A Multimodal Feature Alignment Prompt-Enhanced Method for Multimodal Reasoning

Notebook for the ImageCLEF Lab at CLEF 2025

Qida Wu, Leilei Kong*, Jianzhong Yan and Junyi Li

Foshan University, Foshan, Guangdong, China

Abstract

Multimodal reasoning is a task focused on multilingual visual question answering, aiming to evaluate the reasoning capabilities of modern LLMs on complex inputs presented in various languages and involving diverse subjects. This paper elaborates on the strategy of using prompts that combine image and text features to enhance the image understanding capabilities of multimodal models. By aligning images with descriptive text features and constructing multimodal prompts, the approach aims to improve the model's comprehension of images. The proposed method achieves an accuracy of 74.56% on the multilingual validation set and 56.19% on the multilingual test set, representing a 29.18% improvement in performance on the test set compared to the baseline.

Keywords

Multimodal Reasoning, Prompt-Enhancement, Feature Alignment

1. Introduction

Multimodal reasoning tasks, such as ImageCLEF 2025 [1], require models to comprehensively process and understand information from multiple modalities (e.g., vision, language, audio, etc.) to accomplish complex reasoning and decision-making. These tasks have not only propelled advancements in the field of artificial intelligence but have also fostered progress in technologies such as multimodal learning, cross-modal alignment, and deep learning. Furthermore, through interdisciplinary research, these tasks have contributed to the development of more robust models, thereby enhancing user experience, addressing challenges posed by complex tasks, and yielding significant benefits in social and economic domains [2, 3, 4].

In recent years, with the ongoing development of multimodal pretraining technology, a new array of multimodal models has emerged. Among these, the Qwen model, as an advanced Vision-Language model, provides innovative solutions for multimodal reasoning tasks with its robust multimodal fusion capabilities and efficient inference performance [5, 6]. However, Vision-Language Models (VLMs) still face challenges in deep logical reasoning and inference. They may struggle to answer questions that necessitate reasoning through complex dependencies or hypothetical scenarios [7].

To overcome this limitation, numerous studies have attempted to enhance the models' deep reasoning capabilities through fine-tuning, such as by introducing additional training data or designing task-specific loss functions to bolster the models' reasoning abilities [8]. However, due to the complex model architecture of VLMs, the large scale of parameters, and the scarcity of training data (for instance, in few-shot settings), directly fine-tuning the entire model for downstream tasks is impractical. Such fine-tuning may also lead to the forgetting of useful knowledge acquired during the large-scale pretraining phase and may cause overfitting to the downstream task [9].

Therefore, while fine-tuning is an effective method to enhance model performance, its high cost and low efficiency limit its practical application for VLMs. This has prompted researchers to explore more efficient ways to improve the deep reasoning capabilities of VLMs, such as by designing lightweight

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

*Corresponding author.

1362990744wu@gmail.com (Q. Wu); kongleilei1979@gmail.com (L. Kong); yanmoge520@gmail.com (J. Yan); m13609768252@163.com (J. Li)

0009-0002-9532-937X (Q. Wu); 0000-0002-4636-3507 (L. Kong); 0009-0002-9047-4690 (J. Yan); 0009-0000-7584-8231 (J. Li)

adaptation modules or employing prompt engineering strategies. These approaches aim to enhance the model's reasoning ability on specific tasks while retaining the knowledge acquired during pretraining [10].

Inspired by prompt engineering, we propose a prompt engineering strategy for the Qwen-VL-Max model. By introducing prompts that combine image and text features, our method guides the model to better understand task requirements, thereby effectively handling complex multimodal reasoning tasks. Our approach not only retains the strengths of the Qwen-VL-Max model in multimodal understanding but also significantly enhances its performance in deep reasoning tasks through prompt engineering, providing an efficient and effective solution for multimodal reasoning tasks.

2. Method

The methodology of this study designs the multimodal prompts to enhance the image understanding capabilities of the visual language model., thereby better generalizing to downstream tasks. The Figure 1 illustrates the overall architecture of this method.

Specifically, for an image reasoning question answering dataset $D = \{I_1, I_2, \ldots, I_n\}$, where I_i denotes the images in the dataset D and n represents the length of the dataset D, we first use a VLM to generate formatted descriptive text T_i for each image I_i. These texts are then combined with standardized prompts P_s to form a set of multimodal prompt pairs $MP = \{(P_s, I_1, T_1), (P_s, I_2, T_2), \ldots, (P_s, I_n, T_n)\}$.

