2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Subtle-Diff: A Dataset for Precise Recognition of Subtle Differences Among Visually Similar Objects

Fumiya Matsuzawa, Yue Qiu, Yanjun Sun, Kenji Iwata, Hirokatsu Kataoka, Yutaka Satoh*

***Abstract*— Visual inspection robots used in factories and outdoor environments require the ability to accurately recognize visual differences between similar objects and further verbalize the recognition results to present the differences to humans. Despite the application of Large Language Models (LLMs) and multimodal LLMs across various domains, our research highlights their insufficiency in verbalizing nuanced differences across images. To address this, we leveraged LLMs and image generation AI to develop a dataset aimed at assessing difference recognition capabilities. We introduced two novel tasks using this dataset: selecting images based on their visual differences and a conditional difference captioning task, and evaluated existing Vision-Language Models (VLMs) on these tasks. Our findings reveal that advanced models like GPT-4V can describe subtle differences with comparative expressions, yet they fall short of matching human performance across all attributes. This discrepancy between model and human recognition, especially in identifying easily discernible differences, suggests that most current models lack the ability to directly compare image pairs for difference detection. Consequently, we propose a new model that incorporates an image-text similarity approach in the difference recognition task, showing superior performance over existing models, including GPT-4V. Our dataset and findings will contribute to advancements in differencing objects and improve robotic applications in visual inspection and object picking. The dataset is available at DICTA challenge page.**

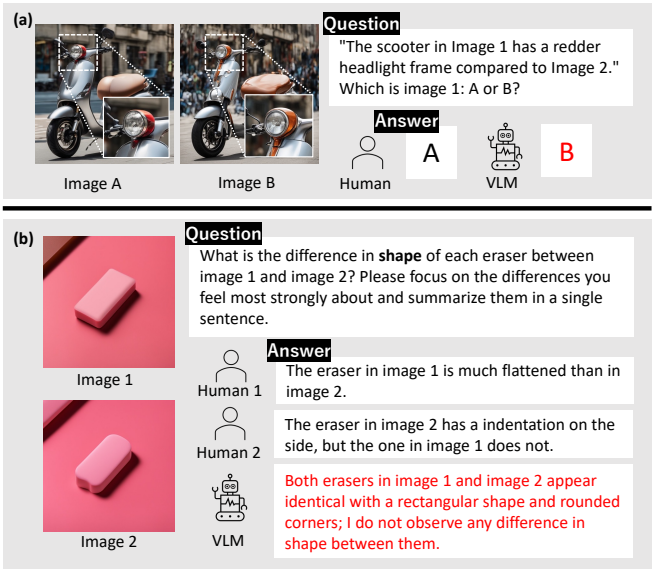

Fig. 1. Illustration of two proposed tasks for subtle difference recognition: (a) the image selection task, and (b) the conditional difference captioning task. These tasks demonstrate scenarios where humans effortlessly outperform current VLMs.

I. INTRODUCTION

The detailed differences among similar objects, such as attributes like the arrangement and quantity of parts, composition of components, shape, size, and texture, are crucial in various applications. For instance, this is particularly important when a picking robot selects an object with the requested features from a vast number of similar objects, or when an automatic inspection robot recognizes temporal changes in an object, such as aging. In these applications, recognizing subtle differences that include continuous variations is important. In human communication, comparative expressions are often used to improve the efficiency of information transmission, rather than using quantitative values of attributes. Examples include phrases like "a rounder cup" or "wood with a rougher surface." If AI can understand these linguistic expressions and continuous differences in the real world, robots can engage in more advanced language interactions. Specifically, operators of picking robots could describe and compare similar objects to control the robots. Also, in visual inspection applications, verbalizing recognized differences could lead to automatic reporting of inspection results or querying a product-specific database of anomalies to improve the quality of the production line.

Existing difference recognition datasets have focused on the addition or removal of objects or people, changes in position, and discrete changes in color or shape between two images. However, these datasets address only simple, discrete changes and not the more complex, continuous differences such as slight shape or texture variations. Models have been proposed for these datasets that can describe differences in language text from the input of two images. However, these models cannot take text as input, making it difficult to recognize or explain detailed differences for specified conditions. Also, the extent to which existing VLMs can recognize differences has not been explored.

