Breaking Language Barriers in Visual Language Models via Multilingual Textual Regularization

Iñigo Pikabea[*1,2], **Iñaki Lacunza**[*1], **Oriol Pareras**[*1],
Carlos Escolano[1,2], **Aitor Gonzalez-Agirre**[1], **Javier Hernando**[1,2], **Marta Villegas**[1],
[1] Barcelona Supercomputing Center, [2] Universitat Politècnica de Catalunya
Correspondence: {inigo.pikabea,inaki.lacunza,oriol.pareras}@bsc.es [*]Core contributors

Abstract

Rapid advancements in Visual Language Models (VLMs) have transformed multimodal understanding but are often constrained by generating English responses regardless of the input language. This phenomenon has been termed as Image-induced Fidelity Loss (IFL) and stems from limited multimodal multilingual training data. To address this, we propose a continuous multilingual integration strategy that injects text-only multilingual data during visual instruction tuning, preserving the language model's original multilingual capabilities. Extensive evaluations demonstrate that our approach significantly improves linguistic fidelity across languages without degradation in visual performance. We also explore model merging, which improves language fidelity but comes at the cost of visual performance. In contrast, our core method achieves robust multilingual alignment without trade-offs, offering a scalable and effective path to mitigating IFL for global VLM adoption.

1 Introduction

Large Language Models (LLMs) have significantly advanced multimodal understanding, leading to the rise of VLMs, which integrate vision encoders into LLM backbones. A widely adopted paradigm is the LLaVA-style architecture (Liu et al., 2023b, 2024a), where a decoder-only LLM is coupled with a vision encoder and an adapter module to align visual representations with textual embeddings.

Despite their success, VLMs exhibit a strong bias toward English due to the predominance of monolingual vision-language training data. Consequently, they often generate English responses regardless of the input language, a phenomenon termed Image-induced Fidelity Loss (or IFL) (Hinck et al., 2024). This issue stems from limitations in the underlying LLM rather than the visual representations.

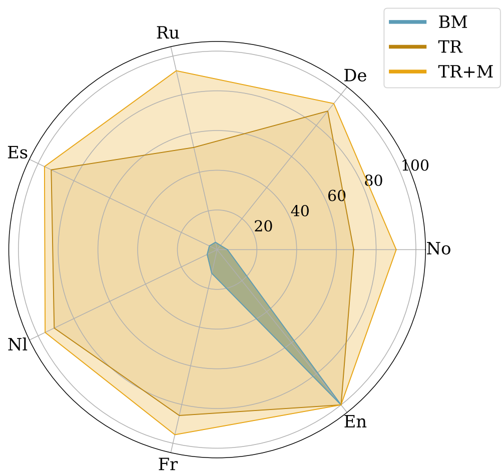


Figure 1: Language Fidelity (LF) accuracy on Crossmodal-3600. (BM: Base Model, TR: model trained with multilingual Textual Regularization, TR+M: TR and merging the final model with the original LLM Backbone)

Ensuring multilingual capability in VLMs is essential for their adoption across diverse linguistic communities, as reliance on English-centric outputs risk erasing cultural and linguistic nuances. Prior work (Qiu et al., 2022; Li et al., 2023b) has explored dataset translation, but this approach incurs high computational costs and introduces translation errors, especially in images with language-dependent elements.

In this paper, we propose an alternative solution by integrating multilingual text-only data during the visual instruction tuning process. Additionally, we explore model merging, combining the visually fine-tuned model with the original multilingual backbone LLM to further preserve linguistic fidelity. As shown in Figure 1, our method effectively prevents the model from defaulting to English in non-English queries. To the best of our knowledge, no previous work has achieved full multilingual competence in VLMs through such a simple and scalable approach.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 299–337
December 20-24, 2025 ©2025 Association for Computational Linguistics

Our contributions are as follows:

- We systematically demonstrate that integrating multilingual text-only data during training significantly reduces IFL bias in LLaVA-style VLMs while maintaining core capabilities.

- We conduct an extensive analysis on the optimal proportion of text-only data required for effective multilingual adaptation.

- We explore a model merging strategy, combining the visually fine-tuned model with the original multilingual backbone LLM, and assess its impact on preserving linguistic fidelity.

Our findings suggest that we can develop high-quality multilingual VLMs that maintain strong performance across multiple languages in a simple and scalable way. By avoiding the need to translate or construct multimodal datasets for each language, our approach lowers the entry barrier for multilingual VLM development. This makes it especially attractive for low-resource settings, where monolingual text is often available but collecting vision-language data is costly or impractical.

2 Related Work

2.1 Multimodal Large Language Models

VLMs typically integrate an image encoder, usually CLIP (Radford et al., 2021; Dosovitskiy et al., 2021), with an LLM backbone. Various strategies exist for combining these components. The predominant approach follows a decoder-only architecture, as seen in the LLaVA series, where an adapter module projects visual representations into the textual embedding space. Other methods include cross-attention mechanisms (Grattafiori et al., 2024), and some models, like NVLM (Dai et al., 2024), adopt a hybrid strategy combining both approaches.

LLaVA-style models tend to default to English due to the scarcity of multimodal training data in other languages (Hinck et al., 2024). This issue arises because the LLM's parameters are updated for a distinct task, which can disrupt its original language capabilities. Llama 3 (Grattafiori et al., 2024) takes a different approach by freezing the LLM during training, which helps preserve its pretrained abilities while incorporating visual information. However, freezing the LLM also limits the model's capacity to learn new visual tasks, creating a trade-off between language preservation and multimodal learning.

2.2 Multilingual Multimodal Learning

A widely adopted approach to improving multilinguality in VLMs is translating existing multimodal datasets. Several works (Song et al., 2024; Hu et al., 2024) have analyzed this strategy and proposed methods to enhance its effectiveness. Several models, such as PALI (Chen et al., 2023), PALI-X (Chen et al., 2024d), mBLIP (Geigle et al., 2024), PALO (Maaz et al., 2024) and Pangea (Yue et al., 2025), have pursued this approach. However, this strategy presents challenges, including computational overhead, translation inconsistencies, and the loss of cultural context in visual-text pairs.

Moreover, recent research (Aggarwal et al., 2024) suggests that continual fine-tuning can harm an LLM's performance. When a model undergoes two consecutive fine-tuning phases with differing task distributions, its ability to perform earlier tasks deteriorates. This raises concerns that direct fine-tuning solely on translated multimodal data may degrade the LLM's original capabilities.

2.3 Catastrophic Forgetting Prevention

In the context of LLMs, the problem of maintaining performance across tasks while integrating new information is known as lifelong learning. This field focuses on a system's ability to acquire, integrate, and retain knowledge without catastrophically forgetting previous information. Visual Instruction Tuning is a case of lifelong learning, and it faces the same challenges. One known mitigation strategy is episodic or experience replay (Zheng et al., 2025), which helps prevent catastrophic forgetting by reintroducing previously learned information.

Several studies (Liu et al., 2022; Ibrahim et al., 2024) have explored ways to incorporate pretraining data during fine-tuning. Bethune et al. (2025) further analyze the impact of this approach and suggest that even a small amount of pretraining data can help retain previously learned knowledge, reducing the risk of performance degradation.

In the case of VLMs, NVLM (Dai et al., 2024) and InternVL 2.5 (Chen et al., 2024e) demonstrate that incorporating high-quality text-only data during Visual Instruction Tuning, not only improves the overall text-generation capabilities, but also multimodal performance. Our approach builds upon these findings by integrating multilingual text-only data throughout VLM training to mitigate IFL, without requiring extensive multimodal multilingual data collection.

2.4 Model Merging

Model merging is a technique that involves combining two or more pre-trained models to create a new model that leverages the strengths of each. By merging a fine-tuned model with its original backbone, this process preserves the model's prior capabilities while incorporating additional refinements from further training. This strategy has been applied in various contexts, such as language transfer, where Alexandrov et al. (2024) demonstrate that model merging facilitates fine-tuning for new linguistic capabilities without compromising the performance of the original LLM.

Building on this insight, we explore model merging as a means of preserving the multilingual competencies of a VLM during the visual fine-tuning process. We adopt the same model merging strategy as Aya Vision (Dash et al., 2025), which has shown strong empirical results, and combine it with our multilingual textual regularization strategy.

3 Experimental Setup

3.1 Data

Our training framework combines multimodal visual-language data from LLaVA-OneVision (Li et al., 2025) with multilingual text-only instruction data from the Salamandra family of models (Gonzalez-Agirre et al., 2025). This hybrid approach ensures robust visual understanding while addressing IFL through explicit multilingual text supervision. All datasets are documented in Appendix A.

Visual Data We employ LLaVA-OneVision's English-only visual pipeline, consisting of a total number of 9,286,732, which is divided into two main groups:

- **General and Detailed Image Captions**: This dataset comprises both basic and highly detailed image captions. The basic captions align the visual embedding space with the LLM's embedding space, while the detailed captions refine the mapping between the two providing a high-quality understanding of the images. This group comprises $4.4M$ unique instances.
- **Task-Specific, Multi-Image, and Video Data**: This dataset is used to instruct the aligned model on specific tasks, including Optical Character Recognition (OCR), infographic understanding, and math & reasoning. Additionally, multi-image and video data are incorporated to enhance the model's ability to interpret diverse visual inputs. This group comprises $4.9M$ unique instances.

Multilingual Text-Only Data To further enhance the model's multilingual proficiency, we incorporate 315,496 text-only samples drawn from 11 diverse datasets covering domains such as general language tasks, multilingual instructions, conversational QA, and code annotations. These sources include human-annotated datasets (Rajani et al., 2023), multilingual instruction collections (Singh et al., 2024; Costa-jussà et al., 2024) and conversational data (Conover et al., 2023; Köpf et al., 2023).

Notably, the text-only samples cover 21 of the 35 languages used in training Salamandra, ensuring extensive linguistic representation. A significant portion of this dataset is machine-translation data.

Although most of the text-only data is in English, matching the language of the visual data, its inclusion remains important. This alignment reinforces the model's linguistic foundation and facilitates the integration of multilingual supervision, ultimately ensuring balanced performance across modalities. The final distribution of languages in the text-only data, complementing LLaVA-OneVision's training data, is shown in Figure 2.

