Exploring Grounding Potential of VQA-oriented GPT-4V for Zero-shot Anomaly Detection

Jiangning Zhang[1], Xuhai Chen[2], Zhucun Xue[3], Yabiao Wang[1], Chengjie Wang[1], Yong Liu[2]
[1]Youtu Lab, Tencent [2]Zhejiang University [3]Wuhan University

Abstract—Large Multimodal Model (LMM) GPT-4V(ision) endows GPT-4 with visual grounding capabilities, making it possible to handle certain tasks through the Visual Question Answering (VQA) paradigm. This paper explores the potential of VQA-oriented GPT-4V in the recently popular visual Anomaly Detection (AD) and is the first to conduct qualitative and quantitative evaluations on the popular MVTec AD and VisA datasets. Considering that this task requires both image-/pixel-level evaluations, the proposed GPT-4V-AD framework contains three components: ***1)*** Granular Region Division, ***2)*** Prompt Designing, ***3)*** Text2Segmentation for easy quantitative evaluation, and have made some different attempts for comparative analysis. The results show that GPT-4V can achieve certain results in the zero-shot AD task through a VQA paradigm, such as achieving image-level 77.1/88.0 and pixel-level 68.0/76.6 AU-ROCs on MVTec AD and VisA datasets, respectively. However, its performance still has a certain gap compared to the state-of-the-art zero-shot method, *e.g.*, WinCLIP ann CLIP-AD, and further research is needed. This study provides a baseline reference for the research of VQA-oriented LMM in the zero-shot AD task, and we also post several possible future works. Code is available at https://github.com/zhangzjn/GPT-4V-AD.

Index Terms—GPT-4V(ision), Anomaly Detection, Unsupervised Learning, Zero-shot Learning, Visual Question Answering

1 Introduction

GPT-4V(ision) [1] is a recent enhancement of GPT-4 [2] released by OpenAI. It allows users to input additional images to extend the pure language model, implementing user interaction through a Visual Question Answering (VQA) manner. Recent works [3], [4], [5], [6] have explored its potential in various tasks, demonstrating its powerful generalization capabilities. On the other hand, due to the growing demand for industrial applications and the development of datasets [7], [8], Anomaly Detection (AD) is receiving increasing attention from researchers and practitioners [9], [10], [11], [12], [13], [14]. The recent zero-shot AD is first proposed in WinCLIP [15], a setting dedicated to detecting image-level and pixel-level anomalies in a given image without any positive or negative samples. This setting addresses the pain point of difficulty in obtaining anomaly samples in industrial applications, thus having high practical value. Current approaches [15], [16], [17], [18], [19] distinguishes anomalies based on a large language model, *e.g.*, CLIP [20], which uses pre-trained vision language alignment to achieve anomaly detection. This framework often requires careful design. Unlike the aforementioned approach, this report explores a more general VQA paradigm for the zero-shot anomaly detection task, hoping to bring new ideas to the solution of the zero-shot AD task.

Specifically, this technical report explores the results of VQA-oriented LMM (using GPT-4V in this paper) on the zero-shot AD task for the first time and proposes a general VQA-based AD framework, which includes ***1)*** Granular Region Division (Sec. 2.1), ***2)***Prompt Designing (Sec. 2.2), and ***3)*** Text2Segmentation (Sec. 2.3). We conduct quantitative and qualitative experiments on the popular MVTec AD and VisA datasets (*c.f.*, Sec. 3.2, Sec. 3.3, Sec. 3.4). The results show that GPT-4V has certain effects on the zero-shot AD task, and even surpasses the zero-shot SoTA method in some metrics, such as achieving 88.0 AU-ROC on VisA, surpassing SoTA CLIP-AD by +6.8↑. However, given that the anomaly detection task requires pixel-level grounding capabilities, the current performance of GPT-4V still needs to be further improved. We hope this technical report can promote more zero-shot AD researches, especially the VQA-oriented paradigm.

2 Methodology

As a VQA model, LMM GPT-4V excels at comprehending input images at the semantic level but lacks pixel-level location perception. Anomaly detection tasks require pixel-level segmentation, so this section explores how to unleash the potential of GPT-4V's grounded vision-language capabilities in the AD task. Specifically, we propose a GPT-4V-AD framework that includes three components: Granular Region Division, Prompt Designing, and Text2Segmentation, as shown in Fig. 1.

