Zero-Shot Anomaly Detection in Battery Thermal Images Using Visual Question Answering with Prior Knowledge

Marcella Astrid Abdelrahman Shabayek Djamila Aouada
Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg
Email: marcella.astrid@uni.lu, abdelrahman.shabayek@uni.lu, djamila.aouada@uni.lu

***Abstract*—Batteries are essential for various applications, including electric vehicles and renewable energy storage, making safety and efficiency critical concerns. Anomaly detection in battery thermal images helps identify failures early, but traditional deep learning methods require extensive labeled data, which is difficult to obtain, especially for anomalies due to safety risks and high data collection costs. To overcome this, we explore zero-shot anomaly detection using Visual Question Answering (VQA) models, which leverage pretrained knowledge and text-based prompts to generalize across vision tasks. By incorporating prior knowledge of normal battery thermal behavior, we design prompts to detect anomalies without battery-specific training data. We evaluate three VQA models (ChatGPT-4o, LLaVa-13b, and BLIP-2) analyzing their robustness to prompt variations, repeated trials, and qualitative outputs. Despite the lack of fine-tuning on battery data, our approach demonstrates competitive performance compared to state-of-the-art models that are trained with the battery data. Our findings highlight the potential of VQA-based zero-shot learning for battery anomaly detection and suggest future directions for improving its effectiveness.**

***Index Terms*—anomaly detection, zero-shot, visual question answering, thermal image, battery**

I. INTRODUCTION

Batteries have become an essential part of modern technology, supporting a wide range of applications. In the automotive industry, they serve as the core energy source for electric vehicles, replacing traditional gasoline-powered engines known for their adverse effects on human health and the environment [1]. In renewable energy systems, batteries store electricity generated from wind and solar power [2]. Given their widespread use, ensuring battery safety is crucial. Anomaly detection plays a key role in the early identification of potential battery failures [3] while also contributing to efficiency [4]. Since temperature is one of the key parameters in battery monitoring [5], our work focuses on anomaly detection in battery thermal images.

Anomaly detection with deep learning has been widely used in various vision applications [6], [7]. However, collecting data, especially anomalous data, can be challenging, particularly in battery applications where it may pose safety risks. As a result, supervised learning methods that require anomalous data are difficult to implement. To address this, recent methods have focused on training with only normal data [8]. However, even collecting normal/non-anomalous data is expensive [9], [10], such as the time cost of cycling through the battery charge and discharge phases. Therefore, we aim to explore whether zero-shot learning can be a viable alternative to detect anomalies without any battery training data.

Visual Question Answering (VQA) pretrained with a large amount of data has been utilized for zero-shot learning in many vision applications due to its strong generalization capability [11]–[13]. It uses text prompts to perform tasks in vision-based applications. For our purpose, we have prior knowledge of what a normal battery should look like. As mentioned in [8], a normal battery should have a homogeneous thermal distribution and should not overheat (i.e., the temperature stays below a certain threshold). Therefore, we design a prompt exploiting this information as a context to detect anomalies in a zero-shot manner, as illustrated in Figure 1(a).

In this work, we explore three VQA models (ChatGPT-4o [14], LLaVa-13b[1] [15], and BLIP-2[2] [16]) for zero-shot anomaly detection in battery thermal images. We analyze their variability across different rephrased prompts, repeated trials, and qualitative behaviors to identify their strengths and weaknesses. Our simple yet effective approach is compared with state-of-the-art (SOTA) models specifically trained on battery data, and we find that our method remains competitive despite not being fine-tuned on battery data. Finally, we conclude with key findings and discuss potential future research directions based on our observations.

Paper organization: We discuss related work in Section II. The method is detailed in Section III. Section IV covers the experimental setup and results. Finally, our conclusion and future work are presented in Section V.