Evaluation Data For evaluation, we use both monolingual and multilingual multimodal datasets. To assess visual performance, we include AI2D (Kembhavi et al., 2016), which tests understanding of diagram-based questions; RealWorldQA[1], a real-world image dataset with open-ended and multiple-choice questions; MMMU (Yue et al., 2024), a diverse multimodal reasoning benchmark; and MMStar (Chen et al., 2024b), which aggregates vision-language tasks for broad multimodal evaluation.

The first two benchmarks primarily assess the exact match accuracy, quantifying the proportion of responses that exactly match the predefined ground truth. These targets are typically short-form text

[1]https://huggingface.co/datasets/visheratin/realworldqa

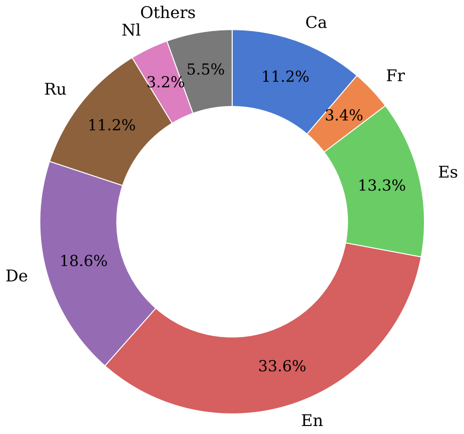


Figure 2: Distribution of the multilingual text-only data used for Textual Regularization. Languages with a volume smaller than 3% are grouped under *Others*, which collectively account for 5.5% of the data. The most frequent languages in this group are Portuguese (2.1%), Italian (0.7%), Polish (0.47%), Swedish (0.42%), Irish (0.39%), Lithuanian (0.29%), Galician (0.22%), Greek (0.20%), and Ukrainian (0.17%).

or multiple-choice answers. On the other hand, MMMU and MMStar are classification tasks that are measured using accuracy.

For multilingual multimodal performance, we have selected Crossmodal-3600 (Thapliyal et al., 2022), a geographically diverse multilingual multimodal dataset for image captioning. It is particularly well-suited for our experiments as it covers the highest number of overlapping languages with Salamandra while allowing for image captioning with multiple reference targets per instance. The dataset comprises approximately 3,600 samples in 36 languages, from which we evaluate on German (De), Russian (Ru), Spanish (Es), Dutch (Du), French (Fr), and English (En). Appendix C discusses the prompts used for caption generation during evaluation and the rationale behind their selection.

3.2 Model Framework

Our implementation also follows the LLaVA-OneVision framework [2], adapted for enhanced multilingual capabilities. The architecture consists of three main components: `SigLIP-SO400M` [3] (Zhai et al., 2023) as the visual encoder, a 2-layer MLP projection module (with GELU activation functions (Hendrycks and Gimpel, 2016)), and `Salamandra-7b-instruct` [4] (Gonzalez-Agirre et al., 2025) as the backbone LLM. Salamandra has been chosen for its high multilinguality, as it has been trained with 35 European languages.

We also adopt LLaVA-OneVision's curriculum learning training strategy, which progresses through four distinct stages:

- **Stage 1 (Language-Image Alignment)**: In this initial phase, only the MLP projector is trained, while both the visual encoder and LLM remain frozen. General image captions are employed to establish basic cross-modal connections.
- **Stage 1.5 (Full Model Training)**: At this stage, all model components are unfrozen to enable end-to-end training. A high-quality set of detailed image captions is used in conjunction with an increased image resolution to enhance visual detail processing.
- **Stage 2 (Single-Image Instruction Tuning)**: Once the model has achieved a deep understanding of images, it is fine-tuned for a diverse set of visual tasks. The image resolution is further increased to support fine-grained visual analysis.
- **Stage 2.5 (Multi-Image and Video Training)**: In the final stage, multi-image and video data are incorporated to enable reasoning across multiple visual inputs. Additionally, single-image data from the previous stage is also utilized.

The key innovation in our approach lies in the strategic injection of multilingual text-only data throughout these training stages detailed in §3.4. After the Visual Instruction Tuning, the model is merged with the baseline LLM weights using linear interpolation.

3.3 Metrics

To evaluate language fidelity and consistency, we employ a common metric established in prior multilingual multimodal evaluation work (Hinck et al., 2024; Schneider and Sitaram, 2024):

Language Fidelity We use GlotLID (Kargaran et al., 2023) to obtain the accuracy of whether the language of the generated captions over

[2] https://github.com/LLaVA-VL/LLaVA-NeXT
[3] https://huggingface.co/google/siglip-so400m-patch14-384
[4] https://huggingface.co/BSC-LT/salamandra-7b-instruct

Crossmodal-3600 images is the same as the user prompt [5]. We named this metric LF, and we observed that, in many cases, it considered as correct samples that had single words in English, or with minor code-switching errors. To address this issue, we extend this metric (LF+) by using `Llama-3.1-8B-Instruct`[6] (Grattafiori et al., 2024) as an LLM-as-a-judge, evaluating if the samples already classified by GlotLID are entirely in the same language or not. Nevertheless, due to a majority voting strategy in its implementation (see Appendix B), the LLM-as-a-judge does not work perfectly, as it sometimes misclassifies correct samples as non-consistent in language. For this reason, this metric can be interpreted as a statistical lower bound of language fidelity.

Visual Performance To evaluate visual performance, we use the English-only multimodal benchmarks detailed in Section 3.1: AI2D, RealWorldQA, MMMU, and MMStar. AI2D and RealWorldQA are evaluated using exact match accuracy, measuring the proportion of responses identical to ground-truth answers (typically short text or multiple-choice). MMMU and MMStar, however, are treated as classification tasks and evaluated via accuracy.

To evaluate multilingual multimodal performance, we also use the same approach used in Hinck et al. (2024); Schneider and Sitaram (2024), and evaluate the captioning quality across different languages with chrF++ (Popović, 2016, 2017) over Crossmodal-3600 samples.

Further discussion on metric selection can be found in Appendix D.

3.4 Experiments

We focus on testing different text-only integration strategies, analyzing the influence of data quantity, examining generalization capabilities on languages not contained during textual regularization, and assessing the effect of model merging on overall performance. To be able to quantify the results obtained with these experiments, we also trained a baseline model (BM) by only conducting the Visual Instruction Tuning, without textual regularization.

[5] Crossmodal-3600 does not include a predefined reference generation prompt. For completeness, we present and explain the employed prompt in Appendix C.

[6] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

Multilingual Data Integration Strategies We explore three distinct strategies for incorporating the text-only multilingual data (315,496 instances) during the visual instruction tuning process:

- **Textual Regularization across Three Stages (TR-3S)**: Multilingual text data was distributed proportionally across the final three training stages (1.5, 2, and 2.5).
- **Textual Regularization across Two Stages (TR-2S)**: Multilingual text data was integrated proportionally only in the last two stages (2 and 2.5).
- **Textual Regularization at a Single Stage (TR-1S)**: Multilingual text data was added exclusively during the final stage (2.5).

Multilingual Generalization Capabilities To investigate whether regularization with multilingual text data extends to languages not explicitly seen during training, we train a variant of the TR-3S model where German was excluded from the multilingual text dataset. This experiment allows us to evaluate the model's generalization ability to new languages.

Influence of Data Balance We vary the proportion of multilingual text data used for regularization alongside visual–text pairs. Starting from the text-only, multilingual instruction-tuning subset of $315k$ text-only instances, we inject this data following the TR-3S approach. Across these three stages, the total number of visual–text pairs is $\approx 9M$, of which the $315k$ text-only examples account for approximately 3.7% of the entire set. We denote $2x$, $0.5x$, and $0.25x$ as using twice, half, and one quarter of the amount of text-only data, respectively, where x corresponds to the original dataset used in the TR-3S configuration.

Model Merging To explore the potential of further enhancing the multilingual capabilities of our best-performing model (TR-3S), we apply model merging. To do so, we perform a linear interpolation between the weights of the visually instructed model with those of the backbone LLM, maintaining the encoder and MLP layers. This allows us to evaluate whether model merging could combine the model's visual understanding capabilities with the language fidelity of the original model.

As explained in §3.3, we evaluate multimodal performance on a suite of English benchmarks and extend it to multiple languages evaluating chrF++ on Crossmodal-3600. Moreover, we use LF to assess IFL and further analyze its bounds with LF+.

3.5 Implementation Details

Our experiments were conducted on custom NVIDIA H100 GPUs, each with 64GB of memory. We trained each model for 6 days in a distributed setup with 8 nodes, each containing 4 GPUs, totaling 32 GPUs per experiment. As we trained 8 models (excluding the merged model, which did not require separate training), the total compute usage amounted to 36,864 GPU hours.

For evaluation, we assessed 9 models across 6 languages, with each requiring one node for 24 hours, resulting in 5,184 GPU hours.

The training hyperparameters were largely based on those used in LLaVA-OneVision and Salamandra's Instruction Tuning, ensuring consistency with prior work. Further details on the training process, including specific hyperparameters and configurations, can be found in Appendix E.

4 Results

This section presents the outcomes of our experimental investigation into the effectiveness of incorporating multilingual text-only data during the visual instruction tuning process for reducing IFL in VLMs.

4.1 Quantifying the Baseline English Bias

To better understand the starting point of our investigation, we first evaluated the baseline model (BM), trained exclusively on English visual instruction data. As anticipated, this model exhibits a pronounced English-centric behavior, responding predominantly in English even when prompted in other languages. This confirms the strong presence of IFL and underscores the necessity of multilingual regularization.

As shown in Table 1, the model demonstrates a very limited capacity to generate non-English responses. Languages such as German and Spanish, for example, show particularly low consistency, often defaulting back to English. This behavior reveals how the training process strongly anchors the model to English due to the lack of multilingual signals.

Interpreting these results in context, the baseline model's bias highlights a fundamental limitation of current VLM training pipelines, where even models based on multilingual backbones revert to English if not explicitly trained with multilingual supervision.

