ViDi: A Benchmark for the Identification and Captioning of Visual Differences in Image Pairs

Anonymous ACL submission

Abstract

Recent advances in Multimodal Large Language Models (MLLMs) have brought forward capabilities such as handling multiple images or engage in multi-turn conversations involving images. In this paper, we present a challenging *spot the differences* benchmark for evaluating MLLMs, geared toward the identification and captioning of Visual Differences (ViDi)[1]. The benchmark corresponds to a test-only dataset with 200 pairs of images, along with human-annotated descriptions of visually noticeable differences for each pair. ViDi goes beyond the identification of single differences, challenging models to articulate changes using natural language, while pinpointing the subject of each difference and its absolute or relative location. Empirical results reveal that MLLMs are still in an early development stage, showing limited performance in *spot the differences* tasks.

1 Introduction

Both academia and industry have been developing ever more powerful Multimodal Large Language Models (MLLMs) at a very fast pace (Yin et al., 2024), together with new benchmarks and evaluation methods (Fu et al., 2024). MLLMs build on large language models with strong reasoning and language generation capabilities, although their visual perception is still limited (Tong et al., 2024; Zhang et al., 2024a). In the current setting, evaluating fine-detail scenarios and spatial reasoning abilities is highly important, so as to provide insights for future research, in the area.

In this paper, we focus on the problem of analyzing multiple visual inputs for differences, and producing visually grounded natural language outputs. Similarly to humans, MLLMs should have the ability to compare multiple visual inputs and process complex requests that focus on this comparison. The prevalent use of chat-based interfaces with multiple interactions presents a promising avenue for evaluating reasoning capabilities across multiple visual instances, a domain that remains largely unexplored in visual understanding research. In this context, we introduce a benchmark based on *spot the differences* puzzles, to systematically evaluate MLLM abilities related to discerning and describing salient visual differences between natural images. The evaluation focuses not only on identifying differences, but also on articulating them with precision in natural language, encompassing subject identification and spatial localization. Previous studies have significant limitations, by focusing on one difference in each pair of images, and/or by often relying on synthetic images. As a result, current models are not being evaluated to their full potential. To address this gap, we consider a more complex comparison problem, focusing on detecting multiple differences in natural image pairs that were specifically designed to challenge human perception. In summary, our main contributions are:

(1) We propose a new benchmark for the identification and captioning of multiple Visual Differences (ViDi), based on a test-only dataset with 200 *spot the differences* puzzles, which collectively contain 1,097 differences spanning diverse image aesthetics, objects, and difference types (Section 3).

(2) We perform a baseline evaluation on prominent MLLMs, showing a generalized poor performance on the ViDi benchmark (Section 4).

2 Related Work

The evaluation of models that couple vision and language has been the focus of many recent studies, e.g. developing benchmarks to test different capabilities (Thrush et al., 2022; Parcalabescu et al., 2022; Bitton-Guetta et al., 2023), and often relying on protocols based on visual question answering

[1]The dataset is available at https://anonymous.4open.science/r/ViDi-FC45/

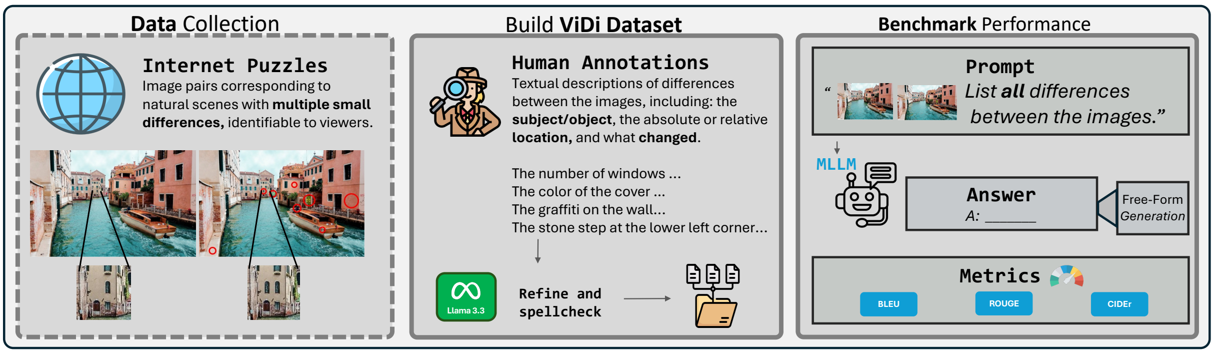

Figure 1: The proposed benchmark targeting the identification and captioning of Visual Differences (ViDi).