2.1 Granular Region Division

We first conduct a naive toy experiment. As shown in Fig. 2, when the original image and prompt are directly fed into GPT-4V, the model can generate judgments about defects, but cannot output accurate location results. We believe this is because GPT-4V can align text and object content at the semantic level, but is not adept at aligning text to pixel level. Therefore, *we attempt to transform the AD problem based on VQA into a text and image region grounding problem*, which is more suitable for GPT-4V(ision). We assume that the anomalous regions have some form of connectivity under certain relationship constraints, such as semantics and local structure. The regions associated with the anomalous area should ideally closely adhere to the anomaly itself, ensuring that the segmentation results have a maximum upper limit.

Region Generation. Fig. 3 shows three region generation schemes we attempted: ***1)*** Grid Sampling that uniformly

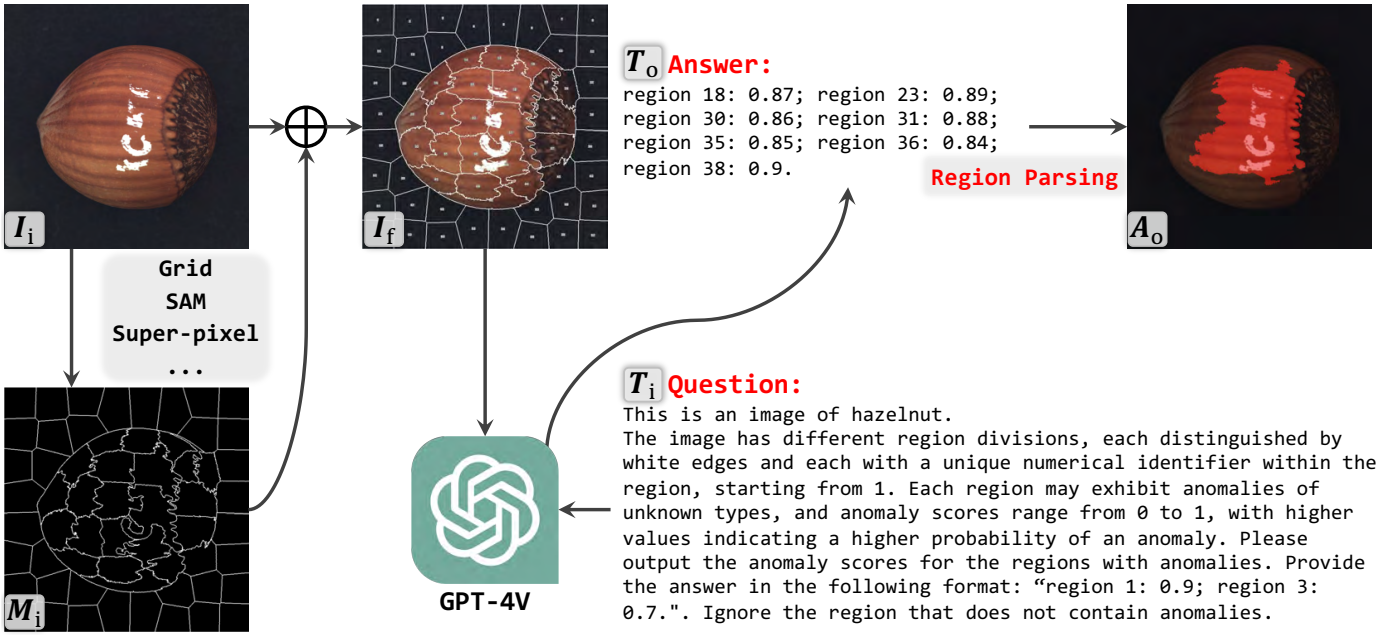

Fig. 1: **Overview of the proposed GPT-4V-AD framework,** which consists of three procedures in tandem: ***1)*** **Granular Region Division** (Sec. 2.1) preprocesses the input image $\boldsymbol{I}_i$, treating pixels that are similar at the structural or semantic level as a common region, resulting in $\boldsymbol{M}_i$. This is then combined with $\boldsymbol{I}_i$ through pixel-wise fusion to obtain the region-divided $\boldsymbol{I}_f$. ***2)*** **Prompt Designing** (Sec. 2.2) designs the suitable prompt $\boldsymbol{T}_i$ for the AD task in conjunction with $\boldsymbol{I}_f$, which is then input into GPT-4V to obtain a formatted output $\boldsymbol{T}_o$. ***3)*** **Text2Segmentation** (Sec. 2.3) combines the regions $\boldsymbol{M}_i$ to parse out pixel-level anomaly segmentation result $\boldsymbol{A}_o$.