For each MP_i, the image I_i is preprocessed to $R \times R$ resolution and divided into N patches, each embedded as a vector $\mathbf{v}_j$. The text T_i is tokenized into n tokens and embedded as vectors $\mathbf{e}_i$. Subsequently, $\mathbf{v}_j$ and $\mathbf{e}_i$ are fed into separate encoders for images and text to obtain the features H_I' and H_T, respectively. These features are then fused into H_{fusion} through modality-specific feature alignment methods. To distinguish between image and text inputs, special tokens $<$ img $>$ and $<$ /img $>$ are used to wrap the image feature sequence, $<$ box $>$ and $<$ /box $>$ are used for bounding box information, and $<$ ref $>$ and $<$ /ref $>$ are used for referenced content. Finally, H_{fusion} is input into the large language model to obtain the results.

The following sections will first introduce the construction of multimodal prompts, followed by an explanation of the multimodal feature alignment methods.

2.1. Construction of Multimodal Prompts

In the context of this study, for each image I_i belonging to the dataset D, a corresponding descriptive text T_i is generated using VLM. This descriptive text is formatted within the model's prompt to standardize the style and structure, adhering to the following requirements:

- **Content Description**: The text must encompass a comprehensive description of the content present in the image.
- **Emphasis on Visual Elements**: Particular attention should be given to describing charts, tables, diagrams, and other illustrative elements that may be present in the image.
- **Problem Statement Clarification**: The text should clearly articulate the problem or question posed within the image.
- **Option Description**: Each option available within the image must be described explicitly.
- **Option Specification**: The range of options (A, B, C, D, E) should be clearly delineated.

By standardizing the descriptive text in this manner, the model is better equipped to understand the questions embedded within the images. Given that the dataset D encompasses question-answer pairs in various languages, with slightly differing option symbols, it is imperative to further standardize the format of the output options within the prompt. This standardized prompt is henceforth referred to as P_s.A specific example is shown in the Figure 2.

Subsequently, the standardized prompt P_s, along with the image I_i and its descriptive text T_i, are combined to form a data pair, creating a multimodal prompt pair $MP_i = (P_s, I_i, T_i)$.