Addressing the lack of datasets for recognizing subtle changes, this study builds a new evaluation dataset, named Subtle-Diff (Figure 1), to assess the performance of existing VLMs in recognizing subtle differences, including continuous variations in attributes and differences in parts. The dataset construction uses LLMs and image generation models to automatically generate similar image pairs containing subtle differences. Specifically, it starts by generating texts describing subtle differences with an LLM, such as "a

*The authors are with Faculty of Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Japan {fumi8.matsuzawa, qiu.yue, yanjun.son, kenji.iwata, hirokatsu.kataoka, yu.satou}@aist.go.jp

darker red car body" or "a top with a rougher texture." Then, the texts generated by the LLM are input into image generation models to automatically generate similar image pairs. Specifically, we utilized two image generation AIs: one generates an object from a given text, and another modifies the generated image based on the difference description. Finally, human annotators describe the differences in the generated image pairs with text. This dataset allows for analysis of differences between human intuitive judgments and VLMs decisions.

Furthermore, this study evaluated the recognition capability of VLMs, including GPT-4V [1], for the recognition of subtle differences using the newly generated dataset, a topic not fully explored in existing research. Specifically, we comprehensively identified weaknesses in existing VLMs related to recognizing subtle changes in attributes including color, shape, and texture. The experimental results highlighted the gap between human difference recognition capabilities and those of models. Considering the shortcomings of existing methods, we proposed a technique to analyze subtle differences from image pairs embedded in feature space by encoders of foundational models trained on large datasets, such as CLIP [2] and ALIGN [3]. The proposed method confirmed higher performance than GPT-4V. Our dataset and proposed method are expected to contribute to the improvement of accuracy and practicality in various robotic applications, such as visual inspection robots.

II. RELATED WORKS

A. Image Differences Recognition

In the field of difference recognition, research has focused on tasks such as judging whether a caption for a pair of images correctly describes their difference [4], [5]. Studies on difference captioning for sets of two or more images include works that use a Computer Graphics (CG) engine to target object movement, deletion, and attribute changes [6], [7], as well as studies targeting differences between images sampled from different frames of a video [8], [9].

Research on recognizing differences between images based on object attributes includes tasks like image-pair conditional similarity evaluation [10], and recognizing continuous attribute changes in groups of images [11], [12], [13]. There are also studies on text-image alignment with minimal changes [14], [15], [16], [17], and image retrieval from images and differential texts [18]. Our research further defines a conditional difference captioning task targeting highly similar image pairs with only subtle differences between images. Our dataset allows for quantitative evaluation regarding subtle differences in specified attributes and assessing the consistency with human intuitive judgment.

B. Zero-shot VLMs

In the context of zero-shot VLMs, models trained on large datasets have achieved high performance in various downstream tasks without specific task training. Models like CLIP [2] and ALIGN [3], which perform contrastive learning between image and language features, and models capable of explaining images have been proposed. Recent models include BLIP-2 [19], which learns only the transformation mechanism to input image features into a pre-trained LLM; OpenFlamingo [20], which trains both the transformation mechanism and the LLM; and LLaVA-1.5 [21], which performs instruction tuning for multiple specialized tasks. GPT-4V [1], used in Chat-GPT, has been reported to recognize differences from image pairs, such as the presence of minor scratches.

However, how well the above methods perform in subtle difference recognition from image pairs has not yet been well addressed. Therefore, this study aims to elucidate the strengths and weaknesses of these zero-shot VLMs in recognizing minor differences through experiments on our dataset.

III. SUBTLE-DIFF DATASET AND TASKS

This dataset is designed to assess whether VLMs can recognize subtle differences between images, going beyond simple class or attribute classification. In detail, the dataset aims to analyze how VLMs perceive continuous differences, such as slight variations in brightness between similar objects, or minor differences in part shapes, compared to human intuitive judgments. To the best of our knowledge, this dataset is the first to focus on subtle changes in object attributes.

A. Generating Similar Yet Distinct Image Pairs

Conventional methods for collecting images that are similar yet distinct have typically involved leveraging the feature similarity of pre-trained image encoders or sampling frames from various moments in a video. In this research, we introduce an image pair generation framework utilizing LLMs and image generation AI to enhance the efficiency of acquiring image pairs with subtle differences (Figure 2).

