Advancing Thermal Fault Diagnostics for Lithium-Ion Energy Storage Systems: An Autonomous Vision Foundation-Based Approach

Huixin Xu[1], Chaoyu Dong[2*], Qian Xiao[1], Yu Jin[1], Hongjie Jia[1]

[1]School of Electrical and Information Engineering, Tianjin University, China

[2]Nanyang Technological University, Singapore

huixin_xu@tju.edu.cn, dong_chaoyu@ntu.edu.sg, xiaoqian@tju.edu.cn, yu_jin@tju.edu.cn, hjjia@tju.edu.cn

Abstract

Thermal fault diagnosis of lithium-ion batteries is vital for safe and reliable energy storage. Leveraging vision foundation models (VFMs) to analyze battery thermal videos provides an end-to-end visual solution that complements signal-based methods. This paper proposes an autonomous vision foundation-based approach for spatiotemporal segmentation and tracking of thermal faults in infrared videos. First, an optimized YOLO detector generates coarse localization of fault regions. These detection boxes then serve as prompts to guide a LoRA-tuned VFM to perform segmentation across the video sequence. Finally, the segmentation masks are used to localize the fault source in the first faulty frame. Both models are trained on a custom thermal video dataset comprising 30 fault scenarios. Experiments show that autonomous YOLO prompts are over 400 times faster than manual annotation while achieving accuracy comparable to standard prompts. LoRA fine-tuning significantly improves the VFM's precision in identifying thermal faults, with the average Jaccard increasing from 0.5771 to 0.9320. Our method also attains high accuracy in fault source localization, with a normalized distance between the tracked and ground-truth locations as low as 0.0004. This work advances the applications of VFMs in thermal fault diagnosis, indicating their potential for intelligent monitoring and maintenance of energy storage systems. Our dataset is released at https://github.com/Huixin-Xu03/LIB-ThermalPerception-Dataset.git

Introduction

Lithium-ion batteries (LIBs) have become the leading rechargeable technology, serving applications from electric vehicles (Hu et al. 2024) to large-scale energy storage (Dong et al. 2020). However, thermal safety remains a critical factor that limits their stable operation (Tian et al. 2024). Localized overheating can cause performance degradation, capacity loss, and increase the risk of thermal runaway (Li et al. 2022). Because temperature directly reflects the thermal state of batteries, monitoring and diagnosing temperatures are essential for preventing thermal runaway and related safety hazards.

Traditional thermal monitoring of LIBs relies on temperature sensors, which can be surface-mounted (Alcock et al. 2021), internally embedded (Wahl et al. 2021), or vertically aligned within electrodes (Yu et al. 2022). However, even in densely instrumented systems, the measurable data from each cell is generally limited to a few discrete points, providing only partial spatial coverage and lacking the capability to capture the whole thermal state. In practical applications, cost and integration constraints further reduce the number of temperature sensors, with the average sensor-to-cell ratio in many battery systems being approximately 1:10 (Zheng et al. 2024). This limited coverage restricts sensor-based diagnostics to only detecting whether a fault exists, without identifying its exact position within the battery pack.

Infrared thermography can map temperature distributions across a wide area. As a visual sensing method, it complements traditional signal-based measurements by providing full-field temperature data. Recent research has used thermal imaging for fault localization and diagnosis in LIBs. (Tian et al. 2024) described a two-stage diagnostic framework that detects faulty cells from thermal images, and (Li et al. 2022) introduced a sequential Transformer-based approach that combines image and signal features to enhance diagnostic precision. While these studies demonstrate the effectiveness of thermal imaging, they mainly rely on image analysis and ignore the temporal changes in thermal anomalies. Video analysis can improve this by enabling more accurate fault localization and tracking how faults develop over time, which helps pinpoint the fault origin.

Recent work in artificial intelligence has explored new approaches for analyzing thermal videos. Vision foundation models (VFMs), trained on large-scale visual datasets, have shown versatility across different visual tasks (Awais et al. 2025). VFMs such as SegGPT (Wang et al. 2023) and

*Corresponding author: Chaoyu Dong

SAM2 (Ravi et al. 2024) process both spatial and temporal information, enabling consistent object segmentation across video frames. Compared to conventional deep learning methods, VFMs allow end-to-end workflows without extensive architectural modification. However, most VFMs are trained on natural videos, which reduces their adaptability for thermal imaging. In addition, thermal faults have unclear boundaries and semantics, making segmentation difficult.