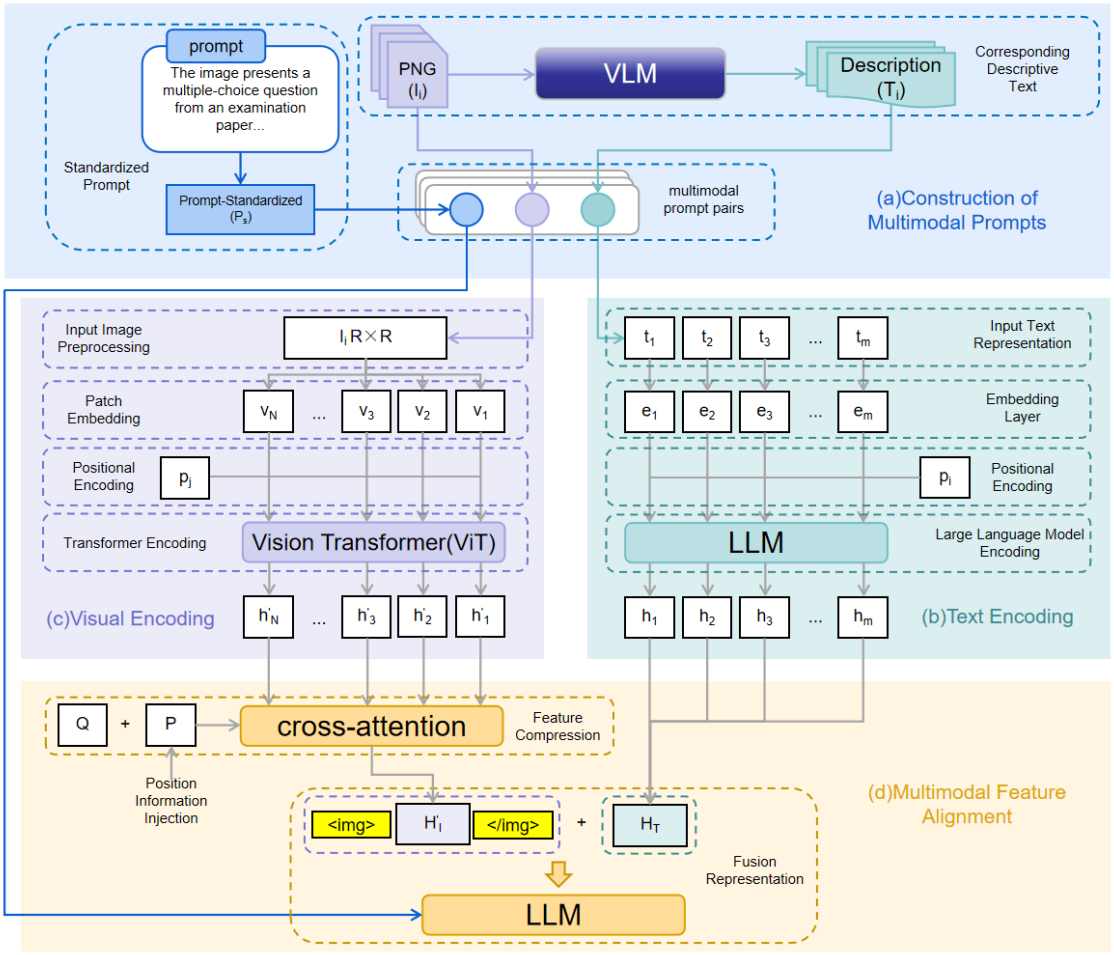


Figure 1: (a)This section describes the creation of multimodal prompt pairs by combining standardized prompts with descriptive texts generated from images.(b)Text is tokenized, embedded, and encoded with positional information before being processed by a Large Language Model (LLM) to produce text feature representations.(c)Images are preprocessed, divided into patches, embedded, and encoded with positional information, then processed by a Vision Transformer (ViT) to generate image feature representations.(d)Image and text features are fused using cross-attention and positional encoding to form a comprehensive multimodal representation for input into the LLM.

2.2. Multimodal Feature Alignment

The multimodal feature alignment method is designed to deeply integrate image and text information for efficient multimodal understanding and generation[11]. This method consists of the following key components:

- **Large Language Model (LLM)**: The foundational component responsible for processing text inputs and generating linguistic responses.
- **Visual Encoder**: Utilizes the Vision Transformer (ViT)[12, 13] architecture to transform image data into feature representations that can be fused with text data.
- **Position-aware Vision-Language Adapter**: Aligns and integrates visual features with textual features, ensuring effective interaction between image and text information.

Text Encoding

The text encoding process is based on a pre-trained LLM and follows these steps:

standardized prompt

You are a complex Visual Language Model (VLM) capable of analyzing images containing multiple-choice questions without language restrictions. To guide your analysis, you can follow the following process:

1. Carefully examine all text and visual information in the image.
2. Identify the question text, even if it is in another language.
3. Extract all answer options (Note: There may be more than four).
4. Look for other visual elements, such as tables, charts, diagrams, or figures.
5. Ensure that any multilingual content present in the image is considered.
6. Analyze the complete context and data provided.
7. Select the correct answer based solely on your analysis.
8. Carefully analyze the language used for the options in the image (Note: The options may be uppercase or lowercase English letters, letters from other languages, or numbers, etc.).
9. Output only the option corresponding to the language in the image to respond, without any additional explanation.
10. If the image is in Arabic and the options are not given in the image, assign four answers in the order of A, B, C, D to the options, and then select the corresponding English letter to answer.
11. If the image is in Bulgarian, the options should be output according to the characters in the image, following the list: ['А','В','Б','а','в','б','Г'].

Figure 2: A sample standardized prompt

- **Input Text Representation**: The input text T_i is tokenized into a sequence of tokens, denoted as $T_i = \{t_1, t_2, \ldots, t_m\}$, where m is the length of the text.
- **Embedding Layer**: Each token t_i is converted into a fixed-dimensional vector $\mathbf{e}_i$ through an embedding layer, i.e., $\mathbf{e}_i = \text{Embed}(t_i)$.
- **Positional Encoding**: To preserve the sequential information of the text, positional encoding $\mathbf{p}_i$ is added to each embedded vector, resulting in $\mathbf{e}'_i = \mathbf{e}_i + \mathbf{p}_i$.
- **Large Language Model Encoding**: The embedded text vector sequence $\{\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_n\}$ is fed into the LLM to generate the text feature representation $\mathbf{H}_T = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_m\}$.

