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

Yuan Liu[1] Saihui Hou[1✉] Saijie Hou[2] Jiabao Du[1]
Shibei Meng[1] Yongzhen Huang[1,3]
[1]School of Artificial Intelligence, Beijing Normal University
[2]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[3]WATRIX.AI

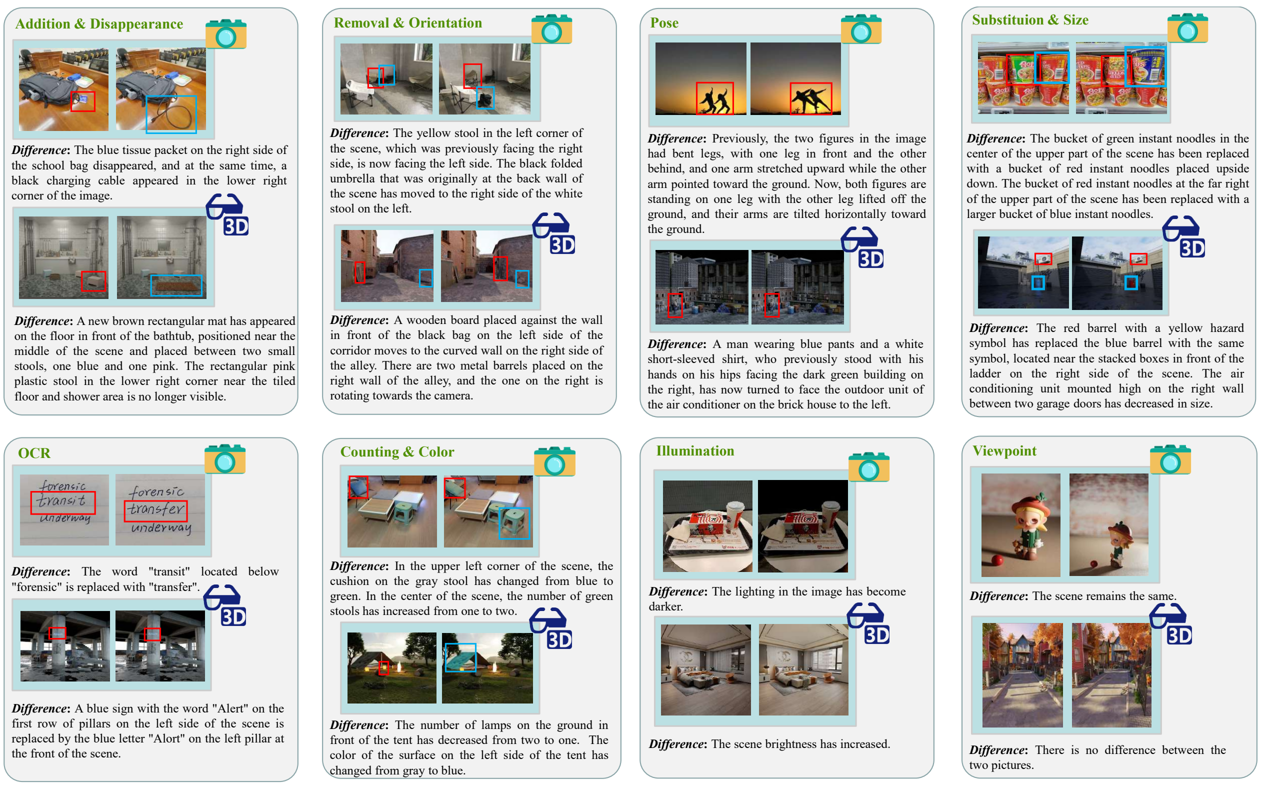

Figure 1. OmniDiff: a comprehensive benchmark comprising 324 complex real-world and 3D synthetic environments, featuring 12 distinct types of variations, with an average difference caption length of 60 words per image pair.

Abstract

*Image Difference Captioning (IDC) aims to generate natural language descriptions of subtle differences between image pairs, requiring both precise visual change localization and coherent semantic expression. Despite recent advancements, existing datasets often lack breadth and depth, limiting their applicability in complex and dynamic environments: (1) from a breadth perspective, current datasets are constrained to limited variations of objects in specific scenes, and (2) from a depth perspective, prior benchmarks often provide overly simplistic descriptions. To address these challenges, we introduce **OmniDiff**, a comprehensive dataset comprising 324 diverse scenarios—spanning real-world complex environments and 3D synthetic settings—with fine-grained human annotations averaging 60 words in length and covering 12 distinct change types. Building on this foundation, we propose **M**[3]**Diff**, a **M**ulti**M**odal large language model enhanced by a plug-and-play **M**ulti-scale **Diff**erential Perception (MDP) module. This module improves the model's ability to accurately identify and describe inter-image differences while maintaining the foundational model's gen-*

✉ Corresponding author.

eralization capabilities. With the addition of the OmniDiff dataset, M³Diff achieves state-of-the-art performance across multiple benchmarks, including Spot-the-Diff, IEdit, CLEVR-Change, CLEVR-DC, and OmniDiff, demonstrating significant improvements in cross-scenario difference recognition accuracy compared to existing methods. The dataset, code, and models will be made publicly available to support further research.

1. Introduction

Image Difference Captioning (IDC) focuses on describing subtle differences between two similar images. Unlike traditional change detection or single-image captioning, IDC requires both precise localization of change regions and accurate semantic expression of these changes, making it a more complex challenge in vision-language understanding. By bridging visual change detection and natural language generation, IDC has significant potential for applications in environmental monitoring and surveillance systems, where articulating visual differences is critical.

The emergence of domain-specific datasets and the refinement of vision-language models significantly enhance deep-learning-driven IDC methods. Despite notable progress in this field, limitations of existing datasets constrain the applicability of IDC methods in complex and dynamic environments. (1) From a **breadth** perspective, existing IDC datasets primarily focus on limited variations of objects in specific scenes, failing to cover the diverse changes encountered in real-world environments. For instance, Spot-the-Diff [17] only covers street surveillance footage from a fixed viewpoint, Birds-to-Words [10] focuses on fine-grained differences among similar bird species, and CLEVR-Change [33] renders simple desktop scenarios featuring five types of object changes. (2) Prior research often lacks **in-depth** discrepancy descriptions, which significantly restricts the model's ability to align visual differences with precise textual representations. For example, the average difference description between image pairs in IEdit [37] comprises merely 8 words, indicating that the captured changes are overly simplistic and fail to reflect the complexity of real-world variations.

