DialBench: Towards Accurate Reading Recognition of Pointer Meter using Large Foundation Models

Futian Wang[1], Chaoliu Weng[1], Xiao Wang[1]*, Zhen Chen[3], Zhicheng Zhao[2], Jin Tang[1]

[1]School of Computer Science and Technology, Anhui University, Hefei 230601, China
[2]School of Artificial Intelligence, Anhui University, Hefei 230601, China
[3]Department of Computer Science and Information Technology, La Trobe University, Bendigo, Australia

{*wft, xiaowang, zhaozhicheng, tangjin*}*@ahu.edu.cn*,
e24201133@stu.ahu.edu.cn, Zhe.Chen@latrobe.edu.au

Abstract

The precise reading recognition of pointer meters plays a key role in smart power systems, but existing approaches remain fragile due to challenges like reflections, occlusions, dynamic viewing angles, and overly between thin pointers and scale markings. Up to now, this area still lacks large-scale datasets to support the development of robust algorithms. To address these challenges, this paper first presents a new large-scale benchmark dataset for dial reading, termed RPM-10K, which contains 10730 meter images that fully reflect the aforementioned key challenges. Built upon the dataset, we propose a novel vision–language model for pointer meter reading recognition, termed MRLM, based on physical relation injection. Instead of exhaustively learning image-level correlations, MRLM explicitly encodes the geometric and causal relationships between the pointer and the scale, aligning perception with physical reasoning in the spirit of world-model perspectives. Through cross-attentional fusion and adaptive expert selection, the model learns to interpret dial configurations and generate precise numeric readings. Extensive experiments fully validated the effectiveness of our proposed framework on the newly proposed benchmark dataset. Both the dataset and source code will be released on https://github.com/Event-AHU/DialBench.

1. Introduction

In industrial and power industries, the reading recognition of pointer meters (such as pressure gauges, ammeters, voltmeters) is critical to ensuring equipment safety and process stability. However, these readings are still predominantly performed through manual, periodic inspections, which suf-

---

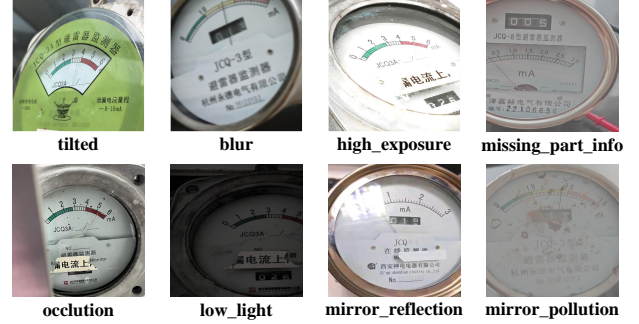*Corresponding Author: Xiao Wang (xiaowang@ahu.edu.cn)

Figure 1. Visualization of dials under diverse environmental conditions, including *low_light*, *high_exposure*, *missing_part_info*, *mirror_pollution*, *tilted*, *blur*, *occlusion*, and *mirror_reflection*. These conditions reflect practical challenges for recognition and detection tasks.

fer from low efficiency, significant subjective errors, frequent omissions or misreadings, inability to enable real-time monitoring, and risks associated with working in hazardous environments. These issues lead to data latency, delayed responses, and a failure to detect transient anomalies, potentially causing equipment overload, energy waste, and even safety accidents, severely hindering the digitalization and intelligent transformation of industrial operations. To address these challenges, it is imperative to develop an AI (Artificial Intelligence)-powered algorithm capable of intelligently and accurately recognizing dial readings. It holds the potential to achieve high-performance automatic recognition, strong generalization capability, end-to-end semantic output, support for multi-modal fusion and contextual understanding, low-cost and rapid deployment, and $7 \times 24h$ continuous monitoring with real-time alerting.

Despite the aforementioned features, AI algorithms specifically designed for this task remain extremely scarce [26], primarily due to the following reasons: 1) The

acquisition of industrial dial images is costly and complex, and often involves enterprise production safety and proprietary information, resulting in an absence of public, large-scale datasets with fine-grained annotations; 2) The diversity of dial types, highly variable industrial environments makes it difficult to develop a unified, generalizable model, hindering scalable deployment and reuse across applications; 3) Critical industries such as power and chemical manufacturing have zero tolerance for misreadings; while traditional rule-based methods offer limited accuracy, they are highly interpretable, whereas the *black-box* nature of AI models often undermines trust and adoption in safety-critical contexts.

Although still scarce, some preliminary efforts have begun to focus on this task. Specifically, Hou et al. [14] proposed a YOLOX-based detection and semantic segmentation framework that accurately localized dial components and achieved a fiducial error below 0.31%. To enhance robustness under image corruption, Wang et al. [37] introduced a Mask Scoring R-CNN (MSC R-CNN) with image corruption augmentation, a balanced aggregation feature pyramid, and a global context block, attaining a 94% successful reading rate even in severely degraded conditions. Fan and Li [12] presented an end-to-end approach integrating YOLOv5 for dial detection and an attention-enhanced U2NET for pointer and tick mark extraction, complemented by CRAFT and CRNN for scale recognition, yielding an average reading error below 5%. More recently, Liu and Shi [20] developed a lightweight YOLOv8S and MC-DeeplabV3Plus-based method that mimics the human reading sequence and incorporates improved attention modules for precise segmentation, achieving fiducial errors of 0.039% in lab settings and 0.733% in real-world scenarios. Collectively, these studies mark important progress toward intelligent meter reading; however, limitations remain in cross-domain generalization, dataset diversity, and model interpretability, which continue to constrain practical large-scale industrial applications.

