoint, and global image properties.

On the evaluation side, most previous IDC benchmarks (Hu et al. 2024; Liu et al. 2025; Jhamtani and Berg-Kirkpatrick 2018) rely on lexical-overlap metrics (e.g., BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE-L (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015)), which are brittle for open-form difference descriptions (Cui et al. 2018; Caglayan, Madhyastha, and Specia 2020). As shown in Figure 2, zero-shot scores of strong MLLMs remain uniformly low on traditional IDC benchmarks (Image-Edit-Request (Tan et al. 2019) and Spot-the-Diff (Jhamtani and Berg-Kirkpatrick 2018)) under these metrics, falling far below the dataset-fine-tuned M^3Diff (Liu et al. 2025) and showing poor separability across models. This is largely because free-form MLLM outputs, which often differ in language style and response format, are poorly matched to n-gram lexical overlap. Moreover, these metrics cannot reliably capture semantic correctness and fail to penalize hallucinations of non-existent differences. Consequently, they cannot provide a reliable and robust evaluation of modern MLLMs on IDC.

To address these issues, we introduce **DiffCap-Bench**, a comprehensive and challenging benchmark for realistic and compositional IDC in modern image editing contexts. DiffCap-Bench is designed with two principles: coverage of diverse real image editing scenarios and semantic reliability of evaluation. For data construction, we combine diverse real-world sampling (e.g., web images, videos, advertisements, posters) with controlled synthetic generation (2D compositing and 3D rendering), yielding 1,075 high-quality image pairs that systematically cover ten difference categories.

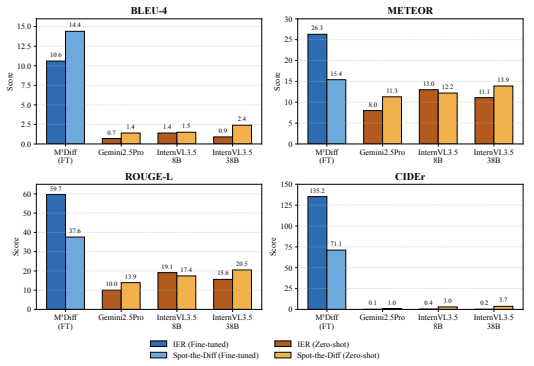


Figure 2: Under lexical-overlap metrics, zero-shot MLLMs score far below fine-tuned model on the traditional IDC benchmarks (Image-Edit-Request, Spot-the-Diff).

Beyond data diversity and challenge, we propose an *LLM-as-a-Judge* (Gu et al. 2024) evaluation protocol grounded in human-validated *Difference Lists* that enables robust assessment of models' ability to capture and describe visual changes. Rather than comparing generated text against a single reference sentence, our protocol operates on human-validated atomic difference items and evaluates completeness and faithfulness through Recall*, Precision*, F_1^*, and Hallucination Rate. This design enables robust assessment under linguistic variation and captures whether a model identifies correct differences while avoiding unsupported claims.

Leveraging DiffCap-Bench, we systematically evaluate 16 proprietary and open-source MLLMs and find: (1) a clear performance gap between proprietary and open-source models; (2) the superiority of *Thinking* variants over their base counterparts, highlighting the role of reasoning in accuracy and faithfulness; and (3) a bottleneck on action, pose, and expression changes, versus stronger performance on background, text, and image-property differences.

To assess reliability, we analyze DiffCap-Bench from three perspectives—cross-judge ranking consistency, alignment with human judgment, and downstream utility. We find highly stable rankings across independent judges, strong agreement between our automatic metrics and human experts, and clear correlation between benchmark scores and downstream image-editing quality under model-generated relabeling data. These findings confirm DiffCap-Bench as a robust IDC evaluation framework and a reliable proxy for real-world utility.

In summary, our contributions are threefold:

- We introduce **DiffCap-Bench**, a comprehensive, challenging and robust benchmark designed for evaluation of MLLMs in fine-grained image difference captioning.
- We propose a semantically grounded evaluation framework based on human-validated Difference Lists, and to our knowledge provide the first systematic evaluation of

MLLMs on IDC, covering 16 state-of-the-art models and revealing key limitations of current systems.

- We validate DiffCap-Bench via cross-judge consistency, human alignment, and downstream correlation, showing it as a robust evaluation protocol and a proxy for MLLM utility in automated image-editing data engineering.

Related Work

Image Difference Captioning Benchmarks.

Early IDC benchmarks mostly focus on constrained scenarios and simple difference types. Spot-the-Diff (Jhamtani and Berg-Kirkpatrick 2018) is built from surveillance videos and shows limited variation patterns, making it less suited to evaluating model robustness in diverse real-world settings. CLEVR-Change (Park, Darrell, and Rohrbach 2019) and CLEVR-DC (Kim et al. 2021) rely on synthetic desktop objects and mainly cover basic operations such as object addition, deletion, and replacement. More recent datasets improve realism and scale, but key limits remain. Image-Edit-Request (Tan et al. 2019) introduces natural images with textual editing requests, yet its difference descriptions are typically short and rarely involve compositional or multi-step changes. OneDiff (Hu et al. 2024) combines real and synthetic images, but suffers from annotation noise and uneven coverage of key difference types, such as text edits and OCR-sensitive changes. DiffTell (Di et al. 2025) covers a wider range of differences (e.g., background, local object edits, text, and style), but most samples still center on single-type changes and do not fully capture complex, multi-aspect differences. OmniDiff (Liu et al. 2025) integrates real scenes with 3D simulation to increase diversity, yet still underrepresents camera-viewpoint and compositional shifts. Overall, existing datasets remain insufficient in both broad coverage and compositional difficulty.

Evaluation Methods for IDC.

Standard IDC evaluation relies on lexical-overlap metrics such as BLEU (Papineni et al. 2002), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), METEOR (Banerjee and Lavie 2005), and ROUGE-L (Lin 2004). However, these lexical-level measures are brittle for open-form descriptions (Cui et al. 2018; Caglayan, Madhyastha, and Specia 2020), as they favor surface similarity over semantic fidelity. As a result, they often fail to detect key factual errors or penalize hallucinations; for example, semantically opposite phrases like *raised left hand* and *raised right hand* may still receive similar overlap scores. To address these limits, recent studies on LLM-as-a-Judge (Gu et al. 2024) have shown better alignment with human judgment in open-ended tasks, indicating their potential for robust evaluation.

DiffCap-Bench

To address the limitations of insufficiently comprehensive and challenging data and the lack of robust evaluation in IDC, we introduce **DiffCap-Bench** as illustrated in Fig. 3. In the following subsections, we first describe diverse image-pair construction and difference list labeling. We then report comprehensive dataset statistics and finally compare DiffCap-Bench with existing IDC benchmarks.

Entry	Count	Definition
Overall		
Samples	1,075	Number of image pairs.
Overall diffs.	6,713	Total verified difference items.
Avg. diffs.	6.25	Mean number of differences per pair.
Difference Type		
Object	2,314	Object-level additions, deletions, or replacements.
Attribute	1,760	Attribute modifications (e.g., size, color).
Action & Pose	1,153	Changes in actions or body poses.
Spatial	1,002	Spatial rearrangements (e.g., moving objects).
Composition	834	Compositional differences, including cropping.
Text & Symbol	831	Addition, removal, or modification of text/symbols.
Background	806	Alterations to the scene background.
Camera/View	571	Camera movement or viewpoint changes.
Image Prop.	425	Global changes (e.g., resolution, saturation).
Expression	333	Changes in facial expressions.

Table 1: DiffCap-Bench statistics and category definitions.

Image Pair Construction

To ensure data diversity and broad coverage of realistic difference types, we construct data through both sampling and synthesis pipelines, with quality filtering applied to ensure the reliability of the resulting image pairs.

