LLaVA-NDiNO: Empowering LLMs with Multimodality for the Italian Language

Elio Musacchio[1,*,†], Lucia Siciliani[2,†], Pierpaolo Basile[2,†] and Giovanni Semeraro[2]

[1]Italian National PhD Program in Artificial Intelligence, University of Bari Aldo Moro, Bari (ITALY)
[2]Dept. of Computer Science, University of Bari Aldo Moro, Via E. Orabona, 4 - 70125 Bari (ITALY)

Abstract

Since their initial inception, large language models have undergone many innovations. One of these innovations concerns multimodality. Several adaptation strategies have been developed to expand LLMs to process multimodal signals. However, the training procedure for these multimodal models is performed on English-only vision-language datasets in the current literature, limiting their capabilities for other languages. This work proposes the first family of LMMs for the Italian language. We trained them using state-of-the-art backbone models and datasets, translated into Italian using the most up-to-date machine translation model available. In support of open science, we publicly release the data, models, and code used to develop these models.

Keywords

NLP, Multimodality, LLM, LMM, LVLM

1. Introduction

Large Language Models (LLMs) have been rising in research interest due to their generalization capabilities, which allow them to solve tasks never seen during training. However, their capabilities are limited to the textual domain. In light of this, researchers have started proposing solutions to bridge the gap between the textual world and the others (e.g. visual or aural). Specifically, instead of pre-training a new model with multimodal capabilities from scratch, these solutions leverage a pre-trained decoder-only LLM. This is both cost-efficient, avoiding the expensive training procedures of full multimodal training, and effective, as many of these solutions reported optimal results.

In this work, we will be focusing on the *vision-language* world, specifically *Large Vision Language Models* (LVLMs). These models are often trained following a traditional two-step approach: *pre-training* followed by *fine-tuning*. However, one notable issue is that the vision-language training mixture often consists of curated and selected datasets that predominantly feature English text, as seen in models like LLaVA [2]. This further propagates an inherent problem of these large models, where the pre-training corpus mainly consists of English data. For example, LLaMA 2 [3], a LLM by META, was pre-trained on a corpus of 89.70% English language and of 8.38% unknown language (e.g. programming code). As a result, even the developers of the models explicitly state that their usage is intended for English use cases only.

Furthermore, there is a significant gap due to the absence of large-scale, multitask and multilingual datasets. While the English vision-language datasets are conceptually diverse and rich (e.g., scientific question answering, OCR), non-English datasets tend to be limited in scope, focusing on specific high-level tasks (e.g., image captioning, visual question answering).

For these reasons, there are currently very few LVLMs in the state-of-the-art for non-English languages. While some models support multilingual and multimodal data, they often fall behind their

NL4AI 2024: Eighth Workshop on Natural Language for Artificial Intelligence, November 26-27th, 2024, Bolzano, Italy [1]

[*]Corresponding author.

[†]These authors contributed equally.

elio.musacchio@uniba.it (E. Musacchio); lucia.siciliani@uniba.it (L. Siciliani); pierpaolo.basile@uniba.it (P. Basile); giovanni.semeraro@uniba.it (G. Semeraro)

0009-0006-9670-9998 (E. Musacchio); 0000-0002-1438-280X (L. Siciliani); 0000-0002-0545-1105 (P. Basile); 0000-0001-6883-1853 (G. Semeraro)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

English counterparts in terms of architecture performance and training data quality. The reasons behind this are twofold: new LLMs are constantly being released, and training data lacks quality, focusing only on high-level tasks due to the lack of data. Furthermore, current multilingual and multimodal benchmarks are not as conceptually rich as English ones, making evaluation of these models more difficult for non-English languages.

Therefore, in this work, we propose an approach to train and evaluate a LVLM for the Italian language. We also release **LLaVA-NDiNO** (*Large Language and Vision Assistant: New Domain integration for Natural Observations*), the first family of openly-available Italian LVLMs trained and evaluated by following the proposed approach. While this approach heavily relies on the use of machine translation, we show that even when using machine-translated datasets at train time it is possible to achieve remarkable performance during evaluation on datasets that are natively in the Italian language. Specifically, the contributions of this work are the following:

- We apply a vision-language adaptation step designed to improve the performance of the model for a specific language. We compare the performance of a model trained using this additional step w.r.t. one without this step;
- We propose a new evaluation suite based on both machine-translated and natively Italian data from state-of-the-art benchmarks;
- We openly release code, data and models that have been obtained from our experiments, in the hope of boosting research in this field and in support of open science.[1]

2. Related Works

LVLMs have begun to see widespread success following the release of **GPT-4V** [4], the OpenAI model which supported vision-language inputs. However, since the model is proprietary, possibilities for research are relatively limited. Because of this, many works proposed open-source solutions, trying to match the performance obtained by GPT-4V on state-of-the-art benchmarks. One of the most popular solutions in this field of research is **LLaVA** [5, 2]. The model uses a projection module (either a projection matrix in its first version or a Multi-Layer Perceptron in version 1.5) to project the visual embeddings extracted from a visual encoder into the latent space of the LLM. This approach is simple and efficient, since it only relies on a single projection module. However, the original LLaVA architecture, as well as other LVLMs, struggled with high-resolution images tasks due to the requirements imposed by vision encoders. This is because vision encoders, like the *Vision Transformer* (**ViT**) [6], are trained on a fixed image size. Therefore, during inference or embedding extraction, the same image size is expected as input. To overcome this limitation, **LLaVA-NeXT** [7] was developed. In this model, the image is split into grids of fixed size and the embeddings for each grid are extracted and concatenated. Finally, the original image is resized and its embeddings are extracted and concatenated to the previous output. This technique allows the model to better understand the overall visual characteristics of the input images.