To address the aforementioned issues, as shown in Fig. 1, we first propose a large-scale benchmark dataset for the reading recognition of pointer meters, termed RPM-10K, which contains 10730 images and comprehensively captures real-world challenges such as *perspective tilted*, *low light*, *blur*, *occlusion*, *missing part information*, *mirror reflection*, *high exposure*, and *mirror pollution*, across more than 300 dial types, while requiring only text-level annotations to reduce labeling cost. Based on this dataset, we retrain and evaluate 18 multi-modal large models, providing a robust data foundation and standardized benchmark for future algorithmic comparison and analysis. More detailed benchmark results can be found in Table 4.

Inspired by the tremendous success of large models in natural language processing [3, 11, 34] and pre-trained

multi-modal models [4, 21, 33], this paper aims to leverage these foundation models to fully enhance dial reading capability. Given an image of a dial and the physical information (i.e., the scale line), we employ a pre-trained large model CLIP to extract feature representations, and then enhance the current image features through cross-attention. Subsequently, we construct a Mixture-of-Expert (MoE) network capable of handling diverse challenging scenarios such as glare and blur. Meanwhile, the visual features are fed into a Q-Former network to further adapt them into the feature space of a large language model. Finally, a large language model decoder is employed to generate the dial reading as output. An overview of our proposed framework can be found in Fig. 4.

To sum up, the key contributions of this paper can be summarized as the following three aspects:

1). We propose a large-scale benchmark dataset for the reading recognition of pointer meter, termed RPM-10K. 18 foundation models are retrained and evaluated on our newly proposed benchmark dataset. The introduction of this benchmark dataset will significantly accelerate the deployment of intelligent dial reading in industrial production.

2). We propose the first vision–language framework for pointer meter reading that explicitly integrates physical relation injection with a Mixture-of-Experts (MoE) architecture, enabling large models to deeply understand about meter physics. This novel paradigm not only significantly improves reading accuracy and robustness under complex conditions but also establishes a new direction for physics-aware multi-modal AI in industrial vision systems.

3). Extensive experiments on the newly proposed benchmark dataset fully validate the effectiveness, robustness, and generalizability of our MRLM framework.

2. Related Work

2.1. Pointer Meter Reading

With the advancement of smart grids, addressing the challenge of analog meter readings has become a significant research topic. Despite the widespread adoption of digital metering systems, many legacy analog meters remain in use, necessitating efficient and accurate reading methods to facilitate integration into modern smart grid infrastructure. Zuo et al. [42] enhanced the automatic reading of pointer meters by integrating an innovative deep learning algorithm. Specifically, it replaced RoIAlign with PrRoIPooling in the Mask R-CNN framework, improving meter type classification while refining the binary mask fitting for the pointer. The final reading was computed using the proposed angle-based calculation method. Li et al. [16] employed deep learning techniques to achieve automatic recognition and reading of pointer meters. Initially, they identified and corrected the scale text in the meter image and applied po-

lar transformation to extract the scale region. Subsequently, secondary search and distance measurement were utilized to accurately locate the pointer position, enabling adaptive detection and reading of various pointer meters. Sun et al. [28] leveraged deep learning for automatic reading of pointer meters by detecting the meter using YOLOv4 and IFF, segmenting the pointer area with Anam-Net, and recognizing scale text and units using CRAFT and E2E-MLT. Additionally, a lightweight CNN was employed to locate the main scale line in the polar coordinate system. Finally, the reading data was computed based on model outputs, ensuring accurate meter reading. Zhang et al. [41] addressed the challenge of automatic pointer meter reading under motion blur by proposing a one-stage network that integrated deblurring and segmentation. It incorporated high-frequency residual attention to refine detail recovery and employed a judgment-reading algorithm to effectively handle 35 types of meters, ensuring accurate recognition and reading. Xiao et al. [38] improved the automatic reading of circular pointer meters by introducing a robust contour-based perspective rectification scheme. It first estimated a rectification matrix by detecting and fitting the deformed meter contour to suppress noise and then applied the matrix to correct the region of interest, thereby enhancing reading accuracy.

2.2. Large Foundation Models

Early foundational work includes Flamingo [1], which pioneered the use of a frozen LLM combined with a perceiver-based visual adapter to process interleaved image-text data, achieving strong few-shot performance on diverse vision-language benchmarks without task-specific fine-tuning. Building on this, LLaVA [17] introduced an instruction-tuned VLM by aligning a vision encoder [25] with an LLM like Vicuna through a projection layer, leveraging GPT-4-generated multimodal instruction-following data for end-to-end fine-tuning. This approach emphasized efficient training and demonstrated superior zero-shot generalization on tasks such as visual question answering (VQA) and image captioning. Subsequent models extended these ideas with specialized adaptations. InstructBLIP [10] enhanced the BLIP-2 framework by incorporating instruction-aware fine-tuning, using a Q-Former to bridge vision and language modalities, resulting in improved performance on instruction-following datasets while maintaining computational efficiency. BLIVA [15] further refined this by integrating BLIP-style bootstrapping with LLaVA-inspired alignment, focusing on better handling of long-context visual inputs and achieving state-of-the-art results in grounded language understanding. The Qwen-VL series [2, 31, 35] from Alibaba advanced multilingual and high-resolution capabilities, employing a transformer-based architecture with dynamic visual token compression to sup-

port diverse languages and detailed image analysis, making it particularly effective for real-world applications like document understanding.