Multisource Sampling We collect image pairs from multiple sources to maximize coverage of real-world differences, including web images, video frames, advertisements, and posters. These sources are rich in human-centric content and naturally contain diverse, realistic transformations. From this pool, we identify candidate image pairs for further filtering.

Synthetic Image Pairs via 2D Compositing and 3D Rendering We generate synthetic image pairs to further improve dataset comprehensiveness and challenge through controlled compositional editing. For 2D synthesis, we use Canva, an online graphic design and layout platform, to perform layered compositing and element-level modifications that induce diverse visual discrepancies, ranging from object manipulation and attribute shifts to background and text alterations. For 3D synthesis, we use Blender, an open-source 3D creation suite, to render scenes with systematically varied elements, enabling structured and controllable difference patterns.

Quality Filtering To ensure both data quality and benchmark challenge, we apply strict filtering to all collected pairs. We employ a strong MLLM, Qwen3VL-Plus (Bai et al. 2025), as the automatic filter with three criteria: (1) *Difference Existence*: the two images must contain real differences rather than being identical; (2) *Visual Similarity*: the pair must remain semantically related rather than depicting unrelated scenes; and (3) *Image Quality*: both images must satisfy quality requirements (e.g., resolution and clarity). All retained pairs are verified by human experts, and pairs deemed insufficiently challenging are removed. Following this pipeline, we curate a final set of *1,075* high-quality image pairs. The MLLM filtering prompts, detailed criteria, and the challengingness filtering procedure are provided in Appendix.

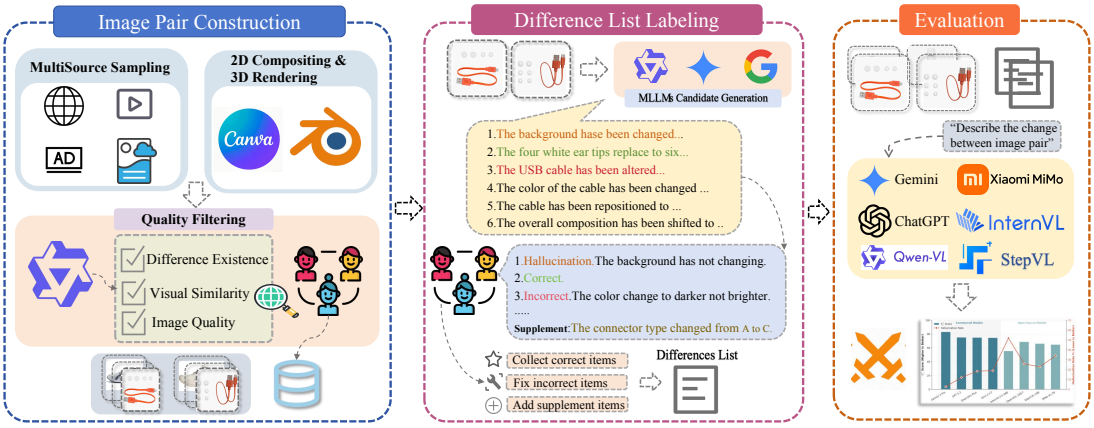


Figure 3: Overview of the DiffCap-Bench construction pipeline, consisting of three stages: *Image Pair Construction* collects and filters image pairs from diverse sources; *Difference List Labeling* builds a human-validated *Difference List* of atomic difference items per pair; and *Evaluation* systematically benchmarks current MLLMs to reveal their IDC capabilities and limitations.

Dataset	Source	Avg. Chg.	Avg. Wd.	Human	Eval.	Metrics	Hall.
CLEVR-Change (Park, Darrell, and Rohrbach 2019)	3D render	1	8	×	Lexical	B/M/R/C	×
Spot-the-Diff (Jhamtani and Berg-Kirkpatrick 2018)	Surv. video	1.86	19	✓	Lexical	B/M/R/C	×
Image-Edit-Request (Tan et al. 2019)	Diverse real	1	8	✓	Lexical	B/M/R/C	×
OneDiff (Hu et al. 2024)	Syn. & real	1	15	×	Lexical	B/M/R/C	×
DiffTell (Di et al. 2025)	Syn. & real	1	9.72	✓	Lexical	B/M/R/C	×
OmniDiff (Liu et al. 2025)	3D & real	2.5	60	✓	Lexical	B/M/R/C	×
DiffCap-Bench (Ours)	2D, 3D & real	6.25	175	✓	LLM-as-a-Judge	$R^*/P^*/F_1^*$/H	✓

Table 2: Comparison of DiffCap-Bench with existing IDC benchmarks (Surv.: Surveillance, Syn.: Synthetic, Chg.: Changes, Wd.: Words; B/M/R/C: BLEU, METEOR, ROUGE-L, CIDEr; $R^*/P^*/F_1^*$/H: Recall*, Precision*, F_1^*, Hallucination Rate).

Difference List Labeling

To enable robust evaluation, we construct the human-validated *Difference List* comprising atomic difference items for each image pair through a three-stage pipeline in Fig. 3. First, multiple MLLMs (Qwen3VL-30B-A3B-Instruct (Bai et al. 2025), Gemini 2.5 Pro (Comanici et al. 2025), and Gemini 3 Pro (Comanici et al. 2025)) are prompted to describe pairwise differences, and the resulting descriptions are segmented into atomic units and merged into a candidate pool (typically 10–30 items per pair) to improve recall. Second, we employ a professional annotation team with extensive image-caption experience and conduct a 10-workday campaign with task-specific training. Each candidate is reviewed in a double-blind manner by at least two annotators and labeled as: (1) *correct*: the difference is real and accurately described; (2) *incorrect-but-fixable*: the difference is real but described inaccurately and requires correction; (3) *hallucinated*: the difference does not exist in the image pair; or (4) *indistinguishable*: the descriptions cannot be reliably determined. Annotators may also add missing differences; in the first round, about 32% of candidates required correction or removal. Finally, all valid differences, including correct, expert-corrected, and expert-added items, are aggregated into a **Key Difference List**, while indistinguishable ones are retained as **Indistinguishable Items**, constituting the human-validated *Difference List* for each pair.

Dataset Statistics

As shown in Table 1, we collect 1,075 image pairs, and after expert annotation, the dataset contains 6,713 difference items, with an average of 6.25 differences per pair and 175 words per *Difference List*. To analyze difference-type distribution, we categorize each difference into predefined edit categories using an LLM-assisted classification process (see Appendix), allowing multi-label assignments as a single difference can involve multiple aspects. All annotations are verified by humans.

Comparisons with Existing Benchmarks

We compare DiffCap-Bench with representative Image Difference Captioning benchmarks in terms of data source, difference complexity, annotation protocol, and evaluation methodology. As shown in Table 2, existing datasets typically contain only a limited number of differences per image pair and rely on lexical-overlap metrics. Moreover, hallucination detection is not explicitly modeled. In contrast, DiffCap-Bench contains more differences per image pair (6.25 on average) and longer descriptions, reflecting higher compositional complexity and benchmark challenge. Furthermore, we adopt a Difference List-based *LLM-as-a-Judge* frame-

work that measures completeness and faithfulness through Recall*, Precision*, F_1^*, and Hallucination Rate, enabling a more robust and discriminative evaluation mechanism.

Experiment

We conduct extensive experiments to evaluate the performance of state-of-the-art MLLMs on our benchmark. This section presents our experimental pipeline and corresponding key findings, which are organized as follows: evaluation protocol and metrics, evaluated models and the results.