However, all of the LLaVA models were trained on English-only vision-language data. Specifically, an instruction-tuning approach over a rich set of vision-language tasks was performed. Therefore, while the LLaVA models perform well on English tasks, the lack of curated multilingual vision-language instruction-tuning datasets makes it challenging to train multilingual LVLMs on a set of conceptually diverse tasks. In light of this, some works focus on multilingual training procedures for LVLMs. Geigle et al. [8] released **mBlip**, a version of the BLIP 2 [9] model trained on an English vision-text dataset machine-translated to 95 different languages To do so, the authors used a neural machine translation model, that is nllb-200-distilled-1.3B [10]. There is also **Pali-X** [11], where the vision and language components are jointly scaled, following the work done in Pali [12]. The model is pre-trained on a rich range of datasets, among which there is **WebLI** [12], a rich corpus consisting of images with alt-texts from the web and OCR annotations obtained from the Google Cloud Vision API, covering a total of

[1] https://github.com/swapUniba/LLaVA-NDiNO

100 languages. Finally, there is **X-LLaVA** [13], where the authors adapted LLaVA 1.5 by expanding its dictionary for English and Korean and performing a language adaptation step based on the one performed by Conneau and Lample [14], that is pre-training on a data corpus extracted from Wikipedia.

Regarding datasets used to train these models, for LLaVA 1.5 a mixture of English only vision-language datasets was used. Specifically, the mixture contained $158,000$ GPT-generated multimodal instruction-following data instances, $450,000$ academic-task-oriented visual question answering data instances and $40,000$ ShareGPT data instances. Laurençon et al. [15] released **The Cauldron**, a collection of 50 different datasets pre-formatted for instruction-tuning. This dataset was used to train **Idefics 2** [15] model. The dataset consists of state-of-the-art vision-language datasets and covers a wide array of conceptual tasks. Specifically, the authors identify the following categories: *general visual question answering*, *captioning*, *OCR*, *document understanding*, *text transcription*, *chart/figure understanding*, *table understanding*, *reasoning*, *logic*, *maths*, *textbook/academic questions*, *differences between two images*, *screenshot to code*.

Despite all this, best practises regarding language adaptation of LVLMs are still unclear.

3. Methodology

We define three different steps in our methodology:

- **Italian vision-language pre-training**: training the model to optimize its general understanding of the Italian language;
- **Italian vision-language instruction-tuning**: fine-tuning the model on task specific vision-language data to improve its performance in following instructions;
- **Italian vision-language long instruction-tuning**: fine-tuning the model to produce long outputs in response to instructions.

We adapt a pre-trained decoder LLM and a pre-trained encoder vision transformer to the Italian language by performing an *Italian vision-language pre-training* approach. This is based on an approach used for LLMs, which consists in further training the model on a wide corpus of generic data of a specific language [14]. In this step, we perform the same approach but using vision-text data instead. Specifically, we directly use an English pre-trained decoder *LLM* and an English pre-trained vision encoder and perform joint language adaptation on both of them, as well as the adaptation module, on a collection of image-text pairs natively in Italian. We expect the model pre-trained on Italian data to perform better in Italian vision-language tasks, thanks to the additional knowledge it has gained.

Furthermore, while the instruction-tuning datasets are often unavailable in multiple languages, vision-language pre-train data is. Thanks to this, the data quality during pre-train is guaranteed since the text would be natively in Italian. However, the situation is different for instruction-tuning. Due to the lack of instruction-tuning Italian datasets, we must rely on machine translation. While the data quality will suffer from this, this approach is the only one that allows us to obtain the large quantity of data needed to achieve the generalization capabilities of LVLMs. Finally, we also perform further instruction-tuning for long response generation. This is because humans tend to prefer long and descriptive answers when interacting with LLMs and LVLMs. We decided to use the LLaVA-NeXT architecture since it is one of the most recent LVLMs available in the state-of-the-art. We detail all the steps we carried out, from data collection to evaluation.

3.1. Dataset Creation

For the Italian language pre-training dataset, following the best practises by Laurençon et al. [15], we setup three conceptually different datasets: **Interleaved image-text documents**, **Image-text pairs** and **PDF documents**. For interleaved image-text documents and image-text pairs, we use the **WIT** [16] dataset, a collection of images and their associated text sections obtained from Wikipedia pages in multiple languages. Specifically, after collecting the Italian portion of the dataset, we use the text of a section where an image appears as interleaved image-text document and the caption of the image as image-text pair. Note that for interleaved image-text documents we only use a single pair of image-text section, rather than multiple sections from the same Wikipedia page. For PDF documents, there are no multilingual datasets fitting this criteria in the literature. In particular, there are no handwritten datasets of this type, but only typewritten. Therefore, we decided to use **MultiEURLEX** [17], a corpus containing European laws in 23 languages. While this corpus is typewritten only, we prefer to include it in the pre-train dataset rather than not covering OCR at all. We retrieve the Italian PDF files associated with the corresponding CELEX_ID and extract the text from each document using Tesseract [18]. We also filter the dataset to control the distribution of these different sets. The pre-train dataset consists of $250,000$ instances, of which $168,000$ are interleaved image-text documents, $72,000$ are image-text pairs, and $10,000$ are PDF documents.