Datasets and models like ShareGPT-4V [5] contributed by providing high-quality, GPT-4V-curated multimodal instruction data, which has been instrumental in fine-tuning VLMs for enhanced visual instruction following, often serving as a backbone for community-driven improvements. The InternVL series [6–8, 36] pushed boundaries with internal scaling laws, combining a large vision foundation model with an LLM for interleaved multimodal processing, achieving competitive results on benchmarks like MM-Vet [40] and LLaVA-Bench through progressive alignment stages. These models collectively represent a progression from early adapter-based integrations to sophisticated instruction-tuned systems, highlighting trends toward scalability, efficiency, and multimodal coherence in VLMs.

3. RPM-10K Benchmark Dataset

To provide an intuitive understanding of the dataset, we present two sets of visualizations. Fig. 2 illustrates representative dial types in the dataset, where subfigures (a)–(f) correspond to the six primary categories of dial types and subfigures (g)–(p) show additional samples collected from web-based sources. Fig. 1 highlights the variability of dial appearances under diverse environmental conditions.

3.1. Protocols

To ensure reproducibility and fair comparison, we define standardized protocols for the usage of RPM-10K. These protocols guarantee consistent and fair evaluation across different methods and emphasize the novelty of our language-based labeling paradigm for automated meter reading.

• **Data Sources:** The dataset consists solely of images paired with language labels, without explicit visual annotations (e.g., bounding boxes or segmentation masks).

• **Diversity of Meter Types:** The dataset covers over 300 types of dials with distinct layouts, scales, and pointer configurations. This diversity significantly increases the difficulty of generalization and requires models to robustly handle heterogeneous dial designs.

• **Environmental Conditions:** Each sample may be affected by one or more of eight complex conditions: low brightness, high exposure, reflection, contamination, blur, tilt, occlusion, and partial information loss. These factors jointly simulate real-world acquisition challenges and stress-test the robustness of dial reading methods.

• **Language Label Annotation Rule:** Instead of conventional visual annotation formats, we directly assign each meter image a textual label representing its reading, e.g., "6.45". For multi-dial mechanical instruments, readings from inner/outer scales, mother–child dials, and multi-
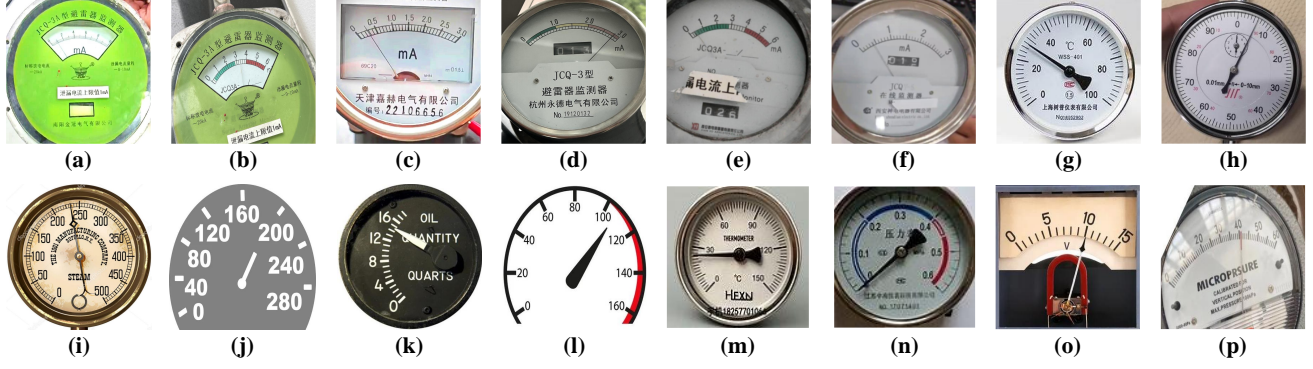
Figure 2. Illustration of the dial types in the dataset. Subfigures (a)–(f) represent the six primary dial categories, whereas (g)–(p) correspond to samples acquired from online sources.

Range	Index Value	#Dials	#Samples
10	0.2	1	2,645
3	0.1	3	502 / 1,155 / 2,056
6	0.2	2	1,290 / 2,182

Table 1. Dial range and index configurations in the dataset.

(a) Distribution of meter types

(b) Distribution of metering environment

Figure 3. Distribution of samples across different dial configurations and environmental conditions.

pointer dials are annotated separately according to their scale or pointer color. Leading zeros and decimal points are preserved to ensure numerical fidelity, enabling end-to-end training of vision–language models without intermediate detection or OCR stages.

• **Train/Test Split:** The dataset is divided into two non-overlapping subsets such that no meter instance appears in more than one split. A typical split ratio of 81%/19% is adopted for training and testing, respectively.