To address the above limitations, this paper proposes an autonomous vision foundation-based approach for thermal fault perception and diagnosis in LIBs. Our method combines task-specific fine-tuning with prompt-based segmentation tailored to thermal video analysis. An optimized YOLO model automatically detects thermal faults and generates coarse localization boxes, which prompts a LoRA-tuned vision foundation model to perform continuous tracking and precise segmentation. The segmentation sequence further enables fault source identification in the initial tracked frame, achieving intelligent spatiotemporal diagnostics without manual intervention. The contributions of this paper are summarized:

1) Autonomous detection and prompt-driven diagnosis: An optimized YOLO model performs initial fault detection, generating bounding box prompts that guide the LoRA-tuned VFM for segmentation and tracking of thermal anomalies across video frames.

2) LoRA-enhanced adaptation for thermal fault perception: Through Low-Rank Adaptation on thermal videos, we enhance the VFM's capability to identify weakly defined and visually ambiguous fault regions in thermal imaging.

3) Thermal video dataset for comprehensive evaluation: The constructed dataset includes 30 internal short-circuit scenarios with varying fault locations, intensities, and diameters, providing a diverse and controlled foundation for effective model training and fine-tuning.

Related Work

In recent years, foundation models (FMs) have emerged as a dominant part of deep learning, demonstrating strong generalization through large-scale pretraining. Inspired by the success of natural language processing, vision foundation models (VFMs) leverage Transformer architectures and attention mechanisms for large-scale visual pretraining (Awais et al. 2025). Representative models include CLIP (Radford et al. 2021) for image-text retrieval, CLIPSeg (Lüddecke et al. 2022) for text-based segmentation, and SAM (Kirillov et al. 2023) for prompt-based image segmentation.

Beyond backbone improvements, SAM introduced the Promptable Visual Segmentation (PVS) framework to improve generality and achieve "segment anything" capability, allowing users to specify targets using points, boxes, or masks for better accuracy and dynamic refinement in various downstream tasks. Using prompts to enable strong cross-domain generalization, SAM has quickly gained popularity in remote sensing (Yan et al. 2023), medical imaging (Ma et al. 2024), and industrial inspection (Wang et al. 2024). SAM consists of an image encoder, a prompt encoder, and a lightweight mask decoder, but its lack of temporal perception limits video applications. SAM2 (Ravi et al. 2024) tackles this using a memory-enhanced architecture that combines current frame features with historical predictions and prompts through the cooperation of a memory encoder, a memory bank, and memory attention mechanism, allowing consistent cross-frame tracking and segmentation. Built on the SAM2 framework, (Zhu et al. 2024) treated 2D image sequences as videos, achieving impressive results in 3D medical image segmentation. (Qiu et al. 2024) used SAM2 for remote sensing change detection, segmenting bi-temporal images to produce stable and accurate boundaries that shape the final change masks. (Zhang et al. 2024) combined SAM2 with the UNI encoder, enabling automated pathology segmentation and setting SOTA on adenoma datasets.

While PVS improves VFMs' adaptability, the automatic generation of effective prompts remains a challenge. The inefficiency and high expense of manual annotation have increased interest in detection-based prompting methods. Recent progress includes combining object detectors with SAM for automated segmentation. Notable examples are GroundingSAM (Ren et al. 2024), which merges text-guided detection with SAM, and medical applications using YOLOv8-driven SAM and HQ-SAM (Pandey et al. 2023).

Problem Statement

Thermal videos provide two main types of information: the precise location of faults in every frame and the progressive development over time. Analyzing thermal videos helps us track how faults evolve and determine when and where they occur. However, VFMs are typically pre-trained on natural images or videos, and their performance often declines when directly applied to thermal data. The blurred boundaries and weak semantics of thermal faults make feature learning more challenging. We use fine-tuned SAM2 within the PVS framework as the foundation of our solution, where autonomous YOLO prompts facilitate accurate segmentation and dynamic tracking. LoRA fine-tuning enables the VFM to quickly adapt to thermal imaging and fault characteristics using only a small amount of video data. The optimized YOLO model, also trained on a tiny dataset, provides real-time analysis and coarse fault localization for video sequences. Our approach focuses on two main tasks: automatic prompt generation and precise fault segmentation.