Designing Base Objects. In this step, the LLM is used to generate information about the components and attributes of objects contained within two images, as well as descriptions of the differences between these objects. Object classes are randomly selected from those used in ImageNet [22], excluding those representing living organisms. GPT-3.5-turbo is employed as the LLM. The Object Design LLM (OD-LLM) is tasked with determining the part segmentation of an object given its class, along with the attributes of each part. Here, a "part" refers to a visually distinguishable unit, such as the backrest or legs of a chair.

Generating Base Object Images. Upon receiving object information from the OD-LLM, the Base Prompt LLM (BP-LLM) generates prompts for a Text-to-Image model to create images of objects. In this process, not only the information about the objects is utilized, but appropriate background settings are also determined by the LLM. Based on the prompts generated by the BP-LLM, object images are then produced using an image generation model. Stable Diffusion XL [23] is used as the image generation model.

Generating Difference Images. In this step, difference prompts are generated by the Difference Prompt LLM (DP-LLM) based on the object information produced by the

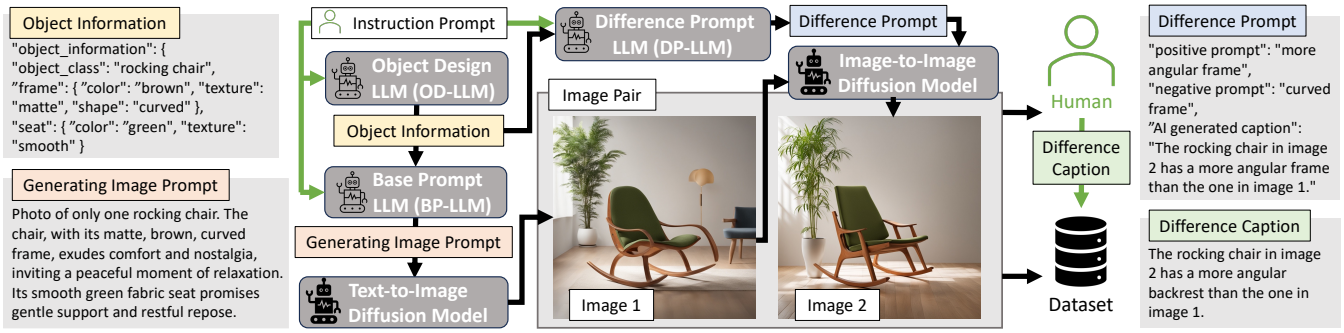

Fig. 2. The generation process of the proposed Subtle-Diff dataset involves three LLMs for generating image generation information, and two diffusion models are used for generating two images with subtle differences, respectively. Better viewed in color.

Color	The china cabinet in Image 2 has a darker look compared to Image 1.	The pencil case in image 1 has darker blue in the zipper area compared to image 2 which is grayish blue.
Shape	The china cabinet in Image 1 has more curved carvings on its base than the Image 2.	The pencil case in image 1 is bulky compared to image 2.
Texture	The china cabinet in 1 has a glossy finish compared to Image 2.	The pencil case in image 1 is glossy compared to image 2.

Fig. 3. Two examples of the Subtle-Diff dataset.

OD-LLM. Initially, the parts and attributes that will exhibit differences are specified from among color, shape, and texture. The LLM defines any minor differences for the attributes described in the object information and converts these into a prompt format suitable for an Image-to-Image diffusion model. Finally, Image 1 (base image) and the difference prompt generated by the DP-LLM are inputted to produce Image 2, which displays differences in the specified attributes, by using an image-editing diffusion model. In image editing using image generation models, it is often the case that the target objects in images become distorted to the extent that class recognition is impossible; such images are excluded as they introduce noise into the difference recognition process.

B. Human Annotation of Differences in Image Pairs

This study aims to examine how existing VLMs' difference recognition results deviate from human intuition and the magnitude of this disparity in their recognition capabilities. To this end, we collected human annotations from two annotators for each image pair generated in the previous step. While data generation involved defining a difference

TABLE I
ANNOTATION DETAILS FOR EACH ATTRIBUTE CONDITION

Attributes	Annotation guideline
Color	Color, hue, brightness, cool or warm, saturation, color scheme
Shape	Roundness, sharpness, curvature, width, narrowness, length, shortness, size of component parts, presence of component parts, etc.
Texture	Roughness, glossiness

in one attribute per image pair as input to an Image-to-Image model, the results often led to changes in multiple attributes. Therefore, each image pair was annotated according to the guidelines in Table I for three attributes: shape, color, and texture. Additionally, changes in the background occurred frequently, but we instructed to disregard such changes (Figure 3). Given that minor differences could be extensively described for each attribute, we asked annotators to document only the most striking difference. For each attribute condition of each image pair, we classified the semantic alignment between difference captions into three patterns—matching, contradiction, and others—using GPT-

4. Captions that expressed agreeing opinions on the same item were classified as matching, those that stated opposing facts were categorized as contradiction, and those focusing on different parts or items were classified as others.

