Evaluating Multimodal Large Language Models for Visual Question-Answering in Italian

Antonio Scaiella[1,2], Daniele Margiotta[1,2], Claudiu Daniel Hromei[1], Danilo Croce[1,*] and Roberto Basili[1]

[1]Department of Enterprise Engineering, University of Rome Tor Vergata, Italy
[2]Reveal s.r.l.

Abstract

Visual Question-Answering (VQA) is a complex multimodal task that requires integrating visual recognition and natural language understanding to answer questions about images. While significant progress has been made in English, resources and models for non-English languages, such as Italian, remain scarce. This paper addresses this gap by evaluating MiniCPM-V 2.6, a state-of-the-art multimodal Large Language Model, on GQA-it, the first large-scale Italian VQA dataset. The primary goal of this work is to investigate the performance of such models when applied off-the-shelf to this task and, if unsatisfactory, to explore how much they can improve with fine-tuning on Italian data. When applied off-the-shelf, MiniCPM-V 2.6 achieves an accuracy of 33.4%. However, after fine-tuning it on the GQA-it dataset, the performance improves significantly, reaching a state-of-the-art accuracy of 59.4%. These findings highlight the importance of language-specific adaptation in multilingual VQA tasks, especially for under-resourced languages like Italian. The trained model is released to the community on a dedicated Huggingface repository: https://huggingface.co/sag-uniroma2/MiniCPM-V-2_6-gqa-it-finetuned.

Keywords

Multimodal Large Language Model, Vision Language Model, Large Language Model, Visual Question-Answering

1. Introduction

Visual Question-Answering (VQA) is a challenging and rapidly evolving task in the field of Artificial Intelligence (AI). It requires a system to provide an accurate answer to a question posed in natural language based on the visual content of an image. The question typically depends on various details within the image, such as objects, relationships, actions, or other visual attributes, and demands that the system integrates both visual recognition and language understanding to generate a correct and contextually relevant response [2, 3]. This task presents significant challenges, as it necessitates the integration of two complex domains—vision and language—while also requiring models to employ reasoning and inference capabilities to arrive at accurate conclusions.

Figure 1 shows an example image, for which a wide range of questions can be asked, each with its respective answer, between brackets, in both English and Italian:

- **Q(A)**$_{en}$: *Is the remote to the right or to the left of the book? (right).*
- **Q(A)**$_{it}$: *Il telecomando è a destra o a sinistra del libro? (destra)*
- **Q(A)**$_{en}$: *How thick is the book to the left of the remote? (thick).*
- **Q(A)**$_{it}$: *Quanto è spesso il libro a sinistra del telecomando? (spesso)*
- **Q(A)**$_{en}$: *What device is to the left of the calculator made of plastic? (charger).*
- **Q(A)**$_{it}$: *Quale dispositivo si trova a sinistra della calcolatrice di plastica? (caricabatterie)*
- **Q(A)**$_{en}$: *What's the charger made of? (plastic).*

NL4AI 2024: Eighth Workshop on Natural Language for Artificial Intelligence, November 26-27th, 2024, Bolzano, Italy [1]
*Corresponding author.
scaiella@revealsrl.it (A. Scaiella); margiotta@revealsrl.it (D. Margiotta); hromei@ing.uniroma2.it (C. D. Hromei); croce@info.uniroma2.it (D. Croce); basili@info.uniroma2.it (R. Basili)
0009-0000-8204-5023 (C. D. Hromei); 0000-0001-9111-1950 (D. Croce); 0000-0001-5140-0694 (R. Basili)

Figure 1: Example of an image from the GQA-it dataset, taken from [4] (image `id`: `n90294`)

- **Q(A)**$_{it}$: *Di cosa è fatto il caricabatterie? (plastica)*
- **Q(A)**$_{en}$: *Are there any phones? (no).*
- **Q(A)**$_{it}$: *Ci sono dei telefoni? (no).*

The task is particularly complex because an image can be rich in information (in this case, a large number of objects), express actions, colors, and sensations, and be the subject of many different questions, as demonstrated by the small number of examples listed above. Some questions (such as those about the color of an object or its position) can be answered by simply observing the image, while others may require prior knowledge (such as understanding the typical materials of a charger). Additionally, some questions could be difficult to answer because it is unclear what the observer's focus is, some may have no answer (for example, if the question concerns an object not present in the image), or certain questions might require follow-up clarifying questions to provide an adequate response, necessitating a dialogical interaction [5, 6].

