Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 7, No. 3, 2026

Multimodal Large Model Guided Diffusion Model for Transformer Oil Leakage Image Generation

Wenqing Zhao [1,*], Cen Yang [1], Xi Chen [2], Jian Shi [3], An Bo [2], Zenghua Ji [4], Xi Chen [2], Congcong Ma [1], Jing Teng [1], Leipeng Zuo [2], Jieshi Qi [2], Shuang Liang [2], Dongyang Zhang [1]

[1] School of Control and Computer Engineering, North China Electric Power University, Beijing, 102206, China

[2] Baoding Power Supply Branch, State Grid Hebei Electric Power Co., Ltd, Baoding, 071000, China

[3] Baoding Tianwei Baobian Electric Co., Ltd, Baoding, 071000, China

[4] Baoding Tianwei Xinyu Technology Development Co., Ltd, Baoding, 071000, China

[*] **Corresponding author**: (Email: zhaowenqing@ncepu.edu.cn)

Abstract: Transformer oil leakage is a critical defect in power equipment inspection, yet its automatic detection remains challenging because real leakage samples are scarce, highly irregular in shape, variable in color, and easily confused with water stains, rust, shadows, and complex structural backgrounds. To alleviate the data scarcity problem and improve downstream segmentation performance, this paper proposes a multimodal semantic-heuristic diffusion framework, termed MSH-Diff, for automatic transformer oil-leakage image synthesis. The proposed framework exploits the visual reasoning capability and prior industrial knowledge of a multimodal large language model to identify potential leakage-prone components, such as flanges, valves, and radiator fin roots, through expert-role prompt engineering. The recognized component semantics are further mapped to spatial Gaussian anchors by analyzing cross-attention responses in the latent diffusion model, enabling leakage-center localization without manually drawn masks or key-point annotations. In addition, MSH-Diff automatically constructs dense scene descriptions involving illumination, viewpoint, and surface material, which are encoded as semantic constraints to enhance structural consistency and physical realism during diffusion sampling. Experimental results demonstrate that MSH-Diff achieves competitive image quality and diversity, with an FID of 38.14 and an IC-L score of 0.29. When the generated samples are incorporated into downstream semantic segmentation training, the mIoU of DeepLabV3+ increases from 60.09% to 62.31%, confirming the effectiveness of the proposed framework for industrial defect data augmentation.

Keywords: Transformer oil leakage, Diffusion model, Industrial defect image generation.

1. Introduction

With the development of power grids toward extra-high voltage and large-capacity operation, oil-immersed transformers, as key components of power systems, are continuously exposed to high thermal-load cycles, mechanical vibration, and external forces. Under these conditions, sealing components in the transformer body, such as flange joints and weld seams, are prone to fatigue failure, which may cause leakage of transformer insulating oil and seriously degrade the insulation level of the equipment. Such leakage may further develop into fire or explosion accidents, resulting in equipment damage and substantial economic losses. Therefore, rapid detection and accurate localization of transformer oil leakage are of great significance for ensuring the safety of power equipment and the normal operation of power grids. In recent years, automated inspection based on image recognition has become an important trend [1, 2]. In particular, semantic segmentation algorithms can accurately obtain the location information of oil-stained regions with sub-pixel-level precision. However, deep learning models are highly dependent on both data quantity and data quality, while the widespread small-sample and long-tailed data characteristics in industrial scenarios limit their performance ceiling.

As high-reliability equipment, transformers have a relatively low failure probability, and real fault samples are therefore extremely scarce. Even in the few available samples, oil leakage, as a non-rigid target, is visually affected by gravity, surface tension, the attached medium, and the curved surfaces of equipment structures, exhibiting highly irregular streamlined or patch-like distributions. Although traditional data augmentation methods, such as geometric transformations (rotation and cropping) or rule-based pixel perturbations (noise injection and color jittering), can be introduced to expand the dataset [3], they cannot generate new semantic information. The generated samples usually lack realism and are unable to simulate illumination and texture variations under complex operating conditions, thus contributing little to improving model generalization.