Evaluation Protocol and Metrics

To provide a robust evaluation of generated difference descriptions, we propose a bidirectional protocol based on the predefined *Difference List*. For a given image pair, the model generates a predicted description a_1 as follows:

$$a_1 = \mathcal{P}(y \mid \text{image pair}, \text{prompt}) \tag{1}$$

The evaluation protocol then prompts an LLM-as-a-Judge to apply a two-stage checking process to ensure comprehensive coverage and factual faithfulness: **Forward Check (Coverage & Precision)** measures coverage and factual correctness with respect to the ground truth. Each item in the *Key Difference List* is cross-referenced with a_1 to determine whether it is captured and correctly described. We define N_{hr} as the number of correctly identified differences and N_{hw} as the number of differences that are mentioned but incorrectly described. **Backward Check (Redundancy & Hallucination)** identifies redundant or erroneous content in a_1. Any predicted element not found in the *Key Difference List* is cross-referenced with the *Indistinguishable Items*. Matches in the latter are treated as ambiguous and excluded from penalties, as even expert annotators cannot provide reliable determinations for these cases. The remaining extra elements are then classified as either valid elaborations or hallucinations, with N_{hallu} denoting the number of identified hallucinations. Based on these statistics, we define the following metrics:

Recall* Quantifies the model's capacity to encompass all ground-truth differences:

$$\text{Recall}^* = \frac{N_{hr}}{|L_{key}|} \tag{2}$$

where $|L_{key}|$ denotes the total number of difference items in the *Key Difference List*.

Precision* Measures the proportion of correct descriptions among all predicted differences:

$$\text{Precision}^* = \frac{N_{hr}}{N_{hr} + N_{hw} + N_{hallu}} \tag{3}$$

F_1^* Score The harmonic mean of Recall* and Precision*, serving as a **holistic metric** to balance coverage and factual integrity:

$$F_1^* = \frac{2 \cdot \text{Recall}^* \cdot \text{Precision}^*}{\text{Recall}^* + \text{Precision}^*} \tag{4}$$

Hallucination Rate Quantifies the frequency of hallucinations relative to the difference density of the scene:

$$\text{Hallucination Rate} = \frac{N_{hallu}}{|L_{key}|} \tag{5}$$

By normalizing hallucinations by the number of ground-truth differences, this metric imposes a stronger penalty on spurious predictions in simpler scenes with fewer visual discrepancies. The prompt for the judge LLM and a representative evaluation case study are provided in the Appendix.

Evaluated Models

We conduct a thorough evaluation across 16 state-of-the-art MLLMs, including both commercial and open-source models. Commercial models include Gemini 3 Pro (Comanici et al. 2025), GPT 5.2 (Singh et al. 2025), Seed1.6-VL (Guo et al. 2025), Seed1.6-VL-Thinking (Guo et al. 2025), Qwen3VL-Plus (Bai et al. 2025), Kimi 2.5 (Team et al. 2026), Kimi 2.5 Thinking (Team et al. 2026), and Grok 4.2 Beta. Open-source models include InternVL3.5 (8B/38B) (Wang et al. 2025), Step3-VL-10B (Huang et al. 2026), MiMo-VL-7B-RL-2508 (Xiaomi and Team 2025), Qwen3VL-Instruct (8B/32B) (Bai et al. 2025), and Qwen3VL-Thinking (8B/32B) (Bai et al. 2025), ranging from 7B to 38B parameters.

For all models, we use the same system prompt and apply the evaluation protocol in section Evaluation Protocol and Metrics, with Gemini 2.5 Pro (Comanici et al. 2025) serving as the judge model. The system prompt and inference configuration for all evaluated models are provided in the Appendix.

Evaluation Results of DiffCap-Bench

Table 3 reports results on DiffCap-Bench, from which we draw the following observations.

A clear performance gap remains between commercial and open-source MLLMs. Among commercial models, Gemini 3 Pro achieves state-of-the-art performance with an F_1^* of 83.1%, substantially surpassing GPT 5.2 (75.1%) and the leading open-source model, Qwen3VL-32B-Thinking (68.7%). While commercial models generally maintain F_1^* above 70%, most open-source models struggle in the 60–70% range. This gap is clearer in Hallucination Rate: Gemini 3 Pro reaches only 2.6%, whereas other models typically exhibit 20%–30%, and the InternVL3.5 series exceeds 50%, indicating a major challenge in factual faithfulness on IDC.

Reasoning ability and Chain-of-Thought significantly enhance IDC performance. Across identical architectures, *Thinking* variants consistently outperform their base counterparts. Seed1.6-VL-Thinking improves F_1^* by 3.4% and reduces Hallucination Rate by 3%; Qwen3VL-8B-Thinking gains 3.8% F_1^* and cuts hallucinations by 6% over Instruct; Qwen3VL-32B-Thinking further yields a 5% F_1^* gain and a 5.5% hallucination drop. Most strikingly, Kimi 2.5 Thinking reduces hallucinations by 19% with a modest F_1^* gain (74.5% vs. 71.3%). These results suggest that CoT reasoning is pivotal for perceiving and describing subtle visual discrepancies. Case studies are provided in the Appendix.

Model	Comp.	Back	Attr.	Text	Cam.	Expr.	Spat.	Img.	Obj.	Act.	R^*	P^*	F_1^*	H
Commercial														
Gemini 3 Pro (Comanici et al. 2025)	88.0	89.4	80.8	88.0	86.5	72.1	**83.5**	87.8	**87.7**	75.5	**76.9**	**93.6**	**83.1**	**2.6**
GPT 5.2 (Singh et al. 2025)	**91.5**	89.0	77.1	86.3	**89.4**	30.0	82.7	**89.1**	83.0	62.9	70.5	85.2	75.1	12.0
Seed1.6-VL (Guo et al. 2025)	68.0	87.2	77.6	85.7	59.3	44.3	70.5	83.3	83.3	63.9	67.4	78.2	69.7	23.9
Seed1.6-VL-Thinking (Guo et al. 2025)	73.6	87.8	79.2	86.3	65.3	55.9	75.6	83.6	86.2	67.5	70.5	80.4	73.1	20.9
Qwen3VL-Plus (Bai et al. 2025)	79.3	87.9	**82.4**	**90.5**	75.1	**79.1**	78.6	85.9	85.3	**75.8**	71.3	83.3	74.9	18.3
Kimi 2.5 (Team et al. 2026)	77.5	**90.0**	80.0	87.5	71.1	48.3	75.0	87.2	85.7	66.4	71.8	76.0	71.3	37.7
Kimi 2.5 Thinking (Team et al. 2026)	77.4	86.9	80.0	88.0	74.5	60.6	78.0	82.1	86.6	66.8	71.9	81.6	74.5	18.9
Grok 4.2 Beta	75.8	82.7	70.3	76.5	72.7	71.6	73.8	72.7	77.4	66.8	63.5	81.1	69.6	10.5
Open-Source														
InternVL3.5-8B (Wang et al. 2025)	62.9	71.9	62.0	78.9	64.9	58.2	63.0	77.2	69.3	59.9	51.1	60.0	51.9	62.6
InternVL3.5-38B (Wang et al. 2025)	59.2	73.8	63.5	78.5	57.2	63.8	63.1	78.9	70.8	62.7	53.3	65.6	55.5	51.8
Qwen3VL-8B-Instruct (Bai et al. 2025)	69.2	77.3	70.4	77.3	70.3	63.7	67.0	76.2	72.4	66.9	56.4	74.7	61.7	26.1
Qwen3VL-8B-Thinking (Bai et al. 2025)	66.7	81.4	73.3	83.7	62.9	48.3	71.1	74.1	79.4	63.3	60.4	77.0	65.5	20.2
Qwen3VL-32B-Instruct (Bai et al. 2025)	72.3	81.3	68.8	76.2	71.7	66.2	71.5	82.9	71.9	65.2	62.7	67.7	63.8	31.3
Qwen3VL-32B-Thinking (Bai et al. 2025)	69.2	85.7	78.6	88.1	63.3	54.2	75.0	75.9	85.0	68.0	65.1	78.7	68.7	25.8
Step3-VL-10B (Huang et al. 2026)	77.2	84.1	70.9	80.9	77.5	37.8	73.9	75.7	79.7	57.7	62.3	76.1	66.1	22.8
MiMo-VL-7B-RL-2508 (Xiaomi and Team 2025)	75.2	83.0	74.9	84.3	72.7	45.0	73.9	75.8	81.4	63.5	62.3	74.8	64.7	34.0

Table 3: Performance of MLLMs on DiffCap-Bench. Comp., Back., Attr., Text, Cam., Expr., Spat., Img., Obj., and Act. denote Composition, Background, Attribute, Text & Symbol, Camera & Viewpoint, Expression, Spatial, Image Property, Object, and Action & Pose, respectively. Category columns report macro-average F_1^*, while other columns report sample-level average F_1^*.

