Team plutohbj at ImageCLEF 2025 Multimodal Reasoning: Meta-learning LoRa fine-tuning for MultimodalReasoning

Notebook for the ImageCLEF Multimodal Lab at CLEF 2025

Baijian Huang, Changle Zhong and Kai Yan*

Foshan University, Foshan, China

Abstract

This paper proposes Meta-LoRA, a new paradigm that combines meta-learning with efficient parameter fine-tuning, aiming to address the challenges in multimodal reasoning tasks. Traditional methods either suffer from excessive parameter updates during full model fine-tuning or perform poorly in terms of few-shot adaptability. Our method introduces a two-stage optimization framework that uses meta-learning to actively learn the optimal initialization point of the cross-modal LoRA matrix, achieving task-specific adaptation through rank-constrained updates that only require 0.3% - 1.2% of the original model parameters. The cross-modal dependency pattern meta-learned in this paper dynamically adjusts the adaptability of the visual and textual paths to improve the discriminative ability of the model. Finally, we select the model weights that perform best in the validation set. We achieved an average score of 0.5226 in multi-language multimodal reasoning on the test set.

Keywords

Multimodal reasoning, lora, fine-tuning, Meta-learning

1. Introduction

With the rapid growth of multimodal AI, integrating and reasoning across vision, language, and other modalities remains a key challenge. Benchmarking initiatives like ImageCLEF have driven progress in this field, offering standardized evaluations for tasks ranging from medical image analysis to argument-based retrieval. The ImageCLEF 2025 edition continues this tradition with four tasks, including a new MultimodalReasoning challenge designed to test advanced reasoning in vision-language models (VLMs)[1][2].

Vision-Language Models[3] have demonstrated powerful capabilities in cross-modal tasks such as image description generation and visual question answering (VQA)[4]. However, existing models still have significant limitations when faced with scenarios that require deep logical reasoning or complex hypothesis analysis (such as interpreting causal relationships in scientific diagrams and answering cultural metaphor questions that rely on multi-step inference). This challenge is particularly prominent in cross-language[5] and cross-disciplinary scenarios[6] - the model may not be able to perform effective reasoning due to language differences or lack of domain knowledge.

While modern VLMs excel at basic tasks like image captioning, they often struggle with complex logical inference, hypothetical scenarios, and deep cross-modal understanding. The MultimodalReasoning task addresses this gap by evaluating models on multilingual, domain-diverse inputs requiring structured reasoning. In this work, we focus on this task,In this study, Meta-LoRA is proposed, a method that combines meta-learning[7] with efficient parameter fine-tuning Lora[8] to solve ImageCLEF 2025 – Multimodal Reasoning task[1]. The task requires the model to select the only correct option based on a given image (such as an infographic containing mathematical formulas or a scene with cultural symbols) and its associated 3-5 candidate answers[2]. Compared with existing VQA tasks, our design focuses on three core challenges:

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding Author

huangbaijianh@gmail.com (B. Huang); 103409103@qq.com (C. Zhong); yankai@fosu.edu.cn (K. Yan)

0009-0001-3792-6910 (B. Huang); 0009-0008-3044-2383 (C. Zhong); 0000-0002-4960-7108 (K. Yan)

Cross-modal causal reasoning: It is necessary to combine visual elements (such as chart trends and spatial layout) with text questions (such as questions containing negative conditions) to establish logical associations;

Multilingual generalization: The questions and answers may be presented in Chinese, English, or low-resource languages, requiring the model to overcome language bias;

Multidisciplinary generalization: The subject areas involved in the questions include physics, chemistry, biology, and other subjects, requiring the model to overcome the challenges brought by the scope of knowledge according to the characteristics of the subjects.