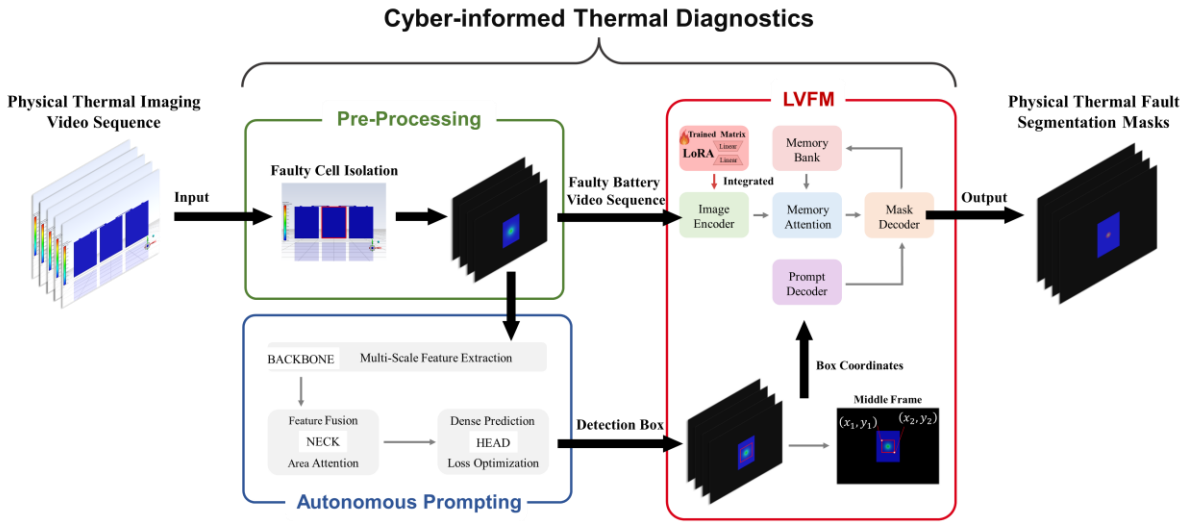

Figure 1: Workflow of the vision foundation-based thermal diagnostic for lithium-ion energy storage systems. (1) Preprocess thermal videos to remove background and isolate faulty cells; (2) Use an optimized YOLO to detect faults and generate box prompts automatically; (3) Apply a LoRA-tuned VFM for precise segmentation, tracking, and fault source localization.

Solution Approach

Our approach's workflow is depicted in Figure 1. The proposed method comprises three stages. First, the thermal video of the battery pack undergoes preprocessing to isolate individual faulty cells, thereby generating a background-removed thermal video sequence of a single faulty cell. Subsequently, the preprocessed video is fed into an optimized YOLO model, which performs frame-by-frame thermal fault detection and generates coarse bounding boxes. The coordinates of the middle frame's box from the detected fault segments are selected as the prompt to guide the next stage's precise segmentation. The middle frame of segments is chosen based on its temporal symmetry, which enhances the prompting representativeness and shortens both forward and reverse inference paths. Finally, a LoRA-tuned visual foundation model (LVFM) integrates the middle frame prompt with each frame's visual features to achieve consistent tracking and precise segmentation of thermal fault regions. The process produces a sequence of thermal fault masks across the entire video and backtracks to the initial fault frame to localize the fault source.

Thermal Fault Dataset of LIBs

In this study, we develop a simulation dataset to train YOLO and the Vision Foundation Model. The dataset comprises video frame sequences of thermal faults, annotated bounding boxes, and label masks for fault regions to facilitate model training and evaluation. The dataset construction adheres to a 7:2:1 ratio across training, validation, and testing.

Modeling and Simulation

We use ANSYS-based finite volume simulations to create our thermal fault dataset, focusing on prismatic $LiFePO_4$ batteries under internal short circuits. The LF50K battery from EVE Energy is modeled in ANSYS Mechanical and meshed in ANSYS Mesh, with detailed parameters provided in Table 1. Nail penetration is simulated using the patch function in ANSYS Fluent to replicate the thermal effects of metallic intrusion.