Visual Encoding

The visual encoding process utilizes the Vision Transformer (ViT) architecture and follows these steps:

- **Input Image Preprocessing**: The input image I_i is resized to a specific resolution $R \times R$, such as 448×448.
- **Patch Embedding**: The image is divided into patches of size $P \times P$. Each patch is flattened and embedded into a fixed-dimensional vector. Suppose the image size is $R \times R$ and the patch size is $P \times P$; the image is divided into $N = \left(\frac{R}{P}\right)^2$ patches. Each patch I_j is embedded into a vector $\mathbf{v}_j$, i.e., $\mathbf{v}_j = \text{PatchEmbed}(I_j)$.
- **Positional Encoding**: To preserve the spatial information of the image, positional encoding $\mathbf{p}_j$ is added to each patch embedding vector, resulting in $\mathbf{v}'_j = \mathbf{v}_j + \mathbf{p}_j$.

- **Transformer Encoding**: The patch embedding vector sequence $\{\mathbf{v}'_1, \mathbf{v}'_2, \ldots, \mathbf{v}'_N\}$ is fed into the Vision Transformer to generate the image feature representation $\mathbf{H}_I = \{\mathbf{h}'_1, \mathbf{h}'_2, \ldots, \mathbf{h}'_N\}$.

Vision-Language Fusion

The fusion of visual and language features is achieved through the position-aware vision-language adapter, following these steps:

- **Feature Compression**: Since the length N of the image feature sequence $\mathbf{H}_I$ is usually much larger than the length m of the text feature sequence $\mathbf{H}_T$, the image features need to be compressed. The adapter uses a single-layer cross-attention module to achieve this. Let the learnable query vectors be $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_K\}$, where K is the number of query vectors. The cross-attention operation is defined as:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{H}_I^T}{\sqrt{d}}\right)$$

$$\mathbf{H}'_I = \mathbf{A}\mathbf{H}_I$$

  where d is the feature dimension, $\mathbf{A}$ is the attention weight matrix, and $\mathbf{H}'_I$ is the compressed image feature representation with length K.
- **Position Information Injection**: To preserve the spatial information of the image, 2D absolute positional encoding $\mathbf{P}$ is injected into the cross-attention operation, i.e.,

$$\mathbf{A} = \text{Softmax}\left(\frac{(\mathbf{Q} + \mathbf{P})\mathbf{H}_I^T}{\sqrt{d}}\right)$$

  where $\mathbf{P}$ is the positional encoding matrix with the same dimension as the query vectors $\mathbf{Q}$.
- **Fusion Representation**: The compressed image features $\mathbf{H}'_I$ and the text features $\mathbf{H}_T$ are fed into the LLM for further integration, generating the final multimodal representation $\mathbf{H}_{\text{fusion}}$.

The above methods can efficiently integrate image and text data, achieving superior performance in multimodal tasks.

3. Experiments

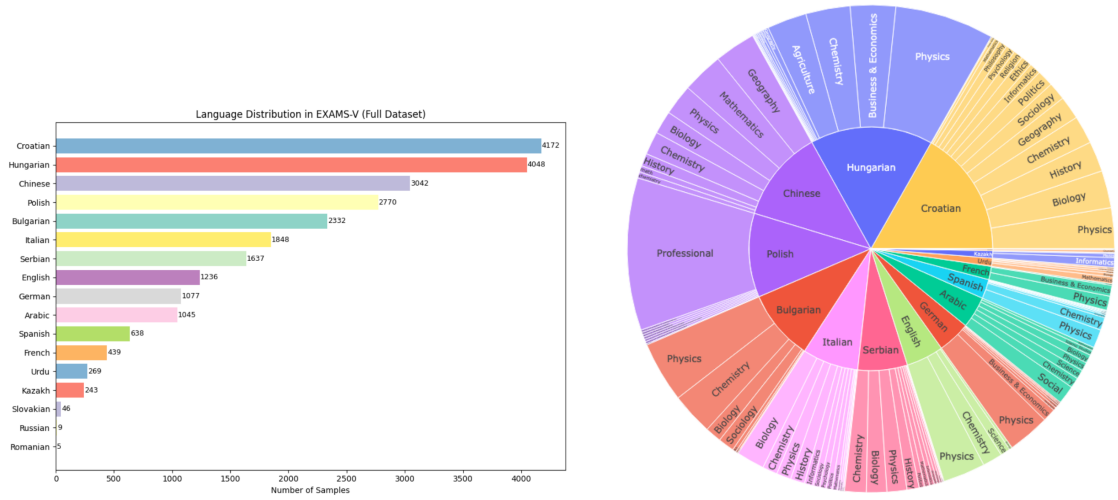
3.1. Data Pre-processing

The EXAMS-V dataset provided by ImageCLEF 2025 for multimodal reasoning tasks consists of 24,856(training set: 16,494, validation set: 4,797, test set: 3,565) multiple choice questions (MCQ) collected from real school exams and other educational sources, presented in the form of images[14]. The data set features:

- **Diverse**: The content covers pure text questions as well as visual elements such as tables, figures, graphs, or scientific symbols.
- **Multilingual**: A multilingual corpus covering 13 different languages, such as English, Arabic, and Chinese.
- **Interdisciplinary**: A wide coverage of academic subjects, including biology, chemistry, physics, and more.