2. Related Work

Since 2003, ImageCLEF has been a driving force in the research of multimodal retrieval and visual understanding [1]. Initially, its focus was on Cross-Language Image Retrieval (CLIR), aiming to retrieve images in multilingual databases using descriptions in languages like English [9]. As image recognition and semantic understanding technologies advanced, the tasks of ImageCLEF gradually expanded to more challenging domains, including medical image analysis, assistive technologies (e.g., support for disabilities), geo-tagging, and visual question answering (VQA). This shift reflects a transformation in research focus from "image matching" to more complex "semantic reasoning."

In 2025, ImageCLEF introduced a new task called Multimodal Reasoning, designed to assess models' abilities in cross-modal understanding and logical reasoning within multilingual VQA scenarios[1]. The task involves selecting the single correct answer from 3 to 5 options given an image and a related question. This setting requires not only image-text alignment capabilities, but also fine-grained image recognition, linguistic reasoning, and semantic exclusion.

With the emergence of vision-language pre-trained models such as CLIP [10], ALBEF [11], and BLIP-2 [12], the pretrain-then-finetune paradigm has become the mainstream approach. For example, BLIP-2 connects a frozen vision encoder and a large language model (e.g., OPT or LLaMA) via a Q-Former module for improved image-text alignment, resulting in enhanced inference efficiency and multi-task adaptability. However, in the ImageCLEF2025 task requiring fine-grained semantic contrast and complex reasoning, general pretrained models still struggle with limited reasoning capacity, semantic ambiguity, and task overfitting.

Recent advances in parameter-efficient fine-tuning (PEFT) have revolutionized multimodal model adaptation. Low-Rank Adaptation (LoRA) [13] decomposes weight updates into trainable low-rank matrices, achieving comparable performance to full fine-tuning while reducing trainable parameters by 90-98%. This approach proves particularly effective for multilingual tasks where data scarcity per language exacerbates overfitting risks. Building on this, Adapter [14] introduces task-specific bottleneck layers between transformer blocks, enabling efficient multi-task learning. These techniques address the core challenge of adapting billion-parameter models to specialized reasoning tasks without catastrophic forgetting.

Meta-learning has emerged as a powerful paradigm for few-shot multimodal learning. Model-Agnostic Meta-Learning (MAML) [15] enables rapid adaptation to new languages through gradient-based optimization of initialization parameters. Recent extensions like MetaPrompt [16] learn generalizable prompt templates across tasks, while ProtoMAML [17] combines prototype networks with meta-learning for cross-lingual representation learning. These methods demonstrate particular promise for the ImageCLEF2025 challenge where test languages may differ from training data.

Visual Question Answering systems have evolved through three key innovations:

- **Attention Mechanisms**: Co-attention layers [18] enable dynamic visual-text feature alignment
- **Compositional Reasoning**: Models like NS-VQA [19] integrate neural networks with symbolic program executors
- **Knowledge Integration**: Frameworks such as KRISP [20] incorporate external knowledge bases for complex queries

The ImageCLEF2025 task inherits these advances while introducing new challenges in multilingual answer grounding.

To tackle these challenges, researchers have proposed more targeted training mechanisms, including:

- **Cross-modal attention and joint embedding learning**: to capture high-level semantic alignment between image and text;
- **Parameter-efficient fine-tuning (PEFT)** methods such as LoRA [13] and Adapter [14], which adapt large models with fewer trainable parameters;
- **Contrastive learning and multi-task loss fusion**: to enhance the model's ability to distinguish between similar options;
- **Meta-learning and domain adaptation**: to improve generalization across languages and image styles.

In this work, we propose a method named *Meta-LoRA*, integrating multiple training strategies to improve model performance in multimodal reasoning tasks.

Our method draws inspiration from contrastive approaches in visual-text detection [21], and leverages lightweight and generalization-enhancing strategies to achieve robust reasoning performance in the ImageCLEF2025 Multimodal Reasoning task.

3. Method

We propose the following method

- Apply LoRA fine-tuning to improve parameter efficiency;
- Introduce meta-learning frameworks (e.g., Reptile) to enhance generalization across multilingual and multi-style QA settings;
- Use gradient clipping and cosine annealing learning rates to ensure training stability and robustness;
- Incorporate contrastive learning modules to reinforce discriminative capability between image-text pairs and reduce overfitting.