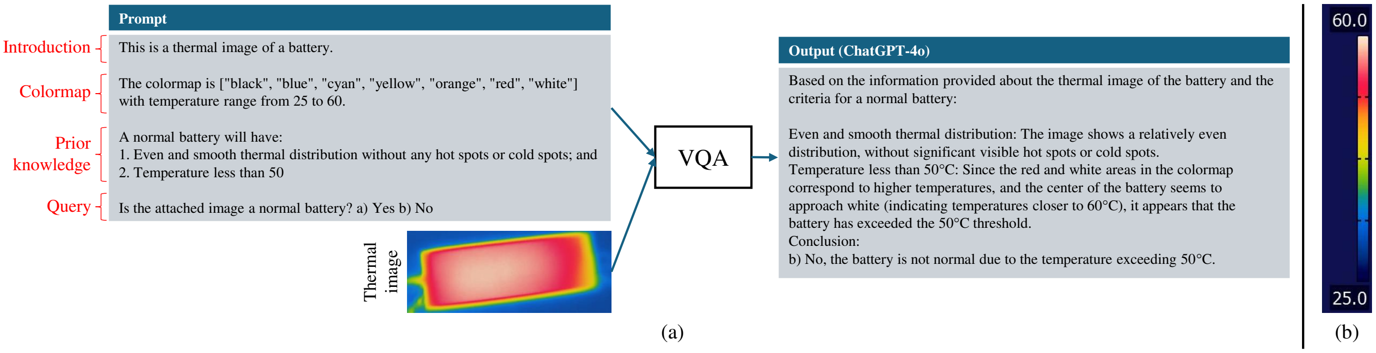
II. RELATED WORK

The two most related works to our study are Zhang *et al.* on zero-shot vision anomaly detection using VQA [12] and Shabayek *et al.* on anomaly detection in battery thermal images [8]. Zhang *et al.* focus on industrial anomaly detection, assuming the presence of anomalous regions that differ from their surroundings. Their approach utilizes super-pixels to define these regions, which is particularly effective for detecting scratches or defects in real-world photographic images.

The research work presented in this article is financially supported by the European Union's Horizon Europe project ENERGETIC (Grant No 101103667).

[1] https://replicate.com/yorickvp/llava-13b

[2] https://replicate.com/andreasjansson/blip-2



This is a thermal image of a battery.

The colormap is ["black", "blue", "cyan", "yellow", "orange", "red", "white"] with temperature range from 25 to 60.

A normal battery will have:
1. Even and smooth thermal distribution without any hot spots or cold spots; and
2. Temperature less than 50

Is the attached image a normal battery? a) Yes b) No

Based on the information provided about the thermal image of the battery and the criteria for a normal battery:

Even and smooth thermal distribution: The image shows a relatively even distribution, without significant visible hot spots or cold spots.
Temperature less than 50°C: Since the red and white areas in the colormap correspond to higher temperatures, and the center of the battery seems to approach white (indicating temperatures closer to 60°C), it appears that the battery has exceeded the 50°C threshold.
Conclusion:
b) No, the battery is not normal due to the temperature exceeding 50°C.

(a) (b)

Fig. 1: (a) Our proposed zero-shot anomaly detection method takes a text prompt and a thermal image as input. The text prompt includes information about the type of image provided (introduction), the colormap range and color, prior knowledge of normal battery characteristics, and the query. (b) Illustration of the colormap used in the thermal image.

TABLE I: Five different prompts are used in this work. The first prompt is handcrafted, while the second to fifth prompts are rephrased using ChatGPT-4o based on the first prompt.

Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5
This is a thermal image of a battery. The colormap is [”black”, ”blue”, ”cyan”, ”yellow”, ”orange”, ”red”, ”white”] with temperature range from 25 to 60. A normal battery will have: 1. Even and smooth thermal distribution without any hot spots or cold spots; and 2. Temperature less than 50 Is the attached image a normal battery? a) Yes b) No	This is a thermal image of a battery. The color scale used is [”black”, ”blue”, ”cyan”, ”yellow”, ”orange”, ”red”, ”white”], representing a temperature range of 25 to 60°C. A normal battery should exhibit: 1. A smooth and even thermal distribution without any hot or cold spots. 2. A maximum temperature below 50°C. Does the attached image indicate that the battery is normal? a) Yes b) No	This thermal image shows a battery. The colormap used includes [”black”, ”blue”, ”cyan”, ”yellow”, ”orange”, ”red”, ”white”], representing temperatures ranging from 25°C to 60°C. A normal battery should exhibit: 1. An even and smooth thermal distribution, free of hot or cold spots; and 2. A temperature below 50°C. Based on the attached image, is this battery normal? a) Yes b) No	The attached thermal image displays a battery with the following colormap: [”black,” ”blue,” ”cyan,” ”yellow,” ”orange,” ”red,” ”white”], corresponding to a temperature range of 25 to 60. A normal battery should meet these criteria: 1. A smooth and even thermal distribution with no hot spots or cold spots. 2. A temperature below 50. Based on this information, is the battery in the image normal? a) Yes b) No	This thermal image represents a battery. The colormap ranges from ”black” to ”white” ([”black”, ”blue”, ”cyan”, ”yellow”, ”orange”, ”red”, ”white”]) corresponding to a temperature range of 25°C to 60°C. For a battery to be considered normal: 1. It should exhibit an even and smooth thermal distribution without any hot or cold spots. 2. The temperature should remain below 50°C. Based on these criteria, does the attached image show a normal battery? a) Yes b) No