Denoising diffusion probabilistic models (Diffusion Models, DDPM) [4] have been widely used for industrial defect image synthesis because of their superior distribution modeling capability and high-fidelity generation performance. However, directly applying a general diffusion model to transformer oil-leakage generation inevitably leads to significant semantic misalignment and spatial misalignment. Conventional diffusion models lack knowledge of industrial equipment and cannot identify leakage-prone locations such as flanges and valves; consequently, the generated leakage locations often violate physical operating rules. Existing controllable generation methods mostly rely on manually drawn hard masks to specify generation regions [5-7]. This not only fails to fundamentally reduce the cost of manual annotation, but also often causes stiff generated boundaries and unnatural blending with the background.

To address these issues, we propose a multimodal

semantic-guided diffusion image synthesis framework, termed the Multimodal Semantic-Heuristic Diffusion Model (MSH-Diff), to achieve a fully automated process from image perception to defect generation. MSH-Diff drives the generation process by automatically extracting key prior parameters. The core idea of the framework is to exploit the strong visual reasoning ability and rich prior knowledge of MLLMs, so that structural risk logic in an image can be inferred and transformed into concrete risk-related semantic labels, enabling the model to act as a virtual industrial expert. The main contributions are summarized as follows:

(1) Prior-knowledge-based risk-region mining: By designing specialized expert-role prompts (Prompt Engineering), the MLLM is guided to finely identify and logically infer the locations of transformer components, such as flanges, valves, and radiator fin roots. Then, with the aid of the spatial response characteristics of cross-attention in the diffusion model, component-level semantic features are converted into explicit spatial coordinates, so that the seepage center can be obtained without manual intervention.

(2) Automatic construction of dense semantic descriptions: The MLLM is used to describe the scene environment, including illumination, material, and viewpoint, from multiple dimensions. The generated dense text provides a semantic reference for consistency constraints, allowing the synthesized samples to adaptively match the structural attributes of the original image while abnormal textures are implanted, thereby improving the physical realism of the generated samples.

2. Related Work

In early industrial defect sample synthesis, image inpainting was the most representative strategy for preserving the background [8]. Owing to their excellent generative capability, diffusion models have gradually become a research focus in generative modeling in recent years, showing unique advantages in both fidelity and diversity of generated samples [9]. Diffusion models were first proposed by Sohl-Dickstein et al. [10] under a nonequilibrium thermodynamics framework, with the construction of a bidirectional Markov chain as the core mechanism. In high-resolution industrial image processing, direct diffusion modeling in pixel space incurs enormous computational cost and memory consumption, as its computational complexity is positively correlated with the number of image pixels. The latent diffusion model (LDM) proposed by Rombach et al. [11] has become a mainstream architecture and effectively alleviates this bottleneck. For transformer images containing tiny cracks or oil-leakage traces. Existing studies attempt to directly guide models to generate anomalous samples by inputting text prompts that describe defect types and appearance characteristics. Although such methods have achieved breakthroughs in generation diversity, they suffer from severe structural uncontrollability in industrial scenarios. To address the structural disorder caused by pure text guidance, recent frontier studies have begun to introduce explicit structural priors or domain knowledge as constraints. For example, Structured-GDM [12] decouples a defect image into three factors: background structure, defect category, and defect morphology, and realizes structured control by introducing contour guidance, category guidance, and shape guidance separately. Anomaly Diffusion [13] achieves a certain degree of location-controllable generation by separating defect appearance embeddings from position embeddings. Although these improvements have made significant progress in plausibility and diversity, they often require additional training of large classification or segmentation networks as auxiliary modules, substantially increasing system complexity. More importantly, these methods still rely on explicit conditional inputs manually provided by users, such as contour maps or bounding boxes, and therefore cannot truly realize spontaneous perception and fully automatic determination of defect locations. In summary, diffusion-based industrial defect generation has achieved promising results, but most existing studies focus on defects on metals, homogeneous materials, or defects with relatively fixed shapes. Transformer oil leakage, in contrast, tends to spread along equipment contours and is affected by transformer geometry and oil fluidity, leading to significant structural dependency and ambiguity. Existing industrial defect generation techniques are therefore difficult to apply directly to this scenario.