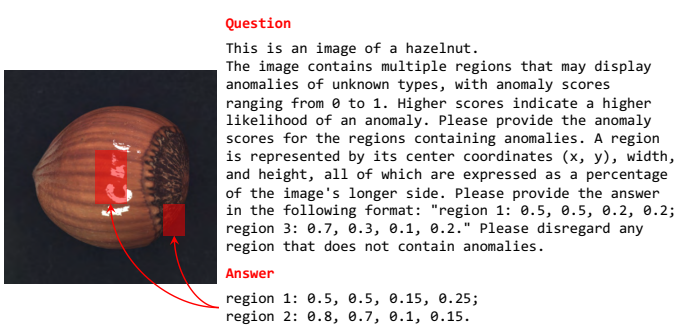

Fig. 2: **A toy experiment with raw image as input and anomaly bounding box (expressed as percentage coordinates) as output for GPT-4V.** This manner leads to uncontrollable and imprecise outputs, and it is challenging to obtains pixel-level segmentation results.

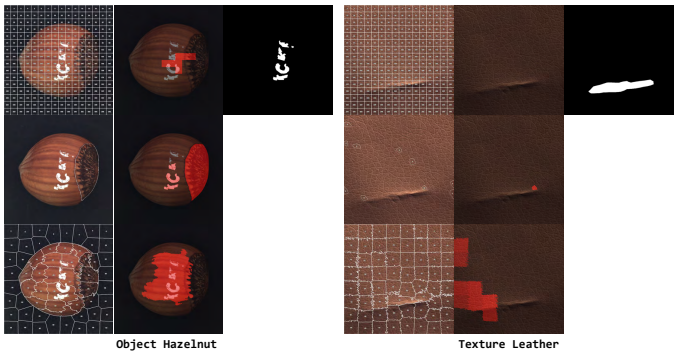

Fig. 3: Ablation study on region division manners, *i.e.*, naive gird, semantic SAM, and structural super-pixel.

samples different areas. ***2)*** Semantic level region division generated by SAM [21]. ***3)*** Super-pixel manner that considers more similarity of layout structure. The qualitative results indicate that the grid manner does not pay enough attention to fine granularity, and the generated regions cannot cover the anomaly area well (*c.f.*, first row of Fig. 3); SAM may overkill by focusing on additional areas with obvious semantics (*c.f.*, second row on the left), or discard areas with unclear semantics (*c.f.*, second row on the right); Super-pixel can relatively stably generate divisions with a high overlap with anomalous ground truth. In addition, the three methods respectively achieve image-level 58.5/71.2/75.4 and 63.7/74.6/77.3 pixel-level AU-ROC results (limited by the number of GPT-4V accesses, experiments are only conducted on object hazelnut and texture leather). Results indicate that grid sampling, which has no structural prior, performs the worst, while the results of super-pixel are the best. Therefore, we recommend super-pixel as the default region generation module for the AD task.

Region Labeling. Similar to recent work [22], the number is used as region labeling, but we outline each region in white without using a color filling. This is because anomaly detection is very sensitive to defect details, and this manner can minimize the impact of excessively small defect areas.

2.2 Prompt Designing

Appropriate prompts are crucial for GPT-4V. For any test image of the popular AD datasets [7], [8], we can obtain the image category. Thus, we design a general prompt description for all categories and then inject the category information to it, *i.e.*, $\boldsymbol{T}_i$ in Fig. 1.

2.3 Text2Segmentation

Through structural output $\boldsymbol{T}_o$, we can easily obtain the final anomaly segmentation result $\boldsymbol{A}_o$ through regular matching combined with preprocessed regions $\boldsymbol{M}_i$.

3 Experiments

3.1 Setup for Zero-shot AD

Task Setting. This investigation discusses the zero-shot AD task recently proposed in WinCLIP [15], which aims to detect image-level and pixel-level anomalies without having seen samples of the anomalous categories. This setting is highly valuable for practical applications, as in many cases, anomalous samples are difficult to obtain or are extremely limited in quantity. Also, due to security reasons, the data may not be transferred externally. *This paper explores the potential of GPT-4V, which is based on the VQA paradigm, in performing this task.*

Dataset. We evaluate GPT-4V(ision) with SoTAs on popular MVTec AD [7] and VisA [8] datasets for both anomaly classification and segmentation. In detail, MVTec AD contains 15 products in 2 types (*i.e.*, object and texture) with 3,629 normal images for training and 467/1,258 normal/anomaly images for testing (5,354 images in total). VisA contains 12 objects in 3 types (*i.e.*, single instance, multiple instance, and complex structure) with 8,659 normal images for training and 962/1,200 normal/anomaly images for testing (10,821 images in total).