(Zhang et al., 2024b). However, most previous work has focused on inputs consisting of a single image, while the problem of comparing images to assess differences between them remains much less explored. Notable exceptions include benchmarks like Spot-the-Diff (Jhamtani and Berg-Kirkpatrick, 2018), CLEVR-Change (Park et al., 2019), CLEVR-Multi-Change (Qiu et al., 2021), Image-Editing-Request (Tan et al., 2019), DiffCap (Hu et al., 2024), or VisMin (Awal et al., 2024), which have been used in the assessment of specialized models focused on image change captions (Park et al., 2019). Appendix A presents a more detailed analysis of the aforementioned previous studies. Still, existing benchmarks feature important limitations, e.g. focusing on the analysis of a single difference, and/or including simplistic/synthetic images that lack aesthetic and object diversity. This paper advances ViDi as a more challenging benchmark, featuring natural image pairs with multiple differences between them, that were specifically designed to challenge human perception.

3 The ViDi Dataset

The ViDi benchmark dataset is designed to examine the capacity of models to interpret similar images and write textual descriptions for multiple subtle differences between them — see Figure 1. Our approach uses *spot the differences* puzzles collected from the Web, specifically created to challenge human visual capabilities, paired with manually curated descriptions of the differences.

In terms of **vision capabilities**, this benchmark increases the complexity in comparison to previous proposals, by using image pairs with multiple subtle differences between them, and ensuring that each variation represents a minor detail within the broader visual context. In terms of **linguistic capabilities**, the benchmark goes beyond the mere generation of grammatically correct sentences, aiming at contextually relevant and semantically coherent descriptions that align with visual inputs. The goal is to precisely identify and locate subjects, e.g. through spatial references, and articulate changes between images. With respect to **reasoning capabilities**, ViDi encompasses multiple cognitive dimensions. These include spatial reasoning to understand object relationships and positions, comparative reasoning to detect visual disparities between images, causal reasoning to infer the nature and rationale of the changes in the scene, or commonsense reasoning to clearly and unambiguously locate the changes. This multi-layered reasoning process demands both precise change detection and understanding of contextual and commonsense cues of significance within the broader scene.

We manually sourced 200 image pairs from the Web, originally in social media sites, professional news and entertainment websites, as well as amateur and hobbyist websites. All images feature natural scenes and we specifically avoided illustrations, although we also included some magazine-type cover posters. Images generated by synthetic processes were avoided as much as possible, and differences pertaining to scene-text in the images also seldom occur. The English annotations for the differences were, in most cases, derived from information available on the website from which the images were collected, but always further curated. The annotations were systematically organized in accordance with the position of the differences in the image, following a reading sequence from top-left to bottom-right in a zigzag manner, as is customary in Western writing. Each sentence describes one difference in the image pair, and conforms to the following guidelines: (a) humans can distinguish the difference in the two images; (b) the sentence should clearly identify the object/-subject of the difference, using simple and natural descriptions for human understanding; (c) the object should be unambiguously located within in the images, either globally (e.g., through object properties) or with respect to other objects (e.g., using spatial relations); and (d) the sentence should clearly describe what was changed (e.g., object color, size, number, etc.).

With a total of 1076 differences, each image pair in ViDi has 3 to 12 differences (average: 5.9). Individual difference descriptions are 3 to 27 words long (average: 12), and total descriptions per pair range from 49 to 140 words (average: 79).

When describing differences, the selection of the anchor for the relative location within the scene can be biased in the annotation process, and automated models can naturally select other anchors to locate the object. To at least partially avoid this type of bias, as well as other particularities of individual writing styles, the descriptions from a first annotation were rephrased through the use of a large language model, and then further curated through a subsequent round of revisions.

Although other existing benchmarks (Awal et al., 2024; Zhang et al., 2024b; Evennou et al., 2024) feature significantly larger datasets, ViDi prioritizes quality by providing 200 meticulously curated samples, ensuring a valuable and reasonable sample size. Despite being small, we argue that the ViDi benchmark supports the precise measurement and effective differentiation of systems with varying capacities, at a fine granularity.

4 Experiments and Results

Experimental Settings. To evaluate model performance, we consider two experimental settings. The *granular alignment* setting evaluates the compatibility of model predictions with individual ground-truth differences, focusing on the closest match. Models are given a pair of images and prompted to return a description of **one** difference. In turn, the *global alignment* setting assesses the similarity of the prediction with the combined and comprehensive context of all ground-truth descriptions for the differences. In this setting, the model is prompted to list **all** the differences in the pair of images. We evaluated seven open-weight MLLMs: MiniCPM-V-2.6 (Yao et al., 2024), Phi3.5 Vision (Abdin et al., 2024), LLaVA-CoT (Xu et al., 2024), NVLM-D-72B (Dai et al., 2024), LLAVA-OV-72B (Li et al., 2024), InternVL2.5-78B (Chen et al., 2024), and Llama-3.2-90B (Dubey et al., 2024). The prompts used to query the models are reported in Appendix B. Each model and setting uses standard metrics for evaluating text generation, namely BLEU (Papineni et al., 2002), ROUGE$_L$ (Lin, 2004), and CIDEr (Vedantam et al., 2015). Across all experiments, we maintained consistent text generation parameters. To account for the probabilistic nature of text generation, we apply sampling with a temperature set at 0.5, execute five generation runs using different random seeds, and average the scores across the runs. We also measure the consistency of the models as the absolute difference of the metrics, when we swap the order of the images. Since the annotations are invariant to the order of the images, the models should show minimal variation in the metrics.