For the Italian language instruction-tuning dataset, we use **The Cauldron** [15], a collection of 50 vision-language datasets already formatted for instruction-tuning. Since the dataset is in English, we use machine translation to Italian. Details regarding the machine translation procedure will be discussed in Section 3.2. However, we first perform a filtering step of the 50 available tasks. This is because many tasks would lose their meaning when translated from English to another language (e.g. extraction of information from the image of a table where the text is in English). Because of this, we remove all tasks which focus on images containing English text (e.g. *docvqa* or *ocrvqa*). After performing this manual filtering step, we have a total of 15 tasks. For each task, we select the first $10,000$ rows of the dataset and perform machine translation on each instance in each row (more than one text-vision pair can be present for each row). Additionally, we also add the train sets of **MTVQA** and **V-EXAMS**, datasets that are natively in Italian. This increases both the quality of the instruction-tuning dataset, as the datasets are not machine translated, and its concept distribution, since two new tasks are added. MTVQA is the only dataset containing Italian visual text extraction and V-EXAMS is the only dataset containing Italian academic visual question answering. In total, the instruction-tuning dataset consists of $260,302$ instances.

For the Italian language long instruction-tuning dataset, we use LLaVA Conversation 58k [5], a subset of the LLaVA Instruct 150K dataset. It consists of 58k conversations, a dataset generated using GPT-4V for conversational purposes. Again, since the dataset is in English, we perform machine translation.

Finally, for evaluation, we collect the OK-VQA, SeedBench and POPE datasets, that are popular benchmarks used in the literature for English LVLMs. We machine translate them to the Italian language as well. We also collect the test sets of MTVQA, V-EXAMS and GQA-it.

We provide an overview of the 15 datasets from The Cauldron used for the instruction-tuning step in Table 1. We also provide the same details for the natively Italian datasets in Table 2 and evaluation datasets in Table 4.

3.2. Translation

To translate the data, we use one of the newest machine translation models openly available, that is **MADLAD-400 3B**[2] [36]. To accomplish this task, we use a cluster equipped with multiple NVIDIA A16 16GB VRAM GPUs. We use 4 GPUs in parallel and perform inference with a batch size per device of 4.

To translate the data from The Cauldron, we directly use the formatted instruction pairs present in the dataset. By doing so, the answer is translated with the context given by the question, reducing the possibility of a translation error. We do the same for closed-ended tasks, where a list of options is given in the question. However, this translation procedure may cause the model to translate text inaccurately. Therefore, some options for closed-ended tasks may not be translated correctly. For example, during translation, some closed-ended options might not align correctly with the original content, causing errors like having more options than in the original text. To avoid this issue, we check via regex matching that: 1) the question or instruction is present at the beginning; 2) the number of

[2] https://huggingface.co/jbochi/madlad400-3b-mt

Dataset	# Train Translated	Description
A-OKVQA [19]	10,107	VQA dataset requiring world knowledge and common sense for a correct answer.
CLEVR [20]	92,670	VQA dataset designed for visual reasoning regarding objects in images.
COCO-QA [21]	16,167	VQA dataset containing descriptive and rich question-answer pairs.
Geomverse [22]	3,324	VQA dataset regarding geometric reasoning.
IconQA [23]	10,980	VQA dataset regarding abstract diagram understanding.
InterGPS [24]	1,498	VQA dataset regarding geometric reasoning, annotated in a formal language.
Localized Narratives [25]	9,178	VQA dataset designed to provide rich descriptions of image contents.
Mimic CGD [26]	16,807	VQA dataset designed to enhance the performance of vision language models in real-life scenarios.
NLVR2 [27]	18,363	VQA dataset regarding truthfulness of a natural language sentence about a pair of photographs.
Raven [28]	9,216	VQA dataset regarding Raven's Progressive Matrices.
Spot the Difference [29]	9,187	VQA dataset regarding differences between two images.
TallyQA [30]	14,024	VQA dataset regarding complex counting questions of objects in images.
Visual7w [31]	43,228	VQA dataset regarding object-level grounding, using questions that start with one of *what, where, when, who, why, how and which*.
VQArad [32]	739	VQA dataset regarding radiology images.
VQAv2 [33]	1,563	VQA dataset requiring understanding of vision, language and commonsense knowledge to answer.

Table 1
Overview of all datasets from The Cauldron used during the instruction-tuning procedure of our models. # Train Translated is the amount of total translated instances obtained from the original first 10k rows of the dataset.

options is the same before and after translation; 3) the answer is present at the end of the translated string. In all cases where a check is not passed, the translated instance is removed from the dataset. We follow this same procedure to translate evaluation benchmarks. Because of this, some of these translated datasets may have a different cardinality w.r.t. original ones.