Parameters	Values
Shape	
Size (mm)	135(L)× 180(H)× 30(T)
Mass (g)	1400
Rating voltage (V)	3.2
Rating capacity (Ah)	50

Table 1: Details of prismatic LiFePO4 batteries.

The electro-thermal behavior is computed based on the MSMD method coupled with the NTGK model. Following the China national standard (National Technical Committee of Auto Standardization 2015), the internal short circuit is modeled as a cylindrical zone vertically penetrating the battery core, with predefined volumetric contact resistance to control fault intensity and location.

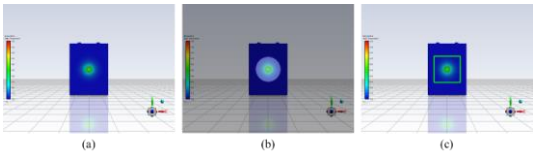

Figure 2: Thermal fault dataset: (a) Thermal image; (b) Label mask; (c) Localization box.

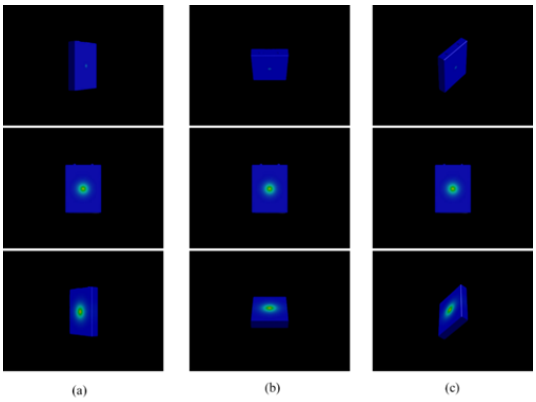

Figure 3: Test videos under dynamic views: (a) Horizontal; (b) Vertical; (c) Diagonal.

The simulation includes 30 scenarios with variations in three short-circuit diameters (6, 8, 10 mm), two resistance values (5×10^{-7} and $1.7\times10^{-7}\,\Omega\cdot m^3$), and five penetration points on the XY plane of the battery (the center point at (0,0) and four offset points at (40,40), (-40,40), (-40,-40), and (40,-40) mm). It begins with the battery at 0% charge and 300 K temperature, applying a surface convection coefficient of 5 W/(m²·K). We perform a 15-second 1C discharge followed by a 75-second short-circuit simulation, recording thermal changes over 90 frames at 480×360 pixels.

Thermal Fault Dataset

Each thermal image is processed to create a binary fault mask and a standard localization box. The mask is generated through temperature contrast filtering, which utilizes the color difference between high-temperature fault regions and low-temperature background areas. Bounding rectangles are then drawn as standard prompts. The dataset includes 30 frontal-view fault scenarios, each with a 90-frame thermal video, containing 75 faulty frames per sequence, totaling 2250 annotated masks and boxes. Figure 2 shows a representative frame with the overlaid mask, and localization box.

To test the model's robustness against viewpoint changes and ensure it can generalize to different perspectives, three videos with horizontal panning, vertical tilting, and diagonal movement are simulated, as shown in Figure 3. Along with the frontal perspective videos, the final dataset includes 21 sequences for training, 6 for validation, and 6 for testing.

Thermal Fault Detection with Optimized YOLO

The initial stage of automated fault diagnosis locates the faulty region and generates a bounding box for the LVFM. YOLOv12 (Tian et al. 2025) is employed to improve prompt quality and efficiency.

YOLOv12 is trained on our thermal dataset. Each image is annotated in YOLO format with standard boxes as labels for fault regions. After training, the optimized YOLO model processes thermal video frames to detect faults and generate localization boxes. The middle frame of each detected segment is selected as the prompt for the LVFM.

Thermal Fault Diagnosis with LoRA-Tuned VFM

Upon receiving automatic YOLO prompts, LVFM leverages prompt information and frame-wise visual features to segment fault regions, track their progression, and identify the fault source.

Low-Rank Adaptation

LoRA (Hu et al. 2022) is a parameter-efficient fine-tuning method that inserts low-rank matrices into the original weights to approximate updates. It preserves the model architecture and input format while training only a small number of parameters.