Metric. Following prior works [15], [16], we use threshold-independent sorting metrics: ***1)*** mean Area Under the Receiver Operating Curve (AU-ROC), ***2)*** mean Average Precision [23] (AP), and ***3)*** mean F_1-score at optimal threshold [8] (F_1-max) for both image-level and pixel-level evaluations. And ***4)*** mean Area Under the Per-Region-Overlap [24] (AU-PRO) is also employed.

Implementation Details. The input image resolution is set to 768×768 to maintain consistency with GPT-4V's input. In the region divisions, areas smaller than 600 or larger than 120K are filtered out. The edge of the region is outlined with a 1-pixel border, and the numerical labeling is placed within the mask and as centrally as possible within the entire region. For SAM [21], we use ViT-H [25] as the region division backbone, and SLIC [26] is chosen as the super-pixel approach with 60 segments and 20 compactness.

3.2 Quantitative Results on MVTec AD

We evaluate the zero-shot generalization ability of GPT-4V(ision) on the MVTec AD dataset [7]. As shown in Tab. 1, the VQA-oriented framework also has category bias, *i.e.*, it cannot maintain consistent performance across different categories, which is also reflected in CLIP-based contrastive zero-shot methods. The bottom of Tab. 1 shows a comparison with the results of the recent zero-shot methods [15], [17], [19]. It can be seen that the VQA-oriented method has similar effects to the most recent SAA [17], but there is still a certain gap compared to the SoTA results, which still needs further research.

TABLE 1: **Quantitative evaluation on MVTec AD dataset.** Top and middle parts shows .

The top and middle parts respectively show the single-category results of GPT-4V on texture and object. The bottom part shows the average results, as well as a comparison with recent SoTA zero-shot AD methods. The attempted VQA-oriented AD paradigm has achieved considerable results, but there is still a certain gap compared to the CLIP-based contrastive framework.

	Category	Image-level			Pixel-level			
		AU-ROC	AP	F_1-max	AU-ROC	AP	F_1-max	AU-PRO
Texture	Carpet	64.9	62.9	73.0	69.9	4.9	15.6	24.0
	Grid	53.7	61.9	72.2	60.7	1.6	5.0	28.6
	Leather	69.3	62.8	70.3	81.3	5.5	12.8	56.8
	Tile	94.1	93.8	88.8	71.7	17.3	30.0	28.5
	Wood	93.2	90.9	90.3	67.8	5.5	18.8	35.8
Object	Bottle	75.8	83.2	79.9	56.2	11.3	20.6	24.4
	Cable	77.9	77.7	71.7	54.6	4.6	8.1	8.9
	Capsule	55.0	60.3	68.8	63.5	1.5	2.6	38.9
	Hazelnut	81.4	81.4	84.2	73.3	10.8	24.4	50.2
	Metal Nut	96.2	94.9	89.7	52.6	7.2	13.3	22.7
	Pill	97.0	34.7	66.4	83.5	12.9	29.9	30.9
	Screw	99.0	38.0	66.1	74.4	1.5	2.1	29.2
	Toothbrush	75.2	75.5	67.4	88.1	2.5	13.1	60.7
	Transistor	70.6	67.8	71.9	56.5	7.4	14.9	10.2
	Zipper	53.4	62.0	66.4	65.9	1.9	8.3	20.7
	Average (GPT-4V-AD)	77.1	69.9	75.1	68.0	6.4	14.6	31.4
Zero-shot SoTAs	WinCLIP	91.8	96.5	92.9	85.1	-	31.7	64.6
	SAA	44.8	73.8	84.3	67.7	15.2	23.8	31.9
	SAA+	63.1	81.4	87.0	73.2	28.8	37.8	42.8
	CLIP-AD	89.9	95.5	91.1	88.7	28.5	35.3	89.9
	CLIP-AD+	90.8	95.4	91.4	91.2	39.4	41.9	85.6

TABLE 2: **Quantitative evaluation on VisA dataset.** .

The top three parts shows the results of GPT-4V on different categories. The bottom part shows the average results, as well as a comparison with recent SoTA zero-shot AD results. The attempted VQA-oriented AD paradigm has achieved highly competitive results on some metrics, but overall, there is still a certain gap compared to the CLIP-based contrastive framework.