• **Evaluation Setting and Usage Protocol:** During evaluation, models receive only raw meter images as input, and their predictions are compared with the ground-truth language labels using the metrics described above. For methods that require fine-tuning, training must be strictly conducted on the designated training split, while the test split is reserved solely for final performance reporting.

3.2. Data Construction and Annotation

The dataset is constructed by extracting frames from meter-reading videos at regular intervals and supplemented with nearly 2,000 web-collected images. All images are manually cropped to retain the dial region and annotated with textual labels. In total, the dataset contains 10,730 meter images, partitioned into 8,730 for training and 2,000 for validation.

3.3. Statistical Analysis

The dataset captures meter readings under controlled variations in dial specifications and environmental conditions,
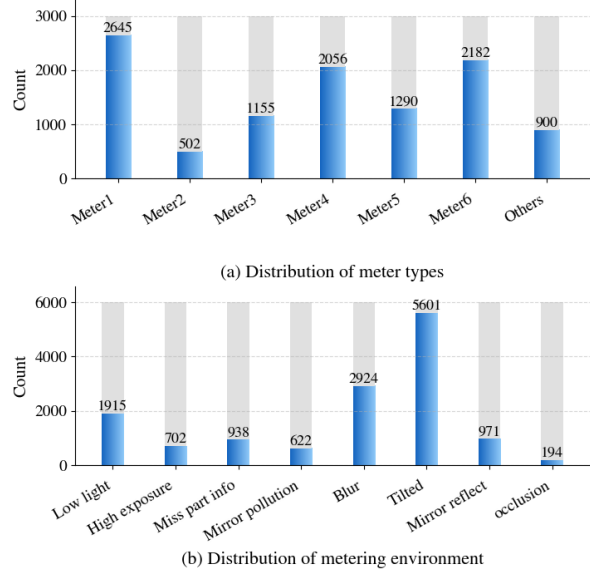
providing a realistic approximation of field scenarios. It covers six major dial types with relatively small measurement ranges, emphasizing fine-grained reading differences.

• **Dial Specifications.** Most dials feature small measurement ranges. Table 1 summarizes the configurations in terms of range, index value, and sample count. These settings reflect the focus on compact-range dials frequently encountered in practice, which require precise discrimination between visually similar readings. Fig. 3(a) visualizes the sample distribution across different range–index combinations.

• **Visual Distribution Analysis.** Fig. 3(b) shows the proportion of samples under various environmental conditions. *tilted* and *blur* dominate, consistent with their prevalence in

real-world deployments. Challenging cases such as *missing_part_info* and *mirror_pollution* are less frequent but introduce significant visual degradation.

Overall, the dataset maintains a balanced composition between common and hard cases, facilitating the development of models robust to diverse real-world conditions.

3.4. Benchmark Baseline

We evaluate a diverse set of state-of-the-art VLMs as baselines on the proposed DIALBENCH benchmark. The evaluated models include Qwen2-VL, InternVL3, LLaMA3.2-Vision, the LLaVA family, BLIVA, and other representative architectures spanning different model sizes and release years. These baselines provide a comprehensive reference for assessing the difficulty of RPM-10K and the effectiveness of future methods.

4. Our Proposed Method

4.1. Overview

As shown in Fig. 4, we propose a unified Physical Relation Injection (PRI) framework for high-precision automated meter reading, termed the MeterReading Large Model (MRLM). Unlike conventional visual recognition systems, MRLM explicitly injects physically grounded relations throughout its multi-modal perception pipeline, enabling consistent and physically informed understanding under diverse real-world conditions. The architecture integrates three key innovations hierarchically: the Key Feature Mining (KFM) module grounds perception in physically meaningful entities such as pointers, scales, and digits, enhancing essential components while suppressing background noise; the Mixture-of-Experts (MoE) module models dynamic interactions between meter components, enabling adaptive reasoning under varying visual conditions; and the language-labeled supervision mechanism linguistically constrains the model's understanding of physical relationships, aligning visual reasoning with symbolic concepts. Together, these components form a comprehensive PRI process that unifies low-level physical grounding, mid-level relational modeling, and high-level symbolic representation to achieve accurate and consistent meter reading.

4.2. Key Feature Mining

The **KFM** module performs the first stage of physical relation injection by grounding the model's perception on physically meaningful entities. Guided by a prior knowledge base of canonical meter archetypes, it isolates and enhances critical dial components, i.e., pointers, scales, and digits, while suppressing irrelevant background.

Template features are first encoded using a pre-trained CLIP model, where early-, mid-, and late-layer features are
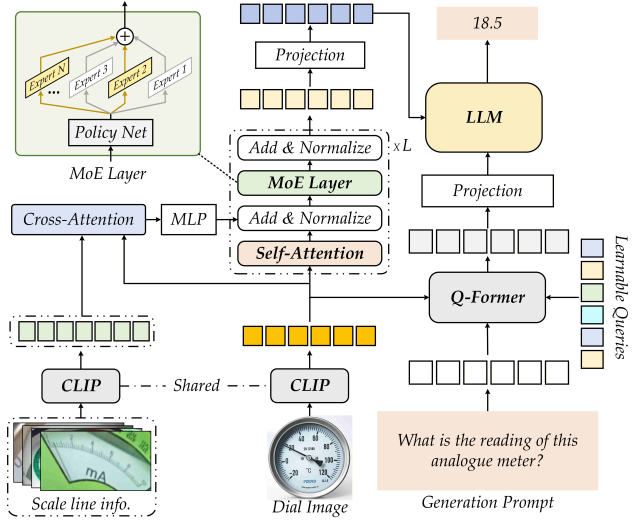