How do models perform at identifying one difference? In the granular alignment setting, the evaluation metrics are calculated considering multiple ground-truth sentences, and the results are presented in Table 1. The InternVL2.5-78B model is clearly the top performer, with the highest scores across all metrics, suggesting better coherence and fluency, and better adherence to our annotations. The small Phi3.5 Vision model is reasonably competitive across metrics given its size, and surprisingly Llama-3.2-90B showed a weaker performance, despite its larger size. Appendix C presents examples of the outputs generated by different models, while Appendix D shows a detailed analysis of the types of differences featured in the dataset, together with the corresponding results.

How do models perform at covering all the differences? For the global alignment setting, the metrics were computed considering the ground truth sentences as a single-paragraph description that models should generate. The results are presented in Table 1. Overall, the Phi3.5 Vision model exhibits the best performance metrics compared to its counterparts. Interestingly, smaller models exhibit notably competitive results, whereas larger models such as LLAVA-OV-72B and Llama-3.2-90B are ranked lower. All models struggle with longer phrase matches, with low BLEU$_3$ and BLEU$_4$ scores. Even the best-performing models have low scores across all metrics, indicating significant room for improvement in generating consistent difference descriptions. The models also exhibit a significantly low CIDEr score, signaling a poor relevance of the response content. Specifically, LLAVA-OV-72B's poor CIDEr score

Granular Alignment Setting						
Model	BLEU$_1$	BLEU$_2$	BLEU$_3$	BLEU$_4$	ROUGE$_L$	CIDEr
Phi3.5 Vision	37.5 ±0.2	21.0 ±0.4	9.9 ±0.3	5.1 ±0.3	29.5 ±0.3	6.9 ±0.4
MiniCPM-V-2.6	23.3 ±0.0	12.3 ±0.0	4.6 ±0.0	2.1 ±0.0	23.2 ±0.0	0.0 ±0.0
LLaVA-CoT	50.0 ±0.4	26.9 ±**0.5**	14.4 ±**0.5**	8.0 ±**0.4**	32.1 ±0.3	12.4 ±0.5
NVLM-D-72B	42.0 ±0.1	23.5 ±0.2	11.8 ±0.2	6.4 ±0.2	32.5 ±0.2	7.9 ±0.1
LLAVA-OV-72B	46.0 ±0.3	27.4 ±0.2	15.7 ±0.3	9.7 ±0.3	33.6 ±**0.6**	15.7 ±**0.8**
InternVL2.5-78B	**53.4** ±**0.5**	**31.4** ±0.1	**17.5** ±0.2	**10.2** ±0.2	**37.5** ±0.0	**17.2** ±0.0
Llama-3.2-90B	34.2 ±0.1	18.4 ±0.1	9.2 ±0.1	5.0 ±0.0	29.9 ±0.0	3.2 ±0.2

Global Alignment Setting						
Model	BLEU$_1$	BLEU$_2$	BLEU$_3$	BLEU$_4$	ROUGE$_L$	CIDEr
Phi3.5 Vision	**27.9** ±0.1	**14.4** ±0.1	**6.6** ±**0.0**	**3.3** ±0.0	**22.6** ±0.0	3.3 ±0.1
MiniCPM-V-2.6	22.4 ±0.1	11.1 ±0.0	4.2 ±0.0	1.7 ±0.0	19.4 ±0.0	1.2 ±0.1
LLaVA-CoT	4.1 ±0.0	2.0 ±0.0	1.0 ±0.0	0.5 ±0.0	13.6 ±0.1	1.0 ±**0.2**
NVLM-D-72B	13.3 ±0.0	6.9 ±0.0	3.1 ±0.0	1.6 ±**0.0**	19.5 ±0.1	0.9 ±0.0
LLAVA-OV-72B	1.5 ±0.0	0.9 ±0.0	0.5 ±0.0	0.3 ±0.0	13.0 ±0.0	0.0 ±0.0
InternVL2.5-78B	16.3 ±**0.3**	8.6 ±0.1	4.6 ±0.0	2.6 ±0.0	18.7 ±0.0	**3.4** ±0.0
Llama-3.2-90B	23.6 ±0.2	11.5 ±**0.1**	5.5 ±0.0	2.9 ±0.0	16.3 ±**0.2**	1.5 ±0.1

Table 1: MLLM performance metrics on the ViDi Benchmark. The values correspond to the average between the two possible image orders, together with the corresponding variation.

can be attributed to its tendency to generate similar and incomplete responses for both granular and global alignment prompts, failing to describe the full range of differences between images. Appendix C presents examples of model predictions.