As a benchmark for evaluating the effectiveness of AI systems, VQA has garnered increasing attention due to its role in assessing how well these systems can perform tasks that involve deeper semantic understanding and cross-modal reasoning between language and vision. In recent years, several models and resources have been developed to address the challenges of VQA, particularly for the English language. These advancements have contributed to significant breakthroughs in the field [7, 8], enabling systems to perform such sophisticated reasoning. Comprehensive surveys, such as [9], provide valuable insights into the diverse methodologies employed in this task, including the integration of vision and language components, which has become an essential area of AI research.

Among the most successful approaches, Multimodal Large Language Models (MLLMs) have shown significant promise [10]. Large Language Models (LLMs) are powerful models designed to generate text based on a given input in an autoregressive manner. These models encode input text by transforming individual symbols in the input sequence into embeddings, which are dense vector representations capturing semantic information [11]. The model then generates output text by predicting the next token in the sequence based on the context provided by the embeddings of the previous tokens. Such models are capable of solving different linguistic tasks, as exemplified in [12]. In the case of MLLMs [10], the input is typically extended to include information from multiple modalities, such as text and images. These models incorporate embeddings not only from textual data but also from visual inputs, which are encoded using specialized encoders such as Convolutional Neural Networks (CNNs) or the recently introduced Vision Transformers (ViT, [13]). By integrating these additional embeddings, multimodal LLMs are able to process and reason over both linguistic and visual information.

Some of the most promising Multimodal LLMs, such as LLaVA [14], CogVLM [15], InternVL2-8B

[16], MiniCPM-V [8], GPT-4 [17], and Gemini [18], extend traditional language models like LLaMA [19], GPT-series [17] or Qwen2 [20] by incorporating visual encoders. These models are designed to handle tasks that require both visual and linguistic input, enabling complex reasoning across modalities. They utilize techniques such as cross-modal attention [21] and visual grounding to map visual information into the LLM's input space, allowing the models to generate more accurate and context-aware outputs in multimodal scenarios [22].

In most cases, these models reuse and specialize existing LLMs that have been pre-trained on large-scale document collections and then fine-tuned within the final multimodal architecture [14]. The quality of the final model is therefore closely tied to the original LLM and may inherit its limitations, such as an unbalanced ability to handle different languages. Models like LLaMA-2 [19] have been pre-trained on datasets containing trillions of words, with more than 90% of the text in English, while languages like Italian account for only 0.1% of the data. This imbalance poses a risk of limiting the effectiveness of such methods in a multilingual context. Furthermore, the fine-tuning of these models for visual reasoning is primarily conducted on English datasets, or in some cases, Chinese or English-Chinese datasets, as seen in CogVLM [15, 23]. The reliance on a limited set of languages and the lack of large-scale multimodal datasets in non-English languages, such as Italian, hinders the performance, robustness, and generalizability of VQA models in multilingual contexts.

This paper aims to answer the following research questions:

- *How do state-of-the-art models for VQA and other multimodal tasks involving both images and text perform when faced with a VQA problem in a non-English language*?
- *Can these models be used off-the-shelf, or would they benefit from—or even require—further fine-tuning on annotated Italian data*?

To explore these questions, we utilize the GQA-it dataset proposed by [4], a large-scale resource specifically designed for Visual Question-Answering in Italian. This dataset provides a valuable benchmark for developing and evaluating VQA systems in non-English contexts, addressing the gap caused by the lack of high-quality multimodal resources in languages beyond English.

In this paper, we evaluate how MiniCPM-V [8], a state-of-the-art open-source multimodal model, performs on the GQA-it dataset. MiniCPM-V was selected for its competitive performance, even in comparison to larger models like GPT-4. A key feature of MiniCPM-V is its multilingual support, including Italian, making it suitable for testing in non-English contexts. Despite its smaller size, MiniCPM-V is optimized for multimodal tasks such as Optical Character Recognition (OCR) and image comprehension, while remaining scalable for more complex datasets and tasks. By testing MiniCPM-V off-the-shelf on the GQA-it dataset, we aim to assess its ability to generalize across languages without additional training. We then compare its off-the-shelf performance to a fine-tuned version to determine how well the model adapts to Italian and whether fine-tuning significantly improves its results on this benchmark.