	Category	Image-level			Pixel-level			
		AU-ROC	AP	F_1-max	AU-ROC	AP	F_1-max	AU-PRO
Complex Structure	PCB1	100.	37.7	65.8	70.6	1.3	1.0	8.0
	PCB2	100.	37.0	65.8	67.2	1.5	1.8	6.2
	PCB3	92.7	85.8	91.3	54.7	0.6	1.2	0.3
	PCB4	73.3	71.9	77.6	96.7	10.5	12.3	78.3
Multiple Instances	Macaroni1	95.9	94.3	89.8	98.7	1.8	2.8	34.2
	Macaroni2	69.2	44.4	66.3	76.6	1.0	0.6	41.3
	Capsules	87.2	38.2	67.2	90.5	0.9	1.8	36.8
	Candle	99.2	100	100	54.8	1.1	0.7	30.8
Single Instance	Cashew	54.0	56.0	67.2	57.0	4.3	12.8	6.3
	Chewing Gum	85.4	85.1	76.2	74.8	12.1	31.1	11.9
	Fryum	99.2	99.5	99.9	81.4	15.9	23.2	28.7
	Pipe Fryum	99.5	99.1	100.	96.0	0.1	0.9	71
	Average (GPT-4V-AD)	88.0	70.7	80.6	76.6	4.3	7.5	29.5
Zero-shot SoTAs	WinCLIP	78.1	81.2	79.0	79.6	-	14.8	56.8
	SAA	48.5	60.3	73.1	83.7	5.5	12.8	41.9
	SAA+	71.1	77.3	76.2	74.0	22.4	27.1	36.8
	CLIP-AD	81.2	83.7	80.0	84.1	9.6	16.0	63.4
	CLIP-AD+	81.1	85.1	80.9	94.8	20.3	26.5	85.3

3.3 Quantitative Results on VisA

We further evaluate the performance of GPT-4V(ision) on the popular VisA dataset [8], which contains more small defects and is more challenging. Surprisingly, unlike the MVTec AD results, the proposed VQA paradigm performs significantly better on this dataset and even surpasses the SoTA method on some metrics, such as achieving an image-level AU-ROC of 88.0 that surpasses SoTA CLIP-AD by +6.8↑.