Table 1 presents the proportion of various data formats, while Figure 3 illustrates the distribution of languages and subjects within the dataset.

This study employs Qwen-VL-Max (7B parameters, 131K-token context window, 448×448 image resolution) as the foundational vision-language model—selected for its state-of-the-art performance with 58.7% accuracy on the MMMU benchmark, representing a 12.3-point improvement over Qwen-VL-Plus(for more information, please visit the official Qwen-VL-Max repository at https://github.com/QwenLM/Qwen-VL.). The raw data is processed through the following steps:



(a) (b)

Figure 3: (a)The following is a histogram showing the distribution of languages in the EXAMS-V dataset. The chart reflects how many samples exist for each language across the full dataset (train, validation, and test).(b)The following sunburst chart shows the distribution of subjects across different languages in the EXAMS-V dataset. The inner ring represents languages, while the outer ring shows the subjects present within each language. This visualization highlights the multilingual and multi-domain nature of the dataset[14].

Table 1
Summary of Question Types and Visual Elements

Category	Visual Qs.	Text Only	Table	Figure	Graph	Total
Count	6,460	18,396	694	4,422	1,266	24,856
Percentage	26.0%	74.0%	2.8%	17.8%	5.1%	100.0%

Table 2
Comparison of Accuracy between Different Methods on the Validation Set

Language	Qwen-VL-Max (Direct)	Qwen-VL-Max (Prompt-Engineering)	Qwen-VL-Max (Prompt-Engineering + Pair)
Multilingual	0.6755	0.7197	**0.7456**
Arabic	0.5493	0.6364	**0.7060**
Bulgarian	0.6813	0.8575	**0.8625**
Chinese	0.6317	0.6750	**0.7200**
Croatian	0.7641	0.7795	**0.8017**
English	0.4017	0.4611	**0.4784**
French	0.8348	0.8348	**0.8438**
German	0.7878	0.7849	**0.7993**
Hungarian	0.6467	0.6766	**0.7140**
Italian	0.7509	0.7633	**0.7954**
Polish	0.5000	0.5600	**0.5800**
Serbian	0.7191	0.7629	**0.7968**
Slovakian	**0.8043**	**0.8043**	**0.8043**
Spanish	0.7100	0.7600	**0.7900**

- **Binary Encoding Conversion**: The binary image encoding is converted into Base64 format, an encoding method that transforms binary data into ASCII strings for convenient transmission and processing in text-based systems.
- **Image Description Generation**: The Qwen-VL-Plus model is utilized to analyze the image and generate a descriptive text for it. The purpose is to extract key information from the image to facilitate better understanding of its content by subsequent models.

Table 3
Comparison of Accuracy between Qwen-VL-Max and Baseline on the Test Set

Language	Qwen-VL-Max(Prompt-Engineering + Pair)	Baseline
Multilingual	**0.5619**	0.2701
English	**0.5312**	0.2480
Bulgarian	**0.7500**	0.2450
Chinese	**0.5799**	0.2678
German	**0.6860**	0.3101
Arabic	**0.3243**	0.2703
Italian	**0.6059**	0.2414
Spanish	**0.6608**	0.3156
Urdu	**0.3569**	0.3011
Serbian	**0.6059**	0.2365
Hungarian	**0.5425**	0.2348
Croatian	**0.6207**	0.2709
Polish	**0.5792**	0.2934
Kazakh	**0.4938**	0.2738

- **Data Pair Construction**: The generated image description text is combined with the Base64-encoded image to form a data pair, which is then passed as input to the Qwen-VL-Max model.

After the model processes the data, the output results are organized in the following format:

- **id**: A unique identifier(matching to a sample from the Test set).
- **language**: The language used in the sample.
- **answer_key**: The identifier for the correct answer option(one of A, B, C, D, or E).