Our experimental results provide interesting insights into the research questions posed. MiniCPM-V, when evaluated in its base form, i.e. zero-shot, achieves an accuracy of 33.4%, demonstrating that while the model possesses general multimodal capabilities, it struggles to effectively handle VQA tasks in Italian without any additional adaptation. This aligns with our first research question, indicating that state-of-the-art models may not perform optimally in non-English contexts when used off the shelf. However, after fine-tuning the model over the GQA-it dataset, its performance dramatically improves, reaching a state-of-the-art accuracy of 59.4%. This significant improvement addresses our second research question, highlighting that fine-tuning seems crucial for enhancing the model's ability to handle Italian, as it allows the model to better understand the linguistic and visual nuances of the target language. The results emphasize that even advanced models require further adaptation to achieve competitive performance in multilingual VQA tasks. These experimental findings underscore the importance of language-specific adaptation in the development of multimodal models, particularly for under-resourced languages like Italian. Even state-of-the-art open-source models, such as MiniCPM-V, struggle to generalize effectively to non-English languages without targeted fine-tuning, as evidenced by the marked performance disparity observed in our GQA-it evaluation. This reinforces the need for

dedicated resources and model adjustments to enable competitive performance in multilingual settings, highlighting that fine-tuning is not just beneficial but often necessary to overcome the inherent bias toward English in many pre-trained multimodal models.

The rest of the paper is organized as follows: Section 2 describes the GQA-it dataset, while Section 3 discusses the investigated Multimodal LLMs (MLLMs). The experimental evaluation is presented in Section 4, and conclusions are drawn in Section 5.

2. GQA-it: Italian Visual Question-Answering Dataset

GQA-it [4] is the first large-scale dataset specifically designed for Visual Question-Answering (VQA) in Italian. It is an adaptation of the GQA dataset [7], which was initially created to assess real-world visual reasoning in English. The original GQA dataset contains over 22 million question-answer pairs across images sourced from the Visual Genome dataset, where each image is annotated with detailed scene graphs that capture objects, attributes, and relationships present in the scene.

GQA-it mirrors this structure, translating the questions and answers into Italian. The GQA dataset itself provides a rich framework for visual question-answering by incorporating multi-step reasoning, compositional questions, and balanced answer distributions to prevent models from exploiting dataset biases [7]. The creation of GQA-it involved translating this extensive set of questions and answers while preserving the complexity and richness of the visual reasoning tasks. In particular, the original English questions and answers were translated into Italian through a semi-automatic process that combines neural machine translation (NMT) and manual validation. This ensures a high-quality resource for training and evaluating VQA models in Italian, addressing the gap of multimodal datasets in non-English languages.

The GQA-it dataset consists of more than 1 million question-answer pairs and preserves the structure and balance of the original GQA dataset. Each question is designed to assess various aspects of visual understanding, including object recognition, spatial reasoning, and scene comprehension, making GQA-it a comprehensive benchmark for evaluating VQA systems. In Figure 1, an example image is displayed along with five question-answer pairs, showcasing the type of visual reasoning tasks that can be posed based on the image content, both in Italian and English. This highlights the dataset's ability to test models across multiple linguistic and visual dimensions.

In the creation of GQA-it, Neural Machine Translation (NMT) was employed to automatically translate the English questions and answers into Italian. Although NMT achieved high-quality translations, manual validation was applied to a subset of 3,000 examples to ensure the reliability of the test data. This validation process involved correcting errors related to gender inflection, lexical ambiguity, and inconsistencies in the translation of answers, which were often sensitive to the context provided by the corresponding question and image.

Dataset Split	#Images	#Question-Answer Pairs
`train`	72,140	943,000
`validation`	10,234	132,062
`test-dev (silver)`	398	12,578
`test-dev (gold)`	398	3,000

Table 1
Statistics of the GQA-it dataset. The gold `test-dev` is a subset of the silver one and has been manually validated.

As shown in Table 1, GQA-it is divided into training, validation, and test sets, with the test set further split into *silver* (automatically translated) and *gold* (manually validated) subsets. The silver test set comprises automatically translated questions and answers, while the gold test set contains manually corrected samples, providing a high-quality evaluation benchmark for VQA models. GQA-it poses significant challenges for models, as it requires a deep understanding of both the Italian language and the visual content. Furthermore, the dataset supports a wide range of visual reasoning tasks, including

object detection, relationship identification, and attribute recognition. This makes GQA-it a valuable resource for advancing research in multilingual multimodal AI.