For LLaVA Conversation 58k we directly translate the user question and the system response. By testing the model, we noticed that translation errors are frequent when a newline character is present in the input. Therefore, we split inputs when two consecutive newline characters are present and further

Dataset	# Train	# Test	Description
MTVQA [34]	2,168	884	VQA dataset of multilingual text scenes. The dataset is manually labelled.
EXAMS-V [35]	1,083	562	VQA dataset of multilingual school exam questions. The dataset is obtained from real exam questions for each language.

Table 2
Overview of all datasets natively in Italian used during the instruction-tuning procedure of our models

split the output when a single newline character is present. The obtained strings are translated and the original newline characters are progressively added for each translated instance, effectively recreating the original formatting of the string but in another language.

4. Experiments

4.1. Training Details

We distinguish between four total train steps:

- **MLP pre-training**: the weights of the MLP module are initialized, following the strategy described by Liu et al. [2];
- **Italian language pre-training**: we optimize the model to the Italian language by further training the English backbones on a mixture of native Italian text-vision data;
- **Italian language instruction-tuning**: we optimize performance of the model in providing meaningful responses by performing instruction-tuning;
- **Italian language long instruction-tuning**: we optimize performance of the model in providing meaningful and descriptive responses by performing instruction-tuning.

For the Multi-Layer Perceptron (MLP) pre-training step, we use the same dataset as Liu et al. [2], that is LCS-558K. It is a subset of the LAION/CC/SBU dataset, filtered with a more balanced concept coverage distribution, and augmented with BLIP synthetic captions. We follow the procedure described in LLaVA 1.5 for this step.

Then, we perform our training using the translated Cauldron dataset on **LLaMA 3 8B base** [37] as LLM and **CLIP ViT large-patch14-338** [38] as vision encoder. This is to follow the configuration used by LLaVA-NeXT, except for the LLM model. We decided to use the *base* version instead of the *instruct* one. Since we have to perform pre-training, we have found the base version of the model to be more fitting for this purpose.

We train all models for a direct response in a single round user-system conversational setting. Specifically, we use two prompt formats: *plain* for the MLP and Italian pre-training, and the *LLaMA 3* instruct format without system prompt for instruction-tuning. These prompt formats are shown in Listing 1 and 2.

A diagram presenting an overview of the entire training pipeline is shown in Figure 1.

For all models, we perform full-parameter training. Regarding additional technical details, we report hyperparameters used in Table 3. The training was run on a cluster with 4 NVIDIA A100 64 GB GPUs per node. Specifically, we use 2 nodes for a total of 8 GPUs. We use a server with 8 NVIDIA A16 16 GB GPUs for evaluation, running the procedure on 4 GPUs.

4.1.1. Instruction-tuning and Evaluation

To assess the performance of the pre-trained model, we perform two different training procedures:

- **LLaVA-NDiNO IT**: only MLP pre-training and instruction-tuning have been performed;

- **LLaVA-NDiNO PT + IT**: MLP pre-training, Italian language pre-training and instruction-tuning have been performed.

```
<|begin_of_text|><image>{text}<|end_of_text|>
```

Listing 1: **Plain Format**, {text} is the text associated with the image

```
<|begin_of_text|><|start_header_id|>user<|end_header_id|>

{user_message}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

{system_message}<|eot_id|>
```

Listing 2: **LLaMA 3 Format**, {user_message} is the message sent by the user, while {system_message} is the model response.

Parameter	**Training Step**			
	MLP pre-train	Italian pre-train	Italian instruction-tuning	Italian long instruction-tuning
batch size	256	128	128	128
lr	1e-3	1e-5	1e-5	1e-5
vision tower lr	-	2e-6	2e-6	2e-6
lr schedule	cosine	cosine	cosine	cosine
lr warmup ratio	0.03	0.03	0.03	0.03
weight decay	0	0	0	0
epochs	1	1	1	500 steps
optimizer	AdamW	AdamW	AdamW	AdamW
max length	8192	8192	8192	8192
DeepSpeed stage	3	3	3	3

Table 3
Hyperparameters used during each training step

To evaluate the models, we distinguish between two different benchmarks:

- **Machine-translated state-of-the-art benchmarks**: we use some of the most popular benchmarks for evaluation of LVLMs translated to the Italian language;
- **Natively Italian benchmarks**: we use benchmarks that include Italian text-vision data instances where the text is originally written in Italian.

For evaluation, we use *lmms-eval*[3] [44] a fork of *lm-eval-harness*[4], a library for evaluation of LLMs, but designed for LVLMs. We create custom tasks to evaluate the models on Italian datasets.

The first set of benchmarks allows us to have somewhat comparable conceptual coverage compared to the state-of-the-art since the datasets that we consider cover the diverse skills of the models. We provide an overview of the tasks alongside their cardinality in Table 4.

Instead, the second set of benchmarks allows us to understand if training on machine-translated data severely affects performance. This is because these datasets are natively in the Italian language. For this purpose we use the test sets of the previously presented MTVQA and V-EXAMS datasets, keeping only the Italian instances of these multilingual datasets.