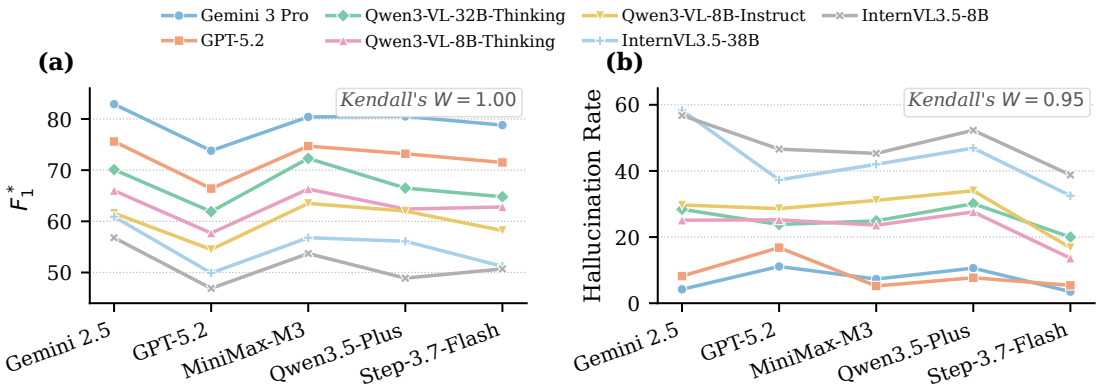


Figure 4: Cross-judge ranking consistency on DiffCap-Bench. We re-evaluate seven representative MLLMs under five independent judges and report (a) F_1^* and (b) Hallucination Rate. Absolute scores vary across judges, but relative rankings remain highly stable (Kendall's W=1.00 for F_1^* and W=0.95 for Hallucination Rate).

Existing models excel at perceiving background, text, and image property changes, yet struggle with action, pose, and expression discrepancies. All open-source models yield macro-average F_1^* below 70% on "action & pose" and "expression". Expression is the hardest: except for Gemini 3 Pro and Qwen3VL-Plus, even leading commercial models perform poorly—e.g., GPT 5.2 reaches only 30.0%, Seed1.6-VL variants stay below 60%, and other open-source models range from 30% to 66%. These results reveal a bottleneck in modeling non-rigid or nuanced visual discrepancies.

Effectiveness Validation

We then validate the effectiveness of DiffCap-Bench from three perspectives: cross-judge robustness, human alignment, and correlation with image-editing data production.

Cross-Judge Ranking Consistency

To test sensitivity to the choice of judge, we randomly sample 200 items and re-evaluate seven representative MLLMs under five independent judges. As shown in Figure 4, absolute F_1^* and Hallucination Rate vary across judges, but rankings remain highly stable (Kendall's W=1.00 for F_1^*; W=0.95 for Hallucination Rate), indicating that conclusions are anchored by the human-validated Difference List rather than a particular judge, and further demonstrating the robustness of DiffCap-Bench.

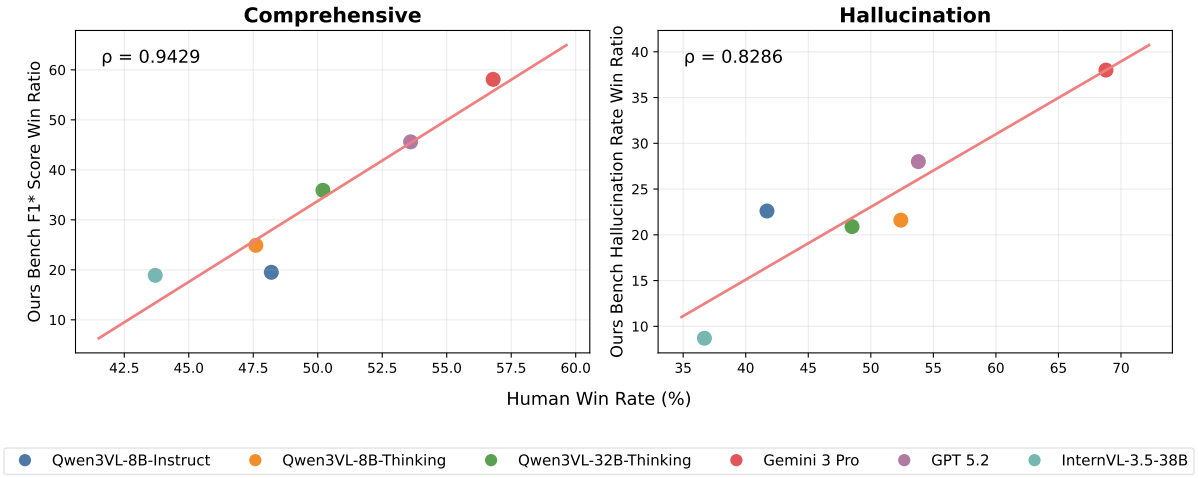


Figure 5: Correlation between DiffCap-Bench metrics and human expert judgments across six representative MLLMs. Per-model win rates are computed under both evaluation protocols, and rank consistency is measured via Spearman's ρ, yielding $\rho = 0.94$ for F_1^* and $\rho = 0.82$ for Hallucination Rate.

Base Model	Relabeling LVLM	ImgEdit Eval					DiffCap-Bench	
		IC	VS	PC	Overall	Rank	F1*	Rank
FLUX.1-Kontext-dev	InternVL3.5-8B	2.15	2.08	2.11	2.11	5	51.9	5
	InternVL3.5-38B	2.27	2.05	2.15	2.15	4	55.5	4
	Qwen3VL-8B-Instruct	2.32	2.23	2.08	2.21	2	61.7	3
	Qwen3VL-8B-Thinking	2.33	2.03	2.12	2.16	3	65.5	2
	Qwen3VL-32B-Thinking	3.12	2.95	2.87	2.98	1	68.7	1

Table 4: Downstream image editing performance of FLUX.1-Kontext-dev fine-tuned on datasets relabeled by different MLLMs. IC, VS, and PC denote Instruction Compliance, Visual Seamlessness, and Physical Consistency, which are ImgEdit scoring dimensions.

Alignment with Human Judgment

We sample 100 instances and evaluate six MLLMs (Gemini 3 Pro, GPT 5.2, Qwen3VL-32B-Thinking, Qwen3VL-8B-Instruct, Qwen3VL-8B-Thinking, and InternVL3.5-38B) under our automatic metrics and human assessment. Experts score holistic quality and hallucination severity (1–5), from which we compute per-model win rates. As shown in Figure 5, automatic metrics align with human judgments (Spearman's ρ=0.9429 for F_1^*; ρ=0.8286 for Hallucination Rate), confirming a reliable, human-aligned protocol for IDC. Win-rate details and scoring rubrics are provided in the Appendix.

Correlation with Data Production

IDC is a key upstream step for constructing image-editing data. To assess whether DiffCap-Bench predicts downstream utility, we use five MLLMs (InternVL3.5-8B/38B, Qwen3VL-8B-Instruct/Thinking, and Qwen3VL-32B-Thinking) to separately relabel 300K pairs from Pico-Banana-400k (Qian et al. 2025) and NanoBanana-150k (Ye et al. 2025a), yielding five training datasets with identical image pairs but different difference captions. The base model `FLUX.1-Kontext-dev` (Labs et al. 2025) is then fine-tuned separately on each relabeled set under the same training setting, and evaluated on 500 held-out pairs following the ImgEdit (Ye et al. 2025b) evaluation protocol. The training details are provided in the Appendix.

