Img-Diff: Contrastive Data Synthesis for Multimodal Large Language Models

Supplementary Material

8. Overview

We provide more details and experiments of this work in the supplementary material and organize them as follows:

- Section 9. **Comparison with Existing Image Difference Datasets**: We compare our IMG-DIFF dataset with existing image difference datasets in terms of characteristics and performance, highlighting the advantages of our dataset.
- Section 10. **Prioritizing Quality Over Quantity**: We clarify that our choice to use 13K samples for testing is motivated by the typical size of task-specific datasets used for MLLM fine-tuning. Furthermore, by expanding the dataset to four times its original size, we confirm that the relationship between data size and performance gains is not linear.
- Section 11. **Expanding Diversity with Lexicons**: We use a lexicon to generate object replacement data and test the new dataset. The results validate the effectiveness of this lexicon-based strategy in enhancing data diversity.
- Section 12. **Performance Based on Contrastive Chain-of-Thought**: We evaluate our dataset using the Contrastive Chain-of-Thought method. The results confirm that our dataset enables the fine-tuned model to more accurately describe image differences, thereby enhancing the model's VQA capability.
- Section 13. **Testing on MLLMs at Different Scales**: We test the performance of our IMG-DIFF dataset across MLLMs of different scales. The results indicate that the performance gains brought by our dataset are not limited by scale.
- Section 14. **Top-Performing MLLMs in Image Difference Detection**: We evaluate the difference detection capabilities of top-performing MLLMs, revealing significant room for improvement among SOTA models.
- Section 15. **Unnatural Images in the Dataset**: We remove unnatural images from the generated data, fine-tune the models and evaluate their performance, revealing that unnatural images do not necessarily degrade model efficacy.
- Section 16. **Impact of our Dataset on Spatial Reasoning Performance**: We evaluate whether our generated data enhances spatial reasoning capabilities in models, confirming its effectiveness.
- Section 17. **Ablation Studies**: We explore the impact of varying filter intensities on the performance of the final dataset. As a result, we identify an optimal threshold that balances data quality and quantity.
- Section 18. **Additional Details of Experiments**: We present additional details, including the preprocessing methods for image pairs, the standard training strategies for MLLMs, the model selection and rationale behind our approach, the filtering thresholds applied throughout the work, and the time consumption for generating data.
- Section 19. **The "Object Removal" Exploration**: We generate an extended dataset that focuses on object removal. Additionally, we experimentally validate its effectiveness.
- Section 20. **Examples**: We present several examples of our "object replacement" data and "object removal" data, highlighting detailed information.

9. Comparison with Existing Image Difference Datasets

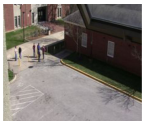
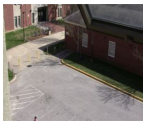
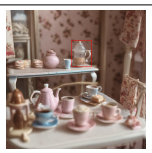
9.1. Characteristics Comparison

Table 4 compares the characteristic differences between our Img-Diff dataset and other existing image difference datasets. The comparison focuses on three key aspects: the "Open-Domain" feature, which refers to whether the dataset covers unrestricted object categories; the "Automatic" feature, which indicates whether the dataset can be generated fully automatically without manual intervention; and the "Region-Focused" feature, which highlights whether the dataset emphasizes image differences in specific detail regions rather than overall image differences.

Specifically, CUB-Birds [69] and Spot-the-Diff [26] are classic examples of traditional datasets where images are collected from the real world and data samples are generated through manual annotations. The former consists of images of various bird species captured in the wild, while the latter is compiled from street-view images taken at different time points by stationary surveillance cameras. In addition, image difference datasets can also be generated using man-made data. For example, Image-Edit-Request [62] collects image pairs consisting of manually edited images and their originals from the web, while MagicBrush [79] employs workers to write editing instructions, which are then used to generate the required image pairs with image editing techniques. These methods are limited by the scarcity of real-world data, as well as the resource and financial costs associated with manual annotation, resulting in limited dataset sizes.

To reduce resource consumption and allow for an unlimited data size, some methods have adopted fully automated generation strategies, such as InstructPix2Pix [6] and MJ-Bench [11]. These methods eliminate the need for manually collected data by using generative models and image

Table 4. Comparison of different image difference datasets. "Open-Domain" refers to whether the dataset has a limited or unrestricted range of object coverage; "Automatic" indicates whether the dataset can be fully generated through automation without human intervention; and "Region-Focused" describes whether the dataset emphasizes detailed regions rather than the overall image.

Datasets	Open-Domain?	Automatic?	Region-Focused?	Size	Source	Target	Text
CUB-Bird [69]	×	×	×	11,788			*"This is a grey bird with a brown and yellow tail wing and a red head. (Select)"*
Spot-the-Diff [26]	×	×	×	13,192			*"The people in the parking lot are no longer there."*
Image-Edit-Request [62]	✓	×	×	3,939			*"Add a sword and a cloak to the squirrel."*
MagicBrush [79]	✓	×	×	10,388			*"Make the man ride a motorcycle."*
InstructPix2Pix [6]	✓	✓	×	UNLIMITED			*"Convert to a realistic photo."*
MJ-Bench [11]	×	✓	×	UNLIMITED			*"Young or Elder. (Select)"*
IMG-DIFF	✓	✓	✓	UNLIMITED			*"The difference is that the teapot in the right image is made of glass, whereas the teapot in the left image is made of porcelain."*

editing techniques to create image pairs. Instead of relying on human-generated annotations, they deploy high-performance VLMs or MLLMs to generate annotations. As a result, the data size is effectively limitless. However, relying on MLLMs for annotation means that these data only describe differences across the entire image. Yet, image pairs generated through image editing involve variations across multiple detailed regions. If the description only describes overall image differences, it may miss important details in fine-grained regions, resulting in inaccuracy.

Unlike the previous datasets, our IMG-DIFF dataset not only employs an automated generation pipeline but also incorporates a segmentation process to identify and capture detailed regions, which are then targeted for precise annotation. Additionally, we employ extensive filtering processes to ensure high data quality. These measures enable our dataset to achieve more comprehensive and accurate difference captions.