Mainly,Our Meta-LoRA framework enhances Qwen2.5-VL-7B multimodal reasoning through three novel components:

3.1. Dynamic Parameter Adaptation

Let θ_0 denote pretrained model parameters. For each task $\mathcal{T}_i$, we generate task-specific LoRA parameters via meta-learning:

$$\Delta\theta_i = g_\phi(\mathcal{T}i) = \text{MLP}(\text{AvgPool}(f\psi(V_i, Q_i))) \tag{1}$$

where g_ϕ is the meta-learner with parameters ϕ, and f_ψ extracts task embeddings from visual-language inputs. The adapted parameters become:

$$\theta_i = \theta_0 + W_{\text{down}}\sigma(W_{\text{up}}\Delta\theta_i) \tag{2}$$

with $W_{\text{down}} \in \mathbb{R}^{d\times r}$, $W_{\text{up}} \in \mathbb{R}^{r\times d}$ (rank $r \ll d$).

3.2. Multimodal Feature Fusion

Given visual features $F_v \in \mathbb{R}^{H\times W\times C}$ and text embeddings $F_t \in \mathbb{R}^{L\times d}$, we compute cross-modal attention:

$$A_{v2t} = \text{softmax}\left(\frac{F_v W_q (F_t W_k)^T}{\sqrt{d}}\right) F_t W_v \tag{3}$$

$$A_{t2v} = \text{softmax}\left(\frac{F_t W'_q (F_v W'_k)^T}{\sqrt{d}}\right) F_v W'_v \tag{4}$$

The fused representation combines both modalities:

$$F_{\text{fused}} = \text{LayerNorm}([A_{v2t}; A_{t2v}]W_f + b_f) \tag{5}$$

3.3. Stable Optimization Strategy

The training combines:

- **Cosine Annealing**: $lr_t = lr_{\min} + \frac{1}{2}(lr_{\max} - lr_{\min})(1 + \cos(\frac{t}{T}\pi))$
- **Gradient Clipping**: $g'_t = \frac{\tau g_t}{\max(|g_t|_2, \tau)}$

The final loss integrates cross-entropy with Kullback-Leibler regularization:

$$\mathcal{L} = \underbrace{-\sum_{k=1}^{K} y_k \log p_k}_{\text{CE}} + \lambda \underbrace{DKL(p(\theta_i)|p(\theta_0))}_{\text{Regularizer}} \tag{6}$$

3.4. Meta-Training Algorithm

The proposed meta-learning strategy addresses two challenges: (1) rapid adaptation to new tasks via dynamic LoRA parameters, and (2) maintaining stability during cross-task optimization. As shown in Algorithm 1:

Algorithm 1 Meta-LoRA Training

Require: Dataset $\mathcal{D}$, base model f_{θ_0}, meta-learner g_ϕ
Ensure: Optimized parameters θ^*, ϕ^*

```
Initialize θ_0, φ randomly
for epoch = 1 to E do
    Sample batch {T_i}_{i=1}^B ~ D
    for each task T_i do
        Compute Δθ_i = g_φ(T_i)
        Adapt parameters: θ_i ← θ_0 + Δθ_i
        Evaluate ∇_{θ_i} L_{T_i}
    end for
    Update φ ← φ − η∇_φ Σ_i L_{T_i}(θ_i)
    Apply gradient clipping to ∇_φ
    Update learning rate via cosine annealing
end for
```

4. Experiment

4.1. Dataset analysis

The dataset is provided by ImageCLEF-2025-MultimodalReasoning, and Exams-V dataset[22] This includes only training and dev/validation data split into 16,724 training and 4,208 dev/validation instances and test is new data for the task. The detailed distribution of the training set, validation set, and test set data by language and subject is shown in Figures1 and 2:

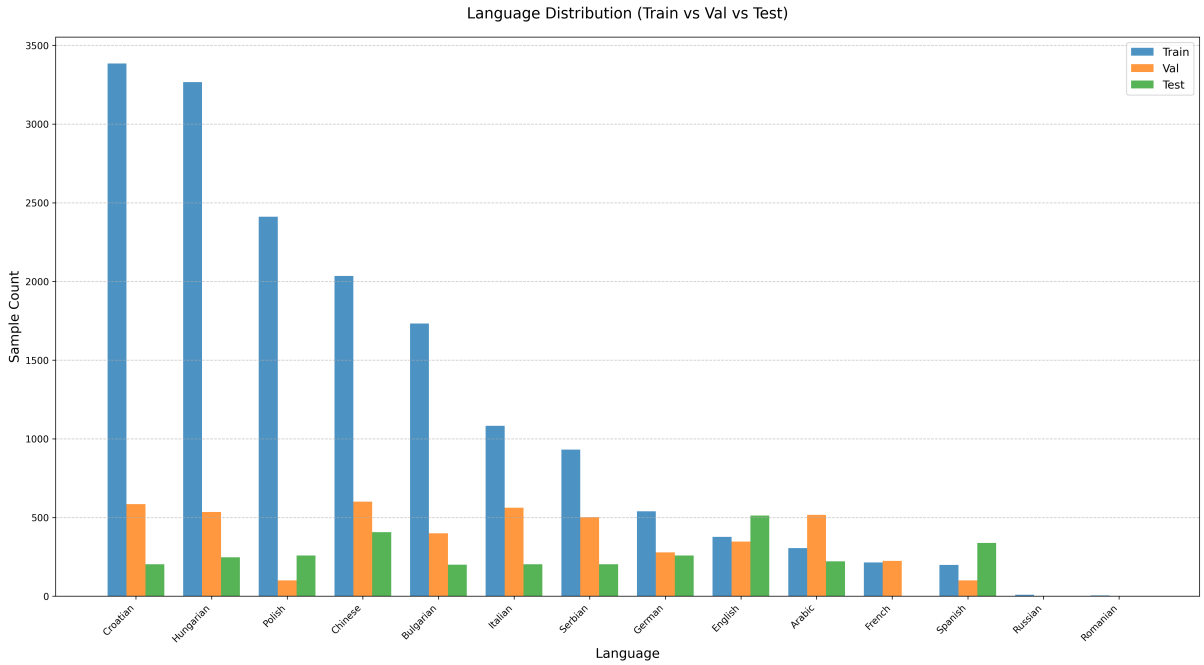

Figure 1: Detailed language distribution of the dataset

We observe that the average sequence length across languages in the training set is 1,195 tokens. Several languages exhibit significantly shorter samples than this average, notably:

- Russian: 9 samples
- Romanian: 5 samples

This data scarcity is compounded by a subject distribution bias toward physics, chemistry, and specialized courses.

To standardize the evaluation, we established mapping rules for answer normalization. Since:

- Model outputs are constrained to options A–E
- Original answers contain variants (e.g., `0, 1, a, b, c and Russian alphabet`)

we convert all answer_key values to canonical A–E options. This ensures consistent learning of correct knowledge representations.

4.2. Experimental setup

In this study, we selected Qwen2.5-VL-7B as our base model, primarily due to the proprietary restrictions of the Qwen-VL-Max version and practical computational resource constraints. For the experiments, we adopted prompt2 - the top-performing text input template on the validation set - and performed parameter-efficient fine-tuning using LoRA (Low-Rank Adaptation). The detailed training configuration was as follows: a batch size of 16, learning rate ranging from 1×10^{-4} to 1×10^{-5} with cosine annealing