To address these challenges, as illustrated in Figure 1, we introduce **OmniDiff**, a high-quality dataset comprising 324 complex real-world and 3D scenes, encompassing 12 distinct types of variations, each annotated with fine-grained human-labeled descriptions. The annotations exhibit an average length of 60 words, ensuring comprehensive and detailed representations of the observed variations. We collect diverse real-world scenes through on-site photography and web crawling, and further expand the dataset by constructing complex 3D scenes using Blender [8] to simulate real-world variations. Unlike CLEVR-series [20, 33, 34] datasets generated with Blender [8], which are confined to simplistic tabletop environments, our approach focuses on creating near-realistic complex settings, effectively enriching the data samples while simultaneously imposing higher demands on the model's 3D spatial perception capabilities.

Leveraging their foundational knowledge and strong generalization capabilities, Multimodal Large Language Models (MLLMs) demonstrate exceptional performance across a wide range of cross-modal tasks, including single-image captioning and visual question answering. This emerging trend motivates the exploration of applying MLLMs to the IDC task, which requires enhanced capabilities in *fine-grained visual perception* and *comprehensive multi-image content understanding*.

In this work, to enhance the fine-grained difference perception capabilities of MLLMs between image pairs, we integrate a plug-and-play **M**ulti-scale **Diff**erential Perception (MDP) Module into the **M**ulti**M**odal Large Language Model framework, establishing **M³Diff**, a strong base model designed for fine-grained IDC. Through our simple yet effective fine-tuning strategy, M³Diff achieves state-of-the-art performance across a variety of benchmarks, including OmniDiff, Spot-the-Diff [17], IEdit [37], CLEVR-Change [33] and CLEVR-DC [20]. These results demonstrate M³Diff's strong generalization capabilities in recognizing differences across varied scenarios, highlighting its adaptability and effectiveness in addressing complex visual understanding tasks.

The primary contributions of this work are summarized as follows:

- We construct **OmniDiff**, a high-quality dataset comprising 324 diverse scenarios, encompassing both real-world complex environments and 3D synthetic settings. The dataset is characterized by fine-grained human annotations, with an average description length of 60 words, and comprehensively covers 12 distinct types of changes.
- We integrate a plug-and-play Multi-scale Differential Perception (MDP) Module into the MLLM architecture to enhance fine-grained difference perception and multi-image content understanding capabilities. This approach facilitates the development of a strong base model, **M³Diff**, specifically tailored for IDC.
- By leveraging the OmniDiff dataset and the M³Diff architecture, we achieve state-of-the-art performance across multiple benchmarks spanning diverse scenarios.

2. Related Works

2.1. Image Difference Captioning

Methods. Research on IDC has evolved through several key methodological advancements. Jhamtani *et al.* [17] pioneer this task by aligning pixel-wise differences between image pairs. Subsequent works [33] [13] [27] attempt to compute differences by directly subtracting image representations. However, these methods exhibit limited generalization capabilities for unaligned image pairs due to pseudo-changes caused by distractors (*e.g.*, viewpoint, illumination). To address robustness against distractors, recent studies introduce specialized learning mechanisms. DIRL [43] stabilizes representations by correlating corresponding channels while decorrelating dissimilar ones, enabling robust difference extraction. Furthermore, multi-task learning frameworks integrate auxiliary objectives to enhance model accuracy. For example, Semantic-CC [54] combines change detection (CD) and captioning (CC) via pixel-level semantic segmentation, improving linguistic precision. With the rise of vision-language pretraining, IDC-PCL [49] and CLIP4IDC [12] propose a two-stage pretrain-finetune framework to effectively model difference representations for change captioning. Recent MLLM-based approaches significantly advance IDC capabilities. FINER-MLLM [51] enhances change captioning through a LoRA [15] fine-tuned MLLM with dual intra- and inter-image constraints, while Hu *et al*. [14] propose a generalist model combining a siamese encoder and Visual Delta Module for detecting and describing subtle differences.

Benchmarks. Benchmarks for change captioning are developed across various domains to address domain-specific challenges. For instance, Spot-the-Diff [17] and STVchrono [36] extract similar video frames from surveillance footage to capture changes in outdoor environments. Birds-to-Words [10] focuses on distinguishing subtle appearance differences among similar bird species, while LEVIR-CC [27] targets changes in urban remote sensing image pairs. Unlike datasets tailored to specific scenes or objects, IEdit [37] comprises approximately 4,000 image pairs capturing diverse daily life scenarios. To address annotation scalability, CLEVR-Change [33] leverages the CLEVR engine [19] to synthesize controlled variations in 3D-rendered tabletop scenes, with CLEVR-DC [20] introducing extreme viewpoint-agnostic shifts and Qiu *et al*. [34] extending the framework through multi-change scenarios and unknown change counts for robust captioning evaluation.

2.2. Multimodal Large Language Models

Multimodal large language models exhibit exceptional capabilities in visual perception and natural language generation by aligning visual and textual features [2, 24] and leveraging visual instruction tuning [9, 28, 29, 52]. Building on these foundations, recent advancements emphasize fine-grained visual understanding. Gromma [31] bridges MLLMs and detection by embedding region-aware visual tokens directly into the language model's latent space. Lai *et al*. [21] propose LISA, an MLLM that integrates segmentation via a token and embedding-as-mask mechanism for complex visual-textual reasoning segmentation. Pushing the boundaries of these advancements, [5, 18, 22, 23] further expand the scope of MLLMs by exploring their ability to understand multi-image and video content, demonstrating their potential for complex multimodal tasks. Despite these significant strides, applying MLLMs to IDC tasks presents unique challenges. IDC requires not only fine-grained visual perception but also the ability to precisely articulate subtle differences between images. Existing MLLMs often struggle to generate accurate and contextually rich descriptions of nuanced visual changes.