9.2. Performance Comparison

In this section, we compare the performance of our IMG-DIFF dataset with existing image difference datasets. We apply two primary dataset configurations for this comparison: the first is the CLEVR-Change [51] dataset, containing 67,600 examples. CLEVR-Change generates random 3D environments with blocks of various shapes, colors, sizes, and positions, which are subsequently altered to create image difference data. The second configuration combines the Spot-the-Diff dataset and the Image-Edit-Request dataset, totaling 13,614 samples.

We conduct experiments on three distinct MLLMs: LLaVA-1.5-7B, MGM-7B, and InternVL2-8B. Specifically,

Table 5. Performance of image difference datasets CLEVR-Change, Image-Edit-Request & Spot-the-Diff, and our Img-Diff dataset on MMVP and 8 MLLM benchmarks.

Model	VQA^{v2}	GQA	POPE	MMB	MMB^{CN}
LLaVA-1.5-7B	78.5	62.0	85.9	64.3	58.3
LLaVA-1.5-7B + CLEVR	79.2	63.1	85.7	65.9	59.2
LLaVA-1.5-7B + ImageEdit + Spot	79.3	63.3	86.4	65.8	58.9
LLaVA-1.5-7B + RP(main page)	79.3	62.8	86.4	66.1	59.8
MGM-7B	80.4	62.6	86.0	69.3	58.9
MGM-7B + ImageEdit + Spot	79.7	61.2	86.8	69.1	62.8
MGM-7B + RP(main page)	80.7	62.7	86.2	68.7	59.6
InternVL2-8B-FT	81.8	62.6	87.7	82.5	81.5
InternVL2-8B + ImageEdit + Spot	81.5	62.0	87.0	82.1	79.8
InternVL2-8B + RP(main page)	81.8	62.6	88.0	82.7	81.4

Model	MM-Vet	SQA^{I}	SEED	△	MMVP
LLaVA-1.5-7B	30.5	66.8	58.6	-	24.0
LLaVA-1.5-7B + CLEVR	29.8	68.0	61.2	+1.30%	28.7
LLaVA-1.5-7B + ImageEdit + Spot	30.5	68.3	61.9	+1.87%	25.3
LLaVA-1.5-7B + RP(main page)	33.2	68.2	61.7	+3.06%	27.3
MGM-7B	40.8	70.6	63.5	-	40.0
MGM-7B + ImageEdit + Spot	41.7	69.3	61.8	-0.19%	39.3
MGM-7B + RP(main page)	44.1	71.7	63.2	+1.28%	50.7
InternVL2-8B-FT	49.2	96.5	69.5	-	38.7
InternVL2-8B + ImageEdit + Spot	51.1	96.8	68.3	-0.28%	40.7
InternVL2-8B + RP(main page)	52.6	96.6	69.9	+1.01%	43.3

Table 6. Performance of image difference datasets CLEVR-Change, Image-Edit-Request & Spot-the-Diff, and our Img-Diff dataset on image difference benchmarks.

Model	Spot-the-Diff			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-7B	8.5	12.0	38.3	30.1
LLaVA-1.5-7B + CLEVR	9.3	12.3	45.2	30.2
LLaVA-1.5-7B + ImageEdit + Spot	9.1	12.9	40.8	30.5
LLaVA-1.5-7B + RP(main page)	9.7	13.0	43.2	30.8
MGM-7B	9.9	12.0	46.3	31.5
MGM-7B + ImageEdit + Spot	7.3	10.3	36.9	28.5
MGM-7B + RP(main page)	10.8	13.1	53.5	33.0
InternVL2-8B-FT	6.6	11.7	26.5	27.3
InternVL2-8B + ImageEdit + Spot	5.7	12.5	24.2	27.8
InternVL2-8B + RP(main page)	8.4	12.8	32.2	28.5

Model	Image-Edit-Request			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-7B	15.1	17.8	60.6	45.2
LLaVA-1.5-7B + CLEVR	15.1	17.8	60.9	45.2
LLaVA-1.5-7B + ImageEdit + Spot	13.0	18.4	56.6	44.7
LLaVA-1.5-7B + RP(main page)	16.2	19.5	60.9	46.7
MGM-7B	16.5	17.7	66.8	44.8
MGM-7B + ImageEdit + Spot	13.9	17.1	55.1	42.5
MGM-7B + RP(main page)	16.6	18.2	68.1	45.7
InternVL2-8B-FT	12.4	14.1	51.5	38.9
InternVL2-8B + ImageEdit + Spot	12.0	16.3	49.0	39.3
InternVL2-8B + RP(main page)	12.5	14.2	56.0	39.4

the first dataset configuration is exclusively tested on LLaVA-1.5-7B, while the second dataset configuration is evaluated across all three MLLMs. To assess performance variation, we incorporate each dataset into MLLMs' fine-tuning data respectively, then fine-tune MLLMs and measure their performance. The results are presented in Table 5 and Table 6.

The tables show that incorporating the CLEVR-Change, Image-Edit-Request, and Spot-the-Diff datasets into the fine-tuning of LLaVA-1.5-7B leads to performance improvements on MLLM benchmarks and image difference benchmarks. However, the performance boost from our Img-Diff dataset is more substantial. Furthermore, introducing the Image-Edit-Request and Spot-the-Diff datasets into MGM-7B and InternVL2-8B leads to noticeable performance degradation.

These results could be attributed to the fact that our dataset's text is specifically generated in the format for instruction-following tasks, which provides a greater benefit for MLLMs. In contrast, existing image differencing benchmarks do not adhere to this format, thereby introducing noise that adversely affects the performance of MLLMs. Moreover, our dataset places more emphasis on the image differences in detailed regions, which enhances the model's ability to capture fine-grained details, thereby improving its overall VQA capabilities and image difference recognition performance more effectively.