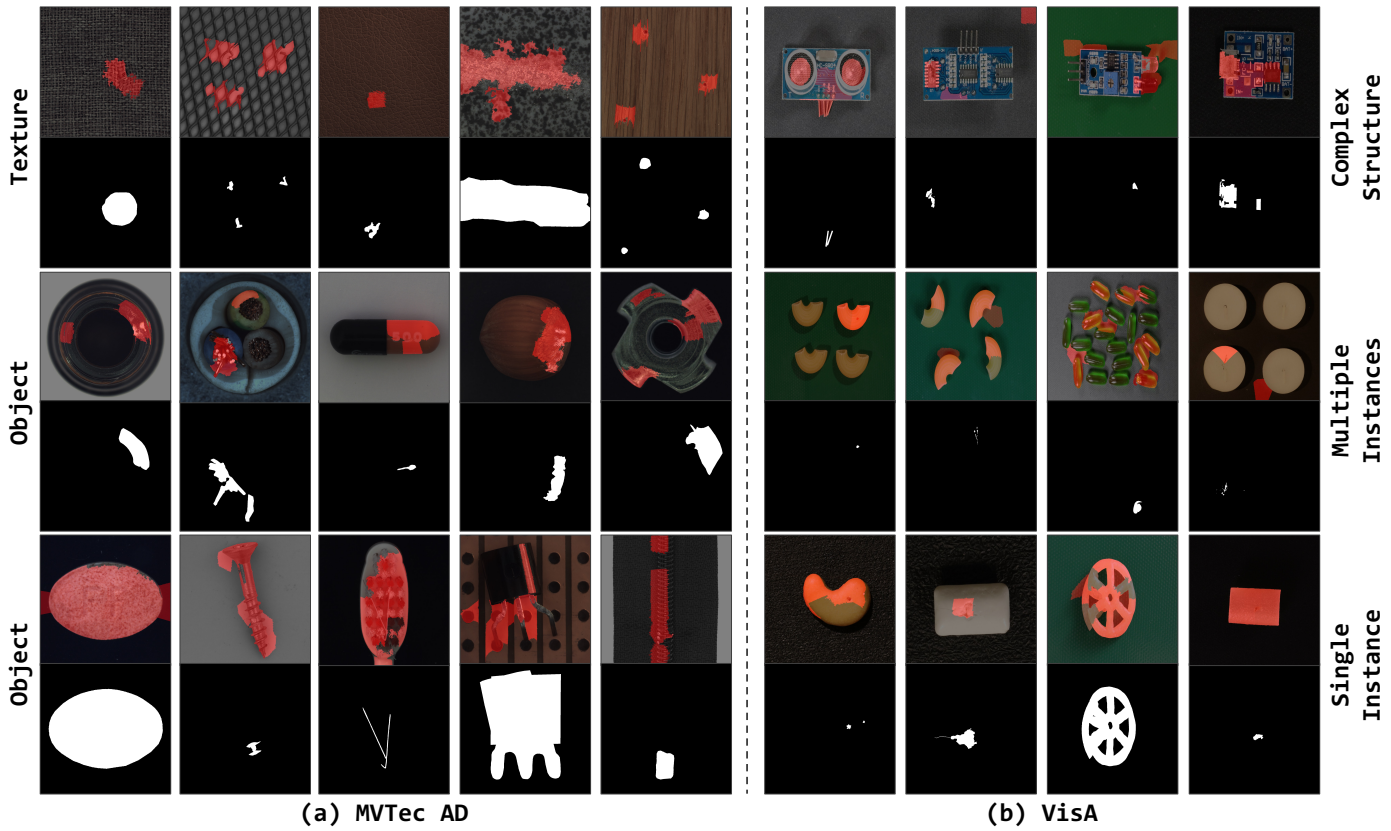

Fig. 4: Non-cherry-picked qualitative results for each category on the MVTec AD (left) and VisA (right) datasets.

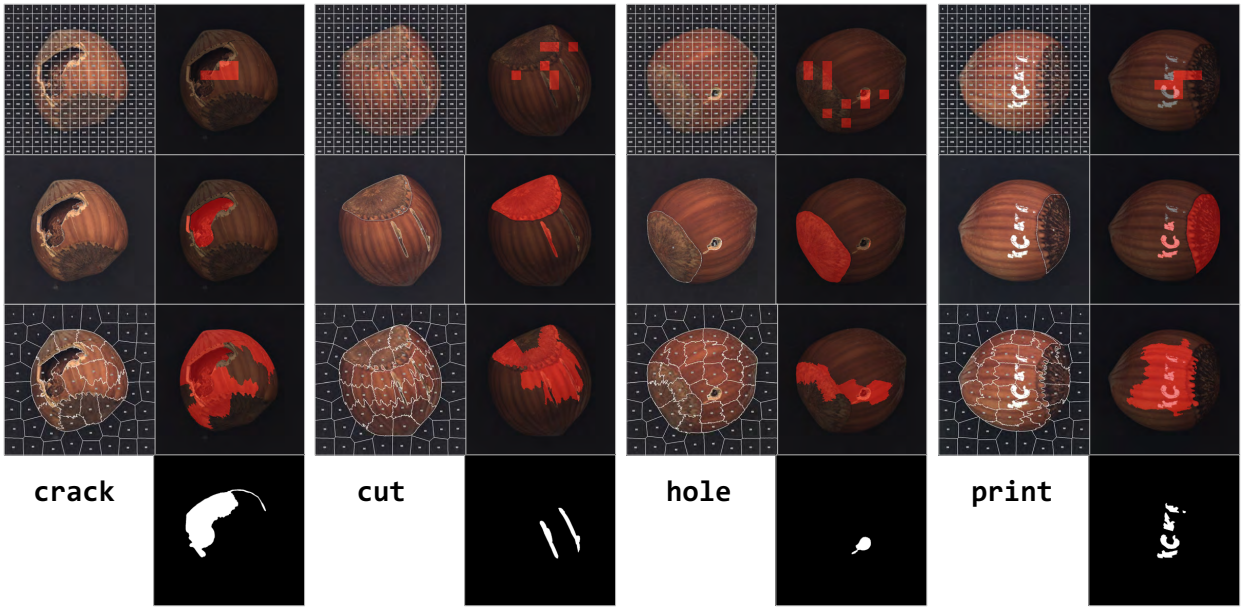

Fig. 5: Qualitative result comparison for different defect categories in the **object hazelnut** on MVTec AD dataset.

