OmniDiff: A Comprehensive Benchmark for Fine-grained Image Difference Captioning

Supplementary Material

In this supplementary material, we provide additional details regarding the **OmniDiff** dataset and the training specifics of the **M^{3}Diff** model. Finally, we further experimentally validate the effectiveness of the OmniDiff dataset and the MDP module, and analyze the qualitative performance of four models on samples from OmniDiff.

7. Dataset Statistics

7.1. Word Distribution

We analyze the word distribution in the difference captions of the OmniDiff dataset, as illustrated in Figure 4. The words "left", "right" and "side" emerge as the most frequently occurring terms. This pattern arises because our annotations consist of two components: the *referring part* and the *change part*. In the *referring part*, these high-frequency words typically describe the spatial location of the changing objects, reflecting the dataset's emphasis on precise positional references.

Figure 4. Wordcloud visualization of the OmniDiff dataset.

7.2. Caption Length Distribution

Figure 5 illustrates the distribution of caption lengths in OmniDiff compared to the IEdit [37], Spot-the-Diff [17], and Birds-to-Words [10] datasets. The average caption length in OmniDiff significantly exceeds that of the other three datasets, demonstrating that OmniDiff provides fine-grained difference annotations for complex and dynamic scenarios. This establishes a new benchmark for fine-grained Image Difference Captioning (IDC) tasks.

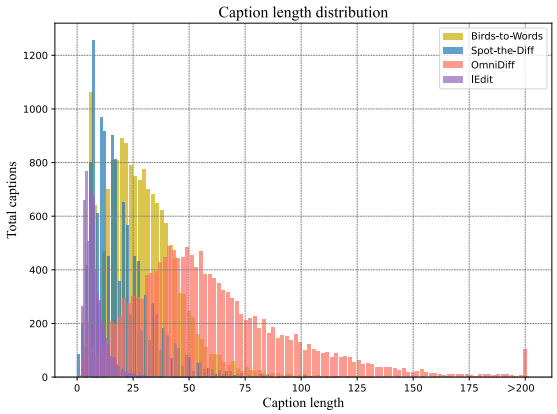


Figure 5. Caption length distribution of the OmniDiff dataset.

8. Training Details

8.1. Training Data

As shown in Table 9, we extend the OmniDiff Dataset by collecting and curating five publicly available IDC datasets to construct a comprehensive instruction-tuning dataset, comprising 145k image pairs and 896k difference captions. The instruction-tuning dataset encompasses not only real-world datasets such as OmniDiff-Real, Spot-the-Diff [17], IEdit [37] and Birds-to-Words [10] but also includes synthetic 3D datasets like OmniDiff-Render, CLEVR-Change [33] and CLEVR-DC [20].

8.2. Hyperparameters

The hyperparameter settings for the fine-tuning process are detailed in Table 10. The global batch size is set to 256, which allows for efficient training while maintaining a balance between computational resources and model performance. The learning rate for the Vision Transformer [50] is configured at 2e-6, while the learning rates for other components of the model (base learning rate) are set to 1e-5. The number of epochs is limited to 1, indicating a single pass through the entire dataset for this specific finetuning task. The optimizer chosen for this process is AdamW [30], known for its effectiveness in handling sparse gradients and regularizing the weight decay. Additionally, the LoRA-rank [15] is set to 128, and the LoRA-alpha [15] is set to 256, parameters that are crucial for low-rank adaptation techniques, enabling efficient and effective model

Table 9. Overview of the training data.

Dataset	Image Pairs	Captions
IEdit [37]	3k	4k
Birds-to-Words [10]	3k	14k
Spot-the-Diff [17]	11k	21k
OmniDiff	14k	28k
CLEVR-DC [20]	43k	385k
CLEVR-Change [33]	71k	444k
All	145k	896k

Table 10. Hyperparameter settings for finetuning.

Hyperparameter	Value
Batch size	256
ViT learning rate	2×10^{-6}
Base learning rate	1×10^{-5}
Epochs	1
Optimizer	AdamW
LoRA rank	128
LoRA alpha	256

customization with minimal additional parameters. These settings collectively contribute to an optimized finetuning strategy tailored to our specific application.