However, due to the gradient nature of thermal images of batteries, this method may not be as effective for our purpose. Additionally, overheating anomalies without distinct hot or cold spots do not conform to the assumption of detectable regional differences. Instead, we leverage prior knowledge of normal thermal image characteristics to identify anomalies.

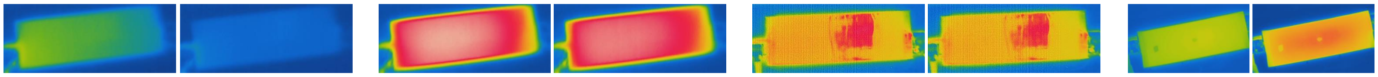
Shabayek *et al*. use normal thermal images to train a model with pseudo-anomaly feature augmentation, simulating artificial defects. However, their method still requires training data, which can be costly and time-consuming to collect. To overcome this limitation, we propose a zero-shot approach that eliminates the need for training data while still enabling effective anomaly detection.

III. METHODOLOGY

The overall method is illustrated in Figure 1(a). The VQA model takes a text prompt and a thermal image as input. Adding context to the prompt has been shown to improve VQA performance in previous works [12], [17]. In our approach, the context includes an introduction to the type of image, the colormap used in the thermal image (as illustrated in Figure 1(b)), and prior knowledge of the normal thermal pattern. The prior knowledge specifies two key characteristics of normal images, as defined in [8]: a temperature below the 50 degree Celsius threshold and a smooth thermal distribution without distinct hot or cold spots. The query is the question related to anomaly detection. Since we utilize prior knowledge of normal data, we formulate the query to ask whether the image is normal or not.

As different prompts may generate different results [18], we create four more prompts rephrased from our first handcrafted prompts. Table I shows all prompts we use in this method. Prompt 1 is our handcrafted prompt. Prompt 2 to 5 are rephrased by ChatGPT-4o from Prompt 1.

This process is performed using a VQA model [14]–[16] pretrained on a large amount of generic data available online. We do not train or fine-tune the model with battery thermal images, making this a purely zero-shot approach. Despite its simplicity, our method remains effective, demonstrating competitive performance compared to specialized models trained on battery data.



(a) Normal (b) Overheating (c) Reflection (d) Spatial tape

Fig. 2: Samples from the test set proposed in [8]. It consists of normal images and three types of anomalies: overheating, reflection, and spatial tape. (a) Normal images show a smooth gradient and a temperature below the threshold. (b) Overheating images exhibit high overall temperatures, even without distinct hot or cold spots. (c) Reflection images display an uneven distribution with hot spots and abnormally high temperatures. (d) Spatial tape cases show cold spots.

TABLE II: Accuracy (Acc.) (%) averaged (Avg.) over multiple trials, range (max - min) (%) of accuracy across multiple trials, and the percentage of unsure predictions across all trials. Tested on different test set splits with various prompts and VQA models. The number of trials is 5, 3, and 3 for ChatGPT-4o, LLaVa-13b, and BLIP-2, respectively.