4.2 Impact of Multilingual Text-Only Data Integration Strategies

The results presented in Table 1 clearly demonstrate that integrating multilingual text-only data substantially mitigates English bias across all evaluated strategies. Notably, proportional integration across the final three training stages (TR-3S) consistently achieves superior LF scores for most non-English languages. This suggests that continuous exposure to multilingual text throughout training stages is most effective in maintaining linguistic fidelity. The strategy of introducing multilingual data exclusively in the final stage (TR-1S) yields the least improvement, indicating that delaying multilingual exposure is insufficient to counteract the English bias ingrained during earlier training phases. The stronger performance of TR-3S can be attributed to its role as a continual regularizer. By more extensively interleaving multilingual text-only data, the model consistently reinforces its multilingual representations, thus more effectively preserving previously acquired language capabilities and reducing IFL.

Lang.	BM	TR-3S	TR-2S	TR-1S
De	2.7	**88.7**	81.3	24.5
Es	4.4	**92.9**	65.4	38.4
Fr	12.2	**85.7**	74.9	29.9
Nl	5.6	**91.8**	91.3	49.2
Ru	3.8	**52.9**	24.8	50.9
En	**100.0**	**100.0**	**100.0**	**100.0**

Table 1: LF accuracy for different integration strategies. The best results are shown in bold.

4.3 Evaluating Multilingual Generalization Capabilities

We obtained a LF scores of 5.4% for German in this scenario. While these results are slightly above the English-biased baseline (2.7% LF), the performance remains very limited. This suggests that the multilingual regularization approach, in the absence of explicit exposure to the target language, does not meaningfully help mitigate IFL. In other words, the model struggles to generalize to unseen languages, and explicit inclusion during training appears necessary for achieving satisfactory multilingual fidelity.

4.4 Analyzing the Influence of Data Quantity

The LF score for these variations are presented on Table 2.

Lang.	$0.25x$	$0.5x$	x	$2x$
De	85.0	**88.9**	88.7	73.3
Es	91.9	92.4	**92.9**	76.4
Fr	**88.6**	83.5	85.7	69.5
Nl	69.8	**96.0**	91.8	93.5
Ru	91.5	81.9	52.9	**93.3**

Table 2: LF accuracy under varying quantities of text-only to visual data. All this variants are trained with the TR-3S approach. The best results are shown in bold.

The results demonstrate a complex relationship between the text-only to visual data quantity and language fidelity, making straightforward interpretation challenging. Increasing the quantity by changing the configuration from $0.25x$ to $0.5x$ generally improves LF scores for most languages, suggesting a positive impact of increased text-only data within this range. However, further increasing the relative quantity of text-only data switching to the $2x$ configuration ($\approx$7% of the total data) does not consistently yield better results and, in some cases, significantly reduces performance, particularly for German, Spanish, and French.

Notably, none of the tested variations drastically degrade LF across all languages compared to the baseline x configuration (3.7% of text-only data from the total amount). This indicates that while the optimal text-only data ratio requires careful consideration, moderate variations around the original amount do not necessarily lead to a substantial loss in language fidelity.

4.5 Evaluating the Effect of Model Merging

	TR-3S (%)	TR-3S + M (%)
De	88.7	**94.1**
Es	92.9	**96.4**
Fr	85.7	**95.5**
Nl	91.8	**96.1**
Ru	52.9	**92.4**

Table 3: LF score on multiple languages before (TR-3S) and after (TR-3S + M) model merging. The best results are shown in bold.

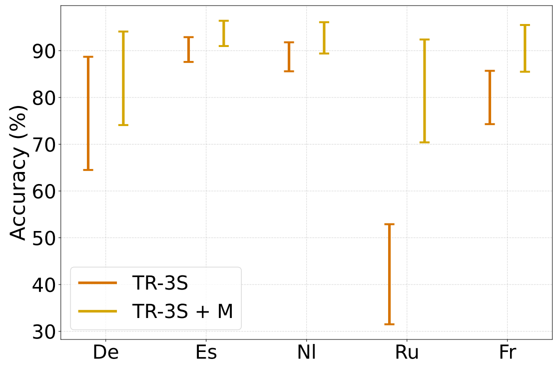


Figure 3: Interval Plot contrasting LF (upper bars) vs. LF+ (lower bars) across languages of our best-performing models.

The results presented in Table 3 demonstrate a clear positive impact of model merging on language fidelity (LF). On average, LF improved by over 12.5 points across the five languages (5.75 points removing Russian). This improvement is consistent across all evaluated languages, indicating a robust effect of the merging strategy in enhancing multilingual fidelity, rather than a language-specific anomaly. Importantly, this analysis focuses solely on language fidelity. The impact of model merging on other VLM capabilities will be discussed in §4.7. These findings support the hypothesis that model merging can be a valuable tool in mitigating IFL.

4.6 Bounding IFL

Figure 3 represents the upper (LF) and lower bounds (LF+) of IFL of our best-performing models. As it can be seen, in most languages we can observe a span of approximately 10% or less, except for De and Ru, which is around 20%. We attribute this difference to the lower performance of the LLM-as-a-judge on these languages (detailed in Appendix B). Nonetheless, it is important to note that this is not a statistical representation, so the interval width is not significant in terms of performance.

The results confirm the validity of our method, as all the lower bounds (excluding TR-3S on Russian) surpass the 65% accuracy.

4.7 Impact on General Tasks Performance

Our analysis shown in Table 4 reveals that the models trained with our proposed regularization techniques, specifically those utilizing proportional multilingual text-only data integration, generally maintain or slightly improve performance on stan-

Model	AI2D$_{EM}$	RealWorldQA$_{EM}$	MMMU$_{Acc}$ (val)	MMStar$_{Acc}$ (avg)
BM	73.96	**56.99**	**34.22**	47.33
TR-3S	**75.39**	54.25	33.56	**48.87**
TR-3S + M	57.19	52.03	34.11	42.25

Table 4: Performance on general VLM benchmarks (only in English). All scores are reported on a 0–100 scale. The best results are shown in bold.

dard VLM benchmarks compared to the baseline English-centric model (BM). For instance, the TR-3S model, which incorporates text-only data across three training stages, exhibits an increase in AI2D and MMStar scores. This demonstrates that our method effectively mitigates IFL without sacrificing the model's core visual-language understanding capabilities. The strategic injection of multilingual text-only data appears to reinforce the LLM's inherent multilingual abilities without disrupting its ability to process and understand visual information.

The evaluation using the chrF++ metric, which measures the quality of text generation by comparing character n-grams, further supports the effectiveness of our multilingual regularization techniques. As shown in Table 5, the TR-3S model demonstrates improved chrF++ scores across all non-English languages compared to the baseline (BM). For instance, German improves from 15.0 to 20.4, and Spanish from 19.1 to 23.7. This indicates that the model not only maintains language fidelity but also generates more accurate and coherent text in multilingual settings.

Model	De	Ru	Es	Nl	Fr	En
BM	15.0	9.9	19.1	16.2	18.1	27.5
TR-3S	**20.4**	**12.5**	**23.7**	**22.0**	**22.8**	**28.2**
TR-3S + M	16.1	10.5	21.6	14.7	18.5	25.5

Table 5: Performance on Crossmodal-3600 by language (chrF++). The best results are shown in bold.

However, a notable observation is the performance degradation observed in the merged model. Despite achieving substantial improvements in multilingual fidelity, the TR-3S M model shows a significant decrease in performance on benchmarks such as AI2D. This decline suggests a potential trade-off between enhanced multilingual capabilities and general task performance when employing model merging techniques. We hypothesize that the merging process, while beneficial for consolidating multilingual knowledge, may introduce conflicts or misalignments in the model's learned visual representations. We further investigated alternative merging methods, including spherical linear interpolation (slerp) and both asymmetric weightings that favor the original backbone (75–25) and the visually instructed model (25–75). These variants, detailed in Appendix F, confirm the trade-offs between language fidelity and multimodal performance, without revealing a universally superior configuration.

Examples of the generation with the TR-3S model can be found in Appendix G.

5 Conclusion

We addressed the challenge of Image-induced Fidelity Loss in VLMs, where models trained on predominantly English data tend to default to English responses. Our approach integrates multilingual text-only data into the visual instruction tuning process, preserving the multilingual abilities of the underlying language model.

Experiments show that proportional multilingual integration (TR-3S) significantly reduces English bias while maintaining core multimodal capabilities. We also analyzed data quantity effects, finding that moderate variations in text-to-visual data ratios do not compromise fidelity, though explicit inclusion of target languages remains necessary. Additionally, we introduced a model merging strategy that further improves language fidelity, albeit with some trade-offs in general task performance, highlighting the need for balance in practical applications.

Overall, our findings demonstrate that multilingual textual regularization is a simple and scalable solution to enhance VLM multilingual competence without large multimodal multilingual datasets. This paves the way for future research on optimizing data integration and refining model merging techniques to balance fidelity and overall performance.

Limitations

Language Coverage

While our approach improves multilingual alignment through text-only supervision, the language coverage remains predominantly European. This raises concerns about the model's ability to generalize to typologically diverse languages, particularly those with non-Latin scripts (e.g., Arabic, Hindi, Chinese). Future work should explore the integration of a wider array of language families and scripts to validate and expand the method's applicability.

Metric Reliability

The fidelity metric (GlotLID and LF+) relies on automatic tools and heuristic judgments, including LLM-as-a-judge assessments that exhibit sensitivity to code-switching and short prompts. Despite efforts to address false positives, such metrics are not infallible and may fail to fully capture semantic fidelity across languages. While the LLM-as-a-judge approach provides a more flexible assessment than rule-based classifiers, it remains inherently limited by the evaluator model's language coverage and internal biases. Its judgments can vary depending on the prompt phrasing, sampling parameters, and the specific LLM used, which may introduce inconsistency across runs. Moreover, the method focuses on surface-level language consistency rather than deeper semantic equivalence, meaning that captions that appear linguistically correct might still deviate in meaning from the intended reference. Code-switching, named entities, and loanwords further complicate the evaluation, as the model can misinterpret these as language errors. Overall, the LLM-as-a-judge should be regarded as an approximate indicator of language fidelity rather than a definitive measure, and we discuss these limitations in more detail in Appendix D.