Figure 4. Overview of the proposed MeterReading Large Model (MRLM) based on the Physical Relation Injection (PRI) framework. The pipeline sequentially injects physical relations at three hierarchical levels: entity grounding (KFM), relational coupling (MoE), and symbolic alignment (language-labeled supervision).

concatenated:

$$F_i = \mathrm{Concat}(F_{i,\mathrm{early}}, F_{i,\mathrm{mid}}, F_{i,\mathrm{late}}), \quad (1)$$

$$F_{\mathrm{template}} = \mathrm{Concat}(F_1, F_2, F_3, F_4, F_5, F_6), \quad (2)$$

with $F_{i,\mathrm{early/mid/late}} \in \mathbb{R}^{256 \times 1408}$, yielding $F_i \in \mathbb{R}^{256 \times 4224}$ and $F_{\mathrm{template}} \in \mathbb{R}^{1536 \times 4224}$. These template features act as the physical query in a cross-attention operation, where the image features $F_{\mathrm{img}} \in \mathbb{R}^{256 \times 4224}$ serve as keys and values:

$$Q = F_{\mathrm{template}}, \quad K = V = F_{\mathrm{img}}. \quad (3)$$

The attended features highlight image regions semantically aligned with physical entities:

$$F_{\mathrm{attn}} = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \quad (4)$$

and the final entity-grounded representation is obtained via element-wise fusion:

$$F_{\mathrm{enhanced}} = F_{\mathrm{template}} + F_{\mathrm{attn}}. \quad (5)$$

This operation constitutes the first stage of PRI—*injecting entity-level physical grounding* into the model's visual understanding.

4.3. Mixture-of-Experts

The MoE module realizes the second stage of physical relation injection by dynamically modeling physical couplings

between meter components. Given the enhanced representation $x = \text{MLP}(F_{\text{enhanced}}) + F_{\text{img}}$, a gating network G routes the information to n specialized experts $\{E_i\}_{i=1}^{n}$:

$$p = \text{softmax}(G(x)), \quad p \in \mathbb{R}^n, \sum_{i=1}^{n} p_i = 1. \quad (6)$$

Each expert captures distinct coupling relations (e.g., pointer–scale orientation, reflection patterns), and the final fused output is:

$$F_{\text{moe}} = \sum_{i=1}^{n} p_i E_i(x). \quad (7)$$

This mechanism enables conditional, physically consistent reasoning under varying conditions—effectively implementing relation-level physical injection across diverse meter types.

4.4. Language-labeled Supervision

The final stage of physical relation injection is achieved through language-labeled supervision, which establishes a symbolic bridge between visual relations and linguistic concepts. By aligning visual embeddings with structured textual prompts, the model establishes a symbolic correspondence between visual and textual representations. This symbolic constraint enforces that the numerical predictions generated by the Large Language Model remain consistent with the underlying physical structure of the meter, achieving symbolic-level physical relation injection and enhancing prediction fidelity.

The MRLM is trained end-to-end with a cross-entropy loss:

$$\hat{y} = H(F_{\text{query}}, F_{\text{moe}}, F_{\text{text}}), \quad (8)$$

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \quad (9)$$

where $H(\cdot)$ denotes the multi-modal fusion network, F_{query} denotes a set of query vectors from the Q-Former module, C is the number of classes, and y_c is the one-hot ground-truth label for class c. This unified training framework ensures that physical relations are injected, preserved, and exploited throughout the full multi-modal inference process.

5. Experiments

5.1. Dataset and Evaluation Metric

For this study, we utilized RPM-10K to train and evaluate our models. The datasets consist of meter readings, where the ground truth values are represented language labels. The Language-based labels allow for end-to-end training, leveraging the model's ability to directly interpret meter readings from visual information.

We use the following evaluation metrics to assess model performance:

- **Ref**: Calculated as $\frac{|y-y^*|}{\text{Range}}$, where y^* is the predicted reading and y is the true reading. It provides a measure of the absolute error normalized by the range of possible values.
- **Rel**: Calculated as $\frac{|y-y^*|}{y}$, where y^* is the predicted reading and y is the true value. This metric quantifies the relative error of the model's predictions for each individual meter reading.
- **Accuracy** ϵ (%): Measures the percentage of predictions where the relative error (ref) is less than or equal to 0.01. This is a crucial metric to assess how close the predicted readings are to the true values within a small error threshold.
- **Accuracy** θ (%): Measures the percentage of predictions where the angular error (rel) is less than 0.05. This metric evaluates the model's performance in terms of the angular deviation in readings.

These evaluation metrics ensure that the model's performance is both accurate and robust, providing a comprehensive understanding of its effectiveness in meter reading tasks.

5.2. Implementation Details

To balance performance and computational efficiency, all input images are resized to 224 × 224 during both the training and inference stages. A batch size of 8 is used, and the model is optimized with the AdamW optimizer at an initial learning rate of 10^{-4}. After 200,000 iterations, the learning rate is reduced to 10^{-6}, followed by an additional 50,000 iterations of fine-tuning to ensure convergence and stability. Our code is implemented using PyTorch [24] based on Python, and all experiments are conducted on a server equipped with A800 GPUs. More details can be found in our source code.