[3] https://github.com/EvolvingLMMs-Lab/lmms-eval
[4] https://github.com/EleutherAI/lm-evaluation-harness

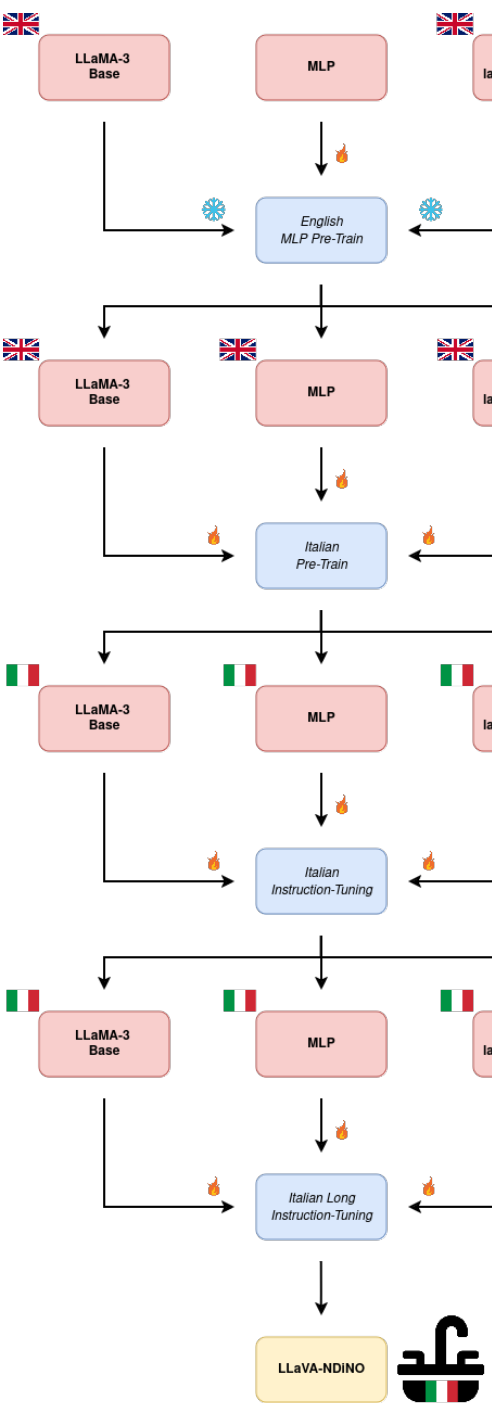

LLaMA-3 Base
MLP
CLIP ViT large-patch14-336
English MLP Pre-Train
LLaMA-3 Base
MLP
CLIP ViT large-patch14-336
Italian Pre-Train
LLaMA-3 Base
MLP
CLIP ViT large-patch14-336
Italian Instruction-Tuning
LLaMA-3 Base
MLP
CLIP ViT large-patch14-336
Italian Long Instruction-Tuning
LLaVA-NDiNO

Figure 1: Overview of the training pipeline, using LLaMA 3 base as LLM and CLIP ViT as vision encoder. There are four total steps: *English MLP Pre-Train*, *Italian Pre-Train*, *Italian Instruction-Tuning* and *Italian Long Instruction-Tuning*. In this figure, all steps of the pipeline are applied.

To understand if our trained models excel in the Italian language, we compare our results with the mBlip T0 [8] model, a multilingual vision-language model which includes Italian as one of the training languages. For the evaluation metrics, in all cases we use *exact match* for open-ended tasks and *accuracy* for closed-ended ones. The only exception is POPE for which we report the *F1* score. All metrics reflect common best practises used for the original datasets in the English language. We followed the same evaluation design for MTVQA and V-EXAMS as well.

Analyzing the results, both our models perform better w.r.t. the baseline in all tasks. Remarkably, while the mBlip model performs very poorly on the MTVQA dataset, both our models show improvements.

Dataset	# Original	# IT MT	Description
GQA-IT [39, 40]	12,578	-	Open-ended VQA dataset regarding compositional questions of real-world images, specifically regarding objects, attributes and relations in the images.
OK-VQA [41]	5,050	5,046	Open-ended VQA dataset regarding questions where the model needs to have external knowledge in order to answer.
SeedBench [42]	18,000	2,496	Closed-ended VQA multiple-choice dataset regarding temporal and spatial questions.
POPE [43]	9,000	9,000	Open-ended VQA dataset regarding object hallucination (answer is expected to be either 'Yes' or 'No').
LLaVA-Bench [5]	60	60	Open-ended VQA dataset to test the abilities of the models in solving challenging tasks, thanks to a highly-detailed and manually-curated description and a proper selection of questions for each instance.

Table 4
Overview of all datasets machine translated to the Italian language used for evaluation. # **Original** and # **IT MT** are the number of instances in the original dataset and in the machine-translated one respectively. For GQA-IT we report the original cardinality

Model	Datasets			
	GQA-IT* ↑	OK-VQA-IT ↑	SeedBench-IT ↑	POPE-IT* ↑
mBlip T0 XL [8]	0.13	0.13	0.51	0.49
LLaVA-NDiNO IT	0.27	**0.19**	0.67	0.84
LLaVA-NDiNO PT + IT	**0.28**	**0.19**	**0.68**	**0.86**