3. Multimodal LLMs

Multimodal input signals enable virtual or physical agents to perceive and interact with their environments in more meaningful ways. One increasingly explored area in this domain is the conjunction of text and images. A seminal contribution to this field is CLIP [24], an architecture that takes a text and an image as input and originally produces a similarity score between the two. The architecture is optimized using a Contrastive learning schema, which defines a specific contrastive loss: this is minimized if the text accurately describes the image and maximized if the two inputs represent entirely different topics. The strength of CLIP lies in its ability to bring similar images and texts closer in a shared latent space while pushing dissimilar pairs far apart. This is achieved in a supervised manner, meaning that the training dataset must be annotated. Building upon CLIP, BLIP [25] was introduced to bootstrap the architecture from a pre-trained model in an unsupervised fashion, leveraging web image data along with their captions. This approach eliminates the need for large amounts of annotated data. Both CLIP and BLIP, however, depend on a global comparison of image-text similarities across a batch, even if the pairs are not directly related. As a solution, Sigmoid loss for Language-Image Pre-training (SigLIP) [26] was developed. Unlike the contrastive learning employed by CLIP, which uses Softmax normalization and relies on global pairwise similarity comparisons, SigLIP introduces a sigmoid loss that operates only on individual image-text pairs. This approach significantly enhances performance while reducing the dependence on large batch sizes for normalization. A major limitation of contrastive loss in CLIP is its assumption of a single correct image-text pairing for each example. However, in real-world scenarios, images and captions often have multiple plausible associations. Sigmoid loss addresses this by computing binary cross-entropy for each potential image-text pair, framing the problem as multi-label classification. This allows the model to score multiple relevant image-text pairs, rather than forcing a single correct match. As a result, Sigmoid loss enables smoother and more flexible alignment between images and texts, effectively handling cases where an image may correspond to several valid captions, or a caption may describe multiple images.

These models are specifically designed to learn the best representations of the two modalities: vision and text. More recently, architectures such as LLaVA [14], CogVLM [15], and CogAgent [22] have demonstrated effective methods for integrating CLIP and/or BLIP with a Large Language Model (often based on variations of LLaMA [19]) to perform complex inference tasks. These tasks include describing images using ad hoc prompts, explaining visual memes found online, and executing visual grounding through bounding boxes that reference the entities present in the image.

LLaVA [14] is an end-to-end multimodal model that connects a vision encoder (CLIP) with a Large Language Model (Vicuna [27]) for general-purpose visual and language understanding. At this point, one challenge arises: how do we reconcile the encoding of images with the encoding of text? To address this, a third component is needed—specifically, a single-layer MLP, known as the Projector, that maps the output of the visual encoder into the input space of the LLM. In LLaVA, this allows the model to seamlessly integrate visual information with textual information for coherent generation.

On the other hand, CogVLM [15] takes a different approach. It bridges the gap between a frozen pre-trained language model and a frozen image encoder by inserting a trainable visual expert module into the attention and feed-forward network (FFN) layers. This modification allows the model to more deeply understand both the text and image inputs. However, a limitation of CogVLM is that it processes images at a lower resolution, which makes it less capable of capturing smaller details in the image. As a follow-up to CogVLM, CogAgent [22] is built on the same architecture but introduces a visual cross-attention module. This module operates between the image-text input pair and a high-resolution version of the input image. By incorporating this high-definition image, the model can boost performance, enabling it to capture finer details and provide more accurate visual understanding. Both CogVLM and CogAgent excel in understanding and describing images, following instructions, and

solving image-dependent tasks such as Visual Question-Answering (VQA). They also have the ability to reference specific entities within an image using bounding boxes, a crucial feature that helps these models focus on particular objects and allows intelligent agents to better understand their surroundings.

Finally, a smaller multimodal model called MiniCPM-V 2.6 [8] has been released, which aims to reduce the number of parameters while maintaining strong OCR and multimodal capabilities. Unlike larger models such as CogVLM and LLaVA, which require considerable computational resources, MiniCPM-V 2.6 is tailored for scenarios where computational efficiency and speed are crucial, such as mobile phones, thanks to its carefully designed training methodology that enables scaling of both model size and data horizons. The architecture of MiniCPM-V 2.6 consists of three main components. The visual encoder, based on the SigLIP model, processes high-resolution images to extract visual tokens. These tokens are then passed through a compression layer, which includes a perceiver resampler (as in Flamingo [28]) that uses cross-attention to reduce the dimensionality of the data while retaining key features. Finally, the compressed visual tokens and text inputs are fed into the LLM, based on the recent Qwen2-7B [20] architecture. Remarkably, it shows competitive performance even when compared to much larger models. This makes it an appealing choice for real-time or resource-constrained environments, as it can process complex multimodal inputs without sacrificing the depth of its understanding or generation quality.

The availability of the above pre-trained models underscores their potential for tackling Visual Question-Answering tasks in languages other than English. To assess their effectiveness in a non-English context, particularly Italian, we leverage the GQA-it dataset. The upcoming section outlines the experimental setup and results, focusing on how MiniCPM-V 2.6 performs in both zero-shot and fine-tuned scenarios. This comparison aims to determine whether fine-tuning significantly enhances the model's ability to handle the linguistic and visual complexities of the Italian language, ultimately addressing the research questions of this work.