10. Prioritizing Quality Over Quantity

10.1. Discussion on Data Quantity for MLLMs

The quality of data is generally more important than its quantity in the domain of MLLMs. As demonstrated by LLaVA-1.5, it uses only a small subset of InstructBLIP's data [14], supplemented with a few small-sized VQA datasets (pre-training data reduced from 129M to 558K, fine-tuning data reduced from 1.2M to 665K), achieving impressive performance and significantly surpassing those of InstructBLIP. Furthermore, a series of MLLM studies [12, 32, 37, 43] validate that enhancing MLLMs' performance requires high-quality task-oriented data rather than merely increasing the volume of data.

The data volume for testing in our paper (13K and 35K) is comparable to that of many mainstream MLLM task-specific datasets, such as AI2D (12K), DocVQA (10K), ChartQA (18K), and OKVQA (9K). Despite not incorporating a large amount of data, our dataset brings appreciable performance improvement to MLLMs with modest training costs, such as elevating MGM-7B from 40 points to 50.7 on the MMVP benchmark, in which the GPT-4V gains a score of 38.7.

Considering the marginal benefits and training costs (fine-tuning 7B MLLMs on 4 A100 would take an additional 2 hours for every extra 50K samples), we aim to pursue a dataset that is small in quantity but high in quality. Recently, more works demonstrate the feasibility of high quality and high data efficiency of visual-text synthesis [74, 83]. Additionally, our paper emphasizes a synthesis method rather than the dataset itself. We can generate any

amount of Img-Diff data as needed, as our dataset is generated automatically.

10.2. Expanding Dataset Does Not Yield Linear Performance Gains

In addition to the 13K "object replacement" samples generated using MSCOCO captions on the main page, we also apply the same process and filtering thresholds to generate 34,583 samples using the captions from the LLaVA pretraining dataset. We compare the MLLM fine-tuned with the 13K samples to the one fine-tuned with the current fourfold larger dataset, aiming to explore the mathematical relationship between dataset expansion and model performance gains. The results are shown in Table 7 and Table 8.

Table 7. Performance comparison on MMVP and 8 MLLM benchmarks (including 35K "object replacement" samples).

Model	VQA^{v2}	GQA	POPE	MMB	MMB^{CN}
LLaVA-1.5-7B	78.5	62.0	85.9	64.3	58.3
LLaVA-1.5-7B + RP(13K)	79.3	62.8	86.4	66.1	59.8
LLaVA-1.5-7B + RP(13K) + RP(35K)	79.2	63.1	86.2	66.9	59.2

Model	MM-Vet	SQA^{I}	SEED	△	MMVP
LLaVA-1.5-7B	30.5	66.8	58.6	-	24.0
LLaVA-1.5-7B + RP(13K)	33.2	68.2	61.7	+3.06%	27.3
LLaVA-1.5-7B + RP(13K) + RP(35K)	33.3	69.0	62.2	+3.40%	31.3

Table 8. Performance comparison on Spot-the-Diff and Image-Edit-Request (including 35K "object replacement" samples).

Model	Spot-the-Diff			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-7B	8.5	12.0	38.3	30.1
LLaVA-1.5-7B + RP(13K)	9.7	13.0	43.2	30.8
LLaVA-1.5-7B + RP(13K) + RP(35K)	**9.8**	**13.1**	**45.3**	**31.0**

Model	Image-Edit-Request			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-7B	15.1	17.8	60.6	45.2
LLaVA-1.5-7B + RP(13K)	16.2	**19.5**	60.9	46.7
LLaVA-1.5-7B + RP(13K) + RP(35K)	**16.4**	19.1	**65.5**	**46.8**

We observe that the average performance gain on the MLLM benchmarks has become 3.40%, while the performance gain from the previous Img-Diff dataset was 3.06%. On the MMVP benchmark, the model fine-tuned with more data achieves further improvement, raising its score from 27.3, obtained with 13K samples, to the current score of 31.3. Furthermore, on Spot-the-Diff and Image-Edit-Request, the additional data also contributes to further performance gains. These results indicate that a moderate increase in data size can further enhance model performance.

Although adding more data can improve the MLLM's performance, it is worth noting that while we quadruple the dataset, the performance improvements do not increase by a factor of four. This aligns with the fact that the relationship between the data size and performance gains is not linear. As we increase the amount of similar data, the performance gains eventually reach a maximum limit. For future work, further investigation can be conducted into the relationship between different data volumes and performance improvements under the same filtering threshold.

11. Expanding Diversity with Lexicons

On the main page, beyond the intrinsic diversity of object names within the caption database, we increase the temperature of the LLM used for object name substitution to enhance the randomness of model outputs. This helps us expand the range of object categories covered by our dataset. Additionally, we experiment with randomly selecting nouns from an object name lexicon to replace original object names in captions, further enriching the dataset's diversity. This section provides a detailed explanation of this "Expanding Diversity with Lexicons" method and the experimental results on LLaVA-1.5-7B.

To construct the object name lexicon, we initially use the NLTK tool to filter all nouns from the WordNet lexicon. Next, we categorize each word based on its synsets entries, labeling them accordingly. Finally, we select object names classified under "machine," "living_thing," "natural_object," "fruit," "vehicle," "container," "clothing," "fixture," "appliance," "furniture," or "food" and form the final object name lexicon. The resulting lexicon comprises 5,526 distinct object names.

Following this, as described on the main page, we generate a test dataset using MSCOCO captions. Specifically, we replace object names in MSCOCO captions randomly with nouns of the same category from the object name lexicon, forming caption pairs that are later used for further generation and filtering processes. This approach resulted in 8,930 high-quality "object replacement" samples. We utilize this data to fine-tune LLaVA-1.5-7B, obtaining the results shown in Table 9 and Table 10.

Table 9. Performance comparison on MMVP and 8 MLLM benchmarks (using data generated with lexicons).