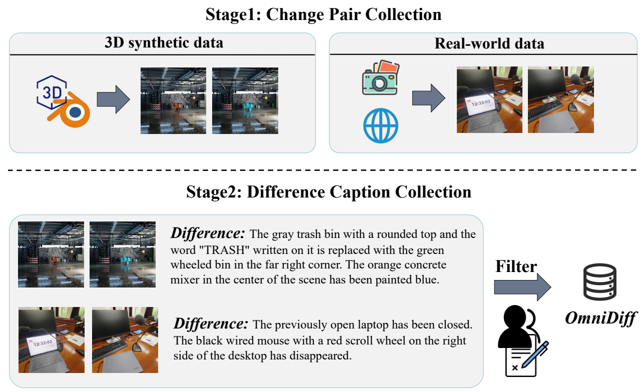

Figure 2. Overview of the OmniDiff construction pipeline.

3. The OmniDiff Dataset

3.1. Overview of OmniDiff

We introduce **OmniDiff**, a high-quality dataset encompassing diverse real-world and synthetic complex scenarios, featuring fine-grained human annotations to ensure comprehensive and accurate labeling. OmniDiff comprises 324 scenes spanning diverse indoor and outdoor environments, collected through a combination of on-site photography, web scraping and 3D rendering. The dataset encompasses 12 categories of fine-grained visual variations, including *viewpoint*, *illumination*, *addition*, *disappearance*, *removal*, *substitution*, *size*, *color*, *orientation*, *pose*, *OCR (textual changes)*, and *counting*. Each similar image pair encompasses one or more types of variations, with the average length of difference descriptions extending to 60 words to ensure a thorough and detailed representation of the observed changes.

3.2. Dataset Construction Process

The construction process of the OmniDiff consists of two key stages: Change Pair Collection and Difference Caption Collection. Figure 2 shows the overall workflow.

3.2.1. Change Pair Collection.

OmniDiff comprises change pairs sourced from 324 diverse indoor and outdoor scenes, encompassing a wide range of everyday environments such as streets, supermarkets,

Table 1. The comparison of OmniDiff with existing image difference caption datasets. OmniDiff excels in both its breadth, encompassing a diverse range of real-world scenarios, and its depth, characterized by detailed and fine-grained annotations. The underlined value indicates a relatively high number of image pairs in CLEVR-Change [33], although these pairs are limited to tabletop environments with constrained object variations. *Distractor change* refers to non-semantic variations (*e.g.*,viewpoint shifts and illumination changes).

Dataset	Environment	Image source	Image pair	Distractor change	Human-labeled caption	Average words per caption	% Long caption (>20 words)
CLEVR-Change [33]	table	3D render	<u>79606</u>	✓	✗	8	0%
Spot-the-Diff [17]	outdoor	surveillance video	13192	✗	✓	19	19%
LEVIR-CC [26]	outdoor	satellite	10077	✓	✓	8	1%
IEdit [37]	**in- & out-door**	diverse real life	3939	✗	✓	8	2%
Birds-to-Words [10]	outdoor	birds	3347	✗	✓	32	25%
OmniDiff (ours)	**in- & out-door**	**3D render & diverse real life**	**15598**	✓	✓	**60**	**37%**

schools, and bedrooms, as illustrated in Figure 1. The change pairs originate from two primary sources: real-world data and 3D synthetic data.

To collect change pairs from real-world scenes, we utilize a combination of on-site photography and web crawling to capture diverse environmental conditions and various types of changes, ensuring comprehensive coverage of real-world scenarios. For on-site photography, we systematically position cameras in diverse indoor and outdoor environments to capture natural scene variations over time, while adjusting camera perspectives to simulate viewpoint changes and enrich the diversity of the collected data. To further enhance scene diversity, we collect online videos to capture environments and change types that are challenging to obtain through on-site photography. We specifically select videos depicting real-life scenarios and manually extract frames with similar scenes but subtle visual differences to construct change pairs. The sources and quality of these videos are rigorously screened to ensure compliance with the dataset's high-quality standards. Through a combination of on-site photography and web crawling, we collect change pairs from 224 distinct real scenes.

In addition to real-world data, we utilize the Blender [8] engine to render complex 3D scenes, simulating real-world variations. Unlike the CLEVR-series [20, 33, 34], which restricts changes to a limited set of objects within a single scene, we source high-quality 3D assets from ArtStation [4], including 50 indoor and 50 outdoor scenes, and manually modify objects within the scenes to emulate realistic and diverse changes.

3.2.2. Difference Caption Collection.

Although state-of-the-art MLLMs such as GPT-4o [1] exhibit strong capabilities in generating textual descriptions, our experiments reveal their limitations in annotating fine-grained image differences in complex scenes, particularly in precisely localizing and describing subtle visual changes. Therefore, we rely on human annotators to ensure the accuracy and reliability of the dataset. Given that the images are captured in complex scenes, we instruct the annotators to accurately describe the location, attributes, and types of changes for the altered objects. To ensure consistency and quality, we provide the annotators with the following annotation guidelines: 1) Focus on describing semantic changes in the scene (*e.g.*, the addition of an object) while ignoring irrelevant variations (*e.g.*, viewpoint shifts). 2) Structure each caption into two parts: the *referring part* (*e.g.*, a man in a blue shirt standing behind a desk on the left side of the scene) and the *change part* (*e.g.*, changed from facing the window on the left to facing the sofa on the right).

Table 2. Statistics of the OmniDiff Dataset.

Statistic	**Value**
Image Change Pairs	15,598
Total Captions	15,598
Avg Words per Caption	60.0
Total Sentences	38,653
Sentences per Caption	2.5
Vocabulary Size	3,793
Splits	Train:Val.:Test = 80%:10%:10%

After completing the annotation process, we review and correct spelling and grammatical errors, resulting in a final collection of 15,598 difference captions with an average length of 60 words per caption.