5.3. Comparison on Public Benchmark Datasets

• **Analysis of the Accuracy ϵ.** As shown in Table 4, under the strict 1% range normalized error threshold with Ref $\leq$ 0.01, MRLM achieves the best Accuracy ϵ at 62.4%. This surpasses the strongest open source baseline LLaVA v1.6 Vicuna 7B at 56.2% by 6.2 percentage points, and also exceeds LLaMA3.2 Vision 11B at 53.6% by 8.8 points and LLaVA 1.5 7B at 49.2% by 13.2 points. Notably, although some baselines exhibit lower average error, for example LLaVA v1.6 Vicuna 7B reports the lowest Ref at 0.038, their hit rate within the tight 1% threshold still trails ours, which indicates that MRLM concentrates more predictions near the true value and mitigates long tail errors. Closed source models such as GPT5 at 7.8% and Gemini2.5 pro at 15.8% lag substantially on this strict metric, which underscores the practical advantage of our approach for high precision meter reading.

Archetype	Accuracyϵ (%)		ΔAccuracyϵ (abs, rel)	Accuracyθ (%)		ΔAccuracyθ (abs, rel)
	with KFM	w/o KFM		with KFM	w/o KFM	
$meter_1$	47.4	35.8	+11.6 (32.4%)	65.6	51.6	+14.0 (27.1%)
$meter_2$	69.6	64.8	+4.8 (7.4%)	84.6	81.0	+3.6 (4.4%)
$meter_3$	72.4	58.6	+13.8 (23.5%)	33.3	25.3	+8.0 (31.6%)
$meter_4$	77.9	68.8	+9.1 (13.2%)	56.3	49.3	+7.0 (14.2%)
$meter_5$	76.1	72.6	+3.5 (4.8%)	83.1	78.5	+4.6 (5.9%)
$meter_6$	65.1	57.8	+7.3 (12.6%)	75.7	68.5	+7.2 (10.5%)
Average	**68.1**	**59.7**	+8.4 (14.1%)	**66.4**	**59.0**	+7.4 (12.5%)

Table 2. Performance of different archetypes on RPM-10K. The original format "with KFM ± w/o KFM" has been split into separate columns. Absolute and relative improvements are highlighted in green.

Environment	Accuracyϵ (%)		ΔAccuracyϵ (abs, rel)	Accuracyθ (%)		ΔAccuracyθ (abs, rel)
	with MoE	w/o MoE		with MoE	w/o MoE	
tilted	66.5	61.6	+4.9 (8.0%)	71.1	66.2	+4.9 (7.4%)
low_light	59.0	55.9	+3.1 (5.5%)	66.5	65.9	+0.6 (0.9%)
blur	62.8	57.1	+5.7 (10.0%)	71.4	67.3	+4.1 (6.1%)
occlusion	61.1	69.4	-8.3 (-12.0%)	61.1	72.2	-11.1 (-15.4%)
missing_part_info	78.3	73.7	+4.6 (6.2%)	80.0	77.1	+2.9 (3.8%)
mirror_reflection	64.1	60.8	+3.3 (5.4%)	55.8	53.0	+2.8 (5.3%)
high_exposure	58.0	48.9	+9.1 (18.6%)	67.2	66.4	+0.8 (+1.2%)
mirror_pollution	58.6	54.3	+4.3 (+7.9%)	64.7	67.2	-2.5 (-3.7%)
Average	**63.6**	**60.2**	+3.4 (+5.6%)	**67.2**	**66.9**	+0.3 (+0.4%)

Table 3. MoE performance across environments on RPM-10K. The original "with MoE ± w/o MoE" results are split into separate columns. Absolute and relative improvements are shown in green (positive) or red (negative).

• **Analysis of the Accuracy θ.** With the angular deviation tolerance with Rel < 0.05, MRLM again ranks first, reaching an Accuracy θ of 70.9%. This yields consistent gains over LLaVA v1.6 Vicuna 7B at 67.4% by 3.5 points, LLaMA3.2 Vision 11B at 64.6% by 6.3 points, and LLaVA 1.5 7B at 61.6% by 9.3 points, as well as over other recent models such as Keye VL 8B Preview at 60.4% by 10.5 points and BLIVA at 60.0% by 10.9 points. While most methods show higher Accuracy θ than Accuracy ϵ, which reflects the relatively looser angular threshold, MRLM maintains the top rank under both criteria, demonstrating robust control of angular errors and strong generalization. Together, the superior Accuracy ϵ and Accuracy θ show that MRLM reliably delivers strict near exact readings and stable performance within practical angular tolerances.