Model	ChatGPT-4o [14]					LLaVa-13b [15]					BLIP-2 [16]				
Prompt	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Avg. Acc. (all)	73.0	82.3	71.0	79.7	82.3	58.3	55.6	68.9	63.3	52.2	55.0	55.0	55.0	55.0	88.3
Avg. Acc. (normal)	42.2	63.0	37.8	60.0	75.6	11.1	2.5	93.8	29.6	6.2	0	0	0	0	74.1
Avg. Acc. (anomaly)	98.2	98.2	98.2	95.8	87.9	97.0	99.0	48.5	90.9	89.9	100.0	100.0	100.0	100.0	100.0
Avg. Acc. (overheating)	98.5	100.0	100.0	98.5	100.0	92.3	97.4	10.3	87.2	97.4	100.0	100.0	100.0	100.0	100.0
Avg. Acc. (reflection)	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	86.1	100.0	100.0	100.0	100.0	100.0
Avg. Acc. (spatial tape)	95.0	92.5	92.5	85.0	50.0	100.0	100.0	33.3	83.3	83.3	100.0	100.0	100.0	100.0	100.0
Range Acc. (all)	5.0	5.0	16.7	6.7	5.0	3.3	1.7	10.0	3.3	10.0	0.0	0.0	0.0	0.0	0.0
%Unsure (all)	1.0	0.0	0.7	2.7	0.0	9.4	1.7	0.0	41.7	0.0	0.0	0.0	0.0	0.0	0.0

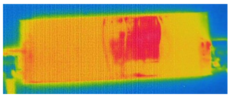
Image	ChatGPT-4o	LLaVa-13b	BLIP-2
	The thermal image of the battery shows a significant hot spot in red and white areas, indicating temperatures well above the normal maximum of 50°C and an uneven thermal distribution. Based on the criteria provided, this battery is **not normal**. Answer: **b) No**	Based on the image, the battery appears to be abnormal, as it exhibits a hot spot in the center. This suggests that the battery might be overheating or experiencing some internal issues, which could potentially lead to performance problems or even damage. The temperature of the hot spot is significantly higher than the surrounding areas, indicating that it is not a normal thermal distribution. Therefore, the answer is b) No.	b)

Fig. 3: Output examples for normal and anomalous data from the three VQA models: ChatGPT-4o [14] (Prompt 2), LLaVa-13b [15] (Prompt 3), and BLIP-2 [16] (Prompt 5). ChatGPT-4o and LLaVa-13b provide explanations in addition to their predictions, while BLIP-2 only generates the final prediction.

IV. EXPERIMENTS

A. *Dataset*

We use the real battery thermal image dataset from [8], evaluating our zero-shot method solely on the test set (Figure 2). The dataset comprises four subsets: (a) normal, (b) overheating, (c) reflection, and (d) spatial tape.

Due to safety concerns, anomalies in (b)–(d) are artificially generated. Overheating images simulate high temperatures (>50°C) with a smooth distribution. Reflection images come from unpainted batteries, creating hot spots and uneven temperatures. Spatial tape images simulate cold spots on unpainted areas. The dataset includes 27, 13, 12, and 8 images for normal, overheating, reflection, and spatial tape, respectively.

B. *Experiment setup*

For each image and text prompt pair, we repeat the experiment five times for ChatGPT-4o and three times for the other models. The temperature for the LLaVa-13b model is set to 0.1. Unsure predictions can occur in a few cases (Section IV-C7 for further details) and are considered anomaly class predictions, as this approach makes more sense in terms of safety.

C. *Results*

We discuss the quantitative and qualitative results of our experiments in this subsection.

1) Output examples: Figure 3 shows output examples from the three VQA models on normal and anomalous data. ChatGPT-4o and LLaVa-13b also provide explanations before making the final anomaly detection prediction. These explanations can be used to further analyze the model, for example, to understand the reasoning behind incorrect predictions.

2) Accuracy across different models: Table II presents the accuracy (%) averaged across multiple trials for different models (Avg. Acc.). Overall, ChatGPT-4o achieves the highest accuracy among the tested VQA models. Averaging the Avg. Acc. across five different prompts on the entire test set, ChatGPT-4o, LLaVa-13b, and BLIP-2 achieve 77.66%, 59.66%, and 61.66%, respectively. These results highlight the importance of selecting an appropriate VQA model.

3) Accuracy across different prompts: Table II shows that Avg. Acc. varies across five prompts, with a range (min-max) across prompts of 11.3% (ChatGPT-4o), 16.7% (LLaVa-13b), and 33.0% (BLIP-2). This highlights the sensitivity of performance to prompt selection, making it a key hyperparameter in model selection.