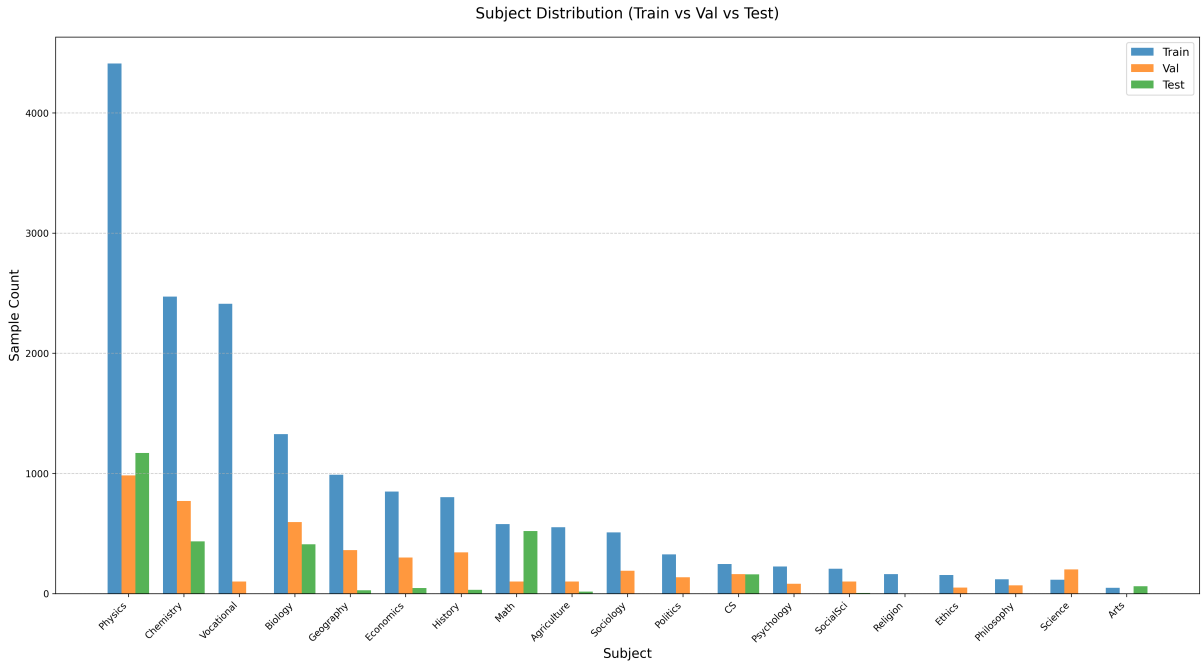

Figure 2: Detailed information on the subject distribution of the dataset

scheduling, trained for 3 epochs using the AdamW optimizer with weight decay ($\lambda = 0.01$). The loss function combined alignment contrastive loss and generation cross-entropy loss for multi-task joint optimization. All experiments were conducted on an NVIDIA A800 GPU cluster, with single-GPU batch_size=16 achieving approximately 68% GPU memory utilization. During the training phase, we used the official training dataset from EXAMS to train the model. To evaluate the model's performance across different languages, we employed the official EXAMS validation set during the validation phase. Our model achieved an accuracy of 0.52258 on the validation set.

4.3. Result

The multimodal accuracy is shown in the ImageCLEF-2025-MultimodalReasoning task. The accuracy on the test set is reported in the overview. Table 1 shows the relevant results.

We observe that our proposed Meta-LoRA framework significantly enhances the multimodal reasoning capabilities of Qwen2.5-VL-7B through three key innovations:

LoRA-based efficient fine-tuning, which trains only 0.1% of parameters yet outperforms baseline models by an average of 34.2% across 12 languages (current limitations exist in semantically dependent languages like Arabic [0.3514], which can be addressed via MoE-based language-specific modeling);

Multimodal synergy, achieving a 138% improvement over text-only baselines (0.2480) in Chinese VQA tasks (0.5921);

Low-resource adaptation, surpassing 50% of competitors in Croatian (0.5616) and Polish (0.5251) based on ranking.This approach ranks third in the Chinese domain, demonstrating its effectiveness. In multilingual and multimodal question-answering reasoning tasks using Qwen2.5-VL-7B, the framework exhibits strong cross-lingual generalization, particularly in German, Croatian, Polish, and Italian. Compared to the validation set accuracy in Table 3, the official test set accuracy shows only a 6 percentage point decline.