In a standard linear layer, an input $x \in \mathbb{R}^k$ is transformed by a weight matrix $W_0 \in \mathbb{R}^{d\times k}$, producing the output:

$$h = W_0 x. \quad (1)$$

LoRA introduces two low-rank matrices $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$, where $r << \min(d, k)$ to approximate updates:

$$\Delta W \approx BA. \quad (2)$$

The forward propagation during LoRA fine-tuning is thus:

$$h = W_0 x + BAx. \quad (3)$$

To ensure training stability and comparability across different rank settings, LoRA introduces a scaling factor α:

$$h = W_0 x + \frac{\alpha}{r} BAx \quad (4)$$

where α is typically equal to r (Hu et al. 2022).

LoRA Fine-tuning of VFM

We use a LoRA-tuned VFM (LVFM) to segment fault regions and locate the fault source in the initial fault frame. LoRA fine-tuning is employed to adapt SAM2 to thermal imaging efficiently. As shown in Figure 4, LoRA matrices are inserted into the image encoder's attention projection layers (QKV).

Fine-tuning dataset is assembled from the simulation dataset. Each faulty frame is an independent prompt, producing one training sample per frame. Dice Loss is used to maximize the overlap between predicted and label masks.

The LVFM tracks the spatiotemporal evolution of thermal faults, produces segmentation masks, and identifies fault initiation and duration. This provides valuable insights for fault classification and cause analysis.

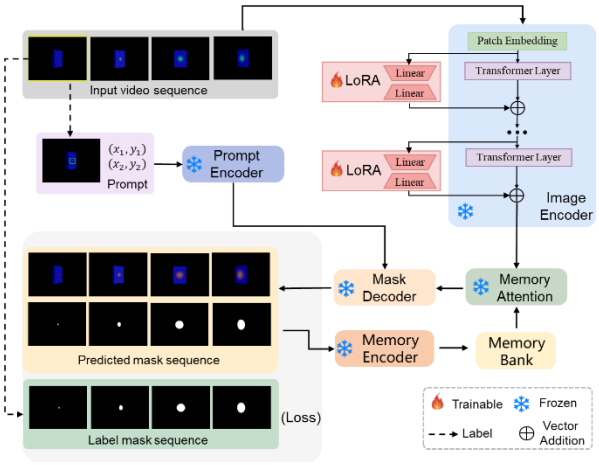

Figure 4: Fine-tuning architecture of LoRA-tuned SAM2.

Experiments

Experimental Setup

YOLOv12 (nano) is trained with CUDA 12.1 on an NVIDIA RTX 4090 GPU. SAM2 (small) is fine-tuned on an NVIDIA A800 GPU, using CUDA 12.1. The configurations are shown in Table 2.

YOLO Training		SAM2 Fine-tuning	
Optimizer	SGD	Optimizer	Adam
Image size	640	Learning rate	5e-6
Batch size	8	Batch size	1
Epoch	50	Epoch	30

Table 2: Parameter configuration for model optimization.

Evaluation Metrics

Metrics for Spatiotemporal Diagnostics of the LVFM

We assess segmentation masks along spatial and temporal axes: spatial segmentation precision using $\mathcal{J}$ (region Jaccard) and $\mathcal{F}$ (boundary F-score), and temporal tracking accuracy with $tIoU$ (time Intersection over Union).

$\mathcal{J}$ measures the overlap between predicted and ground truth masks. For frame t, $M_{(t)}$ denotes the predicted mask and $G_{(t)}$ is the ground truth mask. $\mathcal{J}$ is computed as:

$$\mathcal{J}^{(t)} = \frac{|M_{(t)} \cap G_{(t)}|}{|M_{(t)} \cup G_{(t)}|}. \quad (5)$$

The $\mathcal{J}_{\text{video}}$ is the average over T faulty frames:

$$\mathcal{J}_{\text{video}} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{J}^{(t)}. \quad (6)$$

$\mathcal{F}$ evaluates the boundary consistency between prediction and ground truth. Let $\partial M_{(t)}$ and $\partial G_{(t)}$ denote the predicted and ground truth boundaries at frame t, respectively.