What are the main takeaways from both evaluation settings? Comparing the performance across both settings reveals several important insights. Notably, the task of finding one difference is more manageable by current MLLMs, whereas the more complex task of comprehensively describing differences is generally not accomplished. This finding is particularly important, as most previous benchmarks focused on a single difference (Park et al., 2019; Awal et al., 2024; Zhang et al., 2024b).

The InternVL2.5-78B model shows adaptability, performing at a higher rank in both scenarios. Although Phi3.5 Vision leads in several metrics for listing all differences, it shows comparatively weaker performance in the granular alignment experiment. Conversely, LLAVA-OV-72B is very competitive in recognizing a single difference, but performed much worse when enumerating all differences, often producing very brief responses. Even the reasoning-based approach of the LLaVA-CoT model failed to systematically identify all the differences, and we noticed that the reasoning chain focused on re-interpreting the input textual prompt, instead of drawing inferences from the visual contents. The parameter count also does not appear to be the determining factor for performance in either task. For instance, Llama-3.2-90B consistently performs below the other models, despite having the largest parameter count, because it is unreliable with multiple images (Bhutani, 2025). The small models demonstrate performance equivalent to the large models, particularly in the more complex setting. All models exhibit a notable degree of confabulation when describing the differences. This is exemplified by outputting several repetitions of one change description with minimal variations. The input order of the images has no significant impact on the results. In Appendix D we report experiments by type of difference, with *attribute* differences being more thoroughly described and *count* differences being more challenging.

5 Conclusions

We presented ViDi as a new benchmark for evaluating MLLMs in the task of describing differences between pairs of images, moving beyond previous work by addressing multiple differences in image pairs originally designed to challenge humans. We also evaluated prominent MLLMs and showed that, despite popular enthusiasm, these models remain in the early stages of development, with their visual comprehension significantly trailing their linguistic abilities. Future work can consider extending the benchmark to support visually situated multi-turn dialog evaluation settings (Zheng et al., 2022).

Limitations and Ethical Considerations

The ViDi benchmark leverages resources collected from the Web, and some of the images in our dataset may be subject to copyright restrictions. While we attempted to prioritize the use of Creative Commons licensed images, the complete copyright status of all images could not be definitively verified. We nevertheless complied with the robot exclusion protocol, and we believe that our use of the images fits into the definition of Fair Use, given the objective of non-profit educational/scientific research aimed at public good. Our public GitHub repository gathers all the processed images, along with a text file containing, in the first line, the source URL of the webpage from which the image was collected, followed by the difference annotations. Copyright holders may contact us to request removal or replacement of specific images.

Another limitation of the work reported in this paper is related to the fact that our annotations have used only the English language. Moreover, despite efforts to maintain consistent evaluation standards, relying on a small pool of annotators may have introduced inherent biases, and potentially limited the generalizability across diverse demographic perspectives and interpretation styles.

Finally, the performance levels of the different models may have been impacted by our use of simple and standardized prompts, which we employed across all experiments for consistent evaluation.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. 2024. VisMin: Visual minimal-change understanding. *arXiv preprint arXiv:2407.16772*.

Sanyam Bhutani. 2025. Discussion on Llama 3.2 11B Vision Instruct HuggingFace Model Page. https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct/discussions/43#66f98f742094ed9e5f5107d4. Accessed: 2025-02-01.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: WHOOPS! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. NVLM: Open frontier-class multimodal LLMs. *arXiv preprint arXiv:2409.11402*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gautier Evennou, Antoine Chaffin, Vivien Chappelier, and Ewa Kijak. 2024. Reframing image difference captioning with BLIP2IDC and synthetic augmentation. *arXiv preprint arXiv:2412.15939*.

Chaoyou Fu, Yifan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. 2024. MME-Survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Erdong Hu, Longteng Guo, Tongtian Yue, Zijia Zhao, Shuning Xue, and Jing Liu. 2024. OneDiff: A generalist model for image difference captioning. In *Proceedings of the Asian Conference on Computer Vision*, pages 114–130.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. 2023. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4156–4165.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Workshop on Text Summarization Branch Out*, pages 74–81.

Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhen Xia Shi. 2022. 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.

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

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 8253–8280.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633.

Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. 2021. Describing and localizing multiple changes with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1951–1960.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.

Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. 2019. Expressing visual relationships via language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1873–1883.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

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

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914.

Zhiming Wang, Mingze Wang, Sheng Xu, Yanjing Li, and Baochang Zhang. 2024. CCExpert: Advancing MLLM capability in remote sensing change captioning with difference-aware integration and a foundational dataset. *arXiv preprint arXiv:2411.11360*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. LLaVA-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 12.

Jiarui Zhang, Ollie Liu, Tianyu Yu, Jinyi Hu, and Willie Neiswanger. 2024a. Euclid: Supercharging multimodal llms with synthetic high-fidelity visual descriptions. *arXiv preprint arXiv:2412.08737*.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Jungqi Zhao, Boyang Li, and Lu Wang. 2024b. SPHERE: A hierarchical evaluation on spatial perception and reasoning for vision-language models. *arXiv preprint arXiv:2412.12693*.