TABLE II
DATASET STATISTICS.

Image pairs	Annotations	Objects	Annotators	Vocabulary	Sentence length
2,802	12,828	570	11	1,930	12.78 (average)

TABLE III
STATISTICS PER ATTRIBUTE

	Color	Shape	Texture
Total annotation	3,608	4,960	4,260
No change annotation	29	4	31
Human agreement rate	81.15%	38.47%	73.85%

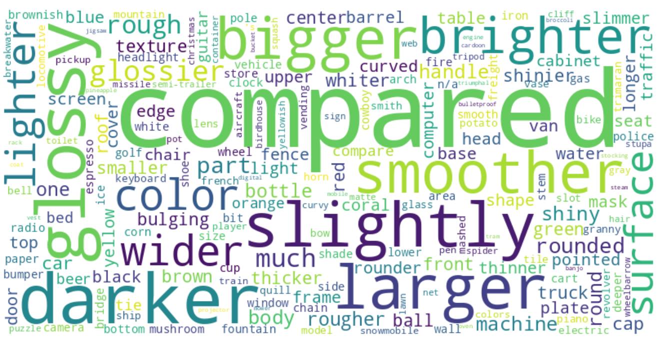

Fig. 4. Wordcloud of the Subtle-Diff dataset.

C. Dataset Statistics

Table II shows the statistics of our dataset. By utilizing LLMs, we were able to generate a diverse set of instances for each class. Annotations of difference caption pairs classified as contradictions by GPT-4 were removed, resulting in a collection of 12,828 captions (6,414 pairs). Table III presents the statistics for each attribute condition. Additionally, the number of annotations marked as no change is shown under "No Change Annotations." The human agreement rate indicates the proportion of pairs classified as "match". The agreement rate is lower for shape conditions as annotators often focus on changes in the shape of different parts.

Figure 4 displays the word cloud of our dataset. Due to the annotation of minute differences, expressions denoting comparison such as "compared", "darker", "larger", and expressions indicating degree such as "slightly" are prevalent, highlighting the dataset's focus on nuanced differences.

D. Task Design

Image Selection Task. To evaluate whether VLMs can integrate and understand human difference captioning and the differences between image pairs, we define a difference-aware image selection task (Figure 1, (a)). In this task, the input consists of two images, A and B, along with a text describing the differences between the image pair. The difference text refers to the images as "Image 1" and "Image 2", and the task is to choose which of images A or B corresponds to "Image 1". This task allows for a quantitative assessment of how well existing VLMs understand comparisons made by humans between objects. It can also serve as an indicator of the extent to which robots can distinguish objects based on human linguistic expressions.

Conditional Difference Captioning Task. The objective of this task is to evaluate to what extent VLMs can articulate minor differences and how the captions they generate differ from human annotations (Figure 1, (b)). Attempting to describe the minor differences between images can lead to an infinite number of answers, many of which may have low informational content. Therefore, this study defines a conditional difference recognition task between images. The term "conditional" refers to describing the most significant difference among those related to specified attribute conditions. By adding this condition, even image pairs containing many minor differences have clear criteria for what should be described, enabling an analysis of the characteristics of difference recognition between humans and VLMs.

IV. METHOD

Baseline Methods. We employ VLMs capable of generating text output from text and image inputs in a zero-shot manner. Specifically, for single image inputs, we use BLIP-2 and LLaVA-1.5, while for multi-image inputs, we utilize Openflamingo and GPT-4V, which are models designed to handle multiple images. For single image models, the input consists of horizontally concatenated image pairs. In the difference image selection task, the prompt includes a text describing the differences between Images 1 and 2, along with Images A and B, and a question text asking which of A or B corresponds to Image 1. The output is designed to be either A or B, represented by a single character. For the difference captioning task, the input includes a question text specifying attribute conditions and the object of interest, along with Images 1 and 2, with instructions to output the most significant difference in a single sentence.