Model	VQA^{v2}	GQA	POPE	MMB	MMB^{CN}
LLaVA-1.5-7B	78.5	62	85.9	64.3	58.3
LLaVA-1.5-7B + RP(main page)	79.3	62.8	86.4	66.1	59.8
LLaVA-1.5-7B + RP(lexicon)	79.2	62.7	86.3	66.2	59.4

Model	MM-Vet	SQA^{I}	SEED	△	MMVP
LLaVA-1.5-7B	30.5	66.8	58.6	-	24.0
LLaVA-1.5-7B + RP(main page)	33.2	68.2	61.7	+3.06%	27.3
LLaVA-1.5-7B + RP(lexicon)	32.2	68.8	61.8	+2.67%	30.0

As shown in Table 9 and Table 10, the current dataset still provides significant performance improvements for LLaVA-1.5-7B. Specifically, the fine-tuned MLLM achieves comprehensive performance improvement across

Table 10. Performance comparison on Spot-the-Diff and Image-Edit-Request(using data generated with lexicons).

Model	Spot-the-Diff			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-7B	8.5	12.0	38.3	30.1
LLaVA-1.5-7B + RP(main page)	**9.7**	**13.0**	**43.2**	**30.8**
LLaVA-1.5-7B + RP(lexicon)	8.9	12.2	41.9	29.9

Model	Image-Edit-Request			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-7B	15.1	17.8	60.6	45.2
LLaVA-1.5-7B + RP(main page)	**16.2**	**19.5**	**60.9**	46.7
LLaVA-1.5-7B + RP(lexicon)	13.9	19.4	60.4	**46.9**

eight MLLM benchmarks, with improvement levels comparable to those on the main page, resulting in an average performance increase of 2.67%. Besides, the current dataset also improves the performance of LLaVA-1.5-7B on image difference benchmarks.

By using a lexicon for object name replacement, we can more effectively enhance the diversity of our Img-Diff dataset. Specifically, we can increase the number of noun samples included in the lexicon, as well as perform multiple rounds of noun replacement on the same caption. As a result, the quality of our data can be further improved.

12. Performance Based on Contrastive Chain-of-Thought

In addition to the standard VQA evaluation, we also assess our IMG-DIFF dataset using the Contrastive Chain-of-Thought (CoCoT [77]) method. This evaluation method involves prompting the model with the instruction, "Please identify the similarities and differences between these two images," and requiring the MLLM to pinpoint the differences before it answers the final VQA question. The differences identified are then used as context-enhanced text to support its own response to the VQA task.

Table 11. Results on MMVP using the CoCoT method.

Model	MMVP	
	w/ CoCot	w/o CoCot
LLaVA-1.5-7B	24.0	22.0
LLaVA-1.5-7B + RP	27.3	29.0

We test the original LLaVA-1.5-7B and our fine-tuned model on the MMVP benchmark using CoCoT. As shown in Table 11, the original model's score drops from 24 to 22, while the score of the model fine-tuned with our data rises from 27.3 to 29. This indicates that, after fine-tuning without IMG-DIFF data, the MLLM demonstrates an enhanced ability to recognize image differences and can generate more accurate descriptive information to support VQA tasks.

13. Testing on MLLMs at Different Scales

On the main page, we primarily conduct experiments on MLLMs with a 7B scale. In this section, we will explore the impact of our dataset on models of different sizes. Specifically, we fine-tune LLaVA-1.5-13B and InternVL2-1B, representing a larger and a smaller model. We then test these models on both MLLM benchmarks and image difference benchmarks.

Table 12. Performance of "object replacement" data on LLaVA-1.5-13B and InternVL2-1B (evaluated on MMVP and 8 MLLM Benchmarks).

Model	VQA^{v2}	GQA	POPE	MMB	MMB^{CN}
LLaVA-1.5-13B	80.0	63.3	85.9	67.7	63.6
LLaVA-1.5-13B + RP	80.3	64.1	86.6	69.2	63.2
InternVL2-1B-FT	77.3	60.2	86.6	68.6	60.7
InternVL2-1B + RP	77.4	60.2	87.1	69.0	60.7

Model	MM-Vet	SQA^{I}	SEED	△	MMVP
LLaVA-1.5-13B	35.4	71.6	61.6	-	24.7
LLaVA-1.5-13B + RP	37.4	71.7	62.9	+1.49%	32.0
InternVL2-1B-FT	31.9	88.5	61.4	-	16.0
InternVL2-1B + RP	33.4	88.7	61.7	+0.84%	18.0

Table 13. Performance of "object replacement" data on LLaVA-1.5-13B and InternVL2-1B (evaluated on Spot-the-Diff and Image-Edit-Request).

Model	Spot-the-Diff			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-13B	9.7	12.3	44.6	31.0
llava-1.5-13b + RP	**9.9**	**13.1**	**45.8**	**31.4**
InternVl2-1B-FT	6.5	11.4	24.7	26.5
InternVl2-1B + RP	6.9	11.5	25.7	26.5

Model	Image-Edit-Request			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-13B	**16.6**	18.0	62.9	46.2
llava-1.5-13b + RP	15.9	**20.1**	**65.3**	**47.2**
InternVl2-1B-FT	7.3	11.6	28.7	35.3
InternVl2-1B + RP	8.4	12.1	30.3	37.4

Table 12 and Table 13 show that our dataset remains effective on LLaVA-1.5-13B and InternVL2-1B, delivering comprehensive performance improvements across eight MLLM benchmarks and the image difference benchmarks. This demonstrates the versatility of our dataset, proving its capability to enhance model performance not only for 7B-scale models but also for smaller or larger models.

14. Top-Performing MLLMs in Image Difference Detection

In this section, we use our generated data to construct an evaluation benchmark to assess the image difference detection capabilities of top-performance MLLMs. Specifically, we compile a new evaluation set consisting of 500 samples (drawn from the version of our IMG-DIFF dataset not used for training) and use our dataset's annotations as the ground truth. We evaluate the ability of two leading MLLMs, InternVL2.5-8B-MPO [73] and Qwen2-VL-7B-Instruct [71], to identify fine-grained differences in specific regions of image pairs. To quantify their performance, we employ InternLM2.5 [7] to assess the alignment between the models' outputs and the annotations, scoring the results based on similarity (on a scale of 0 to 3, with a maximum possible total score of 1500).