4) Accuracy on normal versus anomaly: As seen in Table II, Avg. Acc. tends to be higher for anomalous data, suggesting a bias toward false positives. This can be caused by normal data must satisfy both prior knowledge conditions, so misclassification of either can lead to incorrect predictions (Figure 4(a)-(b)). In Figure 4(a), subtle color changes on the

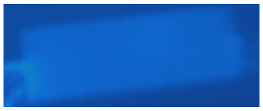
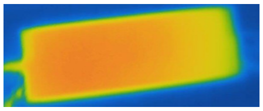
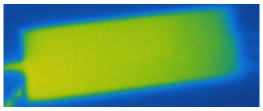
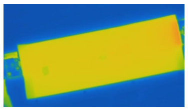
	(a)	(b)	(c)	(d)
Image				
Ground truth	Normal	Normal	Normal	Anomaly
Prediction	Anomaly	Anomaly	Anomaly	Normal
Smooth thermal distribution	No	Yes	No	Yes
Temperature below threshold	Yes	No	No	Yes

Fig. 4: Samples of incorrect predictions in ChatGPT-4o, with the detected condition from the prior knowledge indicated in its explanation. Red text represents incorrect predictions, while green text represents correct predictions.

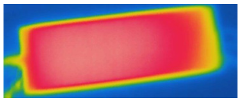
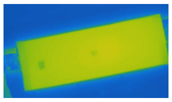
	(a)	(b)
Image		
Ground truth	Anomaly	Anomaly
Prediction	Anomaly	Anomaly
Smooth thermal distribution	No	No
Temperature below threshold	No	No

Fig. 5: Samples of correct predictions in ChatGPT-4o, with incorrectly predicted conditions from the prior knowledge indicated in its explanation. Red text represents incorrect predictions, while green text represents correct predictions.

left part of the image may cause misclassification. In Figure 4(b), a temperature close to 50°C leads to an overheating prediction. In addition to these cases, errors can also arise from misclassifications in both conditions (Figure 4(c)).

Additionally, Figure 5 illustrates cases where reasoning errors still yield correct predictions on anomalous data, leading to higher accuracy on anomalous data. For example, in Figure 5(a), an incorrect anomaly attribution does not affect the final outcome. Similarly, in Figure 5(b), a mistaken high-temperature detection still leads to a correct prediction.

5) Accuracy on different types of anomalous data: Comparing Avg. Acc. (overheating), Avg. Acc. (reflection), and Avg. Acc. (spatial tape) in Table II, the reflection case appears to be the easiest to detect. This is because reflection anomalies exhibit both uneven thermal distribution and overheating conditions, making them more distinguishable compared to overheating and spatial tape cases, which involve only one condition. In particular, the uneven distribution in spatial tape cases can be very subtle, making it more difficult to detect, as seen in Figure 4(d).

6) Accuracy fluctuations across different trials: We report the range (max Acc. - min Acc.) across different trials in Table II. Since accuracy varies between trials, conducting multiple trials helps achieve more reliable predictions and a better understanding of the model's consistency.

7) Unsure cases: Table II reports the percentage of unsure predictions (%Unsure) for each prompt and model, with most coming from LLaVa-13b. Figure 6 illustrates cases where the model either fails to determine conditions (Figure 6(a)) or detects conditions but does not provide a final prediction (Figure 6(b)).

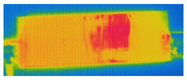
	Image	Response
(a)		I cannot directly analyze the thermal image from your uploaded file. If you can describe the specific features of the image (e.g., where certain colors like red or yellow appear), I can help determine if the battery is normal. Alternatively, you can provide more context or use external software to analyze the image and share the data with me.
(b)		Based on the thermal image, the battery appears to have a temperature of 50, which is within the acceptable range. However, the image also shows a hot spot in the middle of the battery, which could indicate a potential issue or irregularity. To determine if the battery is normal, a more detailed examination of the battery's condition and performance would be required.

Fig. 6: Samples of unsure predictions from the LLaVa-13b model on (a) normal and (b) anomalous data.