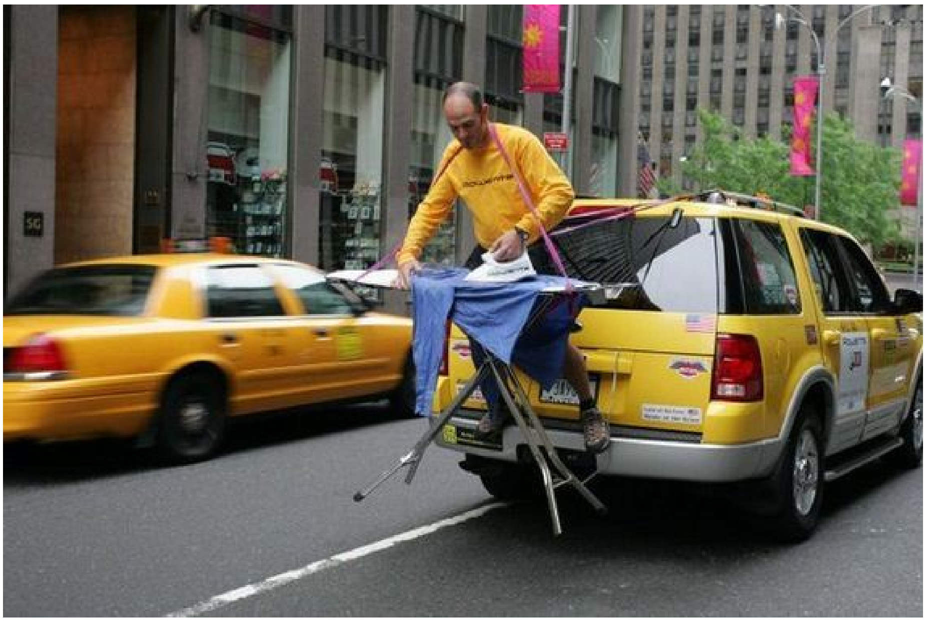
Table 5
Results obtained for evaluation datasets machine translated to the Italian language. <DATASET_NAME>-IT refers to the machine translated version of the original dataset. For GQA-IT, OK-VQA-IT and SeedBench-IT the metric is *exact match*, for POPE-IT the metric is Accuracy. The ↑ indicates that the greater value obtained for the metric of that dataset the better the performance. The asterisk indicates that there is statistical significance between the two LLaVA-NDiNO model results for that dataset

However, for both LLAVA-NDiNO models, average results are fairly similar regardless of the pre-training step. In light of this, we perform statistical testing using **McNemar's test**. The test reveals that for most tasks, the p-value is greater than 0.05; therefore, there are no discernible differences between the two setups. We believe this is due to the nature of the evaluation tasks, since the model only needs to pick the correct option or to generate a simple word or phrase. These tasks are not useful for evaluating the quality of the pre-train. In light of this, we will perform an additional experiment to assess the models' performance on longer and richer textual descriptions.

4.1.2. Instruction-tuning and Evaluation for Long Output Generation

For this step, we further train our models for long response generation. Specifically, we use data taken from LLaVA Conversation 58k extracting user question and system answer pairs to use as single-round interactions. After extracting the single-round instances, we perform training following the same procedure used for instruction-tuning.

We perform four different training procedures:



Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

Short Answer Question: Quante persone ci sono in questa immagine? Rispondi brevemente.
English Translation: How many people are there in the image? Answer briefly.

LLaVA-NDiNO PT + IT Answer: 1.
English Translation: 1.

LLaVA-NDiNO PT + IT + LONG-IT Answer: C'è una persona in questa immagine.
English Translation: There is one person in this image.

Long Answer Question: Cosa c'è di strano in questa immagine?
English Translation: What is strange about this image?

LLaVA-NDiNO PT + IT Answer: Un uomo è seduto su una sedia a rotelle che lava i panni.
English Translation: A man is sitting in a wheelchair washing clothes.

LLaVA-NDiNO PT + IT + LONG-IT Answer: L'immagine è strana perché mostra un uomo che asciuga le camicie mentre è in piedi sulla parte superiore di un camion giallo, che è un modo insolito e non convenzionale per asciugare le camicie.
English Translation: The image is strange because it shows a man drying shirts while standing on top of a yellow truck, which is an unusual and unconventional way to dry shirts.

Figure 2: Example comparing the answers of two different models to two different questions.

- **LLaVA-NDiNO LONG-IT**: only MLP pre-training and long instruction-tuning have been performed;
- **LLaVA-NDiNO PT + LONG-IT**: MLP pre-training, Italian language pre-training and long Italian language instruction-tuning have been performed;
- **LLaVA-NDiNO IT + LONG-IT**: MLP pre-training, Italian language instruction-tuning and long Italian language instruction-tuning have been performed;
- **LLaVA-NDiNO PT + IT + LONG-IT**: MLP pre-training, Italian language pre-training, Italian language instruction-tuning and long Italian language instruction-tuning have been performed.