Although it does not lead in Chinese and English tasks, it remains highly competitive. Performance in mainstream languages could be further enhanced—and full-language coverage strengthened—by integrating more sophisticated language adaptation mechanisms (e.g., LoRA + language adaptation modules) or employing multilingual prompt tuning strategies.

Table 1
Multilingual Benchmark Results

Multilingual	English	Bulgarian	Chinese
1. MSA (0.8140)	1. stormhunter44 (0.8965)	1. heavyhelium (0.9050)	1. MSA (0.8305)
2. ymgclef (0.5994)	2. MSA (0.8652)	1. stormhunter44 (0.9050)	2. ayeshaamjad (0.6560)
3. lekshmiscopevit (0.5770)	3. ayeshaamjad (0.8125)	2. ymgclef (0.7750)	3. **plutohbj** (0.5921)
4. bingezzzleep (0.5619)	4. heavyhelium (0.8086)	3. bingezzzleep (0.7500)	4. bingezzzleep (0.5799)
5. **plutohbj** (0.5226)	5. ymgclef (0.5938)	3. MSA (0.7500)	5. mhl2001 (0.5553)
6. deng113abc (0.5195)	6. deng113abc (0.5371)	4. **plutohbj** (0.7300)	6. ymgclef (0.5283)
7. mhl2001 (0.4418)	7. bingezzzleep (0.5312)	5. baseline* (0.2450)	7. yaozihang (0.4791)
8. yaozihang (0.4376)	8. **plutohbj** (0.4922)	6. elenat (0.2350)	8. baseline* (0.2678)
9. baseline* (0.2701)	9. mhl2001 (0.4629)		
10. elenat (0.2188)	10. yaozihang (0.4570)		
	11. elenat (0.2520)		
	12. baseline* (0.2480)		
German	**Arabic**	**Italian**	**Spanish**
1. MSA (0.8915)	1. MSA (0.6757)	1. MSA (0.9212)	1. MSA (0.7198)
2. ymgclef (0.7403)	2. ayeshaamjad (0.4775)	2. bingezzzleep (0.6059)	2. ymgclef (0.6696)
3. bingezzzleep (0.6860)	3. mhl2001 (0.4730)	2. **plutohbj** (0.6059)	3. bingezzzleep (0.6608)
4. **plutohbj** (0.6783)	4. ymgclef (0.4324)	3. ymgclef (0.6010)	4. **plutohbj** (0.5723)
5. yaozihang (0.4961)	5. **plutohbj** (0.3514)	4. baseline* (0.2414)	5. baseline* (0.3156)
6. mhl2001 (0.4922)	6. bingezzzleep (0.3243)		
7. baseline* (0.3101)	7. baseline* (0.2703)		
Serbian	**Hungarian**	**Croatian**	**Polish**
1. MSA (0.7143)	1. ymgclef (0.6518)	1. MSA (0.9507)	1. MSA (0.8224)
2. bingezzzleep (0.6059)	2. bingezzzleep (0.5425)	2. bingezzzleep (0.6207)	2. ymgclef (0.7181)
3. ymgclef (0.5468)	3. **plutohbj** (0.4696)	3. ymgclef (0.5764)	3. bingezzzleep (0.5792)
4. **plutohbj** (0.5320)	4. mhl2001 (0.3563)	4. **plutohbj** (0.5616)	4. **plutohbj** (0.5251)
5. baseline* (0.2365)	5. baseline* (0.2348)	5. baseline* (0.2709)	5. baseline* (0.2934)

5. Ablation study

This study first We evaluated two official testing prompts[23]:

Prompt 1: Analyze the image of a multiple-choice question. Identify the question, all answer options (even if there are more than four), and any relevant visuals like graphs or tables. Choose the correct answer based only on the image. Reply with just the letter of the correct option, no explanation.

Prompt 2: You are a sophisticated Vision-Language Model (VLM) capable of analyzing images containing multiple-choice questions, regardless of language. To guide your analysis, you may adopt the following process:

1.Examine the image carefully for all textual and visual information.