Table 14. Performance scores of top-performing MLLMs in image difference detection.

	InternVL2.5-8B-MPO	Qwen2-VL-7B-Instruct
$D_{test500}$	836	695
	InternVL2-8B	**InternVL2-8B + RP(main page)**
$D_{test500}$	620	727

The results, as shown in Table 14, reveal that both two top-performing MLLMs struggle to recognize fine-grained differences, achieving scores of only 836 and 695, respectively. This underscores that current SOTA MLLMs have not been specifically optimized for image difference detection tasks, indicating significant room for improvement in this capability.

Additionally, we evaluate the performance of InternVL2-8B and its fine-tuned variant (trained on our IMG-DIFF dataset) on the newly constructed evaluation set. The results, presented in Table 14, show that our dataset effectively enhances the models' performance on image difference detection, highlighting its practical value for advancing this capability.

15. Unnatural Images in the Dataset

Due to the limitations in the generation quality of SDXL and Prompt2Prompt, the production of some unnatural images is inevitable during the creation of our IMG-DIFF dataset. Nevertheless, our work goes beyond the current dataset, emphasizing a novel data synthesis method. Our pipeline allows for the substitution of more advanced text-to-image models and image editing techniques to improve image quality. Additionally, we can also propose the incorporation of filters specifically targeting unnatural images to mitigate their impact.

To investigate the influence of unnatural images on the performance of the fine-tuned models, we conduct a new experiment. We filter and remove unnatural images from the 13K IMG-DIFF dataset and subsequently fine-tune LLaVA-1.5-7B on the remaining data to observe the remaining performance (we have open-sourced the new filter in our code repository). Specifically, we deploy InternVL2.5-8B to quantify unnatural images in IMG-DIFF, identifying 28% as unnatural. The model is then trained on the filtered dataset, and the results are presented in Table 15 and Table 16.

Table 15. Performance of LLaVA-1.5-7B fine-tuned on IMG-DIFF after removing unnatural images (evaluated on MMVP and 8 MLLM Benchmarks). "w/o UNI" means "fine-tuning without unnatural images".

Model	VQA^{v2}	GQA	POPE	MMB	MMB^{CN}
LLaVA-1.5-7B	78.5	62.0	85.9	64.3	58.3
LLaVA-1.5-7B + RP w/o UNI	78.5	62.0	87.0	67.0	59.4
LLaVA-1.5-7B + RP(main page)	79.3	62.8	86.4	66.1	59.8
Model	MM-Vet	SQA^{I}	SEED	Δ	MMVP
LLaVA-1.5-7B	30.5	66.8	58.6	-	24.0
LLaVA-1.5-7B + RP w/o UNI	31.1	68.6	60.3	+1.87%	24.3
LLaVA-1.5-7B + RP(main page)	33.2	68.2	61.7	+3.06%	27.3

Table 16. Performance of LLaVA-1.5-7B fine-tuned on IMG-DIFF after removing unnatural images (evaluated on Spot-the-Diff and Image-Edit-Request). "w/o UNI" means "fine-tuning without unnatural images".

Model	Spot-the-Diff			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-7B	8.5	12	38.3	30.1
LLaVA-1.5-7B + RP w/o UNI	8.6	12.8	38.3	30.3
LLaVA-1.5-7B + RP(main page)	9.7	13.0	43.2	30.8
Model	Image-Edit-Request			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-7B	15.1	17.8	60.6	45.2
LLaVA-1.5-7B + RP w/o UNI	15.4	18.0	59.6	45.6
LLaVA-1.5-7B + RP(main page)	16.2	19.5	60.9	46.7

Surprisingly, after removing the unnatural images and fine-tuning LLaVA-1.5-7B again, we observe that the performance of the resulting model is inferior to that of the model fine-tuned with unnatural images. This finding suggests that the presence of unnatural images does not necessarily degrade model performance and may, in fact, contribute positively to the training process.

16. Impact of our Dataset on Spatial Reasoning Performance

Spatial changes constitute a significant aspect of object variations in images [55]. In this section, we investigate whether our dataset can effectively enhance models' spatial reasoning capabilities. To this end, we newly evaluate our

Table 17. The impact of different filtering thresholds on the performance of our dataset.

Threshold	VQA^{v2}	GQA	POPE	MMB	MMB^{CN}	MM-Vet	SQA^{I}	SEED	△
LLaVA-1.5-7B	78.5	62.0	85.9	64.3	58.3	30.5	66.8	58.6	-
(1) IS 0.9-0.98 + BITM 0.3 + CS 0.9 + CITM 0.3	79.1	62.3	86.0	66.8	59.5	32.7	66.6	61.6	+2.42%
(2) IS 0.9-0.98 + BITM 0.35 + CS 0.9 + CITM 0.3	79.1	62.2	85.9	66.7	59.5	32.7	67.1	61.9	+2.52%
(3) IS 0.9-0.98 + BITM 0.35 + CS 0.85 + CITM 0.4	79.3	62.8	86.4	66.1	59.8	33.2	68.2	61.7	+3.06%
(4) IS 0.85-0.98 + BITM 0.35 + CS 0.85 + CITM 0.4	79.2	62.7	86.3	66.2	57.4	32.2	68.8	61.8	+2.24%

models on SpatialEval [70], a benchmark specifically designed to assess spatial reasoning abilities. Our evaluation focuses on two of its subsets: (1) Spatial-Map, which examines the understanding of spatial relationships between objects in map-based scenarios, and (2) Spatial-Real, which assesses real-world spatial understanding. The results are presented in Table 18.

Table 18. Performance comparison on SpatialEval.