Model	Datasets	
	MTVQA-IT ↑	V-EXAMS-IT ↑
mBlip T0 XL [8]	0.04	0.20
LLaVA-NDiNO IT	0.15	**0.25**
LLaVA-NDiNO PT + IT	**0.17**	0.24

Table 6

Results obtained for evaluation datasets natively in Italian language. <DATASET_NAME>-IT refers to the filtered version of the original multilingual dataset containing only Italian instances. For both MTVQA and V-EXAMS the metric is *exact match*. The ↑ indicates that the greater value obtained for the metric of that dataset the better the performance

Model	Datasets	
	LLaVA-Bench-IT ↓	MTVQA-IT ↓
LLaVA-NDiNO IT	10.22	81.10
LLaVA-NDiNO PT + IT	9.53	**62.92**
LLaVA-NDiNO LONG-IT	4.49	138.68
LLaVA-NDiNO PT + LONG-IT	**4.29**	119.13
LLaVA-NDiNO IT + LONG-IT	4.91	121.77
LLaVA-NDiNO PT + IT + LONG-IT	4.76	107.91

Table 7

Results obtained for Perplexity evaluation of the models. <DATASET_NAME>-IT refers to the machine translated version of the original dataset for LLaVA-Bench and to the filtered version with only Italian instances for MTVQA-IT. ↓ indicates that the lesser value obtained for the metric of that dataset the better the performance. In cases with ◇, Perplexity was always greater than the fixed threshold.

To evaluate the quality of long output generation, we use both the LLaVA-Bench and the MTVQA datasets. LLaVA-Bench is selected for its inclusion of GPT-4V responses, allowing us to evaluate models on long and descriptive answers. Meanwhile, MTVQA is used to extend the previous evaluation on instruction-tuned models.

In this case, we use Perplexity as metric, to understand how certain a model is of the actual answer. The question-answer pairs of the datasets are formatted using the previously presented prompts LLaMA 3 instruct format. We compute the perplexity of the model on the expected answer only, but conditioned on the context of the question (that is, the loss is only computed on the answer tokens). Instances where the Perplexity exceeds 1,000 are treated as outliers and skipped. We expect models trained on multiple steps to have an overall lower degree of Perplexity. The results of this evaluation step, shown in Table 7, align with the expectations: models subjected to long instruction-tuning have better performance on LLaVA-Bench, while instruction-tuned models perform better on MTVQA. Furthermore, while in the previous evaluation step there were no significant differences on the MTVQA dataset, we can assess in these results that the instruction-tuned models have learned a different language distribution. This is important since using a generation strategy different from greedy decoding can lead to notably different outputs.

Finally, we showcase two different examples to further illustrate the difference between models trained on long output generation and others. In Figure 2, we compare two of our models on answering two different questions (one expecting a short answer while the other a long one) for the same image.

5. Conclusions

We introduce and release a family of LMMs trained for the Italian language. Specifically, we train the models considering three different possible steps: *Italian adaptation*, *Italian instruction-tuning* and *Italian instruction-tuning for long responses*. To train the models, we collect a large collection of state-of-the-art datasets for the English language. Specifically, The Cauldron and LLaVA Conversation 58k for instruction-tuning and GQA, OK-VQA, SeedBench, POPE and LLaVA-Bench for evaluation. These datasets are then translated using MADLAD, one of the most recent neural machine translation models. We also collect natively Italian data to boost the quality of both training and evaluation. Specifically, we collect MTVQA and V-EXAMS for both instruction-tuning and evaluation, as well as a rich pre-training corpus consisting of image-text pairs from WiT and MultiEURLEX.

We train several models on different possible configurations, that is multiple train steps using different datasets. An extensive evaluation procedure compared our results with a popular multilingual and multimodal model that is, mBlip. Results are promising against the baseline, but we noticed that for most tasks there were no significant differences on the results of the instruction-tuned models. However, we find relevant differences when evaluating the models using Perplexity.

As future works, we plan to investigate the performance difference between a model instruction-tuned for both short and long answer generation in Italian at the same time w.r.t. proposed pipeline. We also aim to study conversational multi-round multimodal models since, in this work, we focused on single-round conversations.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU. Models are built on the Leonardo supercomputer with the support of CINECA-Italian Super Computing Resource Allocation, class C project IscrC_LLMM (HP10CLKWTP).

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] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 26296–26306.

[3] H. Touvron, et al., Llama 2: Open foundation and fine-tuned chat models, 2023. URL: https://arxiv.org/abs/2307.09288. arXiv:2307.09288.

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

[5] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, Advances in neural information processing systems 36 (2024).

[6] G. Sharir, A. Noy, L. Zelnik-Manor, An image is worth 16x16 words, what is a video worth?, arXiv preprint arXiv:2103.13915 (2021).

[7] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, C. Li, Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, arXiv preprint arXiv:2407.07895 (2024).

[8] G. Geigle, A. Jain, R. Timofte, G. Glavaš, mBLIP: Efficient bootstrapping of multilingual vision-LLMs, in: J. Gu, T.-J. R. Fu, D. Hudson, A. Celikyilmaz, W. Wang (Eds.), Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7–25. URL: https://aclanthology.org/2024.alvr-1.2.

[9] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, PMLR, 2023, pp. 19730–19742.

[10] N. Team, et al., No language left behind: Scaling human-centered machine translation, 2022. URL: https://arxiv.org/abs/2207.04672. arXiv:2207.04672.

[11] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay, et al., Pali-x: On scaling up a multilingual vision and language model, arXiv preprint arXiv:2305.18565 (2023).

[12] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, et al., Pali: A jointly-scaled multilingual language-image model, arXiv preprint arXiv:2209.06794 (2022).

[13] D. Shin, H. Lim, I. Won, C. Choi, M. Kim, S. Song, H. Yoo, S. Kim, K. Lim, X-LLaVA: Optimizing bilingual large vision-language alignment, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2463–2473. URL: https://aclanthology.org/2024.findings-naacl.158. doi:10.18653/v1/2024.findings-naacl.158.