9. Experiments

9.1. Performance Comparison with Identical Data

For a fair comparison with other MLLM-based IDC methods (*e.g.*, FINER-MLLM [51]), we reproduce this model using the same 896k training data as M^{3}Diff. As shown in Table 11, M^{3}Diff demonstrates consistent superiority across all metrics on the OmniDiff and CLEVR-DC benchmarks.

9.2. Extended Analysis of the MDP Module

To further validate the effectiveness of the plug-and-play MDP module, we conduct experiments using two advanced MLLMs, Qwen-2.5-VL-7B-Instruct [5] and InternVL3-8B-Instruct [53], as our base models. Table 12 demonstrates that integrating the MDP module into various backbones consistently enhances the model performance. Additionally, M^{3}Diff with LLaVA-OneVision-7B [22] as the backbone maintains strong performance, potentially due to the increased use of multi-image task data in pre-training.

9.3. Extended Analysis of the OmniDiff Dataset

With exceptional instruction-following and semantic understanding capabilities, LLMs serve as a crucial tool for measuring sentence similarity [11]. Therefore, to comprehensively validate the impact of the OmniDiff dataset on model performance, we design prompts to instruct GPT-4o [1] to evaluate semantic consistency between predictions and annotations, outputting a score ranging from 0 to 100. The prompt is shown in Figure 6. Table 13 shows that the LLM-based evaluation confirms the effectiveness of OmniDiff.

Prompt for LLM-based Evaluation

You are an impartial judge evaluating text similarity.
Your job is to evaluate a model-generated caption for the difference between two images against a human-annotated Ground Truth caption. Follow these guidelines:
(1) Carefully read the Ground Truth caption and model-generated caption.
(2) Identify aspects in the Ground Truth caption and calculate the percentage covered (through exact or partial matches) in the model-generated caption.
(3) Score from 0 to 100, where each aspect contributes equally to the score.
(4) Provide your score (0-100) and a short justification.
Standard answer: {standard_answer}
Assistant's response: {assistant_response}

Figure 6. The prompt for LLM-based evaluation.

10. Case Study

In this section, we aim to compare the difference captions provided by different models (GPT-4o [1], CARD [45], FINER-MLLM [51], M^{3}Diff) for image pairs in our dataset. The focus is on how each model captures the presence and behavior, as well as other changes in the scene.

1) In Figure (a), GPT-4o [1] and FINER-MLLM [51] both fail to capture the presence of a second bird in the background. CARD [45] incorrectly analyzes scene changes, such as variations in lighting and object movement, while failing to accurately identify the presence of two newly introduced birds in the scene. In contrast, our method, M^{3}Diff, excels by providing a comprehensive and accurate description that includes both the primary bird's interaction with the peanuts and the secondary bird in the background, along with all relevant environmental details. This highlights the superior accuracy and thoroughness of M^{3}Diff in analyzing complex scene changes.

2) In Figure (b), GPT-4o [1] correctly notes the absence of a person with a bag and the addition of "Villa" on the ground but lacks detail. CARD [45] hallucinates the presence of non-existent individuals in the scene and fails to recognize the correct text. FINER-MLLM [51] captures some correct details but includes incorrect observations like a missing wheelchair. In contrast, our method, M^{3}Diff, accurately describes the disappearance of a man in specific

Table 11. Performance comparison with FINER-MLLM [51] under identical training data.

Method	OmniDiff-Real				OmniDiff-Render				CLEVR-DC				
	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
FINER-MLLM	9.4	14.5	26.1	20.6	13.8	15.8	30.5	18.0	53.1	34.2	69.5	92.3	17.8
M^{3}Diff (ours)	**14.3**	**18.9**	**32.9**	**31.3**	**15.7**	**19.9**	**35.3**	**28.3**	**60.6**	**37.6**	**73.0**	**109.4**	**21.3**

Table 12. Ablation study of the MDP module based on Qwen2.5-VL-7B-Instruct [5] and InternVL3-8B-Instruct [53]. † indicates the model is evaluated in a zero-shot setting.