TABLE III: AUC (%) comparisons with SOTA methods trained on battery data, using either clean or noisy data. The SOTA method results are taken from [8]. Our method is zero-shot and therefore does not require any training data.

Method	AUC (%) clean train	AUC (%) noisy train	Method	AUC (%) clean train	AUC (%) noisy train
CFLOW-AD [19]	87.3	76.9	STFPM [20]	96.1	87.1
PatchCore [21]	99.0	77.4	CFA [22]	94.2	99.0
FastFlow [23]	100.0	79.3	DRAEM [24]	99.1	92.2
DFM [25]	99.6	80.3	SimpleNet [26]	100.0	97.7
EfficientAD [27]	100.0	81.0	FAUAD [8]	100.0	99.0
PaDiM [28]	99.6	86.3	Ours (zero-shot)	86.6	

8) Comparisons to SOTA: We compare our method with SOTA methods in Table III. Following [8], we evaluate performance using the Area Under the ROC Curve (AUC (%)). We report the AUC of ChatGPT-4o with Prompt 2, averaged across five trials. Despite not using any training data, our method remains competitive with SOTA approaches. Furthermore, since our method does not rely on training data, it is unaffected by noisy training data.

9) Increasing performance on normal data: One weakness of this method is its performance on normal data. We believe that the noisy background may contribute to an uneven thermal distribution. Therefore, we apply rotation and cropping to the test data to remove the background, as illustrated in Figure 7. The results (Table IV) show that in some cases, preprocessing significantly improves performance on normal data. Although performance on anomalous data slightly declines, the substantial improvement in normal data leads to an overall performance increase. Thus, preprocessing thermal images is one of the way to enhance performance.

V. CONCLUSION AND FUTURE WORK

We explore the possibility of zero-shot learning for anomaly detection in battery thermal images using VQA with prior knowledge of the normal data. Our findings suggest that

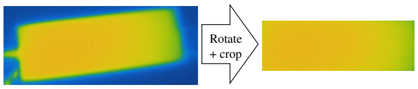


Fig. 7: To remove background noise and enhance performance, especially for normal data, the test data is cropped and rotated.

TABLE IV: Comparison of Avg. Acc. (%) before and after pre-processing. We focus on models and prompts that underperformed on normal data in Table II. The better performance between before and after pre-processing is marked in bold. BLIP-2 performance remains unchanged.

	Before pre-processing						After pre-proprecessing					
Model	ChatGPT-4o		LLaVa-13b				ChatGPT-4o		LLaVa-13b			
Prompt	1	3	1	2	4	5	1	3	1	2	4	5
Avg. Acc. (all)	73.0	71.0	58.3	55.6	63.3	**52.2**	**83.0**	**82.3**	**63.9**	**56.7**	**68.3**	51.1
Avg. Acc. (normal)	42.2	37.8	11.1	2.5	29.6	**6.2**	**65.2**	**63.7**	**49.4**	**7.4**	**50.6**	**6.2**
Avg. Acc. (anomaly)	**98.2**	**98.2**	**97.0**	**99.0**	**90.9**	**89.9**	97.6	97.6	75.8	97.0	82.8	87.9

this approach is promising despite its simplicity as it can compete with SOTA methods even without any training data. However, we identify several weaknesses. First, the method struggles with normal data classification. One simple solution is to preprocess the image to remove unnecessary background. Future work can also incorporate one or a few normal data samples as visual context, making it a one-shot or few-shot approach. Second, predictions can fluctuate across trials, so performing multiple trials is recommended for more reliable results. Third, different prompts can yield varying outcomes, making it necessary to identify the best prompt for each model.

REFERENCES

[1] D. R. Peters, J. L. Schnell, P. L. Kinney, V. Naik, and D. E. Horton, "Public health and climate benefits and trade-offs of us vehicle electrification," *GeoHealth*, vol. 4, no. 10, p. e2020GH000275, 2020.

[2] T. Lehtola and A. Zahedi, "Solar energy and wind power supply supported by storage technology: A review," *Sustainable Energy Technologies and Assessments*, vol. 35, pp. 25–31, 2019.