	LLaVA-1.5-7B	**LLaVA-1.5-7B + RP(main page)**
Spatial-Map-ACC	0.26	**0.29**
Spatial-Real-ACC	0.39	**0.41**
	MGM-7B	**MGM-7B + RP(main page)**
Spatial-Map-ACC	0.38	**0.41**
Spatial-Real-ACC	0.42	**0.46**
	InternVL2-8B	**InternVL2-8B + RP(main page)**
Spatial-Map-ACC	0.47	**0.49**
Spatial-Real-ACC	0.44	**0.52**

As shown in Table 18, our findings reveal that IMG-DIFF indeed contributes to improving models' spatial reasoning performance. Moreover, it is important to note that our current data synthesis pipeline is not explicitly designed to incorporate spatial transformations. In the future work, we will integrate data synthesis strategies that specifically address spatial transformations, further enhancing the dataset's utility for spatial reasoning tasks.

17. Ablation Studies

To investigate the impact of filtering thresholds on our data performance, we set different filtering thresholds and generate various versions of our "object replacement" dataset. We then finetune multiple versions of LLaVA-1.5-7B using these datasets and evaluate their performance on commonly used MLLM benchmarks. Specifically, the threshold for the Image Similarity Filter of the Difference Area Generator is abbreviated as **IS** (Image Similarity). The threshold for the Image-Text Matching Filter of the Difference Area Generator is abbreviated as **BITM** (Bounding Box Image-Text Matching). The threshold for the Caption Similarity Filter of the Difference Captions Generator is abbreviated as **CS** (Captions Similarity). The threshold for the Image-Text Matching Filter of the Difference Captions Generator is abbreviated as **CITM** (Captions Image-Text Matching). The evaluation results are shown in Table 17.

Image Similarity (IS) Based on Table 17, Model (3) adjusts the IS threshold from 0.9-0.98 to 0.85-0.98 compared to Model (4), reducing the filtering intensity for the similarity of image pairs. This adjustment leads to a significant performance decline, indicating that the similarity of image pairs has a substantial impact on data quality. When the similarity is low, the data generation process may introduce more ineffective instances, as segmentation could generate more areas unrelated to the valid objects (i.e., the replaced or replacing objects).

Bounding Box Image-Text Matching (BITM) Model (2), compared to Model (1), increases the BITM threshold, meaning that when filtering to obtain valid bounding boxes, only those more likely to contain valid objects are retained. After raising the threshold, slight improvements in model performance are observed, which demonstrates that only bounding boxes more related to the replaced or replacing objects should be retained.

Captions Similarity (CS) and Captions Image-Text Matching (CITM) Model (3) increases both the CS threshold and the CITM threshold compared to Model (2). Raising the CS threshold implies a greater filtering strength for similar captions, which means that if the two objects corresponding to the same bounding box coordinate in an image pair are similar, the bounding box will be filtered out. As for the CITM threshold, increasing the CITM threshold aims to enhance the alignment between the captions and the objects being described. After raising both the CS and CITM thresholds, the model's performance shows a significant improvement.

Based on Table 17, it can be concluded that the stronger the filtering intensity, the better our dataset's effectiveness. However, due to the increased filtering intensity resulting in a reduced number of final instances, we choose the settings of Model (3) as our optimal threshold to ensure a sufficient number of generated instances. In our future work, we will expand the data sources to generate more pairs of similar images and then evaluate the effects of data obtained with higher filtering intensity.

18. Additional Details of Experiments

18.1. Preprocessing of image pairs before inputting into MLLMs during training and inference

The MLLMs selected in our paper (LLaVA-1.5, MGM, InternVL2) only support single-image input. Therefore, our image pairs need to be horizontally concatenated before being fed into MLLMs' image encoder. Specifically, we horizontally concatenate the images in pairs and add a vertical black dividing line, 20 pixels wide, between the images.

18.2. Training Process for MLLMs

The training process for advanced MLLMs, including LLaVA-1.5, MGM and InternVL2, typically involves two stages: the pre-training stage and fine-tuning stage. During the pre-training stage, the MLLMs keep the backbone LLM and the vision encoder frozen and zero-initialize the learnable projector which is used for semantic mapping and cross-modality alignment. Only the projector is trained using the pre-training dataset. In the fine-tuning stage, we unfreeze the backbone LLM and fine-tune both the backbone LLM and the learnable projector using the visual instruction tuning dataset. Specifically, the pre-training dataset is usually an image captioning dataset, while the visual instruction tuning dataset typically consists of VQA datasets for various tasks. Thus, our Img-Diff dataset is integrated into the visual instruction tuning dataset during the fine-tuning stage and used together with the original dataset to fine-tune the MLLMs.

18.3. Model Selection

18.3.1. Overview

The models used in our project are among the best-performing ones identified for the tasks assigned to them. Besides, they are interchangeable. Therefore, if better model options become available, researchers can replace the current models with those that offer superior performance to achieve a more effective dataset.

18.3.2. Selection of the Semantic Segmentation Model

In our project, we need to use a semantic segmentation model to identify regions containing objects in images. To ensure a diverse range of object categories is covered, we opt for models like SAM [30] instead of traditional semantic segmentation models. Furthermore, to reduce time consumption, we select FastSAM, one of the most efficient and effective models within the SAM-like category, as our segmentation model.

18.3.3. Model Size

Considering the device limitation and time consumption, our paper utilizes the LLM Vicuna-1.5-13B [13] for object name replacement in the image pairs generation process. For semantic segmentation in the Difference Area Generator, the FastSAM-x model is employed. For the CLIP model, we choose "clip-vit-base-patch32", and for the BLIP model, we select "blip-itm-large-coco". In the Difference Captions Generator, we use the MLLM LLaVA-NEXT-13B to generate content captions and difference captions. These models are interchangeable. When resources allow, researchers can substitute them with higher-performance models to achieve datasets with improved performance.

18.4. Filtering Thresholds

During the generation process of "object replacement" data, we employ multiple filtering operations. In this subsection, we will outline the filtering thresholds we use.

In the Difference Area Generator, we use FastSAM to perform semantic segmentation on images and obtain bounding box information for regions where objects might be present. To ensure we gather a sufficient number of candidate regions, we set the confidence score threshold to 0.05, which means that we consider a region to contain objects when its confidence score is greater than 0.05. Additionally, to prevent overlapping regions, we set the Intersection over Union (IoU) threshold to 0.5.