Method	OmniDiff-Real				OmniDiff-Render				Image-Edit-Request				
	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Qwen2.5-VL-7B†	3.8	9.5	19.8	6.2	2.1	6.8	18.3	3.3	16.0	9.8	23.2	32.3	10.8
Qwen2.5-VL-7B-SFT w/o MDP	13.1	16.7	31.8	35.3	14.9	18.7	34.7	25.2	30.0	26.5	58.3	131.2	26.6
Qwen2.5-VL-7B-SFT with MDP	13.8	18.8	**32.9**	**36.7**	15.5	19.7	35.6	27.8	31.2	**26.8**	59.1	132.5	**27.7**
InternVL3-8B†	2.7	12.2	19.2	4.3	4.0	9.2	19.3	5.1	12.1	9.2	17.1	28.6	9.3
InternVL3-8B-SFT w/o MDP	12.5	17.0	31.3	34.6	14.7	19.1	35.1	26.1	29.4	25.7	57.1	128.1	26.3
InternVL3-8B-SFT with MDP	13.5	18.5	32.5	35.7	15.5	**20.1**	**35.7**	28.1	30.8	26.1	58.2	130.5	27.1
M^{3}Diff (ours)	**14.3**	**18.9**	**32.9**	31.3	**15.7**	19.9	35.3	**28.3**	**33.6**	26.5	**59.7**	**136.6**	27.5

Table 13. Ablation study of OmniDiff based on LLM evaluation.

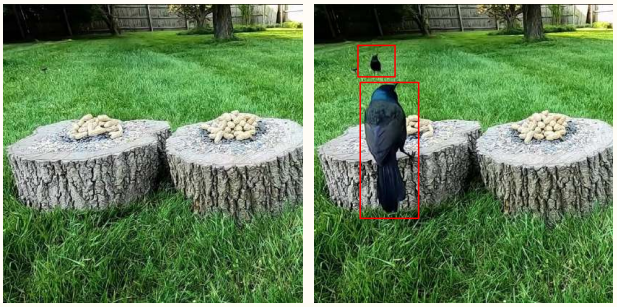
Method	OmniDiff-Real	OmniDiff-Render	Image-Edit-Request
	LLM	LLM	LLM
M^{3}Diff w/o OmniDiff	12.3	15.8	63.8
M^{3}Diff with OmniDiff	**37.3**	**37.0**	**68.5**

clothing and the precise location of the added text "villa", aligning closely with the ground truth.

3) In Figure (c), GPT-4o [1] correctly identifies the addition of "Cautiuos" above the bike but misses the lighting change. CARD [45] notes the brighter ambiance but focuses on irrelevant elements like a treadmill sign. FINER-MLLM [51] captures the brightness and text addition but misspells "cuitieus." Our method, M^{3}Diff, accurately describes the increased light and the placement of the sign but fails to perform accurate OCR, misinterpreting the word as "Caution" instead of "Cautiuos." This demonstrates that while M^{3}Diff excels in identifying environmental changes, it struggles with precise text recognition, indicating a need for improvement in its fine-grained OCR capabilities.

4) In Figure (d), GPT-4o [1] correctly notes the addition of a toy vacuum and a green toy with holes but lacks detail. CARD [45] focuses on irrelevant changes like lighting and misses the new toys. FINER-MLLM [51] incorrectly states that a blue toy car has disappeared. In contrast, our method, M^{3}Diff, accurately describes the appearance of a new toy train on the left and a toy watering can in the center, aligning closely with the ground truth.

The study highlights M^{3}Diff outperforms GPT-4o [1], CARD [45], and FINER-MLLM [51] in dynamic scene analysis, excelling at analyzing object interactions (*e.g.*, flocking birds, toy manipulation) and environmental changes. While M^{3}Diff demonstrates advanced spatial reasoning and scene dynamics comprehension, it occasionally encounters text recognition errors (e.g., misinterpreting "Cautiuos" as "Caution"), underscoring the need for improved OCR integration. In contrast, GPT-4o [1] struggles with fine visual details, CARD [45] generates hallucinations in complex scenes, and FINER-MLLM [51] mislabels objects despite robust reasoning. Overall, M^{3}Diff emerges as the most effective model for fine-grained IDC tasks.



GT: **A black bird** has appeared on the front edge of the left tree stump, facing away and looking into the distance. **Another black bird** has appeared on the grass at the far left of the image, with its head raised, looking toward the front right of the image.

GPT-4o: The contrast between the two images is the presence of **a bird**. In the first image, there are two tree stumps with piles of peanuts on them. In the second image, a bird has landed on the left stump, interacting with the peanuts. This adds an element of life and activity to the scene.

CARD: ~~The lighting has improved.~~ The small black twig, previously placed horizontally, has now moved to the upper right corner of the image. ~~A new large brownish - shaped sculpture~~ has appeared in the center of the scene.