4. Experimental Evaluation

In this section, we address the research questions posed: how well state-of-the-art multimodal models perform on VQA tasks in non-English contexts, and whether they benefit from fine-tuning on Italian-specific data. To verify this, we apply MiniCPM-V to the GQA-it dataset, comparing its off-the-shelf performance to its results after fine-tuning. The aim is to evaluate the extent to which fine-tuning on the target language enhances the model's capabilities, and we conclude by analyzing the most common error types observed in each setup.

4.1. Experimental Setup

The aim of the evaluation is to compare MiniCPM-V 2.6 in two configurations: off-the-shelf (zero-shot) and fine-tuned. For this purpose, the model was assessed using the 3,000 examples from the manually validated (gold) set of the GQA-it test data (Section 2). This choice ensures the highest quality evaluation, as the manually validated set provides more reliable and accurate data than the automatically translated silver set. Additionally, this setup allows us to directly compare the results with LXMERT [21], a BERT-based model introduced in [4], which serves as a strong baseline in this context.

In GQA-it, each question is presented in natural language, and the expected answer is a short expression, typically consisting of one to four tokens. The models are evaluated based on their accuracy, which is the percentage of questions for which the system's response exactly matches the expected answer. This metric allows for a clear and direct measurement of the performance of the model in providing correct answers.

To evaluate the performance of the MiniCPM-V 2.6 model, we conducted experiments in two distinct scenarios, each designed to assess different aspects of the model's capabilities.

First, the model was applied off-the-shelf in a zero-shot setting, meaning no additional training or

Model	Accuracy
Baseline [4]	17.6%
LXMERT-it [4]	51.0%
MiniCPM-V 2.6 (Zero-shot)	33.4%
MiniCPM-V 2.6 (Fine-tuned)	59.4%

Table 2
Performance comparison of various models on GQA-it.

fine-tuning was performed on the target dataset. MiniCPM-V 2.6 was used from Huggingface[1]. While the model was primarily trained on data that may have been predominantly in English, it is expected to possess some understanding of other languages, including Italian. This setting provides insights into the model's ability to transfer its learning to a new language and domain without any task-specific adjustments. The following prompt[2] was used:

> `<IMAGE_HERE>` *Rispondi alla seguente domanda con una sola parola o poche parole ma solo se necessario, non aggiungere ulteriori informazioni:* `<QUESTION_HERE>`

This formulation was designed to be clear and direct, ensuring that the model understood it needed to provide a concise, single-word (or few words) response without adding any extra information.

In the second scenario, we employed a fine-tuned setting, where the MiniCPM-V 2.6 model was further trained on the GQA-it training set before being evaluated. This fine-tuning process involved adapting the model to the specific linguistic characteristics of Italian and the visual question-answering requirements of the dataset. The training was executed using the DeepSpeed framework[3] on a cluster of four A100 GPUs, each equipped with 80 GB of memory. The hyperparameters for the fine-tuning were set as follows: the standard fine-tuning parameters were used[4], with a learning rate of 1e-6, and the model was trained for 1 epoch due to the high number of examples. Both the vision and large language model (LLM) components were fine-tuned. The batch size per device during training was set to 4 to reduce memory usage. During the fine-tuning process, the model was trained using a simplified prompt, where only the original question and the image itself were presented:

```
<IMAGE_HERE> <QUESTION_HERE>
```

It is important to note that the prompt was kept extremely simple during fine-tuning, as additional instructions were unnecessary, given that the same prompt structure would be repeated for all observed questions during both training and model application. This streamlined approach aimed to help the model focus purely on the content of the question without additional instructions. By fine-tuning with this minimal prompt, the model was adapted to the task of responding directly to questions in Italian.

4.2. Experimental Results

Results are presented in Table 2. The first row reports the baseline model, which, due to the significant imbalance in the dataset, simply predicts the most frequent response ("*yes*" or "*si*"). This naive approach achieves an accuracy of 17.6%, offering a minimal benchmark for comparison.

The LXMERT-it model, introduced in [4], is specifically trained on the GQA-it dataset and achieves an accuracy of 51.0%, representing the current state-of-the-art on this benchmark. In contrast, when MiniCPM-V 2.6 is applied off-the-shelf in a zero-shot setting, it reaches an accuracy of 33.4%, which, while higher than the baseline, still lags behind the state-of-the-art. A notable portion of errors in this zero-shot configuration arises from the model's tendency to generate responses in languages other than Italian, as well as its struggle with yes/no questions. Our hypothesis is that since these models are

[1] https://huggingface.co/openbmb/MiniCPM-V-2_6
[2] In English: *Answer the following question with one word or a few words but only if necessary, do not add more information:*
[3] https://github.com/microsoft/DeepSpeed
[4] https://github.com/OpenBMB/MiniCPM-V/blob/main/finetune/finetune_ds.sh