5.4. Component Analysis

Tables 2 and 3 evaluate two modules: KFM and MoE. Each improves the base pipeline in its target setting, and together they add physical grounding and adaptive routing. Across six archetypes, KFM consistently improves **Accuracyϵ** and **Accuracyθ**. On average, **Accuracyϵ** rises from 59.7% to 68.1% and **Accuracyθ** from 59.0% to 66.4%. The largest gains appear on $meter_1$ and $meter_3$. Knowledge guided component isolation strengthens entity grounding and suppresses clutter. Across environments, MoE improves 7 of 8 conditions. Mean **Accuracyϵ** increases from 60.2% to 63.6% and **Accuracyθ** from 66.9% to 67.2%. Excluding occlusion, the gains rise to 5.0 points in **Accuracyϵ** and 1.9 points in **Accuracyθ**. The occlusion drop likely reflects limited supervision, and increasing the number of occlusion samples may improve performance. Overall, KFM

Table 4. Comparison of pointer meter reading accuracy on the RPM-10K dataset. † denotes closed-source visual language model.

No.	Model	Year	Acc $_\epsilon$ (%)	Acc $_\theta$ (%)	Ref↓	Rel↓
01	Qwen2VL-7B	2024	45.8	57.9	0.058	0.364
02	Qwen2.5VL-7B	2025	25.6	31.3	0.31	2.28
03	InternVL3-8B	2025	41.2	54.4	0.053	0.472
04	LLaMA3.2-Vision-11B [22]	2024	53.6	64.6	0.046	0.287
05	LLaVA-1.5-7B [18]	2023	49.2	61.6	0.047	0.311
06	LLaVA-v1.6-Vicuna-7B [19]	2024	<u>56.2</u>	<u>67.4</u>	**0.038**	**0.215**
07	MiniCPM-V-2_6	2024	45.3	57.1	0.063	0.839
08	Keye-VL-8B-Preview [30]	2025	45.5	60.4	0.043	0.239
09	PaliGemma2-10B-224 [27]	2024	18.2	25.4	0.133	1.718
10	Pixtral-12B [13]	2024	36.7	50.4	0.101	1.218
11	Gemma-3-12B-it [29]	2025	33.3	48.2	0.063	0.557
12	Gemma-3n-E4B-Instruct [29]	2025	7.70	9.64	0.878	6.953
13	MiniCPM-V-4 [39]	2024	34.3	46.6	0.073	0.607
14	GLM-4.1V-9B-Thinking [32]	2025	28.8	34.6	0.172	1.282
15	GPT5 [23]†	2025	7.8	15.7	0.361	2.376
16	Gemini2.5-pro [9]†	2025	15.8	23.4	0.213	1.088
17	BLIVA [15]	2023	51.2	60.0	0.114	0.925
18	**MRLM**	2025	**62.4**	**70.9**	0.063	0.535

Variant	Accuracyϵ (%)	Accuracyθ (%)	Ref↓	Rel↓
BASELINE	51.2	60.0	0.114	0.925
w/o KFM (MOE only)	55.4	63.4	0.079	<u>0.469</u>
w/o MOE (KFM only)	<u>58.3</u>	68.0	**0.056**	**0.378**
MRLM	**62.4**	**70.9**	<u>0.063</u>	0.535

Table 5. Ablation study on RPM-10K. The table reports performance across four metrics for different variants.

Variant	Accuracyϵ (%)	Accuracyθ (%)	Ref↓	Rel↓
BASELINE	51.2	60.0	0.114	0.925
4_scale	<u>60.3</u>	<u>68.2</u>	**0.055**	<u>0.486</u>
8_scale	58.7	65.1	0.068	**0.368**
6_expert	57.4	65.3	0.074	0.574
10_expert	59.2	67.2	<u>0.062</u>	0.545
MRLM	**62.4**	**70.9**	0.063	0.535

Table 6. This table presents the results of experiments evaluating the parameters of different model modules.

boosts grounding and archetype discrimination, and MOE improves robustness to tilt, blur, and exposure. Combined, they increase accuracy and stability.

5.5. Ablation Study

To assess the contribution of each component, we conduct single-toggle ablations by selectively removing individual modules. Specifically, we consider the following variants: **w/o MOE**, **w/o KFM**, and the full MRLM model. only the target module is toggled. As shown in Table 5, both KFM and MOE contribute positively to overall performance. Removing either component leads to a noticeable performance drop across all evaluation metrics, confirming the complementary roles of the two modules in enhancing model robustness. We further analyze the parameterization of key

Table 7. This table presents a comparison of different models in terms of parameter size and inference latency. All measurements are conducted with a batch size of 1 on a single NVIDIA A800 GPU to ensure a fair evaluation of efficiency and computational performance.

Method	Efficiency (↓)	
	Params (M)	**Latency (ms)**
BASELINE	7.9B	605.6
w/o KFM (MOE only)	8.6B	656.3
w/o MOE (KFM only)	8.0B	580.1
MRLM	8.7B	688.0

modules using the variants in Table 6. Here, 4/8scale denote the number of templates in the kfm module, and 6/10expert denote the number of experts in the moe module. Our full model (MRLM) uses 6scale and 8expert. Table 7 shows the parameter counts and latency measured with a batch size of 1 on a single A800 GPU.

5.6. Visualization

To further demonstrate the performance of our method, we include a visualization of the model's prediction results in Fig. 5. This qualitative analysis complements the quantitative metrics and highlights the practical applicability of our approach.

Figure 5. Visualization of model prediction results on representative test samples.