3. Method

3.1. Model Architecture

To reduce the dependence of generation-based methods on manual priors, the MSH-Diff framework mainly consists of three components: a multimodal semantic perception module, a prior-parameter mapping module, and a controllable diffusion generation module.

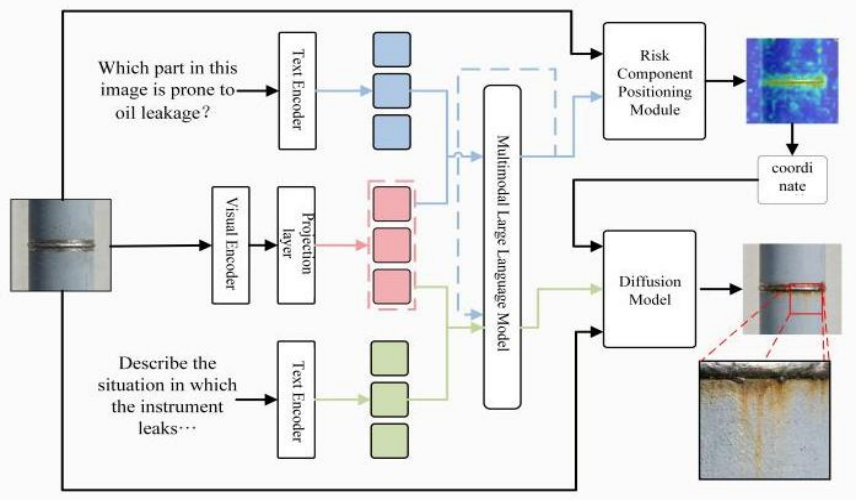

Figure 1. MSH-diff framework

The input of the model is an original transformer inspection image I_{org} and a text prompt. The processing pipeline follows a perception-reasoning-generation paradigm, as shown in Figure 1:

(1) Multimodal semantic perception stage: A pretrained multimodal large language model is used as both the visual encoder and the reasoning agent to process I_{org}. By designing specific prompts, the model is guided to output two types of semantic information. The first is a potential component-risk set K_{parts} (e.g., flanges and valves), which is used in the downstream prior-parameter mapping stage to locate the leakage center. The second is a dense text description T_{dense} containing illumination and material details, which is used to formulate the visual-semantic consistency loss L_{img}.

(2) Prior-parameter mapping stage: This module serves as a bridge between high-level semantics and low-level generation. For spatial localization, the model maps the risk components K_{parts} into the cross-attention space of the diffusion model and automatically calculates the Gaussian anchor P_c representing the leakage center by parsing the attention response maps.

(3) Controllable diffusion generation stage: Using the diffusion model, the automatically extracted P_c and c' are taken as conditional inputs, and reverse diffusion sampling is performed based on I_{org}. The final output is an oil-leakage image with physical plausibility.

3.2. Multimodal Semantic Perception

Unlike traditional convolutional neural networks, which can only assign probability weights to manually annotated categories, the fully automated pipeline for transformer oil-leakage image synthesis requires the MLLM to map visual inputs to natural language expressions. In other words, it must be able to infer equipment and scene attributes from general physical-world knowledge.

Considering both visual-information understanding and language-information reasoning, this study adopts LLaVA (Large Language-and-Vision Assistant) as the base model architecture. Through visual instruction tuning, LLaVA effectively aligns visual features with the semantic space of the language model. Specifically, the reasoning process begins with visual feature encoding. The input image is first fed into a pretrained CLIP-ViT-L/14 visual encoder to extract high-dimensional visual features rich in semantic information. These features are then mapped into a sequence of visual tokens through a linear projection layer, so that their dimensions are consistent with those of text tokens. The resulting visual tokens are used as special image tokens and fed into the language-model backbone (Vicuna-7B). This process leverages the world knowledge stored in the LLM regarding concepts such as transformers, flanges, and oil stains, thereby enabling the transition from pixel-level perception to logical reasoning.