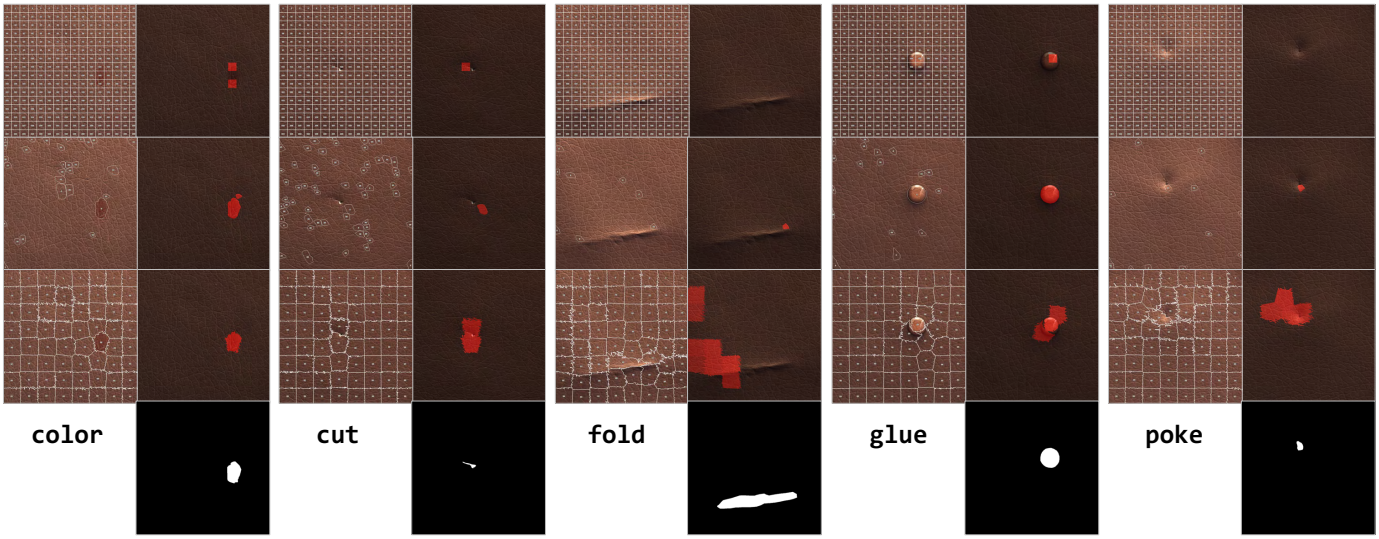

Fig. 6: Qualitative result comparison for different defect categories in the **texture leather** on MVTec AD dataset.

3.4 Qualitative results

Fig. 4 shows the qualitative segmentation results. Even based on the intuitive VQA approach, general-purpose GPT-4V can still obtain segmentation results closer to the ground truth of the anomaly region, demonstrating its powerful visual grounding ability.

3.5 Analyses and Visualization

1) Ablation study on region division manner. This paper adopts three Granular Region Division methods, *i.e.*, grid, SAM [21], and super-pixel [26]. Fig. 5 and Fig. 6 respectively show the qualitative results of the three methods on object hazelnut and texture leather. SAM tends to locate more semantically biased areas, ignoring some weak semantic areas with unclear edges, which has a good effect on defects with obvious semantic classification, *e.g.*, crack in the hazelnut and glue in the leather. However, it may misjudge areas that are too semantically obvious (*c.f.*, bottom of the hazelnut), and miss minor defects (*c.f.*, fold defect in leather). In contrast, super-pixel can pay attention to every area in the image and provide better division results for partial structures than the grid manner.

2) Results with Extra Reference Images. The few-shot setting introduces additional images into the model to improve its performance. Therefore, we further input two extra normal images as reference images into the model. The experimental results in Fig. 7 show that the current version of GPT-4V cannot effectively utilize the additional reference images in the anomaly detection task, but often gets disturbed and cannot output results normally.

3) Repeatability Analysis. We conduct repeated experiments on the consistency of GPT-4V's results. As shown in Fig. 8, when we use the same image and prompt inputs, the output anomaly regions will have slight differences, including both region number and confidence score.

Fig. 7: Analysis of results with extra normal reference images, but GPT-4V often fails to respond as expected.

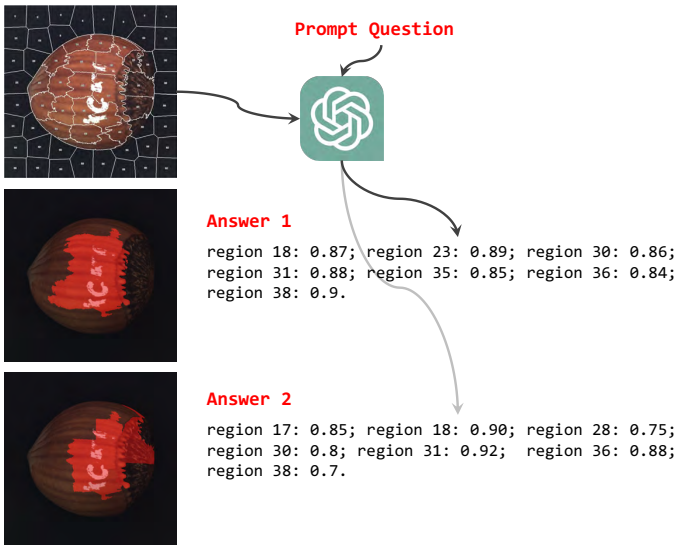

Fig. 8: Stability analysis of VQA-oriented GPT-4V's output with multiple inputs, and there is a slight difference in the results each time.