3.3. Dataset Statistics

As shown in Table 2, OmniDiff contains a total of 15,598 change pairs, consisting of 8,609 pairs of real-world scene images and 6,989 pairs of 3D-rendered scene images, achieving a balanced ratio of 1.2:1 between real-world and synthetic data. The average difference description for each image pair consists of 60 words, ensuring a precise and detailed representation of fine-grained variations between images. We divide both indoor and outdoor scenes into training, validation, and test sets at an 8:1:1 ratio, ensuring that each subset comprehensively encompasses all 12 distinct types of changes.

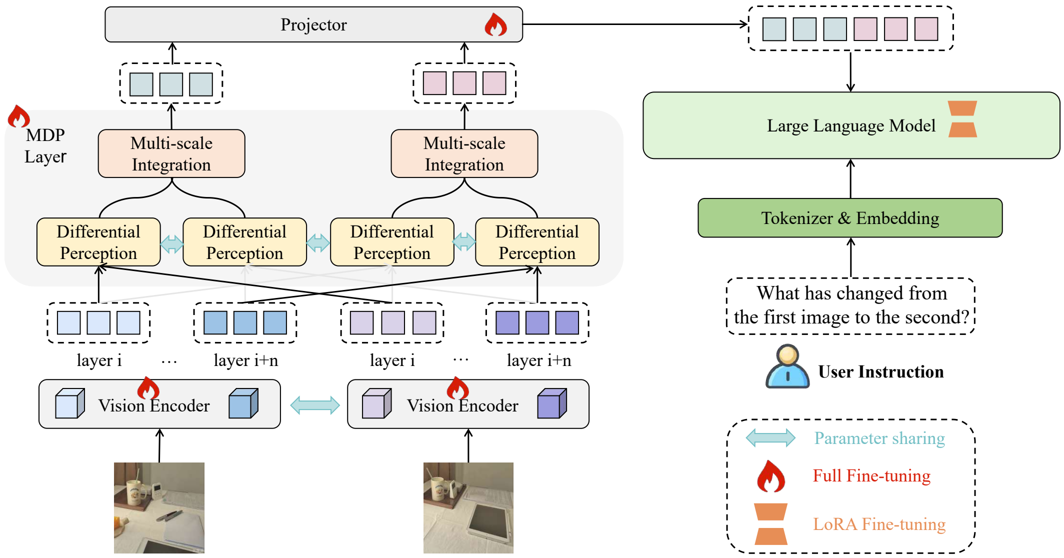

Figure 3. The whole structure of M^{3}Diff with the multi-scale differential perception module.

3.4. Evaluation protocol

For change captioning, we evaluate performance using five standard metrics: BLEU-4 (B) [32], METEOR (M) [6], ROUGE-L (R) [25], CIDEr-D (C) [46], and SPICE (S) [3]. All metrics are computed using the Microsoft COCO evaluation server [7].

3.5. Comparisons with Existing Datasets

Table 1 further highlights the advantages of OmniDiff compared to other datasets. *From the breadth perspective*, existing datasets primarily focus on limited variations of objects within single scenes (*e.g.*, CLEVR-Change [33]), whereas OmniDiff combines real-world and 3D-rendered scenes to construct image pairs in complex and diverse environments, encompassing 12 distinct types of variations. *In terms of depth*, previous datasets typically provide short annotations that lack the granularity needed to clearly describe fine-grained differences between image pairs. In contrast, OmniDiff includes a significant portion of long, detailed descriptions that precisely identify changes in complex scenes, offering a richer and more informative resource for training and evaluating multimodal models.

4. Method

4.1. Overview

Inspired by recent advancements in leveraging MLLMs for change captioning tasks [47, 54], our M^{3}Diff extends the standard MLLM architecture with a multi-scale differential perception module, as illustrated in Figure 3. This dedicated component implements feature-level differencing through channel-wise subtraction between image pairs, followed by adaptive cross-layer fusion operations that establish coherent difference representations prior to linguistic decoding. In this section, we first introduce the MDP module for multi-scale difference modeling, followed by a systematic exposition of the training paradigm of our framework.

4.2. Multi-Scale Differential Perception

MDP module is employed to extract multi-scale discriminative discrepancy features through channel-wise subtraction, then fuses them with original image features via cross-attention mechanisms. By explicitly modeling cross-scale visual disparities and their contextual interactions, this module enhances precise difference localization and descriptive feature representation for captioning tasks.

Differential Perception. The difference feature between image pairs is critical for IDC tasks but prone to distractors (*e.g.*, illumination/viewpoint). The purpose of this operation is to generate a robust differential feature, and then integrate the enhanced difference representation back into the original feature streams for improved discriminability.

Specifically, the differential feature is calculated by:

$$\boldsymbol{\lambda}_1 = \sigma\big(\mathbf{W}_m[\mathbf{F}_1^i \| \mathbf{F}_2^i]\big),\ \boldsymbol{\lambda}_2 = \sigma\big(\mathbf{W}_m[\mathbf{F}_2^i \| \mathbf{F}_1^i]\big), \quad (1)$$

$$\hat{\mathbf{F}}_k^i = \mathbf{F}_k^i \odot \boldsymbol{\lambda}_k,\ k \in \{1, 2\}, \quad (2)$$

$$\Delta\mathbf{F}^i = \mathbf{W}_p\big[\hat{\mathbf{F}}_1^i \| \hat{\mathbf{F}}_2^i \| (\hat{\mathbf{F}}_1^i - \hat{\mathbf{F}}_2^i)\big], \quad (3)$$

where $\mathbf{F}_1^i, \mathbf{F}_2^i$ are image pair features from the i-th layer, $\|$ denotes concatenation, and σ is sigmoid function. The weight $\mathbf{W}_m \in \mathbb{R}^{d\times 2d}$ processes concatenated features through a learnable projection, where d denotes the feature dimension. The resulting attention weights $\boldsymbol{\lambda}_k \in [0,1]^d$ modulate input features via element-wise multiplication ($\odot$), preserving original dimensions. $\mathbf{W}_p \in \mathbb{R}^{d\times 3d}$ projects the concatenated triplet into $\Delta \mathbf{F}^i$, a d-dimensional vector that explicitly encodes feature variations.