As shown in Table 4, DiffCap-Bench F_1^* rankings align closely with Overall editing scores (Spearman's ρ=0.90), and Instruction Compliance (IC) follows a notably consistent trend with F_1^*. In particular, Qwen3VL-32B-Thinking attains the best Overall (2.98) and F_1^* (68.7%). These results show DiffCap-Bench is highly correlated with downstream utility and serves as a proxy for MLLM utility in automated image-editing data engineering.

Conclusion

In this work, we present **DiffCap-Bench**, a comprehensive, challenging, and robust benchmark for Image Difference Captioning, covering ten difference categories across 1,075 compositional image pairs. To enable reliable evaluation, we introduce an *LLM-as-a-Judge* protocol grounded in human-validated Difference Lists, measuring Recall*, Precision*, F_1^*, and Hallucination Rate. Extensive experiments on 16 MLLMs reveal a substantial performance gap between proprietary and open-source models, consistent improvements from reasoning-enabled variants, and persistent challenges in recognizing action, pose, and expression changes. Moreover, stable cross-judge rankings, strong agreement with human

judgments, and high correlation with downstream image-editing performance demonstrate that DiffCap-Bench provides a reliable evaluation framework for IDC and serves as a practical proxy for MLLM utility in automated image-editing data engineering. Beyond leaderboard ranking, these results highlight that faithful difference understanding remains an open challenge for current MLLMs, especially under non-rigid and fine-grained visual changes. We hope DiffCap-Bench will foster more faithful evaluation of MLLMs and facilitate the development of stronger difference-aware data construction pipelines.

References

Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; Ge, W.; Guo, Z.; Huang, Q.; Huang, J.; Huang, F.; Hui, B.; Jiang, S.; Li, Z.; Li, M.; Li, M.; Li, K.; Lin, Z.; Lin, J.; Liu, X.; Liu, J.; Liu, C.; Liu, Y.; Liu, D.; Liu, S.; Lu, D.; Luo, R.; Lv, C.; Men, R.; Meng, L.; Ren, X.; Ren, X.; Song, S.; Sun, Y.; Tang, J.; Tu, J.; Wan, J.; Wang, P.; Wang, P.; Wang, Q.; Wang, Y.; Xie, T.; Xu, Y.; Xu, H.; Xu, J.; Yang, Z.; Yang, M.; Yang, J.; Yang, A.; Yu, B.; Zhang, F.; Zhang, H.; Zhang, X.; Zheng, B.; Zhong, H.; Zhou, J.; Zhou, F.; Zhou, J.; Zhu, Y.; and Zhu, K. 2025. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*.

Banerjee, S.; and Lavie, A. 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*, 65–72.

Caglayan, O.; Madhyastha, P. S.; and Specia, L. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2322–2328.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Cui, Y.; Yang, G.; Veit, A.; Huang, X.; and Belongie, S. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5804–5812.

Di, Z.; Shi, J.; Fan, Y.; Tan, H.; Black, A.; Collomosse, J.; and Liu, Y. 2025. DiffTell: A High-Quality Dataset for Describing Image Manipulation Changes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 24580–24590.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. 2024. A survey on llm-as-a-judge. *The Innovation*.

Guo, D.; Wu, F.; Zhu, F.; Leng, F.; Shi, G.; Chen, H.; Fan, H.; Wang, J.; Jiang, J.; Wang, J.; et al. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Hu, E.; Guo, L.; Yue, T.; Zhao, Z.; Xue, S.; and Liu, J. 2024. Onediff: A generalist model for image difference captioning. In *Proceedings of the Asian Conference on Computer Vision*, 2439–2455.

Huang, A.; Yao, C.; Han, C.; Wan, F.; Guo, H.; Lv, H.; Zhou, H.; Wang, J.; Zhou, J.; Sun, J.; et al. 2026. Step3-vl-10b technical report. *arXiv preprint arXiv:2601.09668*.

Hui, M.; Yang, S.; Zhao, B.; Shi, Y.; Wang, H.; Wang, P.; Zhou, Y.; and Xie, C. 2024. HQ-Edit: A High-Quality Dataset for Instruction-based Image Editing. *arXiv preprint arXiv:2404.09990*.

Jhamtani, H.; and Berg-Kirkpatrick, T. 2018. Learning to Describe Differences Between Pairs of Similar Images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kim, H.; Kim, J.; Lee, H.; Park, H.; and Kim, G. 2021. Agnostic change captioning with cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2095–2104.

Kim, M.; Lee, S.; and Park, D. 2026. VLM-SubtleBench: How Far Are VLMs from Human-Level Subtle Comparative Reasoning? *arXiv preprint arXiv:2603.07888*.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, Y.; Hou, S.; Hou, S.; Du, J.; Meng, S.; and Huang, Y. 2025. OmniDiff: A Comprehensive Benchmark for Fine-grained Image Difference Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21440–21449.

Lu, Y.; Yuan, J.; Li, Z.; Zhao, S.; Qin, Q.; Li, X.; Zhuo, L.; Wen, L.; Liu, D.; Cao, Y.; et al. 2025. Omnicaptioner: One captioner to rule them all. *arXiv preprint arXiv:2504.07089*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Park, D. H.; Darrell, T.; and Rohrbach, A. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4624–4633.

Qian, Y.; Bocek-Rivele, E.; Song, L.; Tong, J.; Yang, Y.; Lu, J.; Hu, W.; and Gan, Z. 2025. Pico-banana-400k: A large-scale dataset for text-guided image editing. *arXiv preprint arXiv:2510.19808*.

Singh, A.; Fry, A.; Perelman, A.; Tart, A.; Ganesh, A.; El-Kishky, A.; McLaughlin, A.; Low, A.; Ostrow, A.; Ananthram, A.; et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Tan, H.; Dernoncourt, F.; Lin, Z.; Bui, T.; and Bansal, M. 2019. Expressing visual relationships via language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1873–1883.

Team, K.; Bai, T.; Bai, Y.; Bao, Y.; Cai, S.; Cao, Y.; Charles, Y.; Che, H.; Chen, C.; Chen, G.; et al. 2026. Kimi K2. 5: Visual Agentic Intelligence. *arXiv preprint arXiv:2602.02276*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wang, W.; Gao, Z.; Gu, L.; Pu, H.; Cui, L.; Wei, X.; Liu, Z.; Jing, L.; Ye, S.; Shao, J.; et al. 2025. InternVL3. 5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*.

Xiaomi, L.; and Team, C. 2025. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 1(2): 5.

Xing, L.; Dong, X.; Zang, Y.; Cao, Y.; Liang, J.; Huang, Q.; Wang, J.; Wu, F.; and Lin, D. 2025. Caprl: Stimulating dense image caption capabilities via reinforcement learning. *arXiv preprint arXiv:2509.22647*.

Ye, J.; Jiang, D.; Wang, Z.; Zhu, L.; Hu, Z.; Huang, Z.; He, J.; Yan, Z.; Yu, J.; Li, H.; He, C.; and Li, W. 2025a. Echo-4o: Harnessing the Power of GPT-4o Synthetic Images for Improved Image Generation. *https://arxiv.org/abs/2508.09987*.

Ye, Y.; He, X.; Li, Z.; Lin, B.; Yuan, S.; Yan, Z.; Hou, B.; and Yuan, L. 2025b. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*.