Proposed Methods. In the task of selecting images based on descriptions of differences, it is necessary to align two images based on specified attribute conditions in response to descriptions of differences that cannot be discerned from a single image alone, such as comparative expressions. However, existing VLMs typically recognize only one image at a time, with limited functionality for comparing between images. Therefore, we propose a new image selection method utilizing VLMs capable of measuring the similarity between text and images in a zero-shot manner. Models like CLIP and ALIGN are trained on the consistency between images and captions but do not directly compare between images. However, it is conceivable that these models, trained on large-scale data, can represent minor differences between images in the feature space. Since using difference descriptions directly for calculating similarity with images is not appropriate, we generate two text pairs from the difference

captions that represent slightly differing attributes between Images 1 and 2 in a contrasting manner using an LLM. For example, for the caption "The traffic sign in Image 1 is darker in color than in Image 2", we generate a text pair such as "dark-colored traffic sign" and "light-colored traffic sign". For this text pair generation, we use GPT-3.5-turbo. The feature vectors of the texts generated for each text pair are denoted as T_1 and T_2, and the feature vectors for Images A and B are denoted as I_A and I_B. The cosine similarity between two feature vectors is represented as $sim()$, with the transformation to feature vectors being conducted using CLIP and ALIGN. We propose a method P_S that uses the similarity of each image to Text 1 and a method P_{SD} that uses the difference in similarity between Text 1 and the images and Text 2 and the images, as shown in following.

$$P_S = \begin{cases} A & \text{if } sim(I_A, T_1) > sim(I_B, T_1) \\ B & \text{else} \end{cases}$$

$$P_{SD} = \begin{cases} A & \text{if } sim(I_A, T_1) - sim(I_A, T_2) \\ & > sim(I_B, T_1) - sim(I_B, T_2) \\ B & \text{else} \end{cases}$$

V. EXPERIMENTS

A. Experimental Settings

The aim of this experiment is to analyze the difference recognition capability of existing VLMs on our dataset. To achieve this, we evaluate the performance of VLMs on two tasks defined in the previous section. CLIP uses openai/clip-vit-base-patch32, ALIGN uses kakaobrain/align-base [24], and for the creation of text pairs, gpt-3.5-turbo-0125 is used. Moreover, for VLMs, Salesforce/blip2-opt-6.7b, OpenFlamingo-9B-vitl-mpt7b, llava-v1.5-13b, and gpt-4-vision-preview were used respectively. As shown in the Figure 1, in the image selection task, VLMs are given a prompt with images A and B, a description of their differences, and the question "which is image 1?", to which the VLMs respond by choosing either A or B. In the difference captioning task, a question specifying the object class name and the attributes describing the difference is input, and VLMs output the difference in a free-form text. GPT-4V conducted experiments by sampling 300 image pairs for each attribute condition of each task.

B. Evaluation Metrics

For the difference image selection task, which is a binary classification task, accuracy was used as the evaluation metric. In the difference captioning task, evaluation was conducted using two human annotations (correct captions) for each dataset instance. As evaluation metrics, BLEU-4 [25] and CIDEr [26], widely used in image captioning tasks, were adopted. Furthermore, the semantic alignment between predicted and correct texts was assessed using GPT-4, facilitating the calculation of recall and precision based on these findings. Recall (Rec.) was defined as the proportion of all captions that semantically matched at least one of the ground truth texts. This method revealed an interesting discrepancy:

TABLE IV
PERCENTAGE OF OUTPUT WITH NO DIFFERENCE

	color	shape	texture
BLIP-2	3.33%	0.33%	21.33%
LLaVA-1.5	0.33%	0.00%	0.00%
Openflamingo	5.00%	1.33%	0.00%
GPT-4V	35.00%	16.33%	14.67%

TABLE V
QUANTITATIVE EVALUATION OF SELECTING IMAGES TASK.