Although general MLLMs have strong generalization ability, they tend to provide broad answers when responding to specific industrial inspection scenarios. Therefore, we propose a role-prompting strategy that constrains the response boundary of the large model by assigning it an expert identity and encourages it to exploit its knowledge of power equipment. Specifically, before the dialogue begins, a high-level prompt is used to instruct the model: "You are an expert in power transformer inspection and are familiar with the appearance and fault characteristics of power transformers." This constraint can effectively modulate the activation of the latent space, encouraging the model to preferentially use terms and knowledge related to mechanical equipment and sealing components during subsequent reasoning, while minimizing the influence of irrelevant information. Considering the algorithmic implementation difficulty of subsequent modules, the overall scene-understanding process is decomposed into two parts: component-risk detection and scene semantic expression, both with strict formatting constraints.

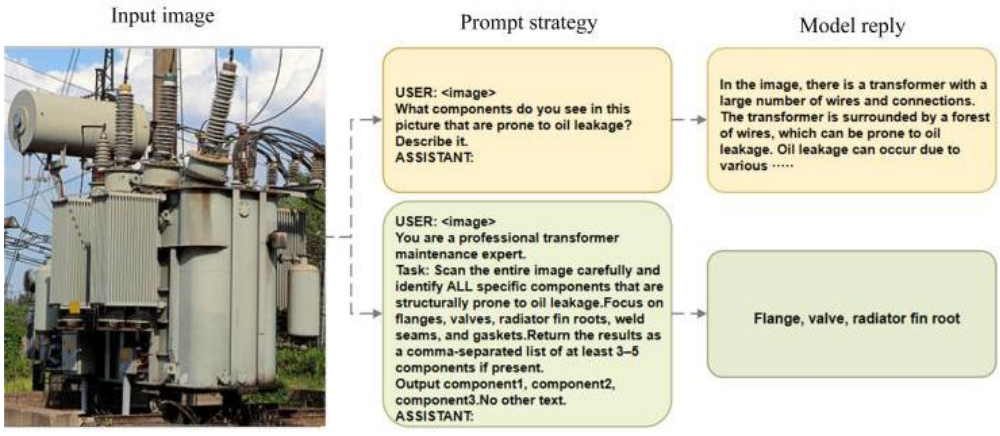

Figure 2. Model response under different prompt strategies

3.2.1. Component-Risk Identification

In the component-risk identification task, the MSH-Diff framework uses a multimodal large language model to deeply decompose image semantics. The design of the perception instruction fully considers the reasoning characteristics of autoregressive models when processing inspection images, as shown in Figure 2.

In general, a large language model tends to greedily search for components in an image, and after identifying a visually salient object in the probability space, generation may be terminated by the end token before the next component is searched for. To ensure the completeness of the risk-component set K_{parts}, we constrain the model through prompts based on full-image search and a specified number of components. This heuristic strategy imposes explicit constraints on the model and performs semantic retrieval across all scales, thereby activating its perception of small failure-prone components such as flanges, valves, and weld seams.

Because the output needs to be connected with subsequent

processing units, the model is required to determine the risk components and to "output comma-separated component names." This requirement removes redundant function words and explanatory statements in natural language, allowing a high-quality sequence of noun phrases to be obtained using a simple regular expression. The resulting clean semantic labels serve as the main input for the final risk-component localization process.

3.2.2. Scene Semantic Description

During controllable diffusion sampling, in addition to controlling the leakage point, the consistency between the generated leakage content and the style of the original background is crucial to the realism of the generated results. Therefore, we use the MLLM to analyze dense semantic information in the background image and obtain the necessary visual conditions for maintaining background style consistency.

In the perception instruction, the model is explicitly guided to focus on the following dimensions:

(1) Illumination conditions: The intensity, direction, and shadow distribution of ambient light are identified to ensure that the reflective properties of the generated oil stains conform to physical optical rules.

(2) Imaging viewpoint: The relative position between the camera and the component, such as a close-up view or a top-down view, is determined to constrain the perspective scale of the leakage flow traces.