$$\text{Precision}^{(t)} = \frac{|\partial M_{(t)} \cap \partial G_{(t)}|}{|\partial M_{(t)}|} \quad (7)$$

$$\text{Recall}^{(t)} = \frac{|\partial M_{(t)} \cap \partial G_{(t)}|}{|\partial G_{(t)}|}. \quad (8)$$

The F-score at frame t is computed as:

$$\mathcal{F}^{(t)} = \frac{2 \cdot \text{Precision}^{(t)} \cdot \text{Recall}^{(t)}}{\text{Precision}^{(t)} + \text{Recall}^{(t)}}. \quad (9)$$

The $\mathcal{F}_{\text{video}}$ is the average over T faulty frames:

$$\mathcal{F}_{\text{video}} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{F}^{(t)}. \quad (10)$$

$tIoU$ assesses temporal alignment between the predicted and ground truth segments. Let $T_p = [t_s^p, t_e^p]$ be the predicted time interval, and $T_g = [t_s^g, t_e^g]$ the ground truth interval. Then:

$$tIoU = \frac{max(0, min(t_e^p, t_e^g) - max(t_s^p, t_s^g) + 1)}{max(1, max(t_e^p, t_e^g) - min(t_s^p, t_s^g) + 1)}. \quad (11)$$

All the above metrics range from 0 to 1, where higher values indicate better performance.

Metrics for Fault Source Localization Accuracy

The normalized distance between the predicted and ground-truth fault centers in the first faulty frame is calculated to evaluate the accuracy of fault source localization.

Given the coordinates of the ground-truth center (x_{gt}, y_{gt}) and the predicted center (x_{pred}, y_{pred}), the Euclidean distance d is computed as:

$$d = \sqrt{(x_{gt} - x_{pred})^2 + (y_{gt} - y_{pred})^2}. \quad (12)$$

The diagonal length of the image normalizes the distance.

$$d_{norm} = \frac{d}{\sqrt{W^2 + H^2}} \quad (13)$$

where, W and H are the width and height of thermal images.

Results and Analysis

Autonomous Prompting Performance

The YOLO model achieved stable training, with precision and recall both surpassing 0.99 and mAP@50 stabilizing around 0.995 on the validation set. As shown in Figure 5 and Table 3, a comparison on testing data reveals that the segmentation masks produced by autonomous prompts are nearly identical to those from standard prompts, with only minimal metric differences.

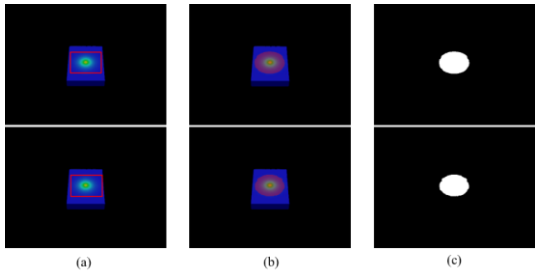

Figure 5: Comparative visualization of segmentation results for a single frame processed by the LVFM (r=8) under different prompting strategies: (a) Prompt box; (b) Segmentation mask; (c) Binary mask. The top row presents the results from the autonomous prompt, while the bottom row shows the corresponding outputs from the standard prompt.

Metrics	Autonomous Prompt	Standard Prompt
$\mathcal{J}_{\text{video}}$	0.9276	0.9320
$\mathcal{F}_{\text{video}}$	0.9955	0.9963
$tIoU$	0.9943	0.9970

Table 3: Test metrics of LVFM of different prompts.

Metrics	Autonomous Prompt	Manual Prompt
Frame Time	0.012 s	5.102 s
FPS	83.333	0.196

Table 4: Inference time of different prompts in testing data.

The results in Table 4 show that autonomous prompting has an average inference time of 0.012 seconds per frame (83.3 FPS), over 400 times faster than manual annotation.

Spatiotemporal Segmentation Performance

The rank r determines the trade-off between expressiveness and parameter efficiency of LoRA matrices, thus influencing fine-tuning performance. SAM2 is compared with four LoRA-tuned SAM2 at r = 8, 16, 32, and 64. The test loss curves in Figure 6 show that all LoRA-tuned models reach lower and more stable losses than SAM2. Among the LoRA configurations, higher ranks generally result in slightly lower losses, suggesting greater adaptation capacity. The quantitative results on the testing data, presented in Table 5, further support the improvement.