Scope of Multilingual Training

While Textual Regularization across Three Stages (TR-3S) effectively mitigates Image-induced Fidelity Loss (IFL) by reinforcing Linguistic Fidelity (LF) without sacrificing core visual performance, our current analysis is limited to the English VLM performance benchmarks. We hypothesize that the textual regularization prevents degradation in multilingual visual reasoning, but a complete assessment would require multilingual multimodal benchmarks that are currently scarce or non-standardized. The scalability of TR-3S to models with significantly larger parameter counts and broader language sets remains a subject for future study.

Practical Trade-offs of Model Merging

The model merging strategy (TR-3S+M) introduces a critical, non-negociable hard trade-off that readers must acknowledge. Merging guarantees a substantial boost in Linguistic Fidelity (LF) by robustly anchoring the model to the original LLM backbone (e.g., over 12.5 point average LF gain). However, this operation simultaneously incurs a cost of visual performance degradation due to the introduction of representational misalignment. This trade-off is directly controlled by the interpolation weighting ratio: selecting the highest LF-favoring ratio (e.g., lerp_075) maximizes LF but accepts the largest visual performance cost, while the lowest ratio (e.g., lerp_025) minimizes this cost but yields a more modest LF improvement. Users must align this ratio, which defines the spectrum of the trade-off, with their primary objective. We evaluated several merging strategies and, although merging does reliably improve LF, the accompanying drop in visual performance is large enough that we do not recommend model merging as a default approach except for very specific applications that explicitly prioritize language fidelity over visual quality. Going forward, we believe the right direction is to develop methods that increase LF while preserving or simultaneously improving visual performance; pursuing such methods is important future work for the field.

Ethical Considerations

Our research tackles key ethical issues related to multilingual representation and inclusivity in visual language models. Enhancing multilingual capabilities promotes accessibility and fairness across diverse linguistic communities.

However, relying on machine-translated datasets may introduce biases or cultural inaccuracies. Ensuring responsible translation and ongoing refinement is crucial.

Real-world deployment also demands cultural sensitivity, especially in sectors like education, health, or governance. We emphasize the need for transparency, continuous monitoring, and collaboration with diverse communities to ensure responsible development and use.

Acknowledgements

This work has been supported and funded by the *Ministerio para la Transformación Digital y de la Función Pública* and the *Plan de Recuperación, Transformación y Resiliencia* – funded by the EU through *NextGenerationEU*, within the framework of the projects Modelos del Lenguaje and Desarrollo Modelos ALIA.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. Tallyqa: Answering complex counting questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8076–8084.

Divyanshu Aggarwal, Sankarshan Damle, Navin Goyal, Satya Lokam, and Sunayana Sitaram. 2024. Towards exploring continual fine-tuning for enhancing language ability in large language model. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Anton Alexandrov, Veselin Raychev, Mark Niklas Mueller, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. Mitigating catastrophic forgetting in language transfer via model merging. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17167–17186, Miami, Florida, USA. Association for Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2022. ScanQA: 3D Question Answering for Spatial Scene Understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19107–19117.

Haoping Bai, Shancong Mou, Tatiana Likhomanenko, Ramazan Gokberk Cinbis, Oncel Tuzel, Ping Huang, Jiulong Shan, Jianjun Shi, and Meng Cao. 2023. Vision datasets: A benchmark for vision-based industrial inspection. *arXiv preprint arXiv:2306.07890*.

Louis Bethune, David Grangier, Dan Busbridge, Eleonora Gualdoni, Marco Cuturi, and Pierre Ablin. 2025. Scaling laws for forgetting during finetuning with pretraining data injection. *arXiv preprint arXiv:2502.06042*.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluis Gomez, Marçal Rusiñol, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4290–4300.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11618–11628.

Jimmy Carter. 2024. Textocr-gpt4v. `https://huggingface.co/datasets/jimmycarter/textocr-gpt4v`.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022a. MapQA: A Dataset for Question Answering on Choropleth Maps. In *NeurIPS 2022 First Table Representation Workshop*.

Yingshan Chang, Guihong Cao, Mridu Narang, Jianfeng Gao, Hisami Suzuki, and Yonatan Bisk. 2022b. WebQA: Multihop and Multimodal QA. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16474–16483.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. ALLaVA: Harnessing GPT4V-synthesized Data for A Lite Vision-Language Model. *Preprint*, arXiv:2402.11684.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online. Association for Computational Linguistics.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2025. ShareGPT4V: Improving Large Multi-modal Models with Better Captions. In *Computer Vision – ECCV 2024*, pages 370–387, Cham. Springer Nature Switzerland.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024b. Are We on the Right Way for Evaluating Large Vision-Language Models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. 2024c. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. In *Advances in Neural Information Processing Systems*, volume 37, pages 19472–19495. Curran Associates, Inc.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2024d. On Scaling Up a Multilingual Vision and Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14432–14444.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

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

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. HiTab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110, Dublin, Ireland. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Erfei Cui, Yinan He, Zheng Ma, Zhe Chen, Hao Tian, Weiyun Wang, Kunchang Li, Yi Wang, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Yali Wang, Limin Wang, Yu Qiao, and Jifeng Dai. 2024. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. NVLM: Open Frontier-Class Multimodal LLMs. *arXiv preprint arXiv:2409.11402*.

Saurabh Dash, Yiyang Nan, Arash Ahmadian, and John Dang. 2025. A deepdive into aya vision: Advancing the frontier of multilingual multimodality. Hugging Face Blog.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. 2019. Neural naturalist: Generating fine-grained image comparisons.

In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 708–717, Hong Kong, China. Association for Computational Linguistics.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in Neural Information Processing Systems*, 36:50742–50768.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2023. G-llava: Solving geometric problem with multi-modal large language model. *CoRR*, abs/2312.11370.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2024. mBLIP: Efficient bootstrapping of multilingual vision-LLMs. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 7–25, Bangkok, Thailand. Association for Computational Linguistics.

Aitor Gonzalez-Agirre, Montserrat Marimon, Carlos Rodriguez-Penagos, Javier Aula-Blasco, Irene Baucells, Carme Armentano-Oller, Jorge Palomar-Giner, Baybars Kulebi, and Marta Villegas. 2024. Building a data infrastructure for a mid-resource language: The case of Catalan. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2556–2566, Torino, Italia. ELRA and ICCL.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, and others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Vincent Cartillier, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Devansh Kukreja, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2024. Ego4D: Around the World in 3,000 Hours of Egocentric Video . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–32.

Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. 2018. Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 598–613.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617, Los Alamitos, CA, USA. IEEE Computer Society.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

Musashi Hinck, Carolin Holtermann, Matthew Lyle Olson, Florian Schneider, Sungduk Yu, Anahita Bhiwandiwalla, Anne Lauscher, Shao-Yen Tseng, and Vasudev Lal. 2024. Why do LLaVA vision-language models reply to images in English? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13402–13421, Miami, Florida, USA. Association for Computational Linguistics.

Jinyi Hu, Yuan Yao, Chongyi Wang, SHAN WANG, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, dahai li, Zhiyuan Liu, and Maosong Sun. 2024. Large multilingual models pivot zero-shot multimodal learning across languages. In *The Twelfth International Conference on Learning Representations*.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. 2024. Hq-edit: A high-quality dataset for instruction-based image editing. *CoRR*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*.

Phillip Isola, Joseph J. Lim, and Edward H. Adelson. 2015. Discovering states and transformations in image collections. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1383–1391.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry Davis. 2017. The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6478–6487, Los Alamitos, CA, USA. IEEE Computer Society.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, Brussels, Belgium. Association for Computational Linguistics.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2024. Geomverse: A systematic evaluation of large models for geometric reasoning. In *AI for Math Workshop@ ICML*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision – ECCV 2016*, pages 235–251, Cham. Springer International Publishing.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36:47669–47681.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3426–3440, Dublin, Ireland. Association for Computational Linguistics.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. 2024. Unlocking the conversion of web screenshots into html code with the websight dataset. *Preprint*, arXiv:2403.09029.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2023a. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. *arXiv preprint arXiv:2308.04152*.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023b. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6322–6331.

Zhuowan Li, Xingrui Wong, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan Yuille. 2023c. Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning . In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14963–14973, Los Alamitos, CA, USA. IEEE Computer Society.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Ziquan Liu, Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, Xiangyang Ji, Antoni Chan, and Rong Jin. 2022. Improved fine-tuning by better leveraging pre-training data. In *Advances in Neural Information Processing Systems*, volume 35, pages 32568–32581. Curran Associates, Inc.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786, Online. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. 2024. Palo: A polyglot large multimodal model for 5b people. *arXiv preprint arXiv:2402.14818*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

U.-V. Marti and H. Bunke. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. InfographicVQA. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2582–2591.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208.

Anand Mishra, Karteek Alahari, and CV Jawahar. 2012. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4623–4632.

Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. 2024. Image textualization: An automatic framework for creating accurate and detailed image descriptions. *arXiv preprint arXiv:2406.07502*.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Chen Qiu, Dan Oneață, Emanuele Bugliarello, Stella Frank, and Desmond Elliott. 2022. Multilingual multimodal learning with machine translated text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4178–4193, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Nazneen Rajani, Lewis Tunstall, Edward Beeching, Nathan Lambert, Alexander M. Rush, and Thomas Wolf. 2023. No robots. `https://huggingface.co/datasets/HuggingFaceH4/no_robots`.

Hareesh Ravi, Kushal Kafle, Scott Cohen, Jonathan Brandt, and Mubbasir Kapadia. 2021. Aesop: Abstract encoding of stories, objects, and pictures. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2032–2043.

Florian Schneider and Sunayana Sitaram. 2024. M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4309–4345, Miami, Florida, USA. Association for Computational Linguistics.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1476, Lisbon, Portugal. Association for Computational Linguistics.

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254. ACM.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10737–10746.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision – ECCV 2016*, pages 510–526, Cham. Springer International Publishing.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Yueqi Song, Simran Khanuja, and Graham Neubig. 2024. What is missing in multilingual visual reasoning and how to fix it. *arXiv preprint arXiv:2403.01404*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. 2019. Expressing visual relationships via language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1873–1883, Florence, Italy. Association for Computational Linguistics.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. VisualMRC: Machine Reading Comprehension on Document Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13878–13888.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, Toronto, Canada. Association for Computational Linguistics.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510.

Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. 2024. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pages 360–377. Springer.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9772–9781, Los Alamitos, CA, USA. IEEE Computer Society.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024a. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. 2024b. Vision-Flan: Scaling Human-Labeled Tasks in Visual Instruction Tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15271–15342, Bangkok, Thailand. Association for Computational Linguistics.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. 2023. UReader: Universal OCR-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858, Singapore. Association for Computational Linguistics.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134.

Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. 2022. Syntax-aware network for handwritten mathematical expression recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4543–4552.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. Pangea: A fully open multilingual multimodal LLM for 39 languages. In *The Thirteenth International Conference on Learning Representations*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5312–5322.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2023a. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, and Hongsheng Li. 2024. Mavis: Mathematical visual instruction tuning. *Preprint*, arXiv:2407.08739.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023b. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Yilun Zhao, Chen Zhao, Linyong Nan, Zhenting Qi, Wenlin Zhang, Xiangru Tang, Boyu Mi, and Dragomir Radev. 2023. RobuT: A systematic study of table QA robustness against human-annotated adversarial perturbations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6064–6081, Toronto, Canada. Association for Computational Linguistics.

Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. 2025. Towards lifelong learning of large language models: A survey. *ACM Comput. Surv.*, 57(8).

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004.

A Data Sources

This section showcases the visual datasets (Table 10) and the text-only datasets (Table 11) used throughout this work.

B Language Consistency Evaluation via LLM-as-a-judge

We have observed cases of code-switched generations, and in some cases, even if most of the sentence is generated in the target language, a few words may still appear in English. The primary goal of the textual LLM-as-a-judge evaluation is to address GlotLID's limitation to classify these cases as incorrect.

To effectively evaluate each sentence, we have designed a prompt (see Figure 4) that instructs the evaluator to perform multiple tasks beyond basic language identification, enabling the computation of a language consistency score:

1. Guess the language of the sentence (to compare with GlotLID, even if we will use GlotLID's outputs).

2. Assign a language consistency score (between 0 and 1).

3. Determine whether the sentence is fully in the target language (a boolean value, where False indicates that at least one word appears in another language).

4. Generate a summary explaining the decisions made by the model.

This method allows us to evaluate the language consistency of our model at the word level from different perspectives, both through a numerical score and a boolean indicator.

Language Consistency Evaluation Prompt Template

Analyze the following text and determine the language it is written in.
- Identify the most likely language.
- Ensure the probability score is a single value, not a range or estimate.
- Determine a language consistency score between 0 and 1.0, where 1.0 means the text is entirely in one language, and 0.0 means it is completely incomprehensible.
- Lower the score proportionally if foreign words are present, but do not assign 0.0 unless the text is nonsensical.
- The language score must be a single number between 0 and 1.0.
- Indicate whether the text is completely written in the identified language (True or False).
- In both language consistency metrics, do not penalize for proper nouns, brand names, or commonly used foreign terms (e.g., 'software', 'email') that do not alter the overall language structure.
- Avoid unnecessary explanations. Summarize the feedback (reason of the mark) in at most 30 words.

Use the exact format below:

- Language: [language_guess]
- Language Score: [single value between 0 and 1.0]
- Fully in Language: [True/False]
- Summary: [Concise explanation (max 30 words)]

Keep your answer short and concise. The sentence to analyze is the following:
<CAPTION GENERATED BY VISUAL SALAMANDRA>

Figure 4: Prompt used to evaluate language consistency via LLM-as-a-judge. The evaluator model assesses the language fidelity of the caption generated by the VLM using multiple criteria. Note that this evaluation focuses solely on language fidelity, not the overall quality of the caption.

For additional robustness, we have performed the text evaluations using three different generation configurations (defined in Table 6) and then applied a majority voting.

Conf.	Temperature	Top_p	Max new tokens
A	0.6	0.7	50
B	0.8	0.6	50
C	1.0	0.5	50

Table 6: Generation parameter settings for LLM-as-a-Judge evaluation.

As shown in Figure 4, a structured response format was explicitly requested to ensure that each field could be reliably extracted from the output. The responses that did not conform to the expected format were replaced with "N/A".

To compute the final scores, we averaged the results across the three different configurations. For numerical scores, when a value was missing, we computed the average using the available values. For boolean scores, we applied a majority voting approach. In cases where one score was missing and the remaining two were True and False, we defaulted to False as the final instance score. This ensures that our results provide a lower bound, making the evaluation more conservative and reliable. All the results are shown in Table 7.

B.1 Judging the Judge: Evaluation of the LLM-as-a-judge

Even though we have implemented measures for robustness (such as using different generation configurations for the LLM-as-a-judge evaluators and requiring evaluators to provide summaries to justify their scores) we aim to further ensure the fairness of the provided scores by evaluating them against the reference captions in the dataset.

To assess the reliability of the language consistency evaluator model, we scored the reference captions from the evaluation dataset (crossmodal-3600) using the same evaluation process applied

Model	Lang.	GL	LLM-L	LLM-S	LLM-B	Model	Lang.	GL	LLM-L	LLM-S	LLM-B
Normal Models											
	De	2.7	2.4	**85.8**	73.5		De	24.5	18.3	84.0	**75.9**
	En	**100.0**	98.4	99.0	99.3		En	**100.0**	98.1	98.7	99.1
	Es	4.4	4.7	91.5	87.2		Es	38.4	37.9	94.4	**95.4**
BM	Fr	12.2	10.8	90.3	88.1	TR-1S	Fr	29.9	26.8	**91.2**	**91.1**
	Nl	5.6	6.02	93.8	92.5		Nl	49.2	48.1	93.4	93.1
	No	5.2	3.84	92.0	89.6		No	46.8	30.3	92.8	91.5
	Ru	3.8	2.75	82.2	58.5		Ru	50.9	28.4	82.2	62.9
	De	81.3	54.4	81.9	73.2		De	88.7	**61.3**	82.8	72.7
	En	**100.0**	98.3	98.9	99.4		En	**100.0**	**98.7**	**99.1**	**99.5**
	Es	65.4	61.8	94.1	93.7		Es	**92.9**	**86.8**	**94.6**	94.3
TR-2S	Fr	74.9	63.8	90.5	89.5	TR-3S	Fr	85.7	71.4	89.6	86.7
	Nl	91.3	84.9	93.4	94.3		Nl	91.8	86.7	92.9	93.2
	No	52.0	40.4	93.7	93.3		No	69.5	49.5	92.6	91.0
	Ru	24.8	13.4	81.6	63.1		Ru	52.9	31.7	**83.1**	59.6
	De	**85.0**	58.4	82.1	75.2		De	88.9	58.9	81.2	72.8
	En	**100.0**	98.1	98.8	99.1		En	**100.0**	98.4	98.7	99.4
	Es	91.9	85.9	94.1	94.8		Es	92.4	85.9	94.0	93.7
TR-3S-0.25x	Fr	**88.6**	**76.3**	90.7	89.8	TR-3S-0.5x	Fr	83.5	71.2	90.0	88.8
	Nl	69.8	65.3	93.8	95.3		Nl	**96.0**	**90.6**	**94.1**	**96.0**
	No	53.2	41.2	**94.0**	**94.6**		No	**82.7**	**62.6**	93.5	93.6
	Ru	91.5	48.5	81.6	63.2		Ru	81.9	43.3	81.1	62.7
	De	73.3	48.6	81.0	72.8						
	En	**100.0**	**98.7**	98.7	99.4						
	Es	76.4	71.7	94.3	95.2						
TR-3S-2x	Fr	69.5	59.1	90.0	88.6						
	Nl	93.5	88.3	93.8	95.9						
	No	67.1	53.7	93.9	94.0						
	Ru	**93.3**	**48.9**	81.3	**64.2**						
Merged Models (+M)											
	De	15.0	12.8	84.1	76.6		De	76.5	58.3	**84.5**	77.4
	En	**100.0**	98.1	**98.9**	99.5		En	**100.0**	**98.8**	98.8	**99.7**
	Es	37.0	36.9	93.4	91.5		Es	86.3	81.9	**93.6**	93.7
BM+M	Fr	47.4	40.0	89.8	86.8	TR-1S+M	Fr	87.9	74.3	90.2	89.4
	Nl	9.3	10.0	**93.1**	92.5		Nl	63.7	65.3	92.0	91.4
	No	14.7	10.9	**91.8**	90.4		No	58.1	47.3	89.5	87.6
	Ru	15.4	10.8	82.2	60.8		Ru	77.7	56.1	85.2	73.1
	De	**95.1**	66.5	84.1	77.1		De	94.1	**69.6**	84.4	**78.7**
	En	**100.0**	97.8	98.8	99.2		En	**100.0**	98.0	**98.9**	99.2
	Es	**97.4**	**89.1**	94.8	**95.3**		Es	96.4	88.9	94.5	94.4
TR-2S+M	Fr	**97.7**	**83.1**	**91.8**	**90.1**	TR-3S+M	Fr	95.5	80.8	91.1	89.4
	Nl	95.6	88.8	92.7	**93.1**		Nl	**96.1**	**90.5**	92.3	93.0
	No	**97.6**	**72.7**	90.1	85.7		No	90.1	70.9	91.7	**90.7**
	Ru	**96.7**	60.8	85.2	74.0		Ru	92.4	**65.7**	**86.7**	**76.2**
Crossmodal-3600 reference samples (Evaluating the LLM-as-a-judge)											
	De	100.0	83.3	85.4	91.7						
	En	99.7	99.2	97.7	99.5						
	Es	99.6	97.2	95.2	99.0						
references	Fr	99.9	90.5	90.8	95.2						
	Nl	99.4	95.8	94.5	98.7						
	No	99.1	91.4	93.5	97.9						
	Ru	99.9	76.0	79.8	83.3						

Table 7: Model comparison showing GlotLID detection percentages for the target language (GL) and scores obtained using LLM-as-a-judge. LLM-L represents the target language detection, LLM-S indicates the numerical language consistency score, and LLM-B denotes the binary language consistency score. The GlotLID+LLM score is calculated as the product of the GlotLID score and LLM-B: GlotLID+LLM $=$ GlotLID $\times$ LLM-B. For each group except the LLM-as-a-judge evaluator evaluation, the highest scores are marked in bold.

to the captions generated by our models. We have evaluated all the available references (up to 3 per image) and only used a single generation configuration (configuration B in Table 6).