Then, we fuse the salient differences with the original image features through cross-attention mechanisms:

$$\Delta \mathbf{F}_{\text{sa}}^i = \text{SA}(\Delta \mathbf{F}^i), \tag{4}$$

$$\Delta \mathbf{F}_{\text{ca}}^i = \text{MLP}\Big(\text{CA}\big(\Delta \mathbf{F}_{\text{sa}}^i,\ \mathbf{F}_1^i \| \mathbf{F}_2^i \| \Delta \mathbf{F}_{\text{sa}}^i\big)\Big), \tag{5}$$

$$\tilde{\mathbf{F}}_k^i = \text{CA}\big(\mathbf{F}_k^i,\ \mathbf{F}_k^i \| \Delta \mathbf{F}_{\text{ca}}^i\big),\ k \in \{1, 2\}, \tag{6}$$

where the cross-attention (CA) establishes correspondences between two streams: (1) using the self-attention (SA) difference representation $\Delta \mathbf{F}_{\text{sa}}^i$ as queries, and (2) employing the concatenation of original features and $\Delta \mathbf{F}_{\text{sa}}^i$ as keys/values. After Multiple Layer Perceptron (MLP) processing, CA further fuses each original feature $\mathbf{F}_k^i$ with the refined difference signal $\mathbf{F}_k^i \| \Delta \mathbf{F}_{\text{ca}}^i$ via query and key/value interactions, preserving spatial dimensions while enhancing channel-wise discriminability.

Multi-Scale Integration. MLLMs typically employ the visual features from the penultimate layer of the visual encoder [22, 28, 29]. However, features from individual network layers face inherent limitations: low-level features lack semantic consistency, while high-level features lose detail perception. Therefore, multi-layer feature fusion is employed to aggregate low-level details and high-level semantics to yield a more comprehensive and discriminative representation. The fusion process can be formulated as:

$$\text{Score}^i = \sigma\Big(\text{MLP}\big(\Phi_{\text{Mean}}(\tilde{\mathbf{F}}_1^i \| \tilde{\mathbf{F}}_2^i \| \Delta \mathbf{F}^i)\big)\Big), \tag{7}$$

$$\mathbf{F}_k' = \sum\nolimits_i \text{Score}^i \odot \tilde{\mathbf{F}}_k^i,\ k \in \{1, 2\}, \tag{8}$$

where Φ_{Mean} stands for the token-wise mean pooling to compress spatial information. Score^i is the fusion weight of the i-th layer's feature. The weights are then used to compute the refined features $\mathbf{F}_1'$ and $\mathbf{F}_2'$ via element-wise multiplication and cross-layer summation.

4.3. Training Strategy

Unlike previous multi-stage approaches [14, 47, 54] adapting MLLMs for IDC tasks, we employ a simple yet effective one-stage fine-tuning strategy. Despite a gap between the MDP module and the well-pretrained MLLM [22], large-scale instruction tuning enables the lightweight MDP module to enhance difference perception without disrupting existing knowledge, achieving plug-and-play functionality.

The dataset utilized for fine-tuning M^{3}Diff consists of 896k question-answer pairs (refer to Section 8.1 of the Supplementary Material for details). In addition to OmniDiff, the dataset integrates several publicly available IDC datasets, including Spot-the-Diff [17], IEdit [37], and Birds-to-Words [10], as well as 3D-rendered samples from CLEVR-Change [33] and CLEVR-DC [20]. During training, we apply LoRA [15] for parameter-efficient fine-tuning to the LLM while fully fine-tuning the rest of the model.

5. Experiments

5.1. Datasets

We evaluate our model on four public datasets and OmniDiff, including the real-world domain datasets Spot-the-Diff [17], Image-Editing-Request [37], and OmniDiff-Real, as well as the 3D-rendered datasets CLEVR-Change [33], CLEVR-DC [20], and OmniDiff-Render.

Spot-the-Diff [17] consists of 13,192 pairs of similar video frames captured from street surveillance footage. Following the standard evaluation protocol used in prior work [17], we evaluate our model in a single-change captioning setting. The dataset is officially split into training, validation, and testing sets with a ratio of 8:1:1.

Image-Edit-Request [37] consists of 3,939 aligned image pairs from real-life scenarios, accompanied by 5,695 editing instructions. Following the official dataset split protocol [37], we use 3,061 image pairs for training, 383 for validation, and 495 for testing.

CLEVR-Change [33] is a large-scale synthetic dataset featuring scene changes involving geometric objects and distractors under moderate viewpoint variations. The dataset contains 79,606 image pairs, split into 67,660 for training, 3,976 for validation, and 7,970 for testing.

CLEVR-DC [20] is a large-scale dataset featuring extreme viewpoint shifts, comprising 48,000 image pairs with the same change types as CLEVR-Change [33]. The dataset is split into 85% training, 5% validation, and 10% testing.

OmniDiff-Real, a subset of the OmniDiff dataset, comprises 8,609 real-world image pairs, split into 7,007 for training, 799 for validation, and 803 for testing. Given the diverse and complex changes in OmniDiff, our model is evaluated in a multi-change captioning setting.

OmniDiff-Render comprises 3D-rendered image pairs from the OmniDiff dataset, with 5,445 pairs for training, 777 for validation, and 767 for testing. Model evaluation is performed under the multi-change setting, consistent with the approach used for the OmniDiff-Real.

5.2. Implementation Details

Architecture. M^{3}Diff initializes the weights of the visual encoder, projector, and LLM from LLaVA-OneVision-7B [22], while the MDP module weights are newly initialized. The visual encoder utilizes the SigLIP model [50], containing 27 transformer layers. Multi-scale features extracted from layers 17, 20, 23, and 26 of the visual encoder serve as input to MDP, which comprises two stacked transformer layers. The projector comprises a two-layer multilayer perceptron, and the LLM is based on the Qwen2 [48].

Table 3. Performance comparison on OmniDiff across all metrics. The best results are highlighted in boldface, while the second-best results are indicated with underlining. † indicates models that are trained without utilizing the OmniDiff dataset.