4 Conclusion and Future Works

This paper explores the potential of VQA-oriented GPT-4V(ision) in the zero-shot AD task and proposes adapted

image and prompt processing methods for quantitative and qualitative evaluation. The results indicate that it has a certain effectiveness on popular AD datasets. Nevertheless, there is still room for further improvement in AD tasks. Moreover, although GPT-4V has achieved epoch-making improvements in human-machine interaction, how to better apply this capability to pixel-level fine-grained tasks remains to be further studied, and the issue of high inference cost needs to be addressed.

Limitations and Future Works. We have summarized some of the current challenges and future works:

1) Due to the limitation on the number of accesses, more suitable image preprocessing methods and more complex prompt designs can be attempted to fully evaluate this task.

2) AD datasets are generally collected from specific scenarios, such as industry, and GPT-4V may have less data from these scenarios in its training set, which could lead to poor generalization performance. Therefore, specific fine-tuning for AD tasks can be studied.

3) This paper only explores the experimental results of VQA-based pure LMM (Large Multi-modal Models), and combining it with current zero-shot AD methods may further improve the model's performance.

4) Prior learning of few-shot normal/anomalous samples should help GPT-4V better understand defects and grounding, which researchers can further explore.

5) Using GPT-4V for semi-automatic data annotation can reduce the cost of manual annotation.

6) The labeling scheme chosen may impact overly small defects, leading to overkill or missed detection. It would be worthwhile to explore more reasonable alternative solutions.

7) The uniqueness of the output from the current paradigm is relatively poor, and there may even be significant gaps among different tests. This paper only provides a preliminary report, and further experiments are warranted for more stable model testing.

REFERENCES

[1] OpenAI. (2023) Gpt-4v system card. Accessed: 2023-11-05. [Online]. Available: https://openai.com/research/gpt-4v-system-card 1

[2] ——. (2023) Gpt-4 research. Accessed: 2023-11-05. [Online]. Available: https://openai.com/research/gpt-4 1

[3] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, vol. 9, 2023. 1

[4] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *arXiv preprint arXiv:2310.11441*, 2023. 1

[5] Y. Wu, S. Wang, H. Yang, T. Zheng, H. Zhang, Y. Zhao, and B. Qin, "An early evaluation of gpt-4v (ision)," *arXiv preprint arXiv:2310.16534*, 2023. 1

[6] Y. Shi, D. Peng, W. Liao, Z. Lin, X. Chen, C. Liu, Y. Zhang, and L. Jin, "Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation," *arXiv preprint arXiv:2310.16809*, 2023. 1

[7] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9592–9600. 1, 2, 3

[8] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 392–408. 1, 2, 3

[9] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 318–14 328. 1

[10] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9737–9746. 1

[11] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le, "A unified model for multi-class anomaly detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4571–4584, 2022. 1

[12] Y. Liang, J. Zhang, S. Zhao, R. Wu, Y. Liu, and S. Pan, "Omni-frequency channel-selection representations for unsupervised anomaly detection," *IEEE Transactions on Image Processing*, 2023. 1

[13] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "Simplenet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 402–20 411. 1

[14] Z. Gu, L. Liu, X. Chen, R. Yi, J. Zhang, Y. Wang, C. Wang, A. Shu, G. Jiang, and L. Ma, "Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 401–16 409. 1

[15] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 606–19 616. 1, 3

[16] X. Chen, Y. Han, and J. Zhang, "A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad," *arXiv preprint arXiv:2305.17382*, 2023. 1, 3

[17] Y. Cao, X. Xu, C. Sun, Y. Cheng, Z. Du, L. Gao, and W. Shen, "Segment any anomaly without training via hybrid prompt regularization," *arXiv preprint arXiv:2305.10724*, 2023. 1, 3

[18] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "Anomalygpt: Detecting industrial anomalies using large vision-language models," *arXiv preprint arXiv:2308.15366*, 2023. 1

[19] X. Chen, J. Zhang, G. Tian, H. He, W. Zhang, Y. Wang, C. Wang, Y. Wu, and Y. Liu, "Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection," *arXiv preprint arXiv:2311.00453*, 2023. 1, 3

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763. 1

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023. 2, 3, 5

[22] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *arXiv preprint arXiv:2310.11441*, 2023. 2

[23] V. Zavrtanik, M. Kristan, and D. Skočaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8330–8339. 3

[24] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4183–4192. 3

[25] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. 3

[26] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012. 3, 5