Overview

- **Image Pair Quality Filtering Details** details the image pair quality filtering procedure, covering the prompt template for MLLM-based automatic filtering and the criteria used in human expert challengingness filtering.
- **Difference Dimension Classification** details the difference dimension classification procedure, including the classification prompt template.
- **Evaluation Details** presents the evaluation details on DiffCap-Bench, including the inference prompt template, inference parameters, judge prompt template, and a representative case study.
- **Benchmark Effectiveness** presents the evaluation details on traditional benchmarks along with qualitative examples, details the human alignment experiment (including the scoring rubrics for each evaluation criterion and the win rate calculation methodology), and the training details for the downstream image-editing correlation experiment.

Image Pair Quality Filtering Details

To ensure both data quality and benchmark challenge, we adopt a two-stage filtering procedure for all collected image pairs: automatic filtering using an MLLM to remove low-quality or invalid pairs, followed by human expert filtering to ensure sufficient challengingness.

MLLM-Based Automatic Filtering. As shown in Table 5, we design a unified prompting protocol for Qwen3VL-Plus (Bai et al. 2025) that evaluates each image pair across three criteria: *difference existence*, *visual similarity*, and *image quality*. A pair is retained only if all three criteria are satisfied simultaneously (*difference existence* = "Yes", *visual similarity* = "High", *image quality* = "Good"); any pair failing on even one criterion is discarded. Pairs passing this automatic stage are then forwarded to human annotators for a second-round review.

Human Expert Challengingness Filtering. Pairs that survive after filtering are further reviewed by human annotators to ensure sufficient challengingness. Annotators first enumerate all clearly observable differences. Let N denote the total number of valid differences identified. If $N > 2$, the pair is directly retained. If $N \leq 2$, annotators additionally evaluate the relative visual area of each difference (i.e., the proportion of image pixels affected). Pairs in which any single difference occupies more than 20% of the total image area are discarded, as such large and salient changes are perceptually easy to spot. This area-based criterion ensures that retained pairs contain only compact, localized differences that are inherently more challenging to detect.

Difference Dimension Classification

We employ Qwen3VL-Plus (Bai et al. 2025) to categorize each annotated difference into predefined dimensions using the prompt template in Table 6, allowing multi-label assignments since a single difference may involve multiple aspects. All dimension annotations are subsequently verified by human annotators to ensure label quality.

Evaluation Details

Inference Details. To evaluate the image difference captioning capability of MLLMs, we apply a unified prompt template (Table 7) across all models. Input images are constrained to a maximum resolution of 1024×1024 and a minimum resolution of 336×336. For all open-source models, we set the temperature to 0.1 and keep all other inference parameters at their default values. For all proprietary models, we use their default inference settings without any additional configuration.

Judge Details. For each sample in DiffCap-Bench, we employ Gemini 2.5 Pro (Comanici et al. 2025) as the judge model, providing it with the image pair, the model's predicted descriptions, and the human-validated difference list as input (Table 8). Based on the judging output, we calculate Recall*, Precision*, F_1^*, and the hallucination rate for each sample.

Case Study. We present a case study comparing Qwen3VL-8B-Instruct (Bai et al. 2025) and Qwen3VL-8B-Thinking (Bai et al. 2025) on the same sample, along with their respective judge outputs. As shown in Fig. 6, Qwen3VL-8B-Instruct (Bai et al. 2025) directly produces a list of differences without any chain-of-thought reasoning, resulting in one missed difference and three hallucinations. In contrast, Qwen3VL-8B-Thinking (Bai et al. 2025) follows a think-before-output CoT paradigm, carefully comparing the two images through iterative reflection and self-checking, ultimately identifying all differences correctly without introducing any hallucinations. This case illustrates that reasoning capability and the CoT paradigm can effectively enhance a model's ability to detect and describe visual differences, while also reducing hallucinations.

Benchmark Effectiveness

Evaluation on Traditional Benchmarks

As shown in Fig. 7, we apply the same prompt template to evaluate MLLMs on two traditional image difference captioning benchmarks, Spot-the-Diff (Jhamtani and Berg-Kirkpatrick 2018) and Image-Edit-Request (Tan et al. 2019). Both GPT 5.2 (Singh et al. 2025) and Gemini 3 Pro (Comanici et al. 2025) correctly identify that the key difference between the two images is the removal of a person in the background, whereas InternVL3.5-8B (Wang et al. 2025) fails to detect this change and instead introduces a hallucination, describing the image as "slightly sharper". Nevertheless, despite the superior descriptive quality of the stronger models, all three models score poorly on lexical-overlap metrics due to linguistic discrepancies with the ground-truth reference sentences, exhibiting low separability across models. This demonstrates that traditional benchmark evaluation protocols are unable to provide robust zero-shot assessment of current MLLMs.

Alignment with Human Judgment

To validate the reliability of our automatic evaluation metrics, we conduct a human alignment study. For each image pair and its corresponding model-generated caption, human

experts independently rate two aspects on a 1–5 scale: *Holistic Quality* and *Hallucination Severity*.

Holistic Quality (1–5) measures the model's ability to accurately, comprehensively, and concisely describe all real differences between two images. A score of 5 (Excellent) indicates that all differences are observed and described correctly, with no errors. A score of 4 (Good) means that most differences are captured, allowing for at most one minor missed difference and at most one inaccurate description. A score of 3 (Fair) reflects that most differences are observed and described, but allows for up to two minor missed differences and up to two inaccurate descriptions. A score of 2 (Poor) indicates that only a few differences are captured or there are three or more significant errors in the descriptions. A score of 1 (Very Poor) means that no differences are captured or all descriptions are incorrect.

Hallucination Severity (1–5) assesses the extent to which the caption introduces content not present in the images, such as false objects, attributes, actions, or spatial arrangements. A score of 5 (None) indicates no hallucinated content. A score of 4 (Minimal) allows for a single hallucinated element that does not affect overall understanding. A score of 3 (Moderate) corresponds to exactly two hallucinated elements that may slightly mislead. A score of 2 (Severe) reflects three to four hallucinated elements that significantly misrepresent the differences. A score of 1 (Extreme) means five or more hallucinated elements, rendering the caption largely unrelated to the actual images.

Evaluation Procedure. We randomly sample 100 instances and collect outputs from six models: Gemini 3 Pro (Comanici et al. 2025), GPT 5.2 (Singh et al. 2025), InternVL3.5-38B (Wang et al. 2025), Qwen3VL-32B-Thinking (Bai et al. 2025), Qwen3VL-8B-Thinking (Bai et al. 2025), and Qwen3VL-8B-Instruct (Bai et al. 2025). Each output is independently scored by six human annotators according to the criteria above.

Win Rate Calculation. For each sample s and criterion c (Holistic Quality or Hallucination Severity), the win rate of model m is computed as:

$$\text{WinRate}_{m,c} = \frac{\sum_{s=1}^{S} \sum_{\substack{m'=1 \\ m' \neq m}}^{M} \mathbf{1}\big(f_{m,c}(s) > f_{m',c}(s)\big)}{S \cdot (M-1)}$$

where S is the total number of samples, M is the total number of models, $f_{m,c}(s)$ is the score of model m on sample s under criterion c, and $\mathbf{1}(\cdot)$ is the indicator function that equals 1 if the condition holds and 0 otherwise. This formulation aggregates all pairwise comparisons across samples, providing a fine-grained measurement of relative model performance across the benchmark.

Training Details

We randomly sample 300K image pairs from Pico-Banana-400k (Qian et al. 2025) and NanoBanana-150k (Ye et al. 2025a), two large-scale image editing datasets, as the training set, and 500 pairs as a held-out test set. Five MLLMs—InternVL3.5-8B (Wang et al. 2025), InternVL3.5-38B (Wang et al. 2025), Qwen3VL-8B-Instruct (Bai et al. 2025), Qwen3VL-8B-Thinking (Bai et al. 2025), and Qwen3VL-32B-Thinking (Bai et al. 2025)—are used to relabel the training data, producing five distinct datasets with identical image pairs but different editing instructions. The base model `FLUX.1-Kontext-dev` (Labs et al. 2025) is then fine-tuned separately on each relabeled dataset using supervised fine-tuning (SFT). To ensure a controlled comparison, we maintain a unified training setting: one epoch, a learning rate of 1×10^{-5}, a total batch size of 32, freezing all other parameters and optimizing only the Diffusion Transformer (DiT) module. The resulting models are evaluated separately on the 500-pair test set following the ImgEdit (Ye et al. 2025b) evaluation protocol.