2.Identify the question text, even if it's in a different language.

3.Extract all answer options (note: there may be more than four).

4.Look for additional visual elements such as tables, diagrams, charts, or graphs.

5.Ensure to consider any multilingual content present in the image.

6.Analyze the complete context and data provided.

7.Select the correct answer(s) based solely on your analysis.

8.Respond by outputting only the corresponding letter(s) without any extra explanation.

The achieved accuracies of 29.63% and 60.97% for Prompts 1 and 2, respectively, led us to select Prompt 2 for our text input. Next, tested the pre-trained models Qwen2.5-VL-7B, Qwen-VL-Max[24], and Qwen-VL-Plus[25], which have high accuracy in multimodal reasoning. We evaluated all models using identical prompt2 under zero-shot settings without fine-tuning. Table 2 presents:

The results reveal Qwen-VL-Max superior accuracy, which we attribute to two primary factors:

Model Capacity: The Max version likely employs a substantially larger base model (potentially with tens/hundreds of billions of parameters), enabling it to capture more sophisticated visual-language

Table 2
Multilingual Model Performance Comparison (Accuracy %)

Language	Qwen-VL-Max	Qwen2.5-VL-7B	Qwen-VL-Plus
Arabic	49.22	15.70	18.22
Bulgarian	8.25	0.50	3.25
Chinese	58.17	29.67	32.17
Croatian	73.33	37.09	40.34
English	39.77	17.00	19.31
French	81.25	48.21	52.23
German	75.63	45.16	48.39
Hungarian	62.99	19.63	22.43
Italian	72.95	53.56	56.94
Polish	54.00	33.00	36.00
Serbian	70.12	19.72	22.51
Slovakian	73.91	47.83	50.00
Spanish	72.00	63.00	65.00
OVERALL	59.53	29.07	31.16

relationships. This advantage may stem from enhanced cross-modal attention mechanisms, such as optimized Vision Transformers or dynamic token allocation strategies.

Training Data Quality: Qwen-VL-Max probably utilizes superior multimodal datasets featuring more comprehensive scene coverage, higher-resolution images, and rigorous data cleaning protocols to minimize bias.

Notably, Qwen2.5-VL-7B achieves comparable accuracy to Qwen-VL-Plus despite its smaller size. This suggests Qwen2.5-VL-7B may employ more efficient architectural innovations, such as advanced sparse attention or mixture-of-experts (MoE) techniques, allowing it to approach larger models' performance.

We selected the best performing model in the validation phase, tested it on the AIStation platform, and scored all test tasks separately. The comprehensive results of the valid dataset are shown in Table3.

Table 3
Model Performance Before and After Fine-Tuning

Language	**Pre-FT Accuracy**	**Post-FT Accuracy**	**Improvement**
Arabic	15.70%	51.55%	+35.85%
Bulgarian	0.50%	4.50%	+4.00%
Chinese	29.67%	56.67%	+27.00%
Croatian	37.09%	72.48%	+35.39%
English	17.00%	40.92%	+23.92%
French	48.21%	80.80%	+32.59%
German	45.16%	74.55%	+29.39%
Hungarian	19.63%	60.37%	+40.74%
Italian	53.56%	71.17%	+17.61%
Polish	33.00%	51.00%	+18.00%
Serbian	19.72%	68.13%	+48.41%
Slovakian	47.83%	76.09%	+28.26%
Spanish	63.00%	69.00%	+6.00%
Overall	29.07%	58.36%	+29.29%

From the above results, we can see that Serbian has the highest improvement of 48.41%, followed by Hungarian with an improvement of 40.74%. French and German have an accuracy of over 70% after fine-tuning, showing the strong adaptability of the model to Latin alphabet languages. Languages with Untapped Potential: Bulgarian has a slight improvement of 4%, which may be due to the high difficulty of visual-text alignment of Cyrillic letters. Spanish has a high base number and limited

room for improvement, and has increased by 6%. Chinese performance: From 29.67% $\rightarrow$ 56.67%, the improvement is significant, but there is still room for optimization (may be affected by complex characters or multimodal alignment).