Duo Zheng, Fandong Meng, Qingyi Si, Hairun Fan, Zipeng Xu, Jie Zhou, Fangxiang Feng, and Xiaojie Wang. 2022. Visual dialog for spotting the differences between pairs of similar images. In *Proceedings of the ACM International Conference on Multimedia*, pages 5698–5709.

A Detailed Analysis on Prior Work

Several recent studies have focused on advancing benchmarks for assessing models that integrate vision with language. For instance, the Winoground dataset (Thrush et al., 2022) tests compositional reasoning, asking models to distinguish subtle differences in image-text relationships. The WHOOPS dataset (Bitton-Guetta et al., 2023) challenges the ability to reason about commonsense and compositionality, by presenting commonsense-defying images. The VALSE benchmark (Parcalabescu et al., 2022) assesses the vision-linguistic grounding capabilities of models on a suite of tests covering various linguistic constructs: existence, plurality, counting, spatial relations, actions, and entity co-reference. The NLVR2 benchmark (Suhr et al., 2019) evaluates the capacity of models to determine the validity of a statement in a visual context. While these studies provide valuable insights into different aspects of vision-language understanding, they primarily focus on single-image scenarios, lacking the ability to evaluate complex multi-image reasoning over visual differences.

A common evaluation paradigm for MLLMs involves following a visual question answering protocol, with questions tailored to evaluate specific skills or comprehension abilities. For instance, the SPHERE dataset (Zhang et al., 2024b) is particularly designed to measure spatial reasoning skills, through basic questions referring to object positions, distances, sizes, and counts. It also features advanced questions that require combinations of spatial and visual skills, and questions that require advanced understanding of a scene as a 3D environment with physical entities. The authors concluded that models still lack the ability to understand distance, to reason from both allocentric and egocentric viewpoints, and to perform physical world reasoning. ViDi is perhaps even more challenging, calling for models to generate coherent textual descriptions that are visually grounded, without involving any question category limitations. Again, analyzing fine-grained image differences is a core component of our new benchmark.

In a *spot the differences* task the images typically share the same perspective, which helps to concentrate on the semantic variations (i.e., alterations such as minor size adjustments or slight shifts are not significant). The task is related to visual semantic understanding and anomaly detection, and has been referred to as image change captioning (Park et al., 2019). Previous work has also looked at this challenge in connection with application areas such as remote sensing (Liu et al., 2022; Wang et al., 2024) or medical imaging (Hu et al., 2023). To produce a textual description of differences, models are required to build on abilities such as object identification, object counting, attribute recognition, and spatial relation reasoning, in addition to language generation. Previous work includes the Spot-the-Diff dataset (Jhamtani and Berg-Kirkpatrick, 2018), created from pairs of images extracted from urban surveillance videos. However, as Awal et al. (2024) highlight, the frames frequently resemble one another, and instances often do not have distinct semantic differentiation. The CLEVR-Change dataset (Park et al., 2019) extends the goal to consider robustness towards distractors, by augmenting a single change fused with a camera angle change. However, the dataset uses synthetic images with only a few object categories. The Image-Editing-Request dataset (Tan et al., 2019) features a large collection of real image pairs with corresponding editing instructions for a single change. The images were obtained from social media websites, specifically from posts aiming to crowdsource a specific change to an original image.

To address the challenges of data scarcity and variability, recent studies build on generative protocols to create synthetic data for model training and evaluation. The DiffCap dataset (Hu et al., 2024) merges existing real-world image difference datasets and synthetic data, resulting in GPT-assisted change captions together with pairs of synthetic images (including subtle and complex changes). Alternatively, Evennou et al. (2024) proposed a synthetic augmentation framework, based on diffusion models, without human or other filtering validation. Change captions are generated from image-text datasets by instructing an LLM with a few change categories, and a diffusion model generates synthetic images based on the intended change descriptions. Similarly, the VisMin dataset (Awal et al., 2024) was also created following a generative protocol, using LLMs and diffusion models. This benchmark requires models to predict the correct image-caption match given two images and two captions, where only one aspect (object, attribute, count, and spatial relation) changes at a time. The authors report that MLLMs exhibit notable deficiencies in understanding spatial relationships and counting abilities. However, approaches leveraging synthetic data are prone to simplify the tasks and introduce confabulations, therefore, requiring a human in the loop

Model	Vision Encoder	# Parameters
Phi3.5 Vision (Abdin et al., 2024)	CLIP ViT-L/14	4.2B
MiniCPM-V-2.6 (Yao et al., 2024)	SigLip-400M	8B
LLaVA-CoT (Xu et al., 2024)	ViT-H/14	11B
NVLM-D-72B (Dai et al., 2024)	InternViT-6B	72B
LLAVA-OV-72B (Li et al., 2024)	SigLip-400M	72B
InternVL2.5-78B (Chen et al., 2024)	InternViT-6B	78B
Llama-3.2-90B (Dubey et al., 2024)	ViT-H/14	90B

Table 2: Summary of MLLMs benchmarked in this study, detailing their vision encoders and parameter counts.

for reliability, or highly engineered data validation pipelines.