Error Type	Example(s)	Our	[4]
Object	*tavola* (‘*table*’) vs *sedia* (‘*chair*’)	39%	30%
Syn or hyp	*persona* (‘person’) vs *donna* (‘*woman*’)	16%	17%
Attributes	blu (‘*blue*’) vs *nero* (‘*black*’); *chiuso* (‘*closed*’) vs *aperto* (‘*open*’)	17%	14%
Morph. feat.	*bella* (‘*beautiful*’) vs *bello* (‘*beautiful*’); *persona* (‘*person*’) vs *persone* (‘*people*’)	5%	3%
Actions	*sta dormendo* (‘*sleeping*’) vs *sta sdraiato* (‘*is lying down*’)	3%	3%
Spatial feat.	*destra* (‘*right*’) vs *sinistra* (‘*left*’)	2%	2%
Binary	*si* (‘*yes*’) vs *no* (‘*no*’)	18%	31%

Table 3
Distribution of errors of MiniCPM-V 2.6 (*Our* fine-tuned) and LXMERT-it on GQA-it gold test set into the predefined classes from [4].

predominantly trained and evaluated on English and Chinese data, they do not yet generalize effectively to Italian.

However, after fine-tuning on the GQA-it dataset, MiniCPM-V 2.6 achieves a substantial improvement, reaching an accuracy of 59.4%, which surpasses both the baseline, the LXMERT-it model, and its own zero-shot performance. This result highlights the critical importance of fine-tuning it on target language datasets to fully leverage a model's potential. Fine-tuning plays a pivotal role in improving a model's performance, especially when dealing with language-specific data, as it allows the model to adapt to the linguistic nuances and complexities of the target language. Moreover, it helps address the specific visual reasoning challenges posed by the dataset, ensuring more accurate and contextually relevant responses.

It is important to recognize that LXMERT-it has at least an order of magnitude fewer parameters compared to MiniCPM-V 2.6. Despite this, MiniCPM-V's architecture, specifically its LLM component based on Qwen2 [20], benefits from being trained on significantly larger and more diverse datasets. This contributes to its superior performance post-fine-tuning, as the extensive training data enables the model to better handle complex multimodal reasoning tasks and cross-lingual understanding.

Additionally, it is worth noting that the fine-tuning process involved adapting both the language model and the vision model. Further experimentation, where fine-tuning is applied selectively or where techniques such as LoRA (Low-Rank Adaptation) [29] are employed, remains an open area of exploration for future research.

4.3. Error Analysis

The error analysis aimed to identify the most frequent types of misclassifications made by the system. The resulting percentages, summarized in Table 3, also include a comparison with the analysis from [4], where errors were analyzed on a random 10% sample of the validated test set. In contrast, our current analysis was manually performed by the authors on the entire test set, offering a more comprehensive overview of the system's performance. An important note is that, while the results are not directly comparable—due to the different samples of misclassified examples analyzed—the error types and their distribution provide a useful indication of how the LXMERT-it model and the fine-tuned version of MiniCPM-V differ in their behavior. These insights offer guidance on how future iterations of the models might be improved.

We will focus on discussing the differences between the two analyses at the end of the section.

The most evident difference appears in binary-type questions, such as those involving yes/no answers. These findings underscore the difficulty in managing seemingly straightforward yet context-dependent queries, where subtle differences in the phrasing of the question can result in incorrect binary decisions, exposing the limitations of the model in reasoning. For instance, in example 2, when asked: “*C'è del vino in questa foto?*”[5], the model incorrectly responds with “*si*” (“*yes*”), when in fact, the image shows wine glasses, but no actual wine. Nonetheless, MiniCPM-V only produces errors in this category in *18%* of the examples, compared to *31%* for LXMERT-it. This could be attributed to the vision and language model

[5]In English: *Is there any wine in this picture*?

Figure 2: Example from the GQA-it gold test set (`image id n28572`) where the fine-tuned MiniCPM-V 2.6 model predicts "*si*" instead of "*no*" for the question *"C'è del vino in questa foto?"*.

components of MiniCPM-V working more effectively together, particularly in this category where more advanced reasoning may be required to respond correctly.

In general, the distribution of the remaining error types follows a similar pattern between MiniCPM-V and LXMERT-it. However, one notable exception is the category of object-related errors. In MiniCPM-V, object-related answers account for 39% of the errors, whereas in LXMERT-it, this percentage is slightly lower at 30%. Despite the lower overall number of errors in MiniCPM-V, this high percentage indicates that the model struggles significantly when distinguishing between similar objects, such as *"tavola"* ("*table*") and *"sedia"* ("*chair*"). In Figure 3, an example is shown where the model incorrectly answers the question *"Quale tipo di mobile è nero?"*[6] by predicting *"tavola"* (table) instead of "*sedia*" ("*chair*"). This type of confusion likely arises from the inherent visual or semantic similarities between objects, suggesting that the model's object recognition or feature differentiation requires further refinement.