5.7. Limitation Analysis

We have utilized large language models (LLMs), but the expected generalization and emergent capabilities of these models were not observed. The model struggles to predict well on meter images it has never seen before. We initially hypothesized that this limitation stems from the dataset design: the dataset contains only the final meter readings without any explicit reasoning process. To address this, we attempted to augment the dataset with reasoning traces (Chain-of-Thought, CoT) using visual-language

models (vLLMs) and prior knowledge. However, the results after supervised fine-tuning (SFT) were even worse. We conjecture that introducing CoT supervision has adverse effects on tasks that are highly sensitive to numerical regression. During training, we observed that the CoT-augmented dataset yielded faster loss convergence and smaller final loss values, which can be attributed to the LLM's strong capability in predicting the textual part. Nevertheless, its numerical predictions were less accurate, whereas our task is solely concerned with the final numeric reading. We thus hypothesize that the superior performance of the pure numeric dataset arises from the model reducing the meter-reading task to a form of image regression, thereby bypassing the unstable linguistic reasoning process.

6. Conclusion

We introduce the MeterReading Large Model, a multimodal framework using Key Feature Mining and a Mixture-of-Experts design for automatic meter reading in complex environments. This approach enhances salient dial features, adapts to diverse meter types, and achieves state-of-the-art performance on our RPM-10K benchmark. We also contribute a challenging dataset and standardized benchmark protocols to found future research. Although limitations in generalizing to unseen cases exist, our work provides a robust baseline and opens new directions for applying large multimodal models to industrial perception.

Future work may involve reinforcement learning (RL), which holds great promise for this task. RL has recently shown significant performance boosts for large models in complex reasoning and mathematical problem solving. Incorporating reinforcement learning, may further enhance model accuracy and generalization. This could enable more reliable deployment in diverse real-world scenarios.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736, 2022.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966, 2023.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[4] Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794, 2022.

[5] Ziyang Chen, Xingwei Qu, Ming Ding, Zihan Liu, Weihan Wang, Xinghan Liu, Qingsong Lv, Yuxiao Dong, and Jie Tang. Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793, 2023.

[6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271, 2024.

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821, 2024.

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24185–24198, 2024.

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, and Sachdeva. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261, 2025. Version v4, 72 pages, 17 figures.

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems, 36:49250–49267, 2023.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186, 2019.

[12] Huahao Fan and Yuan Li. Image recognition and reading of single pointer meter based on deep learning. IEEE Sensors Journal, 24(15):25163–25174, 2024.

[13] Saurabh Garg. Pixtral 12b, 2024. Used Pixtral-12B-2409 variant.

[14] Liqun Hou, Sen Wang, Xiaopeng Sun, and Guopeng Mao. A pointer meter reading recognition method based on yolox and semantic segmentation technology. Measurement, 218:113241, 2023.

[15] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 2256–2264, 2024.

[16] Zhu Li, Yisha Zhou, Qinghua Sheng, Kunjian Chen, and Jian Huang. A high-robust automatic reading algorithm of pointer meters based on text detection. Sensors, 20(20):5946, 2020.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916, 2023.

[18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 26296–26306, 2024.

[19] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

[20] Qi Liu and Lichen Shi. A pointer meter reading method based on human-like reading sequence and keypoint detection. Measurement, 248:116994, 2025.

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32, 2019.

[22] AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Meta AI Blog. Retrieved December, 20:2024, 2024.

[23] OpenAI. Gpt-5 system card. Technical report, OpenAI, 2025. 60 pages, includes evaluations and safeguards.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 2019.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021.

[26] Gabriel Salomon, Rayson Laroca, and David Menotti. Deep learning for image-based automatic dial meter reading: Dataset and baselines. In 2020 International joint conference on neural networks (IJCNN), pages 1–8. IEEE, 2020.

[27] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. arXiv preprint arXiv:2412.03555, 2024.

[28] Junjiao Sun, Zhiqing Huang, and Yanxin Zhang. A novel automatic reading method of pointer meters based on deep learning. Neural Computing and Applications, 35(11):8357–8370, 2023.

[29] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025.

[30] Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl technical report. arXiv preprint arXiv:2507.01949, 2025.

[31] Qwen Team. Qwen2.5-vl, 2025.

[32] V Team, Wenyi Hong, Wenmeng Yu, and Xiaotao Gu. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.

[33] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems, 37:87310–87356, 2024.

[34] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

[35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.

[36] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. arXiv preprint arXiv:2411.10442, 2024.

[37] Zhaolin Wang, Lianfang Tian, Qiliang Du, Yi An, Zhengzheng Sun, and Wenzhi Liao. Robust pointer meter reading recognition method under image corruption. IEEE Transactions on Instrumentation and Measurement, 73:1–16, 2024.

[38] Xiao Xiao, Donghua Hu, Kun Yan, and Hsiao-Chun Wu. Novel robust perspective rectification for automatic circular-meter reading. IEEE Sensors Journal, 2024.

[39] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024.

[40] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490, 2023.

[41] Hongyu Zhang, Yunbo Rao, Jie Shao, Fanman Meng, and Jiansu Pu. Reading various types of pointer meters under extreme motion blur. IEEE Transactions on Instrumentation and Measurement, 72:1–15, 2023.

[42] Lin Zuo, Peilin He, Changhua Zhang, and Zhehan Zhang. A robust approach to reading recognition of pointer meters based on improved mask-rcnn. Neurocomputing, 388:90–101, 2020.