At the beginning stage of the Difference Area Generator, before using FastSAM for segmentation, we employ the Image Similarity Filter to retain only those with similarity between 0.9 and 0.98. This ensures that the image pairs are highly similar but not identical.

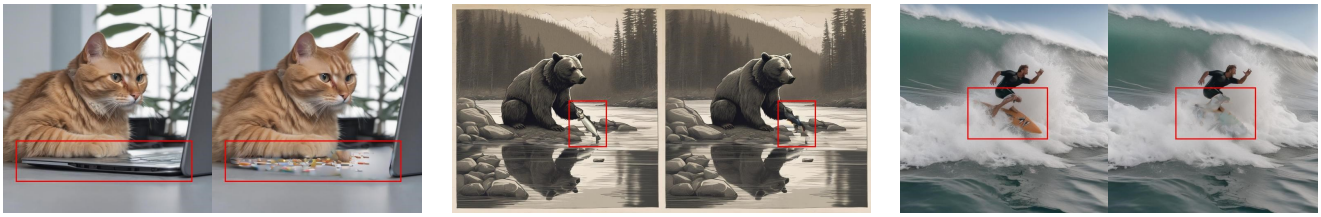
In the Difference Detector stage of the Difference Area Generator, after cropping sub-images based on the bounding box information, we use the Image Similarity Filter to filter the sub-image pairs and consider them to be different only when the similarity score is less than 0.85.

In the mid-stage of the Difference Area Generator, after performing sub-image cropping based on the bounding box information, we use the Image-text Matching Filter to determine whether these sub-images contain valid objects. When the score exceeds 0.35, we consider the sub-image to contain valid objects, and the bounding box is deemed effective.

In the Difference Area Generator, after obtaining all effective bounding boxes, we use the IoU method to filter out the overlapping ones. We set the IoU threshold to 0.5, retaining only the bounding boxes with a higher degree of difference for similar positions.

In stage 1 of the Difference Captions Generator, after cropping the images into sub-images and generating content captions, we use the Image-text Matching Filter to evaluate the matching degree between the sub-images and the captions. We only consider a caption to be correct if the image-text matching score exceeds 0.4.

In stage 1 of the Difference Captions Generator, we use the Captions Similarity Filter to determine whether the two



The cat is sitting on the laptop. Fish in bear's hand. The surfboard is orange.

Figure 7. Three "object removal" examples.

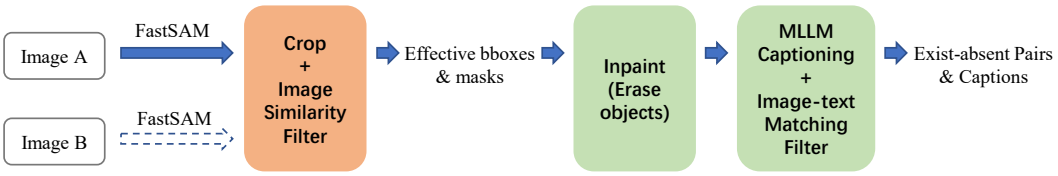


Figure 8. An overview of the generation steps for "object removal" data.

content captions of an image pair, describing the regions of the same bounding box, are different. We use CLIP to obtain text features for the two captions and then calculate the cosine similarity between them. When the cosine similarity is below 0.85, we consider the two captions to be different.

Setting the filtering intensity too high may lead to a reduced number of remaining samples. To ensure that the dataset still has enough samples after filtering, we outline adjustable thresholds as described above. As mentioned in Section 17, higher filtering intensity typically results in better model performance. Therefore, researchers may consider expanding the data sources and increasing the filtering intensity to improve dataset performance.

18.5. Resource and Time Consumption

With four NVIDIA A100 GPUs, it took 4.5 days to synthesize 118K high-quality image pairs. The subsequent filtering and description-generating processes took approximately two days in total.

19. The "Object Removal" Exploration

19.1. Overview

On the main page, we generate pairs of similar images focusing on object replacement. Their bounding box regions generally contain objects. However, the ability to determine the object's presence is also crucial. Thus, we generate another set of image pairs where the difference lies in the presence or absence of objects, to enhance the model's ability to determine object presence. We refer to these image pairs as "exist-absent pairs" and the data as "object removal" data.

19.2. Generation Process

19.2.1. Workflow

"Object removal" involves erasing a specific object from an image and then merging the edited image with the original to form an exist-absent pair. The detailed workflow is as follows: first, FastSAM is used to segment the image, which provides a set of bounding boxes and masks. Next, an Image Similarity Filter is applied to filter the bounding boxes and accompanying masks, keeping only those that contain objects. Then, we use the text-to-image generative model SDXL-turbo[57] to inpaint the images with the remaining masks, erasing specific objects from the images and generating exist-absent pairs. Next, we use an MLLM to describe the removed object for each exist-absent pair, and a filter is employed to verify the accuracy of the description. Finally, we draw red boxes on images based on the bounding box information, and then the object descriptions are converted into multiple-choice questions, such as: "*which image has the object related to 'DESCRIPTION' within the red bounding box? A. the left image B. the right image.*" Here, *DESCRIPTION* refers to the description of the erased objects. After all processing and filtering, we obtain 5,773 pieces of "object removal" data. The general framework is shown in Figure 8.

19.2.2. Image Similarity Filter

In the current process, the function of the Image Similarity Filter is to filter out the bounding box regions that do not contain objects. For each image, we need its corresponding image in the image pair generated in Section 3.2 to de-

termine whether its bounding box regions contain objects. Since the image pairs are generated by replacing objects, the difference areas between the two images are highly likely to be the regions containing valid objects. Therefore, for each bounding box, we crop the sub-images from image A (the current image) and image B (the other image in the pair), and then calculate the similarity of these two sub-images. When the similarity is below 0.9, we consider these two sub-images to be different, indicating that the bounding box region contains an object.