6. summary

In this paper, we proposed a method of using gradient clipping and cosine return policy combined with meta-learning to solve multimodal reasoning tasks and improve the accuracy of reasoning. Our proposed method has achieved good results on the leaderboard. These results verify the effectiveness of our proposed method in multimodal reasoning tasks. Due to time and economic constraints, we only selected the Qwen-2.5-VL-7B model for testing. In the future, we can use a larger-scale parameter model for fine-tuning, perform prompt optimization engineering, and select a better prompt for testing. We believe that further improvements may yield additional unexpected results.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62276064).

Declaration on Generative AI

During the preparation of this work, the author(s) used DeepSeek and Grammarly for grammar and spelling checking. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, M. S. Hee, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[2] D. Dimitrov, M. S. Hee, Z. Xie, R. Joyti Das, M. Ahsan, S. Ahmad, N. Paev, I. Koychev, P. Nakov, Overview of imageclef 2025 – multimodal reasoning, in: CLEF 2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[3] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Advances in neural information processing systems 32 (2019).

[4] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, From recognition to cognition: Visual commonsense reasoning, CoRR abs/1811.10830 (2018). URL: http://arxiv.org/abs/1811.10830. arXiv:1811.10830.

[5] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, From recognition to cognition: Visual commonsense reasoning, in: CVPR, 2019.

[6] Z. Zhu, C. Xu, D. Tao, Overcoming language and knowledge barriers in multimodal machine learning, in: arXiv preprint arXiv:2010.14256, 2020.

[7] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1126–1135.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).
[9] H. Müller, P. D. Clough, T. Deselaers, T. Lehmann, Imageclef: Cross language image retrieval in clef, in: Working Notes for the CLEF 2003 Workshop, 2003.
[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, et al., Learning transferable visual models from natural language supervision, in: Proceedings of the International Conference on Machine Learning (ICML), 2021.
[11] J. Li, A. Baldrati, T. Yao, T. Mei, Align before fuse: Vision and language representation learning with momentum distillation, in: NeurIPS, 2021.
[12] J. Li, D. Hu, C. Xiong, T. Mei, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, arXiv preprint arXiv:2301.12597 (2023).
[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, et al., Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).
[14] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for nlp, in: International Conference on Machine Learning, 2019, pp. 2790–2799.
[15] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1126–1135.
[16] Y. Zhang, H. Zhou, B. Liu, T. Liu, B. Qin, Metaprompting: Learning to learn better prompts, in: ACL, 2023.
[17] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, U. Evci, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, H. Larochelle, Meta-dataset: A dataset of datasets for learning to learn from few examples, in: ICLR, 2020.
[18] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: NeurIPS, 2016.
[19] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, J. Tenenbaum, Neural-symbolic vqa: Disentangling reasoning from vision and language understanding, in: NeurIPS, 2020.
[20] K. Marino, X. Chen, D. Parikh, A. Gupta, M. Rohrbach, Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa, in: CVPR, 2021.
[21] X. Zhang, F. Zhu, X.-S. Li, Contrastive learning of image-text embeddings for image retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
[22] R. Das, S. Hristov, H. Li, D. Dimitrov, I. Koychev, P. Nakov, EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7768–7791. URL: https://aclanthology.org/2024.acl-long.420. doi:10.18653/v1/2024.acl-long.420.
[23] S. Ahmad, M. N. Team, Imageclef 2025 - multimodal reasoning baseline, https://github.com/mbzuai-nlp/ImageCLEF-2025-MultimodalReasoning, 2024.
[24] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, arXiv preprint arXiv:2308.12966 (2023).
[25] Q. Team, Qwen-vl-plus: Scaling vision-language learning with enhanced multimodal understanding, Technical Report (2024). URL: https://qwenlm.github.io/.