Approach	Accuracy (%)			
	Color	Shape	Texture	Ave.
$CLIP_S$	53.87	52.34	55.86	54.02
$CLIP_{SD}$	56.94	51.85	56.59	55.13
$ALIGN_S$	**58.98**	**56.81**	**58.96**	**58.25**
$ALIGN_{SD}$	56.75	56.61	54.61	55.99
BLIP-2	50.01	50.19	50.73	50.31
LLaVA-1.5	52.28	53.67	54.04	53.33
OpenFlamingo	49.17	49.53	50.78	49.83
GPT-4V	57.00	52.67	58.67	56.11

despite the rarity of "no difference" annotations in human-labeled data, VLM experiments produced a relatively high percentage of descriptions stating no difference (Table IV). Given this observation, a manual evaluation was conducted to ensure a comprehensive understanding. Precision (Prec.) was then defined as the proportion of annotations, excluding those described as having no difference, that were accurately matched to the image pairs.

C. Quantitative Evaluation

Image Selection Task. The quantitative evaluation results of the image selection task are shown in Table V, where GPT-4V was evaluated on 300 test examples. When comparing two proposed methods, it was found that using the CLIP model, analyzing the difference in similarity using texts with opposing meanings improved accuracy in color and texture, more so than analyzing similarity with image pairs for a single text. However, the reason for not observing an improvement in shape accuracy is thought to be due to the high proportion of discontinuous differences described in human-annotated disparity descriptions, making it difficult to generate semantically opposing text pairs. In the case of the ALIGN model, it was observed that the accuracy of P_S is higher in all attributes. This could be due to the lower consistency in feature space with texts of opposing meanings compared to CLIP. Furthermore, compared to using VLMs, our proposed method was found to be more accurate in selecting images from difference descriptions.

Conditional Difference Captioning Task. The results of the difference captioning task are shown in Table VI. Evaluations by GPT-4 and manually showed that GPT-4V had higher accuracy. Especially in shape difference captioning, GPT-4V was able to accurately answer differences using diverse expressions, even for image pairs that were difficult for other VLMs to describe. However, GPT-4V often answered that there was no difference (Table IV) in image pairs with

TABLE VI

EVALUATION OF CONDITIONAL DIFFERENCE CAPTIONING, WITH UNDERLINES DENOTING ANALYSIS ON 300 TEST EXAMPLES.

Approach	Color				Shape				Texture			
	BLEU ↑	CIDEr ↑	Rec. (%) ↑	Prec. (%) ↑	BLEU ↑	CIDEr ↑	Rec. (%) ↑	Prec. (%) ↑	BLEU ↑	CIDEr ↑	Rec. (%) ↑	Prec. (%) ↑
BLIP-2	3.82	43.40	19.33	19.66	4.72	31.46	17.67	18.06	1.86	25.70	24.67	33.05
LLaVA-1.5	3.32	**53.01**	**58.00**	62.54	2.93	**42.99**	22.33	36.33	2.47	34.00	62.67	62.67
OpenFlamingo	**11.44**	41.00	54.33	63.16	7.76	38.15	15.33	27.03	9.02	**51.24**	**63.00**	**64.33**
GPT-4V	8.06	10.43	51.33	**85.64**	**8.16**	6.86	**58.00**	**87.25**	**9.87**	11.53	50.00	62.50