FINER-MLLM: **The black bird** that appeared in the center of the image has shifted to the left side of the stump on the right side of the image, a new peanut has appeared near the inner edge of the stump on the left side of the image , a new peanut has appeared.

M³Diff (ours): **A black bird has appeared on the left side of the image**, standing on the ground with its back to the camera and facing the tree stump on the left. The number of peanuts on the tree stump in the lower right corner has decreased. **A black bird has appeared in the background,** positioned behind the tree stump on the left side of the image, with its body facing away from the camera.

(a)

GT: **A person** wearing a gray coat, blue pants, and a red schoolbag on one shoulder, previously present on the left side of the road, **is no longer visible**. A blue sign has appeared on the ground in front of the left side of the central road, **with the text \"Villa\" written on it**.

GPT-4o: In the second image, compared to the first, the modifications observed include: 1. **The person walking with a bag in the street is absent.** 2. **The word \"Villa\" is added on the ground** near the left sidewalk. Other elements in the environment appear unchanged between the two images.

CARD: **The man wearing a red vest, blue pants, and black shoes** who was previously in the middle of the screen, **is no longer visible**. The word ~~\" Shed \"~~ has appeared ~~on the wall behind the woman ' s left hand. The man wearing a white vest and black pants, who was previously facing the camera with his back to the camera , is now facing the camera with his right hand resting on his chest The man in front of the house , who was previously facing the camera with his left hand on his chest , is now facing the right side of the picture , is now facing the camera.~~

FINER-MLLM: ***A*** **man wearing a red vest , gray pants , and black shoes** , who was facing left with his back to the right side of the image and facing the road with his head turned to the right front of the camera **, is no longer visible** ~~the bicycle in the left front of the road in front of the middle house has turned to face the right side of the picture the black wheelchair in the left front of the middle of the road is no longer present~~ the blue sign with the word ~~will e~~ across the road at the center of the scene has increased in size

M³Diff (ours): **A man wearing a hat, a brown coat, and blue pants**, who was standing on the left side of the road, facing the front of the house with his back to the camera and holding a red backpack, **is no longer visible. The word \"villa\" in large blue letters** has appeared on the concrete road near the lower-left corner of the scene, positioned between two sidewalks.

(b)

GT: **The environment has become brighter.** A blue text sign that reads **\"Cautiuos\" has appeared** on the upper right wall of the scene, near the yellow bike and just above a small light fixture.

GPT-4o: **The second image has the word \"Cautiuos\"** (likely meant to be \"Cautious\") added to the wall above the bike. Everything else appears to be the same.

CARD: **The ambiance has brightened** The ~~word \" jordiniene \"~~ has appeared on the wall behind the treadmill The white plastic box on the left side of the scene has changed from yellow to black

FINER-MLLM: **The environment has become brighter** a new sign with ~~the word cuitieus~~ has appeared above the motorcycle on the left side of the scene , above the chair in the staircase

M³Diff (ours): **The light in the scene has increased.** A new sign with ~~the word \"Caution\"~~ appears on the wall to the right of the electric bicycle, positioned above a blue box and next to the graffiti.

(c)

GT: **A children's push toy car**, featuring a yellow lid and teal wheels, has appeared in the front left area. To the right of the push toy car, **a dark teal cone-shaped toy has been newly placed**.

GPT-4o: The two images display a similar setting with children's toys arranged on a wooden floor in front of a cabinet. ### First Image:- Features four items: a green baby seat, a blue rocking toy, a colorful activity walker, and a tricycle.### Second Image:- Displays six items: the same four toys as the first image, with **the addition of a toy vacuum** on the left and **a green toy with holes in the center**. The contrast mainly lies in the additional two toys present in the second image, which increases the overall number of items on the floor.

CARD: ~~The ambient lighting has brightened. The gray - furred dog bottle on the left side of the scene has moved to the right side of the table The number of fans with cartoon dinosaur patterns on the left side of the scene has decreased from one to two.~~

FINER-MLLM: ~~The blue toy car in front of the television on the right side of the scene has disappeared.~~

M³Diff (ours): **A new toy train has appeared on the left side of the scene**, positioned in front of the green chair and to the left of the blue toy car. **A new toy watering can has appeared in the center of the scene**, positioned between the green chair and the blue toy car.

(d)