Errors related to attributes were also prevalent, contributing 17% to the overall misclassifications. Examples like confusing "*blue*" with "*black*" demonstrate that the model struggles with precise attribute identification, potentially due to issues with color or texture recognition. This suggests that the model may require improvements in fine-grained feature extraction. Errors stemming from synonyms (syn) or hypernyms (hyp) followed closely, at 16%. These included confusion between terms like "*donna*" and "*ragazza*" ("*woman*" vs. "*girl*") or "*uomo*" and "*persona*" ("*man*" vs. "*person*"), pointing to challenges in linguistic nuances and semantic hierarchical relations. This kind of confusion may indicate that the language understanding capability of the model is limited in distinguishing between related (e.g. hyperonyms or hyponyms) but distinct concepts.

Other errors stem from genuine ambiguity. In Figure 4, for instance, the answer for the question "*Cosa c'è davanti alla felpa?*"[7] was annotated as "*tappeto*" ("*carpet*") by the human annotator, while the model answered "*scrivania*" ("*desk*"). This discrepancy (both answers could be deemed correct) underscores the challenge of interpreting spatial relationships without additional context. If the model were able to ask clarifying questions, it might inquire, "*Could you clarify what you mean by 'in front' in this context?*". The human annotator may have considered the observer's perspective, interpreting '*in*

[6] In English: *What type of furniture is black*?"
[7] In English: *What is in front of the sweatshirt?*

Figure 3: Examples from the GQA-it gold test set (`image id n283587`) where fine-tuned MiniCPM-V 2.6 model predict "tavola" instead of "sedia" when answering the question *"Quale tipo di mobile è nero?"*

front' as the space between himself and the chair, whereas the model might have interpreted '*in front*' as the position of the sweatshirt on the chair. The ability to seek such clarification could help resolve these ambiguities and enhance the accuracy of spatial understanding in similar scenarios.

Figure 4: Examples of ambiguity from the GQA-it gold test set (`image id n398257`) where fine-tuned MiniCPM-V 2.6 model answered "*scrivania*" instead of "*tappeto*" to the question "*Cosa c'è davanti alla felpa?*"

Another type of ambiguity can arise from differences in attention between the system and the annotator. For instance, in Figure 5, when asked "*Su cosa è sdraiato il gatto?*"[8], the annotator provided the annotation "*tappetino*" ("*mouse pad*"), while the model answered "*scrivania*" ("*desk*"). Once again, the discrepancy highlights a potential divergence in focus: the annotator might be concentrating on a more specific object, such as the mat directly beneath the cat, whereas the model may be considering a larger, more general object, like the desk that the mat is placed on. Such differences in attentional focus

[8]In English: "*What is the cat lying on?*"

can contribute to ambiguities in interpreting spatial relationships and objects in images.

Figure 5: Examples of attention ambiguity from the GQA-it gold test set (`image id n433692`) where fine-tuned MiniCPM-V 2.6 model responded "*tappetino*" instead of "*scrivania*" to the question "*Su cosa è sdraiato il gatto?*"

5. Conclusion

This study underscores the role that fine-tuning plays in enhancing the performance of multimodal models, particularly for under-resourced languages. Our analysis of MiniCPM-V 2.6, a state-of-the-art MLLM, demonstrates significant improvements when fine-tuned on Italian-specific datasets like GQA-it. While the zero-shot performance of the model yields a 33.4% accuracy, fine-tuning nearly doubles its accuracy to a state-of-the-art 59.4%. This emphasizes the importance of adapting models to the linguistic characteristics of the target language to fully unlock their potential. For languages like Italian, where available training data is scarce, the impact of language-specific fine-tuning becomes even more critical.

The challenges posed by multilingual Visual Question-Answering (VQA) tasks further support the necessity of language adaptation. Our findings show that models pre-trained predominantly on English data often exhibit limited generalization capabilities in other languages, reinforcing the presence of an English-centric bias in many multimodal pre-trained models. Addressing this bias through targeted fine-tuning on non-English datasets is crucial for developing AI systems that can operate effectively across multiple linguistic contexts.

Lastly, the comparative analysis highlights that MiniCPM-V 2.6, once fine-tuned, surpasses the previously best-performing Italian VQA model, LXMERT-it, which reached 51.0% accuracy. With a new benchmark of 59.4%, MiniCPM-V 2.6 demonstrates the potential for advanced multimodal LLMs to set state-of-the-art standards for VQA tasks in non-English languages. This success further underscores the value of fine-tuning with relevant language-specific data to achieve competitive performance in diverse linguistic settings.