Visual comparisons in Figure 7 demonstrate the segmentation improvement at r = 8 as an example. The LoRA-tuned SAM2 generates more precise masks.

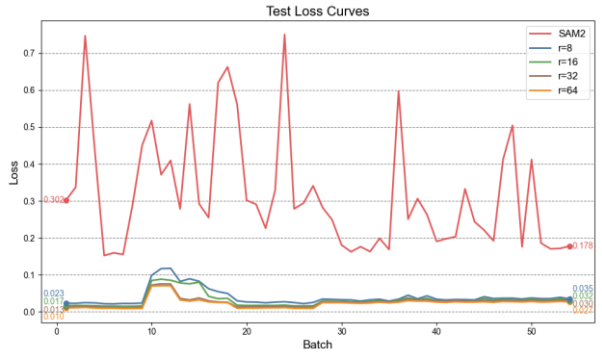

Figure 6: Test loss of SAM2 and LoRA-tuned SAM2.

Metrics	SAM2	r = 8	r = 16	r = 32	r = 64
$\mathcal{J}_{\text{video}}$	0.5571	0.9320	0.9444	0.9527	0.9580
$\mathcal{F}_{\text{video}}$	0.5515	0.9963	0.9969	0.9977	0.9978
$tIoU$	1.0000	0.9970	0.9955	0.9917	0.9977

Table 5: Test metrics of SAM2 and LoRA-tuned SAM2.

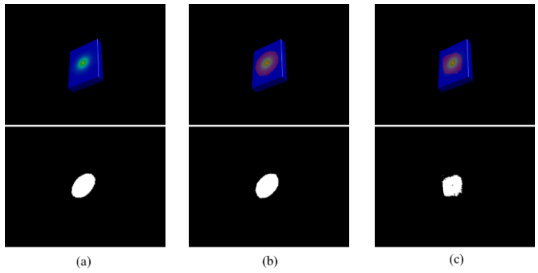

Figure 7: Segmentation results on a test frame: (a) Thermal image and ground truth mask; (b) LoRA-tuned SAM2 prediction; (c) SAM2 prediction.

Experiments show that LoRA fine-tuning significantly improves the VFM's spatial localization and boundary identification accuracy, and that performance gains tend to level off as the rank goes up, with r =16 already providing a good balance between accuracy and parameter efficiency.

Precise Fault Source Localization Performance

The fault source center is localized on the initial fault segmentation mask, and the normalized distance is computed to evaluate localization accuracy on three frontal-view test videos. As summarized in Table 6, the distances between the tracked location and ground-truth location are consistently small, confirming our method's high-precision localization capability for fault sources.

Video ID	Ground-truth Location	Tracked Location	Distance
Video 1	239.50, 181.50	239.38, 181.73	0.0004
Video 2	272.73, 149.50	271.61, 150.14	0.0021
Video 3	206.27, 213.50	207.95, 213.77	0.0028

Table 6: Normalized distance between ground-truth and tracked fault source locations.

Conclusion

This study proposes an autonomous vision foundation-based approach for lithium-ion energy storage systems, an automated method for intelligent diagnosis of thermal faults. By integrating an autonomous YOLO detector with a LoRA-tuned VFM, the method achieves end-to-end automation from fault detection to region segmentation and source localization. Experiments confirm its high accuracy and efficiency in spatial localization of thermal faults, highlighting the potential of VFMs in advancing intelligent diagnostics for thermal safety monitoring and maintenance. Our dataset has been made publicly available to support reproducibility and to benefit the broader research community.

Future work will validate the approach on real-world fault data, optimize models under varied conditions, and benchmark against SOTA models for accuracy and generalization. The visual diagnostic method will be integrated with signal-based techniques for better fault diagnosis and early warning.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. U24B6008).

References

Alcock, K. M.; Grammel, M.; González-Vila, Á.; Binetti, L.; Goh, K.; and Alwis, L. S. M. 2021. An accessible method of embedding fibre optic sensors on lithium-ion battery surface for in-situ thermal monitoring. *Sensors and Actuators A: Physical*, 332: 113061.

Awais, M.; Naseer, M.; Khan, S.; Anwer, R. M.; Cholakkal, H.; and Shah, M. 2025. Foundation models defining a new era in vision: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, early access.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Chen, W. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 1(2): 3.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lüddecke, T.; and Ecker, A. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7086–7096.