19.2.3. Erase Objects

We use the generative model SDXL-turbo[57] to erase objects based on the masks obtained during segmentation. The prompt is "background, nothing, 8k." After inpainting, the object in the masked regions is erased, while the rest of the image remains unchanged. Hence, we obtain exist-absent pairs.

19.2.4. MLLM Captioning

We use the MLLM LLaVA-NEXT to generate descriptions for the erased objects. Specifically, we provide the MLLM with the bounding box coordinates and ask it to describe the corresponding area in the original image. Subsequently, we crop the exist-absent pairs based on the bounding box information and then use an Image-Text Matching Filter to assess the matching degree between the sub-images and the descriptions. If the matching score between the sub-image containing objects and its description is greater than 0.35, and the matching score between the sub-image not containing objects and its description is less than 0.2, we consider the description to be accurate and the exist-absent pair to be valid.

19.3. Evaluation

We merge the "object removal" data with the "object replacement" data, making our dataset focus on both object changes and object presence. To test the performance changes of LLaVA-1.5-7B after adding "object removal" data, we incorporate this combined data into the original visual instruction tuning dataset of the MLLM and conduct fine-tuning. Then, we evaluate the fine-tuned model on image difference benchmarks and eight MLLM benchmarks, similar to what is presented on the main page.

In the tables, "RM" represents "object removal" data.

19.3.1. Results on MLLM Benchmarks

Table 19 shows the performance of LLaVA-1.5-7B fine-tuned with additional "object removal" data on commonly used MLLM benchmarks. With the assistance of "object removal" data, LLaVA-1.5-7B achieves further improvements across various benchmarks compared to the model that only uses "object replacement" data, with an average increase of 3.91%.

Table 19. Performance comparison on 8 MLLM benchmarks (including "object removal" data).

Model	VQA^{v2}	GQA	POPE	MMB	MMB^{CN}
LLaVA-7B	78.5	62.0	85.9	64.3	58.3
LLaVA-7B + RP	**79.3**	62.8	86.4	66.1	59.8
LLaVA-7B + RP + RM	79.2	**62.9**	**86.8**	**67.9**	**61.3**

Model	MM-Vet	SQA^{I}	SEED	△	MMVP
LLaVA-7B	30.5	66.8	58.6	-	24.0
LLaVA-7B + RP	**33.2**	68.2	61.7	+3.06%	27.3
LLaVA-7B + RP + RM	33.1	**68.8**	**61.9**	+3.91%	**28.7**

Table 20. Results on image difference benchmarks (including "object removal" data).

Model	Spot-the-Diff			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-7B	8.5	12.0	38.3	30.1
LLaVA-1.5-7B +RP	9.7	**13.0**	43.2	30.8
LLaVA-1.5-7B +RP +RM	**9.8**	**13.0**	**46.5**	**31.5**

Model	Image-Edit-Request			
	BLEU	METEOR	CIDEr-D	ROUGE-L
LLaVA-1.5-7B	15.1	17.8	60.6	45.2
LLaVA-1.5-7B +RP	16.2	**19.5**	60.9	**46.7**
LLaVA-1.5-7B +RP +RM	**16.8**	18.6	**63.9**	45.7

19.3.2. Results on Image Difference Benchmarks

Table 20 shows the performance of LLaVA-1.5-7B fine-tuned with our "object removal" data on image difference benchmarks. With "object removal" data, LLaVA-1.5-7B shows further improvements in its performance on both the MMVP benchmark and the Spot-the-Diff benchmark, surpassing the results achieved with "object replacement" data alone. Besides, its scores fluctuate on the Image-Edit-Request benchmark.

19.3.3. Further Analysis

The results indicate that the "object removal" data has a comprehensive positive impact on LLaVA-1.5-7B, leading to performance improvements in both MLLM benchmarks and image difference benchmarks. However, during our analysis of sample quality, we notice that some of the generated "object removal" samples exhibit subpar image quality, with certain image pairs showing inadequate object removal effects. In light of this, employing a more robust inpainting model or applying additional filters to enhance the quality of these image pairs could further optimize the performance of this dataset.

20. Examples

{"**conversations**": [{**"from": "human", "value":** "Analyse the left image and the right image (separated by the black vertical bar). What is the difference between the red bounding box area in each image? Answer the question in a few concise sentences."},
{"from": "gpt", "value": "The left image shows a truck in the foreground with cows in the background, while the right image shows a tractor in the foreground with cows in the background. The red bounding box in the left image highlights the front of a truck, while the red bounding box in the right image highlights a yellow tractor. The difference is the type of vehicle highlighted in the bounding boxes."}],
"**bbox**": "[0.86, 0.41, 0.98, 0.54]",
"**captions1**": "The front of a truck.",
"**captions2**": "A yellow tractor.",
"**path**": "./filtered_new_edit_data/2_2311_0"}

{"**conversations**": [{**"from": "human", "value":** "Analyse the left image and the right image (separated by the black vertical bar). What is the difference between the red bounding box area in each image? Answer the question in a few concise sentences."},
{"from": "gpt", "value": "The left image shows a woman holding a black camera, while the right image shows the same woman holding a stack of white plates. The difference is that the woman is now holding plates instead of a camera."}],
"**bbox**": "[0.40, 0.07, 0.64, 0.18]",
"**captions1**": "A black camera.",
"**captions2**": "A stack of white plates.",
"**path**": "./filtered_new_edit_data/1_10505_0"}

Figure 9. Examples of “object replacement” data, including the image pair and the text content in JSON format.

{"**bbox**": [0.63, 0.35, 0.77, 0.44],
"**conversations**": [{**"from": "human", "value":** "Analyse the left image and the right image (separated by the black vertical bar). Which image has the object related to \"A red frisbee.\" within the red bounding box?\nA. the left image\nB. the right image\nAnswer with the option's letter from the given choices directly."},
{**"from": "gpt", "value":** "B"}],
"**path**": "./inpaint/2_17718_img0_0_2"}

Figure 10. An example of “object removal” data, including the image pair and the text content in JSON format.