Most of the aforementioned previous studies focus on tasks in which only one difference is evaluated. With ViDi, we advocate for multiplicity because it increases the search space, and fine-grained understanding becomes harder with confounding factors, thus requiring robust feature representations. Real-world relevance also requires handling multiplicity. Exceptionally, the CLEVR-Multi-Change dataset (Qiu et al., 2021) consists of synthetic image pairs that contain multiple changes, change captions, and bounding boxes of the changed regions. However, this specific benchmark again only features synthetic images that lack aesthetic and object diversity.

B Implementation Details

In this study, we used a computing node equipped with four NVIDIA A100 80GB GPUs, which enabled us to run all models efficiently, eliminating the need for model quantization. To truly explore visual perception on models with high-resolution capabilities, we processed the smaller images with a recent super-resolution model (Wang et al., 2021), ensuring a minimum resolution of 1024x1024 pixels. The models used in the experiments are summarized in Table 2 . We obtained their released versions on HuggingFace.

The prompts used to instruct the models in the two evaluation scenarios (i.e., one difference and all differences) are presented next.

Granular Alignment Prompt:

<image>
<image>
Describe one single difference between the two images.
Use one sentence that adheres to the following guidelines:

Identify the main subject/object of the change;
Describe its location relative to other objects or within the image (e.g., using directional terms and nearby reference points);
State what has specifically changed (e.g., color variation, quantity difference, presence/absence, or text modification);
Describe both versions of what has changed, separated by the word "or" (e.g., "red or blue", "5 or 7", or "facing up or looking down").

Present the difference on a single individual English sentence, without any additional context.
Do not reference explicitly which image shows which version of the change/subject/object.

Global Alignment Prompt:

<image>
<image>
List all the differences between the two images.
For each difference, use one sentence that adheres to the following guidelines:

Identify the main subject/object of the change;
Describe its location relative to other objects or within the image (e.g., using directional terms and nearby reference points);
State what has specifically changed (e.g., color variation, quantity difference, presence/absence, or text modification);
Describe both versions of what has changed, separated by the word "or" (e.g., "red or blue", "5 or 7", or "facing up or looking down").

Present each difference as a single individual English sentence, without any additional context.
Do not reference explicitly which image shows which version of the change/subject/object.

In the case of models that only support one image together with each request, we suggest using a multi-round conversational prompt that is presented next. We show the version corresponding to the global alignment setting, but a similar prompt can also be used when on the granular alignment setting. In our experiments, all the models that were considered support multi images in the prompt, and therefore this multi-round conversational prompt was not used.

Chat Interaction Prompt:

User:
<image>
Analyze the image and provide a detailed description.

Assistant: ...

User:
<image>
Analyze this other image and describe the differences in comparison to the previous image.
For each difference, use one sentence that adheres to the following guidelines:

Identify the main subject/object of the change;
Describe its location relative to other objects or within the image (e.g., using directional terms and nearby reference points);
State what has specifically changed (e.g., color variation, quantity difference, presence/absence, or text modification);
Describe both versions of what has changed, separated by the word "or" (e.g., "red or blue", "5 or 7", or "facing up or looking down").

Present each difference as a single individual English sentence, without any additional context.
Do not reference explicitly which image shows which version of the change/subject/object.

Assistant: ...

C Dataset Examples

Two examples of natural scene image pairs featured in the ViDi dataset, together with the corresponding annotations for the differences and also with the results generated by different models, are presented in Figures 2 and 3. The images from these two instances were originally made available online by Zack Eckley[2] under a Creative Commons license.

The color of the sky is blue or orange.
The tower at the centre, with or without an antenna.
The color of the logo of the coffee shop on the left, either green or purple.
The traffic lights on the left are either green or red.
The arrowhead on the light bulb sign at the construction site has 4 or 5 light bulbs.
The logo at the construction site is either red or green.
The crosswalk at the bottom, with the left stripe more or less faded.

Model	Granular Alignment Outputs
Phi3.5 Vision	The color of the sky changes from blue in the first image to a pinkish hue in the second image, or the time of day appears different, or the lighting conditions are altered.
Llava-CoT	The color of the Starbucks sign, which is facing up or looking down on the street, is red or blue.
InternVL2.5-78B	The pedestrian crossing the street is facing the camera or is facing away from the camera.
LLAVA-OV-72B	The sky has changed from a hazy orange color to a clear blue, indicating a change in the time of day or weather conditions.