[14] A. Conneau, G. Lample, Cross-lingual language model pretraining, Advances in neural information processing systems 32 (2019).

[15] H. Laurençon, L. Tronchon, M. Cord, V. Sanh, What matters when building vision-language models?, 2024. arXiv:2405.02246.

[16] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, M. Najork, Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning, arXiv preprint arXiv:2103.01913 (2021).

[17] I. Chalkidis, M. Fergadiotis, I. Androutsopoulos, Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021. URL: https://arxiv.org/abs/2109.00904.

[18] R. Smith, D. Antonova, D.-S. Lee, Adapting the tesseract open source ocr engine for multilingual ocr., in: V. Govindaraju, P. Natarajan, S. Chaudhury, D. P. Lopresti (Eds.), MOCR '09: Proceedings of the International Workshop on Multilingual OCR, ACM International Conference Proceeding Series, ACM, 2009, pp. 1–8. URL: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/35248.pdf. doi:http://doi.acm.org/10/1145/1577802.1577804.

[19] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, R. Mottaghi, A-okvqa: A benchmark for visual question answering using world knowledge, in: European conference on computer vision, Springer, 2022, pp. 146–162.

[20] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, R. Girshick, Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, in: CVPR, 2017.

[21] M. Ren, R. Kiros, R. Zemel, Exploring models and data for image question answering, Advances in neural information processing systems 28 (2015).

[22] M. Kazemi, H. Alvari, A. Anand, J. Wu, X. Chen, R. Soricut, Geomverse: A systematic evaluation of large models for geometric reasoning, arXiv preprint arXiv:2312.12241 (2023).

[23] P. Lu, L. Qiu, J. Chen, T. Xia, Y. Zhao, W. Zhang, Z. Yu, X. Liang, S.-C. Zhu, Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning, arXiv preprint arXiv:2110.13214 (2021).

[24] P. Lu, R. Gong, S. Jiang, L. Qiu, S. Huang, X. Liang, S.-C. Zhu, Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning, in: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), 2021.

[25] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, V. Ferrari, Connecting vision and language with localized narratives, in: ECCV, 2020.

[26] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, Z. Liu, Mimic-it: Multi-modal in-context instruction tuning, 2023. URL: https://arxiv.org/abs/2306.05425. arXiv:2306.05425.

[27] A. Suhr, M. Lewis, J. Yeh, Y. Artzi, A corpus of natural language for visual reasoning, in: Annual Meeting of the Association for Computational Linguistics, 2017. URL: https://api.semanticscholar.org/CorpusID:19435386.

[28] C. Zhang, F. Gao, B. Jia, Y. Zhu, S.-C. Zhu, Raven: A dataset for relational and analogical visual reasoning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[29] H. Jhamtani, T. Berg-Kirkpatrick, Learning to describe differences between pairs of similar images, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.

[30] M. Acharya, K. Kafle, C. Kanan, Tallyqa: Answering complex counting questions, in: AAAI, 2019.

[31] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, Visual7W: Grounded Question Answering in Images, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[32] J. J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, Scientific data 5 (2018) 1–10.

[33] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, VQA: Visual Question Answering, in: International Conference on Computer Vision (ICCV), 2015.

[34] J. Tang, Q. Liu, Y. Ye, J. Lu, S. Wei, C. Lin, W. Li, M. F. F. B. Mahmood, H. Feng, Z. Zhao, Y. Wang, Y. Liu, H. Liu, X. Bai, C. Huang, Mtvqa: Benchmarking multilingual text-centric visual question answering, 2024. arXiv:2405.11985.

[35] R. J. Das, S. E. Hristov, H. Li, D. I. Dimitrov, I. Koychev, P. Nakov, Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, 2024. arXiv:2403.10378.

[36] S. Kudugunta, I. Caswell, B. Zhang, X. Garcia, D. Xin, A. Kusupati, R. Stella, A. Bapna, O. Firat, Madlad-400: A multilingual and document-level large audited dataset, Advances in Neural Information Processing Systems 36 (2024).

[37] A. Dubey, et Al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[38] 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.

[39] D. Croce, L. C. Passaro, A. Lenci, R. Basili, Gqa-it: Italian question answering on image scene graphs, Computational Linguistics CliC-it 2021 (2022) 92.

[40] D. A. Hudson, C. D. Manning, Gqa: A new dataset for real-world visual reasoning and compositional question answering, Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[41] K. Marino, M. Rastegari, A. Farhadi, R. Mottaghi, Ok-vqa: A visual question answering benchmark requiring external knowledge, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[42] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, Y. Shan, Seed-bench: Benchmarking multimodal llms with generative comprehension, arXiv preprint arXiv:2307.16125 (2023).

[43] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, J.-R. Wen, Evaluating object hallucination in large vision-language models, in: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023. URL: https://openreview.net/forum?id=xozJw0kZXF.

[44] B. Li, P. Zhang, K. Zhang, F. Pu, X. Du, Y. Dong, H. Liu, Y. Zhang, G. Zhang, C. Li, Z. Liu, Lmms-eval: Accelerating the development of large multimoal models, 2024. URL: https://github.com/EvolvingLMMs-Lab/lmms-eval.