Acknowledgments

Claudiu Daniel Hromei is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XXXVII cycle, course on *Health and life sciences*, organized by the Università Campus Bio-Medico di Roma. We acknowledge financial support from the PNRR MUR project PE0000013-FAIR and support from Project ECS 0000024 Rome Technopole, - CUP B83C22002820006, NRP Mission 4 Component 2 Investment 1.5, Funded by the European Union - NextGenerationEU.

References

[1] G. Bonetta, C. D. Hromei, L. Siciliani, M. A. Stranisci, Preface to the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024), 2024.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[3] Y. Srivastava, V. Murali, S. R. Dubey, S. Mukherjee, Visual question answering using deep learning: A survey and performance analysis, 2020. arXiv:1909.01860.

[4] D. Croce, L. C. Passaro, A. Lenci, R. Basili, Gqa-it: Italian question answering on image scene graphs, in: E. Fersini, M. Passarotti, V. Patti (Eds.), Proceedings of the Eighth Italian Conference on Computational Linguistics, CLiC-it 2021, Milan, Italy, January 26-28, 2022, volume 3033 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-3033/paper42.pdf.

[5] C. D. Hromei, D. Margiotta, D. Croce, R. Basili, MM-IGLU: Multi-modal interactive grounded language understanding, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 11440–11451. URL: https://aclanthology.org/2024.lrec-main.1000.

[6] F. Borazio, C. D. Hromei, E. Passone, D. Croce, R. Basili, MM-IGLU-IT: Multi-Modal Interactive Grounded Language Understanding in Italian, in: A. Artale, G. Cortellessa, M. Montali (Eds.), AIxIA 2024 - Advances in Artificial Intelligence - 23th International Conference of the Italian Association for Artificial Intelligence, Springer, 2024.

[7] D. A. Hudson, C. D. Manning, Gqa: A new dataset for real-world visual reasoning and compositional question answering, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6700–6709.

[8] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, et al., Minicpm-v: A gpt-4v level mllm on your phone, arXiv preprint arXiv:2408.01800 (2024).

[9] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, Z. Ge, Medical visual question answering: A survey, Artificial Intelligence in Medicine 143 (2023) 102611. URL: http://dx.doi.org/10.1016/j.artmed.2023.102611. doi:10.1016/j.artmed.2023.102611.

[10] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen, A survey on multimodal large language models, 2024. URL: https://arxiv.org/abs/2306.13549. arXiv:2306.13549.

[11] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2023. URL: https://arxiv.org/abs/2303.18223. arXiv:2303.18223.

[12] C. D. Hromei, D. Croce, V. Basile, R. Basili, ExtremITA at EVALITA 2023: Multi-Task Sustainable Scaling to Large Language Models at its Extreme, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[13] 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, 2021. URL: https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

[14] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, 2023.

[15] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, J. Tang, Cogvlm: Visual expert for pretrained language models, 2023. arXiv:2311.03079.

[16] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, J. Dai, Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, arXiv preprint arXiv:2312.14238 (2023).

[17] OpenAI team, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[18] Gemini Team at Google, Gemini: A family of highly capable multimodal models, 2024. URL: https://arxiv.org/abs/2312.11805. arXiv:2312.11805.

[19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. URL: https://arxiv.org/abs/2307.09288. arXiv:2307.09288.

[20] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Fan, Qwen2 technical report, arXiv preprint arXiv:2407.10671 (2024).

[21] H. Tan, M. Bansal, LXMERT: Learning cross-modality encoder representations from transformers, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5100–5111. URL: https://aclanthology.org/D19-1514. doi:10.18653/v1/D19-1514.

[22] W. Hong, W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji, Y. Wang, Z. Wang, Y. Dong, M. Ding, J. Tang, Cogagent: A visual language model for gui agents, 2023. arXiv:2312.08914.

[23] W. Hong, W. Wang, M. Ding, W. Yu, Q. Lv, Y. Wang, Y. Cheng, S. Huang, J. Ji, Z. Xue, et al., Cogvlm2: Visual language models for image and video understanding, 2024. arXiv:2408.16500.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020.

[25] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL: https://arxiv.org/abs/2201.12086. arXiv:2201.12086.

[26] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, 2023. URL: https://arxiv.org/abs/2303.15343. arXiv:2303.15343.

[27] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

[28] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a visual language model for few-shot learning, 2022. URL: https://arxiv.org/abs/2204.14198. arXiv:2204.14198.

[29] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.