Model	Global Alignment Outputs
Phi3.5 Vision	The sky has a gradient of colors in the second image, ranging from a warm orange near the horizon to a cooler blue at the top, or the sky is a uniform light blue in the first image.
Llava-CoT	The Starbucks sign is red or blue. The street sign is black or white. The skyscraper is a tall building or a tall building with a black top. The Starbucks sign is facing up or looking down. The street sign is facing up or looking down. The skyscraper is facing up or looking down. The Starbucks sign is on the left or on the right. The street sign is on the left or on the right. ...
InternVL2.5-78B	A person is walking in the crosswalk or a person is not walking in the crosswalk. The flag on the boat is present or absent.
LLAVA-OV-72B	The sky has changed from blue to orange.

Figure 2: Images associated to sample #14 in the ViDi dataset, together with the ground-truth annotations (top), model outputs for the granular setting (in the middle), and model outputs for the global alignment setting (at the bottom), when providing the image in the left as the first input to the model.

[2]https://www.tes.com/teaching-resource/spot-the-difference-starters-photoshop-set-1-12038773

The number of windows on a house in the center.
The color of the cover hiding a construction site in the center is green or red.
The pink house in the center, with or without a small statue of a saint on its wall.
The color of the door on the brick wall, either brown or green.
The color of the shirt of the person driving the boat is blue or purple.
The number written on the side of the boat is 9246 or 92466.
The graffiti on the orange wall of the house on the right is missing.
The stone step at the lower left corner, with or without a docking ring.

Model	Granular Alignment Outputs
Phi3.5 Vision	The text "VV 27285" on the boat in the image has a slight variation in the second image, where the numbers "27285" are repeated twice.
Llava-CoT	The color of the boat is brown or blue.
InternVL2.5-78B	The boat in the canal has a person wearing a purple shirt or a black shirt.
LLAVA-OV-72B	The person in the boat is facing forward or looking down.

Model	Global Alignment Outputs
Phi3.5 Vision	The main subject/object of the change is the boat. In the first image, the boat is located closer to the foreground and is positioned on the left side of the canal, while in the second image, the boat is further away and positioned on the right side of the canal. The boat in the first image has the text "UV 2788" on its side, which is not visible in the second image.
Llava-CoT	The boat is white or brown and has the text "V8 92466" or "V8 92465" on it. The boat is near the canal or the buildings. Graffiti is on the buildings or not.
InternVL2.5-78B	The boat's color is a light beige or a darker tan. The text on the boat's side is "VE 9246" or "VE 92466." The person driving the boat is wearing a dark shirt or a light shirt. There is graffiti on the wall near the right side of the image or there is no graffiti. The flag on the boat is present or absent.
LLAVA-OV-72B	The person on the boat is wearing a different shirt, blue or purple.

Figure 3: Images associated to sample #15 in the ViDi dataset, together with the ground-truth annotations (top), model outputs for the granular setting (in the middle), and model outputs for the global alignment setting (at the bottom), when providing the image in the left as the first input to the model.

Difference Category	% in annotations	% in image pairs
Attribute	42.8	95.5
Object	28.6	80.0
Count	12.5	49.0
Symbolic	9.4	31.0
Spatial Relation	6.8	29.5

Table 3: Distribution of difference categories in the ground truth descriptions. The columns correspond to the percentage of sentences per category, and the percentage of image pairs, containing at least one sentence of each category.

D Analysis of Results by Difference Category

With the goal of improving the understanding of the model's responses, we used an LLM to classify each ground truth difference description into a predefined set of change types. The chosen categories are as follows: Object (presence/absence), Attribute, Count, Spatial Relation (position/direction), and Symbolic (textual/conceptual). The classification was performed by employing Llama 3.3 70B, instructed with the following prompt that features a definition for each of the classes. Upon manual examination of a sample of sentences featuring 178 instances, we estimate the classification accuracy to be around 86%.

Category Classification Prompt:

The input sentence represents a description of one difference detected between pairs of images. Your task is to classify the sentence into one of the following categories:

Object - Refers to the presence or absence of an object in the scene. This includes cases where an object appears in one image but is missing in the other (e.g. "a tree is present in one of the images but absent in the other.").

Attribute - Describes a change in the characteristics of an object, such as its color, texture, size, shape, or material (e.g. "the car is red in one of the images and blue in the other.").

Spatial Relation - Captures differences in the position, orientation, or arrangement of objects relative to each other or within the scene (e.g. "the chair is near the table in one of the images but far from it in the other.").

Count - Refers to a change in the number of instances of an object or group of objects in the scene (e.g. "there are three apples in one of the images but only two in the other.").

Symbolic - Refers to changes in the meaning, purpose, or interpretation of an object. These differences may involve symbols, text (e.g. "the sign reads 'Stop' in one of the images and 'Yield' in the other.").