(3) Surface material: The physical properties of the attached medium, such as rusty metal or a clean painted surface, are described so that the synthesized oil-stain texture can adapt to the background and produce a corresponding wetting effect.

To satisfy the above requirements, we propose a scene-element-based prompting method. The prompt explicitly requires the model to provide concrete descriptions of illumination, viewpoint, and material, thereby avoiding semantic drift caused by stochastic sampling in general models. For example, the dense description T_{dense} generated by the model, such as a close-up of a rusty flange under strong sunlight, can accurately capture the low-frequency style information of the image, thus providing a precise semantic reference for maintaining background visual consistency through the CLIP text encoder. Specifically, the detailed description T_{dense} output by the MLLM is input into the pretrained CLIP text encoder Φ_{txt} to extract a semantic embedding vector containing rich environmental context information:

$$e_{struct} = \Phi_{txt}(T_{dense}) \tag{1}$$

The vector is then directly used as the target text embedding. The visual-semantic consistency loss L_{img} is obtained as follows:

$$L_{img} = 1 - \cos(\Phi_{img}(D(z_t)), e_{struct}) \tag{2}$$

Where $\Phi_{img}(D(z_t))$ denotes the visual feature of the generated image. Through this automatic mapping mechanism, the MLLM serves as a semantic anchor. For example, when the MLLM describes the image as a rusty flange, L_{img} forces the generation model to preserve rusty texture details while rendering the oil stain. This not only eliminates the cost of manually writing detailed prompts, but more importantly ensures that the synthesized oil stains can automatically adapt to the physical material of the original image, thereby improving the physical realism and structural consistency of the generated samples.

3.3. PerceptionPrior-Parameter Mapping

3.3.1. Semantic Mapping of Risk Components

After the text set of potential risk components K_{parts} is obtained using the multimodal large model, these abstract semantic labels need to be mapped to specific image-space coordinates, so that the center anchor P_c for generating oil leakage corresponding to each risk component can be automatically determined. This process exploits the inherent spatial response property of the cross-attention mechanism in latent diffusion models, namely, the attention-weight map of a specific text token can highlight the corresponding object region in the image.

Let the set of risk components output by the MLLM be $K_{parts} = \{k_1, k_2, \ldots, k_N\}$, where k_i denotes the text description of a single component, such as a flange. To extract the spatial distribution of these components in image I, we encode image I into a latent representation and feed it into a frozen U-Net for a single forward inference pass.

In the l-th cross-attention module of the U-Net, the attention map $M_l^{(i)} \in \mathbb{R}^{H_l \times W_l}$ between the spatial feature Q_l and the text embedding k_i is defined as

$$M_l^{(i)} = softmax\left(\frac{Q_l \cdot k_i^T}{\sqrt{d}}\right) \tag{3}$$

Where d denotes the feature dimension.

Since U-Net contains multiple resolution levels, attention maps from different layers capture low-frequency object structures and high-frequency details, respectively. To obtain localization results, we select intermediate-layer attention maps with resolutions ranging from 16×16 to 64×64 and average them for fusion. This multi-scale aggregation strategy effectively balances semantic completeness and localization accuracy. The aggregated raw response map A_{raw} can be expressed as:

$$A_{raw} = \frac{1}{N}\left(\sum_{l \in L} \mathrm{Resize}(M_l^{(i)})\right) \tag{4}$$

The Resize operation uniformly interpolates attention maps at different scales to 512×512, and the resulting attention response map can represent the spatial distribution of the component.

3.3.2. Coordinate Extraction of Risk Components

Although the raw response map A_{raw} already provides an initial spatial distribution of the component, directly extracting the extremum point is susceptible to background texture noise and may cause the localization result to deviate from the object core. Therefore, we design a coordinate extraction algorithm based on Gaussian smoothing and response enhancement.

Given a two-dimensional Gaussian kernel G_σ, the smoothed global attention map A_{global} is computed as follows:

$$A_{global} = A_{raw} \cdot G_\sigma \tag{5}$$

By searching for the pixel location with the maximum response intensity in the global attention map A_{global}, the center coordinate for oil-leakage generation is calculated as follows:

$$P_c = (x_c, y_c) \text{ where } (x_c, y_c) = \mathrm{argmax}_{(u,v)} A_{global}(u, v) \tag{6}$$

This coordinate is then used as the leakage center.