Method	Real				Render			
	BLEU-4	METEOR	ROUGE-L	CIDEr	BLEU-4	METEOR	ROUGE-L	CIDEr
VARD-LSTM (TIP'23) [40]	5.5	12.8	24.2	6.7	4.2	10.2	24.7	3.7
SCORER (ICCV'23) [42]	7.2	11.1	23.5	7.0	9.9	13.3	27.1	4.1
DIRL (ECCV'24) [43]	7.5	11.0	23.8	5.4	11.7	13.8	27.6	5.9
CARD (ACL'24) [45]	9.1	12.6	25.1	9.2	11.3	13.4	27.3	7.3
MLLM								
†GPT-4o [1]	3.1	13.6	21.0	5.2	4.6	10.7	21.4	5.6
†LLaVA-OneVision-7B [22]	0.2	4.1	13.6	1.7	0.3	4.7	15.6	1.6
†Qwen-2.5-VL-7B [5]	3.8	9.5	19.8	6.2	2.1	6.8	18.3	3.3
FINER-MLLM (MM'24) [51]	8.9	13.8	25.9	11.7	13.6	15.6	29.9	14.0
M^{3}Diff (ours)	**14.3**	**18.9**	**32.9**	**31.3**	**15.7**	**19.9**	**35.3**	**28.3**

Table 4. Performance comparison on Spot-the-Diff across all metrics. The best results are highlighted in boldface, while the second-best results are indicated with underlining.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
DUDA (ICCV'19) [33]	8.1	11.8	29.1	32.5	-
M-VAM (ECCV'20) [35]	10.1	12.4	31.3	38.1	-
M-VAM+RAF (ECCV'20) [35]	11.1	12.9	33.2	42.5	17.1
IFDC (TMM'22) [16]	8.7	11.7	30.2	37.0	-
DUDA+TIRG (CVPR'21) [13]	8.1	12.5	29.9	34.5	-
VACC (ICCV'21) [20]	9.7	12.6	32.1	41.5	-
SRDRL (ACL'21) [39]	-	13.0	31.0	35.3	18.0
R^{3}Net (EMNLP'21) [38]	-	13.1	32.6	36.6	18.8
MCCFormers-D (ICCV'21) [34]	10.0	12.4	-	43.1	18.3
CLIP4IDC (AACL'22) [12]	11.6	14.2	35.0	47.4	-
VARD-Trans (TIP'23) [40]	-	12.5	29.3	30.3	17.3
SCORER (ICCV'23) [42]	10.2	12.2	-	38.9	18.4
SMARL (TPAMI'24) [44]	-	13.1	32.8	40.0	19.5
SMART (TPAMI'24) [44]	-	13.5	31.6	39.4	19.0
DIRL (ECCV'24) [43]	10.3	13.8	32.8	40.9	19.9
MLLM					
FINER-MLLM (MM'24) [51]	12.9	14.7	35.5	61.8	22.1
OneDiff (ACCV'24) [14]	12.8	14.6	35.8	56.6	-
M^{3}Diff (ours)	**14.4**	**15.4**	**37.6**	**71.1**	**24.9**

Table 5. Performance comparison on Image-Edit-Request across all metrics. The best results are highlighted in boldface, while the second-best results are indicated with underlining.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
DUDA (ICCV'19) [33]	6.5	12.4	37.3	22.8	-
MCCFormers-D (ICCV'21) [34]	8.3	14.3	39.2	30.2	-
CLIP4IDC (AACL'22) [12]	8.2	14.6	40.4	32.2	-
NCT (TMM'23) [41]	8.1	15.0	38.8	34.2	12.7
VARD-Trans (TIP'23) [40]	10.0	14.8	39.0	35.7	-
SCORER (ICCV'23) [42]	10.0	15.0	39.6	33.4	-
SMARL (TPAMI'24) [44]	10.4	15.1	40.3	34.6	-
SMART (TPAMI'24) [44]	10.5	15.2	39.1	37.8	-
DIRL (ECCV'24) [43]	10.9	15.0	41.0	34.1	-
MLLM					
FINER-MLLM (MM'24) [51]	14.1	15.9	40.4	53.5	15.9
OneDiff (ACCV'24) [14]	29.6	25.1	55.6	109.6	-
M^{3}Diff (ours)	**33.6**	**26.5**	**59.7**	**136.6**	**27.5**

Training setting. Building upon the OmniDiff, we collect and curate publicly available real-world datasets, including Spot-the-Diff [17], IEdit [37], and a subset of Birds-to-Words [10], alongside 3D-rendered datasets such as CLEVR-Change [33] and CLEVR-DC [20]. These datasets are transformed into a question-answer format, yielding a fine-tuning dataset of 896K samples. For the CLEVR-Change [33] and CLEVR-DC [20], we utilize each caption per image pair as an independent training sample. During the fine-tuning phase, we employ LoRA [15] for parameter-efficient tuning of the LLM to minimize computational costs while preserving the LLM's inherent knowledge. The LoRA configuration includes a rank of 128 and an alpha value of 256. Concurrently, the visual encoder, projector, and MDP module undergo full-parameter fine-tuning. The AdamW optimizer sets the initial learning rate to 2e-6 for the visual encoder and 1e-5 for the remaining modules. During training, a cosine annealing scheduler with a linear warm-up dynamically adjusts the learning rate. Our experiments utilize 8 NVIDIA A100 (40G) GPUs with a global batch size of 256, completing in 26 hours.

5.3. Performance Comparison

To validate the effectiveness of M^{3}Diff in complex and dynamic scenarios, we fine-tune state-of-the-art IDC methods (VARD [40], SCORER [42], DIRL [43], CARD [45]) and the MLLM-based approach FINER-MLLM [51] on the OmniDiff dataset. We then conduct a comprehensive evaluation on the test set to compare their performance. As demonstrated in Table 3, M^{3}Diff achieves superior performance across all metrics in both real-world and rendered scenarios. Additionally, we perform zero-shot evaluations of leading MLLMs (GPT-4o [1], LLaVA-OneVision [22], and Qwen-2.5-VL [5]) on OmniDiff, demonstrating their limitations in analyzing complex scene differences despite their general multimodal capabilities.