3.2. Experimental Results

The official evaluation metric for this task is accuracy. In this experiment, we use Prompt 2 provided by ImageCLEF 2025 (a step-by-step reasoning prompt encouraging deeper analysis of textual and visual cues) as the standardized prompt. Table 2 shows the accuracy of Qwen-VL-Max on the validation set using the following three methods:

- **Qwen-VL-Max (Direct)**: This method directly applies the Qwen-VL-Max model without any prompt engineering or additional data pairing.
- **Qwen-VL-Max (Prompt-Engineering)**: This method adjusts the prompt to guide the model towards more accurate reasoning.
- **Qwen-VL-Max (Prompt-Engineering + Pair)**: This method combines the adjusted prompts with multimodal data pairs to form multimodal prompts.

Table 3 presents the comparison of accuracy between Qwen-VL-Max and the baseline methods on the test set.

The experimental results show that by introducing multimodal prompts, the Qwen-VL-Max model has achieved enhanced performance in multimodal reasoning tasks. On the validation set, the model's accuracy across all languages has surpassed both the direct use of the model and the use with adjusted prompts, reaching 74.56% in multilingual settings. On the test set, compared to the baseline methods, Qwen-VL-Max with prompt Engineering and data pairing has seen a comprehensive improvement in accuracy across all languages, with a 29.18% increase in multilingual accuracy, reaching 56.19%. This indicates that the proposed method in this paper can effectively enhance the model's ability to understand and reason with complex multimodal inputs.

4. Conclusion

This paper presents a multimodal prompting strategy for the Qwen-VL-Max model, focusing on enhancing the performance of Vision-Language Models (VLMs) in multimodal reasoning tasks. The core objective of this study is to enhance the model's comprehension and reasoning abilities for both image and text information through meticulously designed multimodal prompts and feature alignment methods, thereby effectively addressing complex multimodal reasoning tasks. The research findings and experimental results on the EXAMS-V dataset provided by ImageCLEF 2025 are detailed in this paper. The experiments demonstrate that the introduction of multimodal prompts can significantly enhance the image understanding capabilities of VLMs.

However, this method, which solely relies on prompting to guide model learning, is highly efficient and easy to implement but has limitations in enhancing the image understanding and reasoning capabilities of VLMs. Future research may further explore the design and optimization of prompts and integrate prompt learning with model fine-tuning to improve the models' reasoning abilities in complex multimodal tasks.

Acknowledgments

This work is supported by the Quality Engineering Projects for Teaching Quality and Teaching Reform in Undergraduate Colleges and Universities of Guangdong Province (No.20251067).

Declaration on Generative AI

During the preparation of this work, the author(s) used kimi in order to: Grammar and spelling check. After using these tool(s)/service(s), 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] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, D. Batra, Vqa: Visual question answering, International Journal of Computer Vision 123 (2015) 4 – 31. URL: https://api.semanticscholar.org/CorpusID:3180429.

[3] A. Taleb, C. Lippert, T. Klein, M. Nabi, Multimodal self-supervised learning for medical image analysis, in: Information Processing in Medical Imaging, 2019. URL: https://api.semanticscholar.org/CorpusID:209202500.

[4] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, I. Posner, Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks, 2017 IEEE International Conference on Robotics and Automation (ICRA) (2016) 1355–1361. URL: https://api.semanticscholar.org/CorpusID:2017183.

[5] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A frontier large vision-language model with versatile abilities, ArXiv abs/2308.12966 (2023). URL: https://api.semanticscholar.org/CorpusID:263875678.

[6] 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, 2023. URL: https://api.semanticscholar.org/CorpusID:261101015.

[7] 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.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: North American Chapter of the Association for Computational Linguistics, 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[9] M. U. Khattak, H. A. Rasheed, M. Maaz, S. H. Khan, F. S. Khan, Maple: Multi-modal prompt learning, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022) 19113–19122. URL: https://api.semanticscholar.org/CorpusID:252735181.

[10] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 3045–3059. URL: https://aclanthology.org/2021.emnlp-main.243/. doi:10.18653/v1/2021.emnlp-main.243.

[11] Qwen Team, Introducing Qwen-7B: Open foundation and human-aligned models (of the state-of-the-arts), https://github.com/QwenLM/Qwen-7B, 2023. Accessed: 2025-05-28.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ArXiv abs/2010.11929 (2020). URL: https://api.semanticscholar.org/CorpusID:225039882.

[13] G. Ilharco, M. Wortsman, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, L. Schmidt, Openclip, https://doi.org/10.5281/zenodo.5143773, 2021. doi:10.5281/zenodo.5143773, version 0.1.

[14] 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.