Show only the category name.
Sentence: <sentence>

Table 3 provides a summary of the classification results, detailing the percentage of sentences linked to each category, and the percentage of instances with at least one sentence in that category. *Attribute* and *object* differences are the most prevalent. Conversely, *spatial relation* and *symbolic* differences occur less frequently in the annotations, but are present in at least 29.5% of the dataset instances.

Table 4 present the CIDEr score for each of the considered models in the granular alignment scenario, along with the percentage of instances where the model response achieves the highest CIDEr score for a ground truth sentence of each difference category, averaged across both image orderings. InternVL2.5-78B showed the best performance in the task, and distinguishes itself describing many *attribute* changes, at the same time seldom reporting *spatial relation* changes. The reasoning model LLaVA-CoT achieved a similar performance to InternVL2.5-78B, distinguishing itself in describing *count* changes. MiniCPM-V-2.6 registered a poor CIDEr score, mostly describing *object* category differences. Phi3.5 Vision and NVLM-D-72B had a similar performance. LLAVA-OV-72B only stands out as better at producing *symbolic* differences, while Llama-3.2-90B generated more *spatial relation* differences compared to other models, albeit having a low CIDEr score.

Additionally, in Table 5 we present the CIDEr score when we prompt the model with a few-shot

Model	CIDEr	%				
		Object	Attribute	Count	Spatial Relation	Symbolic
Phi3.5 Vision	6.9	32.0	41.7	7.5	7.0	11.8
MiniCPM-V-2.6	0.0	**40.3**	40.3	6.3	8.1	5.0
LLaVA-CoT	12.4	28.9	45.2	**10.2**	5.5	10.2
NVLM-D-72B	7.9	33.2	40.2	8.8	7.2	10.6
LLAVA-OV-72B	15.7	29.6	42.5	8.0	6.5	**13.6**
InternVL2.5-78B	**17.2**	29.0	**45.6**	8.2	5.8	11.3
Llama-3.2-90B	3.2	33.1	41.0	8.3	**8.7**	8.8

Table 4: Model performance in terms of CIDEr scores for the granular alignment setting, along with the percentage of instances where the highest score value corresponds to each difference category.

instruction that was derived from the prompt template used in the main experiments. The specific prompt is presented above. Particularly, the instruction includes, in context, all but one of the differences of the corresponding image pair, and we executed multiple evaluations with each instance, in order to cover all possibilities as the missing difference that should be identified. This approach helps the model adhere to the annotation style and facilitates the clear identification of differences that the model perceives or is unable to recognize. The models NVLM-D-72B and LLAVA-OV-72B generally perform the best across categories, and the larger models tend to perform better than the smaller ones. Category-wise, the *attribute* differences are more easily perceived by most models, with high scores by the NVLM-D-72B and InternVL2.5-78B models. The higher performance score of the LLAVA-OV-72B model in *object* differences supports its suitability for detection tasks. The lower performance score of the *count* differences category suggests that this is a more challenging task, and inherently the reasoning-based approach of the LLaVA-CoT model seems to be beneficial. We reviewed the LLaVA-CoT outputs and confirmed that the reasoning sequence is homogeneous, starting with an interpretation of the user prompt followed by an image caption section, then a planning section that highlights what the model identifies as meaningful to fulfill the task, and finally a conclusion, but lacking backtracking and self-validation. Despite using this strategy, the model often produces confabulations, including errors in the planning section that propagate to the final answer. The model also elaborates much more on the interpretation of the input prompt, instead of the interpretation of the images and their differences. Finaly, the *symbolic* differences category shows the most variability between models.

Few-shot Prompt:

<image>
<image>
This is a list of differences between these two images:

Difference annotation example 1
Difference annotation example 2
...

Describe one missing difference between the two images.
Use one sentence that adheres to the following guidelines:

Identify the main subject/object of the change;
Describe its location relative to other objects or within the image (e.g., using directional terms and nearby reference points);
State what has specifically changed (e.g., color variation, quantity difference, presence/absence, or text modification);
Describe both versions of what has changed, separated by the word "or" (e.g., "red or blue", "5 or 7", or "facing up or looking down").

Present the difference as a single individual English sentence, without any additional context.
Do not reference explicitly which image shows which version of the change/subject/object.

Model	(All categories)	Object	Attribute	Count	Spatial Relation	Symbolic
Phi3.5 Vision	18.8	17.4	22.8	16.5	17.6	13.1
MiniCPM-V-2.6	5.5	6.8	5.3	4.1	7.3	5.8
LLaVA-CoT	20.9	18.5	24.6	**22.9**	19.4	16.5
NVLM-D-72B	**38.1**	23.4	**37.4**	21.2	22.9	26.3
LLAVA-OV-72B	31.1	**27.8**	35.5	19.8	**28.4**	**37.2**
InternVL2.5-78B	31.1	27.3	37.4	22.4	26.4	34.3
Llama-3.2-90B	11.8	8.9	14.6	10.6	16.2	13.8

Table 5: Model performance in terms of CIDEr scores in the *few-shot* evaluation setting.