Through the above algorithm, the model can automatically identify the structural region with the highest probability of leakage in transformer images without manually drawing masks or annotating key points, thereby realizing automatic mapping from semantic understanding to spatial localization.

4. Experimental Results and Analysis

4.1. Dataset and Evaluation Metrics

Because there is currently no public dataset for transformer oil-leakage images, we constructed an experimental dataset by ourselves. First, 50 transformer images were collected from publicly available Internet image resources as base scenes for diffusion-model generation, and an additional 10 oil-stain texture images were collected to provide oil-stain material characteristics and guide texture refinement during generation. In addition, 2,368 transformer inspection images were collected by unmanned aerial vehicles in real substation environments, and the oil-leakage regions in the images were manually annotated to form a test dataset with real defect annotations. This dataset is used to evaluate the influence of generated samples on downstream task performance.

We use the FID metric and the intra-cluster pairwise LPIPS distance (IC-L) metric to evaluate image generation quality and diversity. Meanwhile, three pixel-level metrics, namely mean intersection over union (mIoU), mean precision (mP), and mean recall (mR), are used to evaluate the performance of the downstream segmentation model and to verify the practical value of the generated synthetic data in industrial applications.

4.2. Experimental Environment and Parameter Settings

The experiments were conducted on Red Hat Enterprise Linux Server 7.9 (Maipo) (RHEL 7.9). The deep learning framework was PyTorch 2.5.1, and the CUDA version was 12.1. The experimental platform was equipped with three NVIDIA Tesla V100-PCIE-32GB GPUs.

LLaVA-v1.5-7B provides a favorable balance between accuracy and speed. It consists of a ViT-L/14 (CLIP) image encoder with 303 million parameters and a Vicuna-7B conversational language model with 6.74 billion parameters. A simple MLP is used as the projector to map the two embedding spaces into a shared space. The image size is set to 336×336 pixels. After passing through the MLP projection layer, the image is converted into 576 visual tokens, which are then concatenated with text embeddings and used as the input to the Vicuna-7B backbone model.

In the risk-component identification inference stage, to ensure deterministic output, the temperature is set to 0 and top_p sampling is disabled. The maximum length of text generation is limited to 512 tokens. Float16 half precision is used to optimize memory utilization and accelerate inference.

4.3. Comparative Experiment on Different Prompting Strategies

To verify the necessity of the proposed expert-level prompt, we take risk-component identification as an example and conduct a comparative experiment on base transformer scene images. A general strategy, a coarse expert-guidance strategy, and the full-domain expert heuristic strategy adopted in this chapter are used for comparison and are denoted as prompts A, B, and C, respectively, as shown in Figure 3.

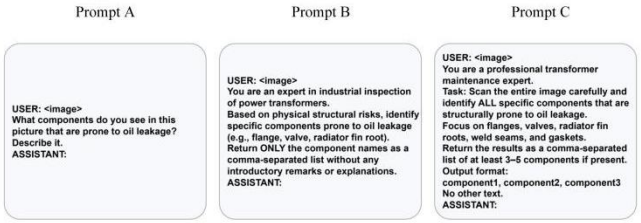

Figure 3. Different strategies for prompts

Table 1. Three Scheme comparing

Prompt Strategies	Model output	Scheme 2	Scheme 3
Prompt A	In the image, there is a transformer with...	11	143
Prompt B	Radiator fin root	45	143
Prompt C	Flange, valve, radiator fin root	126	143

As shown in Table 1, the experimental results show that when prompt A is used, the output of the MLLM contains a large amount of redundant and low-relevance information. In addition, because it lacks prior knowledge of the power industry, the MLLM tends to generate descriptive natural-language paragraphs, and only 11 components are identified.