As shown in Table 7, most of the obtained scores are above 90%, demonstrating the effectiveness of the chosen model as an evaluator. However, some minor errors are present, which can be attributed to multilingual limitations. `Llama-3.1-8B-Instruct` officially supports seven languages in addition to English (French, German, Hindi, Italian, Portuguese, Spanish, and Thai). While this allows it to handle most European languages, it is expected that the model may occasionally struggle with languages outside its primary training set, leading to some misclassifications.

In terms of language consistency, we have discarded selected the binary score due to its higher scores. The evaluator classifies the samples as correct more than the 95% of times in the majority of languages. The lower performance in German (91.7%) can be attributed to the fact that it is a Germanic language that shares a large amount of words with English, what may induce classification errors. In the case of Russian (83.3%), the lower performance may be explained by the limited support for languages using the Cyrillic alphabet in the LLM.

C Caption Generation Prompt

The Crossmodal-3600 dataset does not specify an explicit prompt for caption generation. However, in their work they provide instructions for generating captions, which we used as a guideline. Based on these instructions, we formulated a simplified captioning approach. The prompts used for generating image captions in our evaluation are presented in Figure 5. The same prompt was applied consistently across all selected languages.

D Metrics Discussion

In this study, we chose not to use teacher forcing loss or perplexity as evaluation metrics due to their inherent limitations in interpretability and comparative analysis across models.

We selected chrF++ as our primary evaluation metric rather than BLEU or ROUGE due to its suitability for multilingual assessments. BLEU relies heavily on exact n-gram matching, often penalizing legitimate linguistic variations common in multilingual contexts, while ROUGE primarily measures recall and is optimized for summarization tasks, making it suboptimal for assessing generative multilingual output quality. In contrast, chrF++ evaluates based on character-level n-gram overlaps, accommodating linguistic diversity and morphological richness across multiple languages, thus providing a more robust and linguistically sensitive assessment for multilingual visual language models.

Additionally, during the study, we employed VLM-as-a-judge to evaluate the quality of multilingual generations. However, we found that chrF++ effectively addressed the limitations related to multilingual performance evaluation inherent in other metrics, thereby serving as a comprehensive solution for our assessment needs.

While the LLM-as-a-judge procedure helped mitigate a class of errors produced by purely rule-based detectors (e.g., short or code-switched captions that GlotLID misclassifies), it carries a distinct set of limitations that affect its interpretation as an automatic metric. First, the evaluator's outputs are sensitive to prompt wording and sampling hyperparameters (temperature, top-p, maximum tokens). Small changes in prompt phrasing or generation configuration can lead to different language-consistency scores for the same sentence, introducing non-determinism at the per-example level even when using conservative aggregation. Second, the evaluator primarily assesses surface-level language consistency (presence/absence of out-of-language tokens) rather than deeper cross-lingual semantic fidelity: a caption can receive a high language-consistency score while still exhibiting mistranslation, subtle semantic drift, or pragmatic errors relative to the reference. Third, model-internal language coverage and training data biases create systematic variability across languages, scripts, and dialects: evaluators trained or optimized on Latin-alphabet European languages will typically perform worse on low-resource languages or non-Latin scripts, producing both false positives and false negatives that are not readily corrected by simple calibration.

E Training Hyperparameters

The training hyperparameters used during the training of the models evaluated in this work are detailed in Table 8.

		Stage-1	Stage-1.5	Stage-2	OneVision
Vision	**Resolution** # Tokens	384 729	AnyRes Max 5 Max 729 $\times$ 5	AnyRes Max 9 Max 729 $\times$ 10	AnyRes Max 9 Max 729 $\times$ 10
Data	**Dataset** # Vision Samples	Single-Image 558K	Single-Image 3.8M	Single-Image 3.1M	Single/Multi-Image, Video 1.6M
Model	**Trainable** 7.8B LLM	Projector 20.0M	Full Model 8.2B	Full Model 8.2B	Full Model 8.2B
Training	**Batch Size**	128	64	64	64
	LR: ψ_{vision}	1×10^{-3}	2×10^{-6}	2×10^{-6}	2×10^{-6}
	LR: $\{\theta_{proj}, \phi_{LLM}\}$	1×10^{-3}	1×10^{-5}	1×10^{-5}	1×10^{-5}
	Epoch	1	1	1	1
	Warmup Ratio	0.03	0.03	0.03	0.03
	LR Scheduler	Cosine	Cosine	Cosine	Cosine
	Grad. Accum.	1	2	2	2

Table 8: Detailed configuration for each training stage of the LLaVA-OneVision model. For a detailed explanation of AnyRes Max, refer to (Li et al., 2025). Anyres Max 5: 384 $\times\{2 \times 2, 1 \times \{2, 3\}, 2, 3 \times 1\}$. AnyRes Max 9: 384 $\times\{\{1 \times 1\}, ..., \{6 \times 6\}\}$.

F Alternative Merging Strategies

To better understand the trade-offs involved in model merging, we conducted a series of additional experiments comparing different interpolation methods and weight ratios. In particular, we investigated:

Linear Interpolation (lerp) This method interpolates model weights using the standard formula $w = (1 - \alpha)w_1 + \alpha w_2$, where w_1 and w_2 are the weights of the visually instructed and backbone models, respectively, and α is the interpolation ratio.

Spherical Linear Interpolation (slerp) Unlike `lerp`, slerp (Shoemake, 1985) interpolates weights along a great arc on the hypersphere, preserving the norm and relative directionality. It is computed as:

$$\text{slerp}(w_1, w_2, \alpha) = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)} w_1 + \frac{\sin(\alpha\theta)}{\sin(\theta)} w_2 \tag{1}$$

where θ is the angle between the two weight vectors. This method can yield smoother transitions in weight space, especially when the models differ significantly.

We evaluated merged models using both interpolation methods under three weighting scenarios:

- **50–50**, giving equal weight to the visually instructed and original backbone models.
- **75–25**, favoring the original backbone to preserve pretrained language capabilities.
- **25–75**, prioritizing the visually instructed model to reinforce vision-language alignment.

The results are presented in Table 9, showing both language fidelity and downstream task performance.

Discussion. As shown, asymmetric merges favoring the backbone (e.g., `lerp_075` and `slerp_075`) achieve near-perfect language fidelity but show weaker performance in multimodal benchmarks. Conversely, merges favoring the visually instructed model (e.g., `lerp_025` and `slerp_025`) lead to substantially improved task performance, but at the cost of lower fidelity in certain languages. The `slerp_050` model—corresponding to our main `TR-3S+M`—offers a more balanced trade-off.

Overall, no single merging configuration yields a clearly optimal trade-off. The best strategy depends on the intended use case: 75% visually instructed weights are preferable when multilingual fidelity is critical, while 25% weights better support general multimodal performance.

G Generation Examples

Figures 6-9 present examples generated using the TR-3S-0.05x model across various languages and diverse tasks.

Story Generation In Figure 6, the model is prompted to generate a story from an image. It accurately reads text within the image to identify characters and establish the setting, demonstrating its ability to craft diverse narratives consistently

Table 9: Merged model results: language fidelity (left) and task performance (right). lerp_050 and slerp_050 correspond to 50–50 merges using linear and spherical interpolation, respectively; slerp_050 corresponds to our main model TR-3S+M. lerp_075 and slerp_075 are asymmetric 75–25 merges favoring the base text model. lerp_025 and slerp_025 invert this ratio to prioritize the visually instructed model.

Model	DE	ES	FR	NL	RU	Avg.	AI2D	MMMU	MMStar	RWQA	Avg.
lerp_075	98.75	99.89	99.92	99.64	98.66	99.37	41.84	27.00	32.56	37.39	34.70
slerp_075	99.08	99.86	99.89	99.64	99.02	99.50	42.16	27.33	32.66	36.99	34.79
lerp_050	92.83	96.42	95.83	94.78	91.11	94.19	57.03	34.44	42.06	52.03	46.39
slerp_050	94.10	96.40	95.50	96.10	92.40	94.90	57.19	34.11	42.25	52.16	46.43
lerp_025	89.32	93.06	91.06	92.83	63.56	85.97	72.51	34.89	47.13	55.95	52.62
slerp_025	90.11	93.53	90.81	94.44	66.30	87.44	72.38	34.89	47.27	56.08	52.66

across different languages—even when the text is in English.

Image Description Figure 7 presents a brief image description task. Although the image shows a salamander perched on a person's hand, some language outputs mistakenly label it as an insect or a predator. Despite these inaccuracies, the descriptions remain largely appropriate.

OCR and Translation Figure 8 showcases a task combining OCR with translation. The model extracts text from an image and then translates it into a target language. This two-step process: OCR followed by translation, highlights the model's ability to merge visual analysis with its linguistic capabilities. Minor errors do occur, particularly in languages not extensively represented during training, resulting in slightly erroneous translations or defaulting to English.

Multi-Image Reasoning Finally, Figure 9 illustrates a multi-image scenario where the model must comprehend the content of several images and reason to provide an appropriate answer. This example further confirms the model's effectiveness in real-world applications.

In general, these examples demonstrate how the VLMs instructed via our approach perform optimally across a wide range of tasks, especially for languages where text-only data was incorporated during the visual instruction process.

G.1 Code Switching in Caption Generation

Figure 10 presents examples of code switching observed during caption generation for the Crossmodal-3600 dataset. We could identify two primary patterns emerge:

- **Independent Words**: Certain technical or less common words are generated in English.
- **Language Alternation**: In some cases, once a word is switched to English, all subsequent words continue in English.