You are an expert visual analyst. Your task is to evaluate whether a pair of images should be retained for a visual difference benchmark. Base your judgment strictly on visible evidence; do not speculate beyond what can be observed.

Evaluate the following criteria:

1. Difference Existence: determine whether there are real, observable differences between the two images. - Output "Yes" if there are clear, meaningful visual differences (e.g., object addition/removal, attribute change, spatial change). - Output "No" if the images are identical or only differ in negligible ways (e.g., compression noise, minor color fluctuation, tiny pixel-level variations).
2. Visual Similarity: determine whether the two images are semantically related. - Output "High" if they depict the same scene or highly similar content with localized modifications. - Output "Low" if they depict different scenes or are completely unrelated (e.g., different objects, environments, or contexts).
3. Image Quality: evaluate whether both images are of sufficient quality for reliable visual comparison. - Output "Good" if both images satisfy all of the following: (1) sufficient resolution (no extreme downsampling or tiny size), (2) clear and sharp content (no heavy blur or defocus), (3) objects are recognizable and not obscured, (4) no severe artifacts (e.g., compression blocks, corruption, overexposure). - Output "Poor" if either image violates any of the above conditions (e.g., blurry, low-resolution, heavily noisy, occluded, or visually ambiguous).

Guidelines:

1. Focus only on visible content; do not assume hidden or implied information.
2. Ignore insignificant pixel-level differences when judging Difference Existence.
3. Consider images similar only if they share the same scene or core semantic content.
4. Treat completely unrelated image pairs as "Low" in Visual Similarity regardless of quality.
5. Reject the pair if either image is difficult to interpret due to poor clarity or artifacts.
6. Provide a concise explanation to justify your decision.

Output Format:

Return your answer strictly in JSON format:
{ "difference_existence": "Yes/No",
"visual_similarity": "High/Low",
"image_quality": "Good/Poor",
"reason": "brief explanation" }

Table 5: Prompt template for MLLM-based quality filtering of image pairs.

You are an expert visual analyst. Your task is to classify a single described visual difference between two images into one or more of the following ten categories, allowing multi-label assignments when a difference involves multiple aspects. Base your judgment strictly on the provided description.

Difference Categories:

1. **Object**: Additions, deletions, replacements, or changes in the number or type of objects and entities (e.g., a phone is removed; a person is added).
2. **Attribute**: Changes in a visual property of an object or region, including color, size, shape, texture, material, clothing, or hair (e.g., the bottle becomes white; the cap is shorter).
3. **Action & Pose**: Changes in a subject's body motion or static posture (e.g., the person raises their arm; the subject shifts from sitting to standing).
4. **Spatial**: Changes in the position, orientation, or relative arrangement of objects within the scene (e.g., the object moves left; two items swap positions).
5. **Composition**: Changes in overall framing, subject scale, or cropping that alter the spatial layout of the scene (e.g., the subject is enlarged; the image is cropped more tightly).
6. **Text & Symbol**: Additions, deletions, or modifications to any text, logo, or symbolic element (e.g., the title text is changed; a logo is removed).
7. **Background**: Changes to background content or the scene environment (e.g., the background becomes a solid color; the setting changes from indoors to outdoors).
8. **Camera/View**: Changes in camera angle, shooting perspective, or lens parameters (e.g., the viewpoint shifts to overhead; the lens zooms in).
9. **Image Property**: Global image-level changes not tied to any specific object, including resolution, aspect ratio, brightness, or saturation (e.g., overall saturation increases; the aspect ratio changes).
10. **Expression**: Changes in a subject's facial expression or emotional state (e.g., the person smiles; the expression becomes serious).

Input: A single difference description: "`<difference description>`"

Output Format: Return your answer strictly in JSON format: {"types": ["<category name>", ...], "reason": "<brief justification>"}

Table 6: Prompt template for difference dimension classification.

You are a visual change description model. Your task is to describe the observable changes required to transform an original image into a target image. Base your description only on visible evidence; do not invent objects, actions, or changes. You will also be provided with the resolution of each image. If the resolution has changed, you must mention it in your description.

Include meaningful changes such as:

1. Objects: added, removed, repositioned, resized, recolored, or otherwise altered.
2. Spatial and composition: layout, alignment, framing, cropping, or perspective changes.
3. Camera and viewpoint: angle, zoom, or focal length changes.
4. Image properties: resolution, sharpness, blur, noise, aspect ratio.
5. Color, lighting, and style: tone, contrast, shadows, highlights, or overall style.
6. Background and environment: visible changes in scene context or background.

Guidelines:

- Describe only changes, not what stays the same.
- Use clear, factual language in an editing-instruction style.
- Mention resolution changes if they occur.
- Do NOT refer to the images as "Image A" or "Image B".
- Enumerate multiple changes if necessary.
- Do not assume hidden content or speculate on intent.
- If a change is not clearly observable, do not mention it.

Input:

Original Image Resolution
{}

Target Image Resolution
{}

Table 7: Prompt template for evaluating image difference captioning in DiffCap-Bench.

You are a **strict, rational evaluator whose only source of truth is the visual evidence in the images**. Your task is **NOT** to generate descriptions, but to evaluate how accurately the model-predicted change descriptions reflect the real visual changes from the original image to the target image.

Task Background:
The visual editing change description task is defined as follows: Given an original image and a target image, a model must describe, in natural language, the visual editing operations required to transform the original into the target. You will receive two types of input:

1. Model-predicted change descriptions (Prediction)
2. Expert-annotated reference answers (Keypoint List)

Your goal is to **systematically verify the prediction under the constraint of the reference answers and produce structured outputs for automatic evaluation**.

Input Description:
Model-Predicted Change Descriptions: May be completely correct; May mention real changes with incorrect details; May include additional valid changes; May miss some real changes; May contain hallucinations (non-existent or unchanged content)
Keypoint List (Expert Annotation): (1) Key Change List: certain and important changes; one item may include multiple equivalent expressions separated by "/", all referring to the same underlying change. (2) Indistinguishable Items: changes annotators cannot fully confirm or are subjective; hitting these changes is **not** considered hallucination.
Your Evaluation Task (Very Important):
Evaluation consists of **Forward Checking** and **Backward Checking**.
1. Forward Checking (Key-change-centered):
Iterate through the Key Change List and evaluate each key change item. Assign one of the following hit statuses: (1) Hit and Correct: prediction mentions the corresponding change accurately (object, attribute, direction, quantity, spatial relation, etc.) (2) Hit but Incorrect: refers to the same change category but contains incorrect attributes (e.g., wrong color, spatial direction, object, quantity) (3) Not Hit: prediction does not mention the change or describes unrelated changes *Note: exact wording is not required; semantic equivalence is sufficient.*
2. Backward Checking (Prediction-centered):
Examine predicted descriptions that do **not** hit any key change item. Perform semantic decomposition: (1) Treat each independent change as a minimal evaluation unit (2) Split multiple changes in one description into separate units
3. Extra Description Status Determination:
For each extra change, assign: (1) Matches Indistinguishable Items: corresponds to an indistinguishable item (2) Extra Correct Description: truly exists from original to target but not recorded in keypoints (3) Hallucination: describes non-existent or unchanged changes, or clearly contradicts visual facts
Important: "Not in keypoint list" does not imply hallucination. Resolution-related unchanged descriptions are also not hallucination unless indicated.