As shown in Tables 4, 5, 7, and 8, M^{3}Diff consistently achieves SOTA performance across various publicly IDC benchmarks, including real-world scenarios such as Spot-the-Diff [17] and IEdit [37], as well as rendered environments like CLEVR-Change [33] and CLEVR-DC [20].

Table 6. Ablation Study of the OmniDiff dataset and MDP Module on OmniDiff and IEdit Benchmarks.

Settings	OmniDiff-Real				OmniDiff-Render				IEdit				
	B	M	R	C	B	M	R	C	B	M	R	C	S
w/o OmniDiff & MDP	0.1	3.3	11.5	1.1	0.0	2.9	10.0	0.7	31.5	26.0	59.6	133.5	25.7
w/o OmniDiff	0.1	3.8	12.7	1.9	0.1	3.7	11.9	1.1	32.3	26.3	59.7	132.8	27.0
w/o MDP	12.2	17.1	32.3	**35.3**	15.5	18.6	34.3	25.4	33.5	26.3	59.7	135.2	26.5
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 7. Performance comparison on CLEVR-Change across all metrics. The best results are highlighted in boldface, while the second-best results are indicated with underlining.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
DUDA (ICCV'19) [33]	47.3	33.9	-	112.3	24.5
M-VAM (ECCV'20) [35]	50.3	37.0	69.7	114.9	30.5
M-VAM+RAF (ECCV'20) [35]	51.3	37.8	70.4	115.8	30.7
IFDC (TMM'22) [16]	49.2	32.5	69.1	118.7	-
DUDA+TIRG (CVPR'21) [13]	51.2	37.7	70.5	115.4	31.1
VACC (ICCV'21) [20]	52.4	37.5	-	114.2	31.0
SRDRL (ACL'21) [39]	54.9	40.2	73.3	122.2	32.9
R^{3}Net (EMNLP'21) [38]	54.7	39.8	73.1	123.0	32.6
MCCFormers-D (ICCV'21) [34]	52.4	38.3	-	121.6	26.8
CLIP4IDC (AACL'22) [12]	<u>56.9</u>	38.4	**76.4**	**150.7**	-
NCT (TMM'23) [41]	55.1	40.2	73.8	124.1	32.9
VARD-Trans (TIP'23) [40]	55.4	40.1	73.8	126.4	32.6
SCORER (ICCV'23) [42]	56.3	<u>41.2</u>	74.5	126.8	33.3
SMARL (TPAMI'24) [44]	55.7	40.6	73.8	126.5	33.4
SMART (TPAMI'24) [44]	56.1	40.8	74.2	127.0	<u>33.4</u>
DIRL (ECCV'24) [43]	54.6	38.1	71.9	123.6	31.8
MLLM					
FINER-MLLM (MM'24) [51]	55.6	36.6	72.5	<u>137.2</u>	26.4
M^{3}Diff (ours)	**57.1**	**41.9**	<u>75.3</u>	130.5	**33.8**

Table 8. Performance comparison on CLEVR-DC across all metrics. The best results are highlighted in boldface, while the second-best results are indicated with underlining.

Method	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
DUDA (ICCV'19) [33]	40.3	27.1	-	56.7	16.1
DUDA+CC (ICCV'19) [33]	41.7	27.5	-	62.0	16.4
M-VAM (ECCV'20) [35]	40.9	27.1	-	60.1	15.8
M-VAM+CC (ECCV'20) [35]	41.0	27.2	-	62.0	15.7
VA (ICCV'21) [20]	44.5	29.2	-	70.0	17.1
VACC (ICCV'21) [20]	45.0	29.3	-	71.7	<u>17.6</u>
MCCFormers-D (ICCV'21) [34]	46.9	31.7	-	71.6	14.6
NCT (TMM'23) [41]	47.5	32.5	65.1	76.9	15.6
VARD-Trans (TIP'23) [40]	48.3	32.4	-	77.6	15.4
SCORER (ICCV'23) [42]	49.4	<u>33.4</u>	66.1	83.7	16.2
DIRL (ECCV'24) [43]	<u>51.4</u>	32.3	<u>66.3</u>	<u>84.1</u>	16.8
MLLM					
M^{3}Diff (ours)	**60.6**	**37.6**	**73.0**	**109.4**	**21.3**

These results demonstrate the model's strong generalization ability in cross-scenario difference identification.

5.4. Ablation Study and Analysis

Effects of OmniDiff Dataset. Table 6 illustrates the impact of fine-tuning M^{3}Diff with the OmniDiff dataset, demonstrating its performance improvements on both the OmniDiff benchmark and IEdit [37]. The experiments reveal that fine-tuning M^{3}Diff solely on existing public datasets fails to effectively describe differences between image pairs in OmniDiff, which contains a wide range of complex and dynamic scenarios. By incorporating the OmniDiff dataset into the fine-tuning process alongside existing public datasets, the model not only achieves significant performance improvements in complex scenarios within the OmniDiff benchmark but also demonstrates enhanced capabilities on IEdit [37]. These results validate that the diverse and dynamic samples in OmniDiff effectively strengthen the generalization ability of M^{3}Diff across varied scenarios.
Effects of Multi-Scale Differential Perception Module. Table 6 demonstrates that introducing the MDP module to enhance difference perception capabilities, while preserving the inherent knowledge of MLLM, leads to further improvements on both the OmniDiff and IEdit [37] test set. Removing MDP leads to consistent degradation in text quality and semantic coherence, highlighting its role in harmonizing multi-scale feature integration. While partial metric fluctuations occur (*i.e.*, the CIDEr variation on OmniDiff-Real), the full model demonstrates balanced superiority, outperforming ablated versions in 12/13 metrics. This holistic enhancement confirms MDP's necessity for fine-grained IDC tasks.

6. Conclusion and Future Outlook

In this work, we tackle the limitations of IDC by introducing **OmniDiff**, a comprehensive dataset featuring 324 diverse scenarios spanning real-world and 3D synthetic environments. With fine-grained human annotations averaging 60 words and covering 12 distinct change types, OmniDiff sets a new benchmark for breadth and depth in IDC datasets. Building on this, we propose **M^{3}Diff**, a MultiModal Large Language Model enhanced by a plug-and-play Multi-scale Differential Perception (MDP) module. Extensive experiments show that M^{3}Diff achieves state-of-the-art performance on multiple benchmarks, demonstrating significant improvements in cross-scenario difference recognition.