Caption Generation Prompts

[English]
Give me a brief summary of the following image, without too many details. The description should be general and have a maximum of 10 words. To do this, identify the most relevant object or person in the image, the main relationship between the highlighted objects, the most important activity represented, the most outstanding attributes of the main object or person, and the context in which the scene takes place. Then, synthesize everything into a single descriptive and concise sentence, without including additional text.

[Spanish]
Dame un breve resumen de la siguiente imagen, sin dar muchos detalles. La descripción debe ser general y tener un máximo de 10 palabras. Para ello, identifica el objeto o persona más relevante en la imagen, la relación principal entre los objetos destacados, la actividad más importante representada, los atributos más sobresalientes del objeto o persona principal y el contexto en el que ocurre la escena. Luego, sintetiza todo en una sola frase descriptiva y concisa, sin incluir texto adicional.

[French]
Donne-moi un bref résumé de l'image suivante, sans trop de détails. La description doit être générale et contenir un maximum de 10 mots. Pour cela, identifie l'objet ou la personne la plus importante dans l'image, la relation principale entre les objets mis en avant, l'activité la plus significative représentée, les attributs les plus marquants de l'objet ou de la personne principale et le contexte dans lequel la scène se déroule. Ensuite, synthétise tout en une seule phrase descriptive et concise, sans ajouter de texte supplémentaire.

[German]
Gib mir eine kurze Zusammenfassung des folgenden Bildes, ohne zu viele Details. Die Beschreibung sollte allgemein sein und maximal 10 Wörter umfassen. Identifiziere dazu das relevanteste Objekt oder die wichtigste Person im Bild, die Hauptbeziehung zwischen den hervorgehobenen Objekten, die wichtigste dargestellte Aktivität, die auffälligsten Merkmale des Hauptobjekts oder der Hauptperson und den Kontext, in dem die Szene stattfindet. Fasse dann alles in einem einzigen prägnanten und beschreibenden Satz zusammen, ohne zusätzlichen Text hinzuzufügen.

[Italian]
Dammi un breve riassunto della seguente immagine, senza troppi dettagli. La descrizione deve essere generale e avere un massimo di 10 parole. Per farlo, identifica l'oggetto o la persona più rilevante nell'immagine, la relazione principale tra gli oggetti evidenziati, l'attività più importante rappresentata, gli attributi più evidenti dell'oggetto o della persona principale e il contesto in cui si svolge la scena. Quindi, sintetizza tutto in un'unica frase descrittiva e concisa, senza includere testo aggiuntivo.

(Continues on next page...)

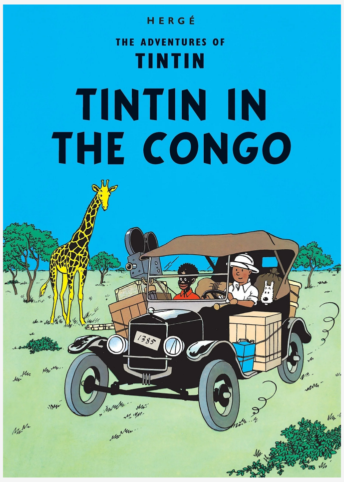
[Dutch]
Geef me een korte samenvatting van de volgende afbeelding, zonder te veel details. De beschrijving moet algemeen zijn en maximaal 10 woorden bevatten. Identificeer hiervoor het meest relevante object of de belangrijkste persoon in de afbeelding, de hoofdrelatie tussen de uitgelichte objecten, de belangrijkste weergegeven activiteit, de meest opvallende kenmerken van het hoofdobject of de belangrijkste persoon en de context waarin de scène zich afspeelt. Vat vervolgens alles samen in één beschrijvende en beknopte zin, zonder extra tekst toe te voegen.

[Portuguese]
Dá-me um breve resumo da seguinte imagem, sem muitos detalhes. A descrição deve ser geral e ter um máximo de 10 palavras. Para isso, identifica o objeto ou a pessoa mais relevante na imagem, a principal relação entre os objetos destacados, a atividade mais importante representada, os atributos mais marcantes do objeto ou da pessoa principal e o contexto em que ocorre a cena. Depois, sintetiza tudo numa única frase descritiva e concisa, sem incluir texto adicional.

[Russian]
Дайте мне краткое описание следующего изображения без лишних деталей. Описание должно быть общим и содержать не более 10 слов. Для этого определите наиболее важный объект или человека на изображении, основную связь между выделенными объектами, главное изображенное действие, самые заметные характеристики главного объекта или человека и контекст, в котором происходит сцена. Затем объедините все это в одно описательное и краткое предложение, не добавляя лишнего текста.

Figure 5: Prompts used to evaluate via LLM-as-a-judge the language consistency of the caption provided by the model.

Story Generation



[English]
User: Create a short story using the following image as source.
VLM: Once upon a time, in the heart of the African savannah, there was a young boy named Tintin. Tintin was an adventurous soul, always eager to explore new places and meet new people. One day, he heard about a mysterious place called the Congo, a land of exotic animals and uncharted territories. Tintin decided to embark on an adventure to discover the secrets of the Congo.
As he traveled through the dense jungle, he encountered a variety of animals, including a majestic giraffe and a wise elephant. He also met a group of explorers who were searching for a hidden treasure. Tintin joined their team and together they set out on a thrilling journey through the jungle.
They faced many challenges and obstacles, but Tintin's curiosity and bravery never faltered. Finally, they reached the heart of the Congo and discovered a treasure—a collection of rare artifacts from different cultures, each with its own story to tell.
Tintin and his friends returned home with a newfound appreciation for the wonders of the world and the importance of exploration and discovery. And so, the legend of Tintin in the Congo began.

[Spanish]
User: Crea una historia corta usando la siguiente imagen como fuente.
VLM: Había una vez un aventurero llamado Tintín que viajaba por el Congo en un coche antiguo. Lo acompañaban un guía y un perro, ambos llamados Milú. Tintín y sus amigos se encontraron con un grupo de gorilas y decidieron ayudarlos a escapar de los cazadores furtivos. Juntos, salvaron a los gorilas y regresaron a la ciudad, convirtiéndose en héroes en el Congo y continuando sus aventuras por el mundo.
[English] *VLM:* Once upon a time there was an adventurer named Tintin who traveled through the Congo in an antique car. He was accompanied by a guide and a dog, both named Snowy. Tintin and his friends came across a group of gorillas and decided to help them escape from poachers. Together, they saved the gorillas and returned to the city, becoming heroes in the Congo and continuing their adventures around the world.

(Continues on next page...)

[Catalan]
User: Crea una història curta utilitzant la següent imatge com a font.
VLM: En Tintín i el seu gos Milú es troben en un safari a l'Àfrica. El cotxe d'en Tintín es trenca i han de demanar ajuda a un grup de nadius. Aquests reparen el cotxe i, a canvi, en Tintín i Milú els ensenyen una pel·lícula que és tot un èxit. Els nadius, encantats, els demanen que tornin l'any següent per mostrar-ne una altra. Finalment, en Tintín i Milú decideixen quedar-se a l'Àfrica i muntar un cinema per als habitants, començant així una nova vida plena d'amistat i intercanvi cultural.
[English] *VLM:* Tintin and his dog Snowy are on a safari in Africa. Tintin's car breaks down and they have to ask a group of natives for help. They repair the car and, in return, Tintin and Snowy show them a film that is a huge success. The natives, delighted, ask them to come back the following year to show another one. Finally, Tintin and Snowy decide to stay in Africa and set up a cinema for the inhabitants, thus beginning a new life full of friendship and cultural exchange.

[French]
User: Créez une histoire courte en utilisant l'image suivante comme source.
VLM: Il était une fois, dans la savane africaine, un petit garçon nommé Tintin, passionné d'aventures et rêvant d'explorer le continent. Un jour, il partit à la recherche de son idole, le célèbre explorateur Tintin, et se lança dans une aventure à travers la savane, rencontrant girafes, lions et éléphants. Finalement, il rencontra son idole dans le Congo, et ensemble, ils découvrirent de nombreux secrets et trésors, réalisant ainsi le rêve d'explorer l'Afrique.
[English] *VLM:* Once upon a time, in the African savannah, a little boy named Tintin had a passion for adventure and dreamed of exploring the continent. One day, he set off in search of his idol, the famous explorer Tintin, and embarked on an adventure across the savannah, encountering giraffes, lions and elephants. Finally, he met his idol in the Congo, and together they discovered many secrets and treasures, fulfilling the dream of exploring Africa.

[German]
User: Erstelle eine kurze Geschichte, indem du das folgende Bild als Quelle verwendest.
VLM: Tintin und sein Hund Milú begeben sich auf ein Abenteuer im Kongo. Während sie mit einem alten Auto durch die Savanne fahren, begegnen sie einem riesigen Giraffen, das sie fasziniert. Nachdem sie weitere Tiere wie Elefanten und Löwen getroffen haben, erleben sie eine spannende Reise, bei der sie schließlich einen Schatz entdecken und stolz auf ihre erlebten Abenteuer zurückblicken.
[English] *VLM:* Tintin and his dog Milú go on an adventure in the Congo. While driving through the savannah in an old car, they come across a giant giraffe that fascinates them. After meeting other animals such as elephants and lions, they experience an exciting journey during which they finally discover a treasure and look back proudly on their adventures.

[Italian]
User: Crea una breve storia utilizzando la seguente immagine come fonte.
VLM: Tintin e il suo cane Milú sono in un'avventura in Africa. Durante il viaggio, mentre guidano il loro veicolo, notano una grande giraffa che si avvicina e decidono di fermarsi per osservarla da vicino, vivendo un momento unico nel cuore della savana.
[English] *VLM:* Tintin and his dog Milú are on an adventure in Africa. On the way, while driving their vehicle, they notice a large giraffe approaching and decide to stop and observe it closely, experiencing a unique moment in the heart of the savannah.

(Continues on next page...)

[Dutch]
User: Maak een kort verhaal met behulp van de volgende afbeelding als bron.
VLM: Tintin in de Congo is een stripverhaal van Hergé waarin Tintin als journalist naar het Congo reist, vergezeld door zijn beste vriend Milú, een geredde chimpansee. Ze beleven tal van avonturen in de jungle, ontdekken de schoonheid van de natuur en maken kennis met de lokale bevolking, wat leidt tot een onvergetelijke reis vol humor en spanning.
[English] *VLM:* Tintin in the Congo is a comic strip by Hergé in which Tintin travels to the Congo as a journalist, accompanied by his best friend Milú, a rescued chimpanzee. They have numerous adventures in the jungle, discover the beauty of nature and meet the local people, leading to an unforgettable journey full of humor and excitement.