Li, M.; Dong, C.; Xiong, B.; Mu, Y.; Yu, X.; Xiao, Q.; and Jia, H. 2022. STTEWS: A sequential-transformer thermal early warning system for lithium-ion battery safety. *Applied Energy*, 328: 119965.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

National Technical Committee of Auto Standardization. 2015. *Safety requirements and test methods for traction battery of electric vehicles: GB/T 31485-2015*. Beijing, China: Standards Press of China.

Pandey, S.; Chen, K.-F.; and Dam, E. B. 2023. Comprehensive multimodal segmentation in medical imaging: Combining YOLOv8 with SAM and HQ-SAM models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2592–2598.

Qiu, J.; Liu, W.; Zhang, X.; Li, E.; Zhang, L.; and Li, X. 2024. DED-SAM: Adapting segment anything model 2 for dual encoder-decoder change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, early access.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:* 2408.00714.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:* 2401.14159.

Dong, C.; Gao, Q.; Xiao, Q.; Chu, R.; and Jia, H. 2020. Spectrum-domain stability assessment and intrinsic oscillation for aggregated mobile energy storage in grid frequency regulation. *Applied Energy*, 276: 115434.

Tian, L.; Dong, C.; Wang, R.; Mu, Y.; and Jia, H. 2024. Anti-interference lithium-ion battery intelligent perception for thermal fault detection and localization. *IET Energy Systems Integration*, 6(4): 593–605.

Tian, Y.; Ye, Q.; and Doermann, D. 2025. YOLOv12: Attention-centric real-time object detectors. *arXiv preprint arXiv:* 2502.12524.

Wahl, M. S.; Spitthoff, L.; Muri, H. I.; Jinasena, A.; Burheim, O. S.; and Lamb, J. J. 2021. The importance of optical fibres for internal temperature sensing in lithium-ion batteries during operation. *Energies*, 14(12): 3617.

Wang, H.; Li, C.; Li, Y.-F.; and Tsung, F. 2024. An intelligent industrial visual monitoring and maintenance framework empowered by large-scale visual and language models. *IEEE Transactions on Industrial Cyber-Physical Systems*, 2: 166–175.

Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; and Huang, T. 2023. SegGPT: Segmenting everything in context. *arXiv preprint arXiv:* 2304.03284.

Tian, L.; Dong, C.; Mu, Y.; Yu, X.; and Jia, H. 2024. Online lithium-ion battery intelligent perception for thermal fault detection and localization. *Heliyon*, 10(4).

Yan, Z.; Li, J.; Li, X.; Zhou, R.; Zhang, W.; Feng, Y.; and Sun, X. 2023. RingMo-SAM: A foundation model for segment anything in multimodal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Yu, Y.; Vergori, E.; Maddar, F.; Guo, Y.; Greenwood, D.; and Marco, J. 2022. Real-time monitoring of internal structural deformation and thermal events in lithium-ion cell via embedded distributed optical fibre. *Journal of Power Sources*, 521: 230957.

Hu, F.; Dong, C; Tian, L.; Mu, Y.; Yu, X.; and Jia, H. 2024. CWGAN-GP with residual network model for lithium-ion battery thermal image data expansion with quantitative metrics. *Energy and AI*, 16: 100321.

Li, M.; Dong, C.; Mu, Y.; Yu, X.; Xiao, Q.; and Jia, H. 2022. Data-model alliance network for the online multi-step thermal warning of energy storage system based on surface temperature diffusion. *Patterns*, 3(2).

Zhang, M.; Wang, L.; Chen, Z.; Ge, Y.; and Tao, X. 2024. Path-SAM2: Transfer SAM2 for digital pathology semantic segmentation. *arXiv preprint arXiv:* 2408.03651.

Zheng, Y.; Che, Y.; Hu, X.; Sui, X.; Stroe, D.-I.; and Teodorescu, R. 2024. Thermal state monitoring of lithium-ion batteries: Progress, challenges, and opportunities. *Progress in Energy and Combustion Science*, 100: 101120.

Zhu, J.; Hamdi, A.; Qi, Y.; Jin, Y.; and Wu, J. 2024. Medical SAM 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:* 2408.00874.