[3] M. Wu, S. Zhang, F. Zhang, R. Sun, J. Tang, and S. Hu, "Anomaly detection method for lithium-ion battery cells based on time series decomposition and improved manhattan distance algorithm," *ACS omega*, vol. 9, no. 2, pp. 2409–2421, 2023.

[4] S. N. Haider, Q. Zhao, and X. Li, "Data driven battery anomaly detection based on shape based clustering for the data centers class," *Journal of Energy Storage*, vol. 29, p. 101479, 2020.

[5] B. Wang, J. H. Fernandez, and A. Massoud, "A wireless battery temperature monitoring system for electric vehicle charging," in *2019 IEEE SENSORS*. IEEE, 2019, pp. 1–4.

[6] J. Hyun, S. Kim, G. Jeon, S. H. Kim, K. Bae, and B. J. Kang, "Reconpatch: Contrastive patch representation learning for industrial anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2052–2061.

[7] M. Astrid, M. Z. Zaheer, D. Aouada, and S.-I. Lee, "Exploiting autoencoder's weakness to generate pseudo anomalies," *Neural Computing and Applications*, vol. 36, no. 23, pp. 14 075–14 091, 2024.

[8] A. Shabayek, A. Rathinam, M. Ruthven, D. Aouada, and T. Amietszajew, "Ai-enabled thermal monitoring of commercial (phev) li-ion pouch cells with feature-adapted unsupervised anomaly detection," *Journal of Power Sources*, vol. 629, p. 235982, 2025.

[9] A. Herle, J. Channegowda, and D. Prabhu, "Overcoming limited battery data challenges: A coupled neural network approach," *International Journal of Energy Research*, vol. 45, no. 14, pp. 20 474–20 482, 2021.

[10] Y. Liu, S. Chen, P. Li, J. Wan, and X. Li, "Status, challenges, and promises of data-driven battery lifetime prediction under cyber-physical system context," *IET Cyber-Physical Systems: Theory & Applications*, vol. 9, no. 3, pp. 207–217, 2024.

[11] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, and S. Hoi, "From images to textual prompts: Zero-shot visual question answering with frozen large language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 10 867–10 877.

[12] J. Zhang, H. He, X. Chen, Z. Xue, Y. Wang, C. Wang, L. Xie, and Y. Liu, "Gpt-4v-ad: Exploring grounding potential of vqa-oriented gpt-4v for zero-shot anomaly detection," in *International Joint Conference on Artificial Intelligence*. Springer, 2024, pp. 3–16.

[13] M. Castrillón-Santana, E. Sánchez-Nielsen, D. Freire-Obregón, O. J. Santana, D. Hernández-Sosa, and J. Lorenzo-Navarro, "Visual question answering models for zero-shot pedestrian attribute recognition: A comparative study," *SN Computer Science*, vol. 5, no. 6, p. 680, 2024.

[14] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[15] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.

[16] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[17] Z. Hu, P. Yang, B. Li, and Z. Wang, "Multi-agents based on large language models for knowledge-based visual question answering," *arXiv preprint arXiv:2412.18351*, 2024.

[18] J. P. Wahle, T. Ruas, Y. Xu, and B. Gipp, "Paraphrase types elicit prompt engineering capabilities," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 11 004–11 033. [Online]. Available: https://aclanthology.org/2024.emnlp-main.617/

[19] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 98–107.

[20] G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for anomaly detection," *arXiv preprint arXiv:2103.04257*, 2021.

[21] 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.

[22] S. Lee, S. Lee, and B. C. Song, "Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization," *IEEE Access*, vol. 10, pp. 78 446–78 454, 2022.

[23] J. Yu, Y. Zheng, X. Wang, W. Li, Y. Wu, R. Zhao, and L. Wu, "Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows," *arXiv preprint arXiv:2111.07677*, 2021.

[24] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107706, 2021.

[25] N. A. Ahuja, I. Ndiour, T. Kalyanpur, and O. Tickoo, "Probabilistic modeling of deep features for out-of-distribution and adversarial detection," *arXiv preprint arXiv:1909.11786*, 2019.

[26] 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.

[27] K. Batzner, L. Heckler, and R. König, "Efficientad: Accurate visual anomaly detection at millisecond-level latencies," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 128–138.

[28] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *International conference on pattern recognition*. Springer, 2021, pp. 475–489.