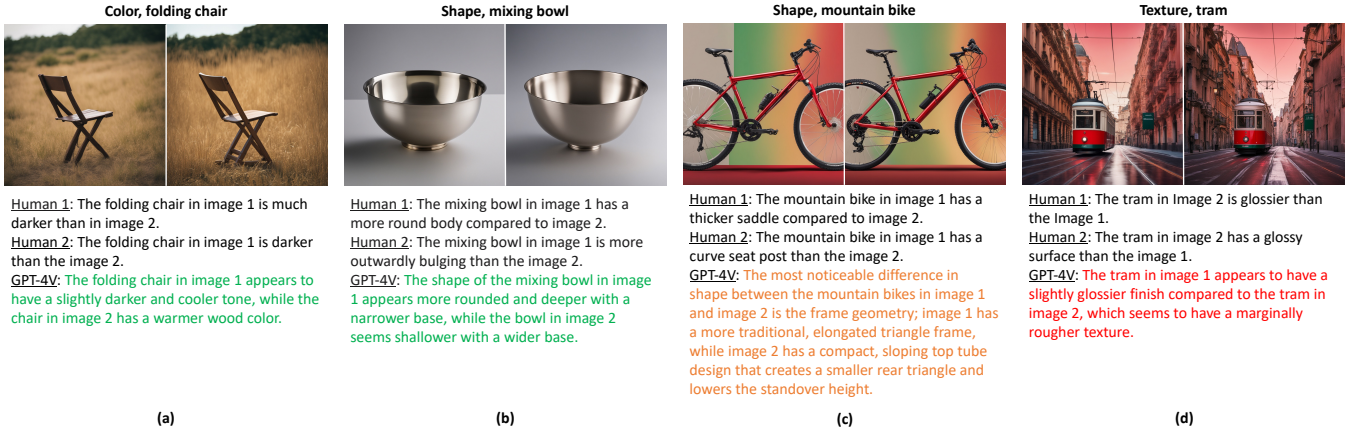

Fig. 5. Comparative results of humans and GPT-4V, highlighting GPT-4V's outcomes in green (aligned with human decisions), orange (misaligned with human decisions), and red (contrary to human decisions). In each image pair, refer to the left image as image 1 and the right image as image 2.

notable differences where human opinions were consistent, resulting in a Recall of about 50%. On the other hand, when differences were correctly answered, Precision was high. GPT-4V tends to produce longer sentences compared to human annotations, and in similarity evaluation using CIDEr, it recorded significantly lower values compared to other models.

D. Qualitative Evaluation

Examples of difference captioning results are shown in Figure 5. Here, four examples from the proposed Subtle-Diff dataset are presented, featuring annotations by human annotators and results by GPT-4V. Overall, GPT-4V tends to produce slightly longer sentences compared to human annotators. In examples (a) and (b), where the object occupies a significant portion of the image, GPT-4V tends to accurately describe differences in the overall color of the object (the color of the chair in example (a)) or the shape of the object (the shape of the bowl in example (b)). However, in situations like examples (c) and (d), where the background of the image is confusing or the object is represented in a relatively small proportion, or when there are differences in parts of the object, humans can easily recognize the differences, whereas GPT-4V may perceive them differently from humans (as in example (c)) or even in the exact opposite way (as in example (d)). Thus, even with VLMs like GPT-4V, which are reported to be highly accurate, there is a gap between human perception and the AI's recognition of subtle differences.

VI. CONCLUSIONS

The recognition of detailed differences between similar objects is important in many robotic applications such as visual inspection and robot picking. However, current VLMs focus primarily on recognizing single images, with little consideration given to the recognition of relationships between multiple images. Furthermore, existing research on difference recognition has mainly examined discrete changes such as object removal or movement, without adequately addressing subtle, continuous distinctions like minor attribute differences. Therefore, this study proposes a new dataset, Subtle-Diff, to address these issues and comprehensively evaluates the ability of VLMs to recognize differences between visually similar objects. Experiments using existing VLMs for the proposed task showed that recent models, like GPT-4V, can explain minor differences requiring comparative expressions in language. However, it was found that these models do not match human recognition capabilities across all attribute conditions. Additionally, discrepancies between VLMs and human recognition can occur even for differences that humans can easily explain.

Since existing VLMs do not explicitly address the relationships between images, we propose a model that uses image-text similarity models to explicitly quantify the relationships between images, achieving performance beyond that of GPT-4V. Further melding these models with LLMs and large-scale pretraining may boost these capabilities. Additionally, applying our research to real-world applications, such as visual inspection robots, represents an intriguing direction for future exploration.

REFERENCES

[1] OpenAI. GPT-4 Technical Report. *arXiv e-prints*, page arXiv:2303.08774, March 2023.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[3] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul 2021.

[4] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[5] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs, 2019.

[6] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

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

[8] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[9] Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022.

[10] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6862–6872, June 2023.

[11] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015.

[12] Manuel S Drehwald, Sagi Eppel, Jolina Li, Han Hao, and Alan Aspuru-Guzik. One-shot recognition of any material anywhere using contrastive learning with physics-based rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23524–23533, 2023.

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

[14] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.

[15] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10910–10921, June 2023.

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

[17] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11998–12008, 2023.

[18] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134, October 2021.

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[20] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023.

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[23] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

[24] Boogeo Yoon, Youhan Lee, and Woonhyuk Baek. Coyo-align. `https://github.com/kakaobrain/coyo-align`, 2022.

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, pages 311–318, 2002.

[26] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.