After the expert role is introduced, prompt B achieves an initial semantic focusing effect and no longer produces divergent scene descriptions; however, it selects only one risk component. This indicates that role assignment can effectively activate the model's knowledge of transformer morphology and remove certain background noise. Nevertheless, because the model generates autoregressively and uses a greedy strategy, inference stops after the first salient object is found, resulting in only 45 identified components.

The final prompt C further introduces full-image search and a quantity constraint on the basis of the above strategy, thereby alleviating model bias. In this case, 126 components are detected, which basically covers the most common leakage-prone locations of transformers. The heuristic strategy can not only identify structural risk points under unknown equipment morphologies, but also theoretically endow the generation model with adaptive adjustment capability for scenarios involving different illumination reflections and complex background interference.

4.4. Comparative Experimental Analysis

To evaluate the contribution of the MLLM to prior-parameter extraction, we compare manual annotation with MLLM-based automatic heuristics and verify the effectiveness of the system under a fully automatic generation pipeline.

Table 2. MSH-diff comparative experimental results

Method	FID↓	IC-L↑
NSA	72.15	0.16
RealNet	69.43	0.19
AnoDiff	48.67	0.21
AnomalyAny	43.42	0.25
MSH-Diff	38.14	0.29

As shown in Table 2, The quantitative comparison of different generation methods in terms of image quality and

diversity shows that models based on traditional anomaly-synthesis strategies perform relatively poorly, with FID values of only 72.15 and 69.43, respectively. This is mainly because traditional methods often rely on simple texture overlay or region replacement strategies, making it difficult to ensure structural plausibility and semantic consistency of defects under complex backgrounds. In contrast, diffusion-based baseline methods achieve substantial improvements on both metrics, demonstrating the superiority of diffusion models for anomaly generation in high-dimensional semantic spaces. Among the diffusion-based methods, the proposed MSH-Diff model can not only maintain diversity but also improve image quality after introducing MLLM-based semantic heuristics.

In terms of generation quality, MSH-Diff eliminates manual intervention while still achieving high-level performance. This indicates that the MLLM has reliable reasoning ability and can spontaneously mine high-risk leakage components such as flanges and valves; the automatically solved Gaussian generation centers are fully consistent with the physical topology of transformers. Meanwhile, the scene semantic descriptions automatically extracted by the MLLM are more objective and dense than manually written short prompts. After being converted into semantic embeddings, these dense texts provide a more precise reference for the visual-semantic consistency constraint L_{img} of the diffusion model, thereby ensuring high fidelity and structural consistency of the generated samples without manual intervention.

In terms of sample diversity, MSH-Diff performs slightly better than the manual mode. From the perspective of the algorithmic mechanism, manual annotation is easily constrained by subjective cognitive inertia, such as a tendency to click similar specific locations or to use repeated descriptive words. In contrast, MSH-Diff exploits the autoregressive sampling property of the language model and a global scanning strategy, enabling it to dynamically output differentiated risk-component combinations and contextual descriptions according to the environmental variables of different input images. After this natural semantic-level variance is mapped into the latent space of the diffusion model, it effectively mitigates mode collapse during generation and further enriches intra-class variations in oil-stain morphology and spatial distribution in the synthetic data.

4.5. Influence of Generated Data on Downstream Task Performance

To verify the application value of the multimodal semantic-heuristic generated data proposed in this chapter for practical engineering tasks, the synthesized anomaly samples are introduced into a downstream semantic segmentation model for performance-gain testing.

To comprehensively verify the practical gains of different generation strategies for downstream defect detection, three independent datasets are constructed for in-depth comparative analysis. The specific data configurations are as follows:

(1) Original training set: This set contains only real collected transformer oil-leakage samples.

(2) Pure MSH-Diff dataset: This set consists entirely of samples generated by the proposed MSH-Diff and contains 2,500 generated samples.

(3) MSH-Diff augmented dataset: This set is constructed by combining the original training data with MSH-Diff generated samples and contains 2,500 generated samples and 1,894 real samples.

During the evaluation of the above three datasets, all validation experiments are conducted using the DeepLabV3+ segmentation network to ensure objectivity and fairness in model performance comparison.