Output Requirements (Strict):
Output **ONLY the final JSON result**. No reasoning or extra text.
The JSON structure must be:

```
{
  "Key_Change_List": [
    {"Description": "...", "Type": "...", "Hit_Status": "...", "Reason": "..."}
  ],
  "Extra_Description_List": [
    {"Description": "...", "Status": "...", "Reason": "..."}
  ],
  "Hallucination": {"Count": ..., "Reason": "..."}
}
```

Model-Predicted Difference Descriptions: <prediction>

Keypoint List: <difference list>

Table 8: Prompt template used by the judge model to assess model-predicted difference descriptions.



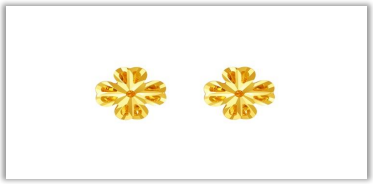


Image Pair

Key Differences List:
1. Adjust the earring to a front-facing view / change the perspective to straight . (object, spatial, camera)
2. Align the earrings horizontally / replace with a horizontally symmetrical arrangement. (spatial)
3. Hide the earring post. (object)

Indistinguishable Items: None

Qwen3VL-8B-Instruct

1. Reposition the left earring so that it is aligned horizontally with the right earring, both facing forward.
2. Rotate the left earring to show its front view, matching the orientation of the right earring.
3. Adjust the lighting to enhance the highlights and reflections on the surface of both earrings.
4. Increase the overall sharpness and clarity of the image.
5. Remove the visible shadow beneath the left earring.

Judging Result for Qwen3VL-8B-Instruct

1. Adjust the earring to a front-facing view.
- Status: Correct
- Reason: 'Reposition the left earring so that it is aligned horizontally with the right earring, both facing forward.
2. Align the earrings horizontally.
- Status: Correct
- Reason: 'Reposition the left earring so that it is aligned horizontally with the right earring, both facing forward.
3. Hide the earring post.
- Status: Missed

Hallucination: 1. Adjust the lighting… 2. Increase the overall sharpness ..
3. Remove the visible shadow…

Qwen3VL-8B-Thinking

So, let's compare the two images. First, check the objects. In Image A, there are two earrings: one is a clover-shaped earring with a post (the back part), and the other is a clover-shaped earring with a post (the front part).Wait, no, looking at Image A: left side is an earring with the post (the back part, so the clover is the decorative part, and the post is the stem), and the right side is the front view of the clover earring (so the post is attached to the clover)…**Wait, maybe I'm overcomplicating. Let's list the changes clearly**:1. Remove the posts (stems) from both earrings, resulting in the posts no longer being visible. 2. Reposition the left earring to display the front view (previously the back view), aligning it with the right earring's front view orientation. 3. Adjust the spatial arrangement so that both earrings are positioned side by side with identical front-facing orientations. **Check if there are other changes**. The background is white in both, so no background change. The resolution is the same (19:13), so no resolution change. **The color and lighting are consistent, so no color or lighting changes**. So the final answer should list these changes.
<answer>
1. Remove the posts (stems) from both earrings, making the posts no longer visible.
2. Reposition the left earring to display the front view (previously the back view), aligning it with the right earring's front view orientation.
3. Adjust the spatial arrangement so that both earrings are positioned side by side with identical front-facing orientations.
</answer>

Judging Result for Qwen3VL-8B-Thinking

1. Adjust the earring to a front-facing view.
- Status: Correct
- Reason: 'Reposition the left earring to display the front view…'
2. Align the earrings horizontally.
- Status: Correct
- Reason: 'Adjust the spatial arrangement so that both earrings are …'
3. Hide the earring post.
- Status: Correct
- Reason: 'Remove the posts (stems) from both earrings....'

Hallucination: None

Figure 6: Qualitative case study comparing Qwen3VL-8B-Instruct and Qwen3VL-8B-Thinking on the same sample, evaluated using the judge model on DiffCap-Bench. Qwen3VL-8B-Instruct outputs a list of differences directly, missing one true difference and producing three hallucinations. In contrast, Qwen3VL-8B-Thinking employs a think-before-output chain-of-thought (CoT) reasoning process, iteratively reflecting and self-checking to accurately identify all differences without hallucinations. This illustrates how CoT reasoning can improve the model's detection and description of visual differences.

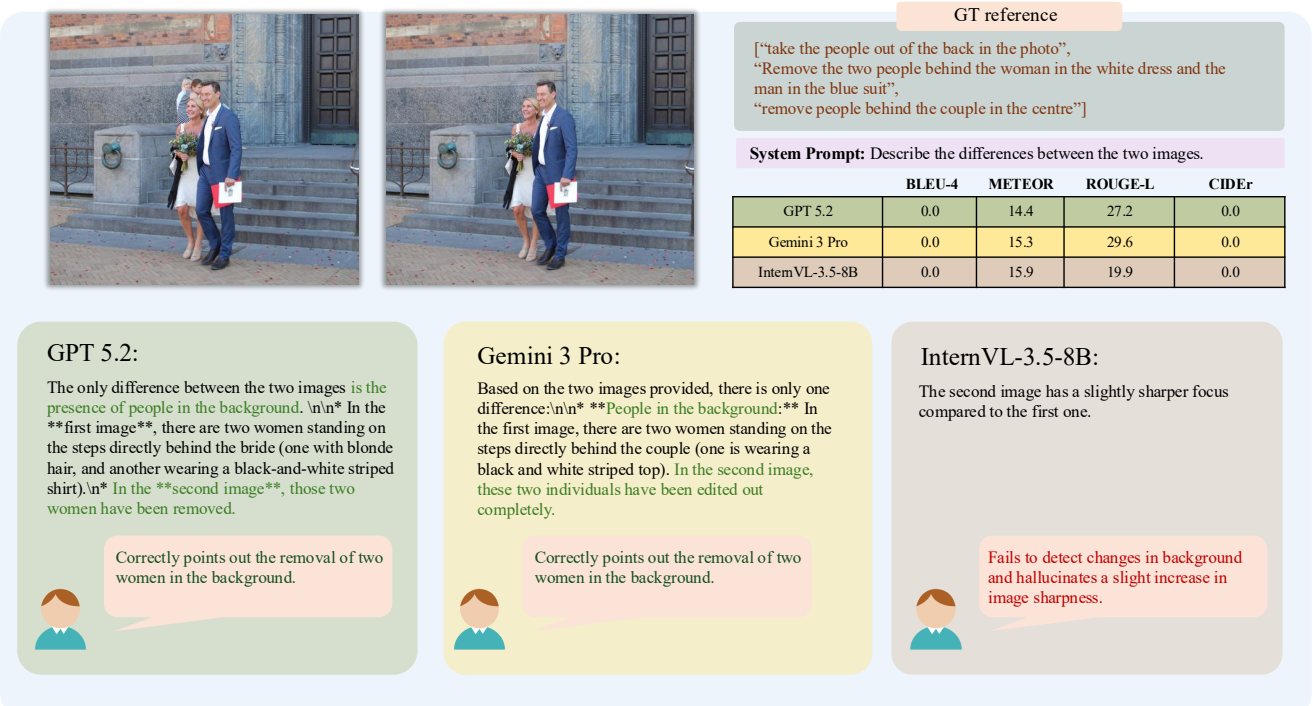


Figure 7: Qualitative evaluation of MLLMs on the Image-Edit-Request benchmark using the same prompt template. GPT 5.2 and Gemini 3 Pro correctly identify the key difference—a person removed from the background—while InternVL3.5-8B fails to detect this change and introduces a hallucination by describing the image as "slightly sharper". Despite the stronger models producing more accurate and descriptive outputs, all models score poorly on traditional lexical-overlap metrics due to linguistic discrepancies with the reference captions, highlighting the limitations of conventional benchmarks for robust zero-shot assessment of MLLMs.