In future work, OmniDiff can be expanded to include image sets capturing continuous real-world changes, enabling more comprehensive modeling of dynamic scenarios. Building on this, M^{3}Diff can further be extended to video-based difference captioning, significantly broadening its applicability to temporally evolving environments.

Acknowledgement

This work is jointly supported by National Natural Science Foundation of China (62206022, 62276025, 62476027) and the Fundamental Research Funds for the Central Universities (2253200026).

References

[1] Josh Achiam, Steven Adler, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 7, 12, 13

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. *SPICE: Semantic Propositional Image Caption Evaluation*, page 382–398. 2016. 5

[4] ArtStation. Artstation - leading showcase platform for art and design. https://www.artstation.com/. 4

[5] Shuai Bai, Keqin Chen, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 7, 12, 13

[6] Satanjeev Banerjee and Alon Lavie. 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, 2005. 5

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5

[8] Blender Community. Blender - a 3d modelling and rendering package, 2018. 2, 4

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 3

[10] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: Generating fine-grained image comparisons. *arXiv preprint arXiv:1909.04101*, 2019. 2, 3, 4, 6, 7, 11, 12

[11] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024. 12

[12] Zixin Guo, Tzu-Jui Julius Wang, and Jorma Laaksonen. Clip4idc: Clip for image difference captioning. *arXiv preprint arXiv:2206.00629*, 2022. 3, 7, 8

[13] Mehrdad Hosseinzadeh and Yang Wang. Image change captioning by learning from an auxiliary task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2725–2734, 2021. 2, 7, 8

[14] Erdong Hu, Longteng Guo, Tongtian Yue, Zijia Zhao, Shuning Xue, and Jing Liu. Onediff: A generalist model for image difference captioning. In *Proceedings of the Asian Conference on Computer Vision*, pages 2439–2455, 2024. 3, 6, 7

[15] Edward J Hu, Yelong Shen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3, 6, 7, 11

[16] Qingbao Huang, Yu Liang, Jielong Wei, Yi Cai, Hanyu Liang, Ho-fung Leung, and Qing Li. Image difference captioning with instance-level fine-grained feature representation. *IEEE transactions on multimedia*, 24:2004–2017, 2021. 7, 8

[17] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018. 2, 3, 4, 6, 7, 11, 12

[18] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 3

[19] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3

[20] Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. Agnostic change captioning with cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2095–2104, 2021. 2, 3, 4, 6, 7, 8, 11, 12

[21] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 3

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3, 6, 7, 12

[23] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 3

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[26] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. 4

[27] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. 2, 3

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3, 6

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3, 6

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 11

[31] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pages 417–435. Springer, 2024. 3

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001. 5

[33] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633, 2019. 2, 3, 4, 5, 6, 7, 8, 11, 12

[34] Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. Describing and localizing multiple changes with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1971–1980, 2021. 2, 3, 4, 7, 8

[35] Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *European conference on computer vision*, pages 574–590. Springer, 2020. 7, 8

[36] Yanjun Sun, Yue Qiu, Mariia Khan, Fumiya Matsuzawa, and Kenji Iwata. The stvchrono dataset: Towards continuous change recognition in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14120, 2024. 3

[37] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689*, 2019. 2, 3, 4, 6, 7, 8, 11, 12

[38] Yunbin Tu, Liang Li, Chenggang Yan, Shengxiang Gao, and Zhengtao Yu. R ˆ3 net: Relation-embedded representation reconstruction network for change captioning. *arXiv preprint arXiv:2110.10328*, 2021. 7, 8

[39] Yunbin Tu, Tingting Yao, Liang Li, Jiedong Lou, Shengxiang Gao, Zhengtao Yu, and Chenggang Yan. Semantic relation-aware difference representation learning for change captioning. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 63–73, 2021. 7, 8

[40] Yunbin Tu, Liang Li, Li Su, Junping Du, Ke Lu, and Qingming Huang. Adaptive representation disentanglement network for change captioning. *IEEE Transactions on Image Processing*, 32:2620–2635, 2023. 7, 8

[41] Yunbin Tu, Liang Li, Li Su, Ke Lu, and Qingming Huang. Neighborhood contrastive transformer for change captioning. *IEEE Transactions on Multimedia*, 25:9518–9529, 2023. 7, 8

[42] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2805–2815, 2023. 7, 8

[43] Yunbin Tu, Liang Li, Li Su, Chenggang Yan, and Qingming Huang. Distractors-immune representation learning with cross-modal contrastive regularization for change captioning. In *European Conference on Computer Vision*, pages 311–328. Springer, 2024. 3, 7, 8

[44] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, and Qingming Huang. Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4926–4943, 2024. 7, 8

[45] Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. Context-aware difference distilling for multi-change captioning. *arXiv preprint arXiv:2405.20810*, 2024. 7, 12, 13

[46] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[47] Zhiming Wang, Mingze Wang, Sheng Xu, Yanjing Li, and Baochang Zhang. Ccexpert: Advancing mllm capability in remote sensing change captioning with difference-aware integration and a foundational dataset. *arXiv preprint arXiv:2411.11360*, 2024. 5, 6

[48] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 6

[49] Linli Yao, Weiying Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3108–3116, 2022. 3

[50] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 6, 11

[51] Xian Zhang, Haokun Wen, Jianlong Wu, Pengda Qin, Hui Xue', and Liqiang Nie. Differential-perceptive and retrieval-augmented mllm for change captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4148–4157, 2024. 3, 7, 8, 12, 13

[52] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

[53] Jinguo Zhu, Weiyun Wang, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 12, 13

[54] Yongshuo Zhu, Lu Li, Keyan Chen, Chenyang Liu, Fugen Zhou, and Zhenwei Shi. Semantic-cc: Boosting remote sensing image change captioning via foundational knowledge and semantic guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3, 5, 6