Table 3. Experimental results of MSH-diff downstream task

Training set	mIOU (%)	mP (%)	mR (%)
basic data	60.09	78.66	70.96
Pure MSH-diff data	36.55	39.79	37.20
MSH-diff enhanced data	62.31	79.83	71.43

As shown in Table 3, The downstream segmentation results under different dataset configurations show that the baseline model achieves an mIoU of 60.09% when trained only on the original data. A purely generated dataset is introduced as a control, and the results show that when the model is trained only on pure MSH-Diff data, its performance is poor, with an mIoU of approximately 36%. This phenomenon indicates that although the diffusion model can generate realistic defect textures, purely synthetic data cannot fully cover the complex background interference and illumination variations in real industrial scenarios, leading to a certain domain shift.

When the generated samples are mixed with the original data, however, the data augmentation effect becomes significant. All metrics of the MSH-Diff augmented dataset outperform those of the baseline group, and the mIoU increases to 62.31%. This demonstrates that the generated samples effectively supplement the scarce defect morphologies in real scenarios and alleviate the limitations caused by the long-tailed distribution of industrial defect data. These results indicate that the proposed automatic framework can ensure comparable downstream performance enhancement while eliminating the cost of manually drawing masks and writing descriptive text.

5. Conclusion

In this study, we designed a method based on MLLM-driven semantic guidance and automatic parameter mining. The method exploits the image reasoning capability and physical knowledge base of the MLLM and combines them with expert prompt-engineering techniques to automatically mine potentially risky components in images and generate dense scene information. In implementation, the semantic information of risk components is mapped to spatial Gaussian-coordinate representations according to the cross-attention responses of the diffusion model, enabling localization without masks. By replacing manual priors, the entire process can automatically perceive images and generate high-quality defect results.

References

[1] Wang, Q., Gao, C., Zhang, Z., et al. (2023). SIRN: An iterative reasoning network for transmission lines based on scene prior knowledge. Engineering Applications of Artificial Intelligence, 125. https://doi.org/10.1016/j.engappai.2023.107168

[2] Freitas-Gutierres, L. F., Maresch, K., & Quatrin, A. D. N. (2025). Advancing substation inspection: The Hilbert-Huang transform approach for partial discharge recognition and assessment. Measurement, 116846. https://doi.org/10.1016/j.measurement.2025.116846

[3] Krizhevsky, A., Sutskever, I., & Hinton, E. G. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386

[4] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (Vol. 33, pp. 6840–6851).

[5] Li, Y., Liu, H., Wu, Q., et al. (2023). GLIGEN: Open-set grounded text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22511–22521).

[6] Zavadski, D., Feiden, J. F., & Rother, C. (2024). ControlNet-XS: Rethinking the control of text-to-image diffusion models as feedback-control systems. In Proceedings of the European Conference on Computer Vision (pp. 343–362). Springer Nature Switzerland.

[7] Zhao, S., Chen, D., Chen, Y. C., et al. (2023). Uni-ControlNet: All-in-one control to text-to-image diffusion models. In Advances in Neural Information Processing Systems (Vol. 36, pp. 11127–11150).

[8] Lugmayr, A., Danelljan, M., Romero, A., et al. (2023). Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11461–11471).

[9] Song, J., Park, D., Baek, K., et al. (2025). DefectFill: Realistic defect generation with inpainting diffusion model for visual inspection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18718–18727).

[10] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., et al. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning (pp. 2256–2265). PMLR.

[11] Rombach, R., Blattmann, A., Lorenz, D., et al. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684–10705). https://doi.org/10.1109/CVPR52688.2022.01042

[12] Xie, Y., Pi, X., Zhang, Y., et al. (2025). Structured guided diffusion models for industrial defect image generation. Knowledge-Based Systems, 114642. https://doi.org/10.1016/j.knosys.2025.114642

[13] Hu, T., Zhang, J., Yi, R., et al. (2024). AnomalyDiffusion: Few-shot anomaly image generation with diffusion model. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 8, pp. 8526–8534). https://doi.org/10.1609/aaai.v38i8.28627