[Portuguese]
User: Crie uma curta história usando a seguinte imagem como fonte.
VLM: Era uma vez um explorador chamado Tintin, que se encontrava numa aventura no Congo. Ao volante de um carro vintage, acompanhado pelo seu amigo e animal de estimação, Milú, Tintin estava em busca de um tesouro escondido. No caminho, tiveram de improvisar para contornar um enorme elefante que bloqueava a passagem, mas conseguiram seguir em frente e, no final, encontraram o tesouro que buscavam.
[English] *VLM:* Once upon a time there was an explorer called Tintin, who was on an adventure in the Congo. At the wheel of a vintage car, accompanied by his friend and pet, Milú, Tintin was looking for hidden treasure. I don't walk, you had to improvise to get around a huge elephant that was blocking the passage, but we managed to keep going and, finally, we found the treasure we were looking for.

[Galician]
User: Crea unha historia curta utilizando a seguinte imaxe como fonte.
VLM: Había unha vez un explorador chamado Tintin no Congo. Durante a súa viaxe, atopouse con animais exóticos como un gorila, un elefante e un león, pero o que máis chamou a súa atención foi un gran xaguar no medio do bosque. Cansado, o xaguar foi axudado por Tintin a chegar a un río para beber, e xuntos descubriron unha cova misteriosa con un tesouro oculto que decidiron compartir.
[English] *VLM:* Once upon a time there was an explorer named Tintin no Congo. During his journey, he encountered exotic animals such as a gorilla, an elephant and a lion, but what caught his attention the most was a large jaguar in the middle of the forest. Tired, the jaguar was helped by Tintin to reach a river to drink, and together they discovered a mysterious cave with a hidden treasure that they decided to share.

[Polish]
User: Stwórz krótką historię, używając poniższego obrazu jako źródła.
VLM: Tytuł: "Tintin w Kong" Autor: Hergé Data wydania: 1931
[English] *VLM:* Title: "Tintin in the Congo" Author: Hergé Publication date: 1931

[Swedish]
User: Skapa en kort berättelse med hjälp av följande bild som källa.
VLM: Tintin och Milú är på en safari i Afrika. De sitter i en old-fashioned bil och ser på en giraffe som är i vägen. Tintin och Milú tar sig över till den andra sidan av vägen och ser på en elefant som är i vägen. De tar sig över till den andra sidan av vägen och ser på en elefant som är i vägen.
[English] *VLM:* Tintin and Milú are on a safari in Africa. They sit in an old-fashioned car and look at a giraffe that is in the road. Tintin and Milú cross to the other side of the road and look at an elephant in the road. They cross to the other side of the road and look at an elephant in the road.

(Continues on next page...)

[Russian]
User: Создайте короткий рассказ, используя следующее изображение в качестве источника.
VLM: Однажды, в далекой Африке, в самом сердце Конго, в 1920-х годах, произошло уникальное приключение, которое стало известно как "Тинтин в Конго". Тинтин, знаменитый репортер и путешественник, был отправлен в Африку, чтобы рассказать о жизни и культуре этого загадочного континента. Он и его верные спутники, Микель и Нельсон, сели в свой старенький, но надежный автомобиль и отправились в путешествие. Они пробирались через дикие джунгли, где их окружали дикие животные и таинственные племена. Тинтин и его команда были не только reporters, но и смелые путешественники, которые хотели узнать больше о мире, который был для них загадкой. Они встретили множество людей, которые были для них новыми и интересными, и узнали о их культуре и традициях. Они также узнали о том, что Конго был не только домом для диких животных, но и для людей, которые жили в гармонии с природой. В конце концов, Тинтин и его команда достигли своей цели и рассказали о своем путешествии в Европе, где их история вызвала большой интерес. "Тинтин в Конго"стал одним из самых известных и любимых приключений Тинтина, и он всегда будет помнить об этом
[English] *VLM:* Once upon a time, in faraway Africa, in the heart of the Congo, in the 1920s, a unique adventure took place that became known as "Tintin in the Congo." Tintin, a famous reporter and traveler, was sent to Africa to report on the life and culture of this mysterious continent. He and his faithful companions, Mikel and Nelson, got into their old but reliable car and set off on their journey. They made their way through a wild jungle where they were surrounded by wild animals and mysterious tribes. Tintin and his team were not only reporters, but also brave travelers who wanted to learn more about a world that was a mystery to them. They met many people who were new and interesting to them and learned about their culture and traditions. They also learned that the Congo was not only home to wild animals, but also to people who lived in harmony with nature. In the end, Tintin and his team reached their goal and told their journey in Europe, where their story generated a lot of interest. "Tintin in the Congo" became one of Tintin's most famous and beloved adventures, and he will always be remembered for it

[Norwegian]
User: Lag en kort historie ved å bruke følgende bilde som kilde.
VLM: Tintin og Milú er på en safari i Afrika. De er i en bil og ser på en giraffe.
[English] *VLM:* Tintin and Milú are on a safari in Africa. They are in a car looking at a giraffe.

[Hungarian]
User: Írj egy rövid történetet a következő kép felhasználásával forrásként.
VLM: Tintin a Kongóba indul, hogy megmentse a királyt és a királynőt, akiket a gorilla rabolt el. A királynő és a királyt a gorilla foglyul ejtette, és a Kongóba menekült. Tintin és a barátai elindulnak, hogy megmentsék a királyt és a királynőt.
[English] *VLM:* Tintin goes to the Congo to rescue the king and queen, who have been kidnapped by the gorilla. The king and queen have been captured by the gorilla and have fled to the Congo. Tintin and his friends set out to rescue the king and queen.

Figure 6: Multilingual generation examples with Text Regularization and merged with the original backbone LLM.

Image Description

[English]
User: Briefly describe the following image, without many details and in a few words.
VLM: A person's hand holding a black and yellow striped salamander.

[Spanish]
User: Describe brevemente la siguiente imagen, sin muchos detalles y en pocas palabras.
VLM: Una mano sosteniendo a un insecto, específicamente un insecto de color negro y amarillo.
[English] *VLM:*A hand holding an insect, specifically a black and yellow insect.

[Catalan]
User: Descriu breument la següent imatge, sense molts detalls i en poques paraules.
VLM: A la mà d'una persona hi ha una salamandra de color negre i groc.
[English] *VLM:* In a person's hand there is a black and yellow salamander.

[French]
User: Décrivez brièvement l'image suivante, sans trop de détails et en quelques mots.
VLM: Un petit animal noir et blanc est dans la main d'une personne.
[English] *VLM:* A small black and white animal is in a person's hand.

[German]
User: Beschreiben Sie das folgende Bild kurz, ohne viele Details und in wenigen Worten.
VLM: In der Hand einer Person liegt ein Frillengecko. Der Gecko hat eine schwarze und gelbe markierte Haut. Die Person hat einen Arm mit einem Armband.
[English] *VLM:* A frilled gecko lies in the hand of a person. The gecko has black and yellow marked skin. The person has an arm with a bracelet.

[Italian]
User: Descrivi brevemente la seguente immagine, senza molti dettagli e in poche parole.
VLM: Un insetto in una mano.
[English] *VLM:* An insect in a hand.

[Dutch]
User: Beschrijf kort de volgende afbeelding, zonder veel details en in een paar woorden.
VLM: De hand van een mens met een salamander in het palm.
[English] *VLM:* The hand of a human with a salamander in the palm.

(Continues on next page...)

[Portuguese]
User: Descreva resumidamente a imagem seguinte, sem muitos pormenores e em poucas palavras.
VLM: Uma pessoa está a segurar um sapo em sua mão. O sapo é de cor preta e amarela. O sapo está vivo.
[English] *VLM:* A person is holding a frog in his hand. The frog is black and yellow in color. The frog is alive.

[Galician]
User: Describe brevemente a seguinte imaxe, sen moitos detalles e en poucas palabras.
VLM: A man sostén unha larva de salamandra. A larva é de cor negra e amarela. A man que sostén a larva é branca.
[English] *VLM:* A man holds a yellow and black lobster nail.

[Russian]
User: Кратко опишите следующее изображение, без подробностей и в нескольких словах.
VLM: В руке человека, который держит змею.
[English] *VLM:* In the hand of a man holding a snake.

[Polish]
User: Opisz krótko poniższy obraz, bez wielu szczegółów i w kilku słowach.
VLM: Oto krótki opis tego obrazu:"Czarna i żółta salamandra w dłoni." To jest fotografia salamandry, która jest trzymana w ręce. Salamandra ma czarne i żółte plamy na swym ciele.
[English] *VLM:* Here is a short description of this image: 'A black and yellow salamander in hand.' This is a photograph of a salamander that is being held in hand. The salamander has black and yellow spots on its body.

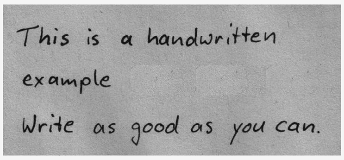
[Swedish]
User: Beskriv kort nedanstående bild, utan många detaljer och med några få ord.
VLM: En hand med en salamander i den.
[English] *VLM:* A hand with a salamander in it.

[Norwegian]
User: Beskriv kort det følgende bildet, uten mange detaljer og med noen få ord.
VLM: En <u>hand holding</u> en salamander.
[English] *VLM:* A hand holding a salamander.

[Hungarian]
User: Röviden írja le a következő képet, sok részlet nélkül, néhány szóban.
VLM: A képen egy ember tartja a kezében egy fekete és sárga ragadozó.
[English] *VLM:* In the picture, a person holds in their hand a black and yellow predator.

Figure 7: Multilingual generation examples with the model trained with Text Regularization and merged with the original backbone LLM. Code switched words are underlined.

OCR & Translation



[English]
User: Read the text that appears in the image and translate it into English.
VLM: This is a handwritten example. Write as good as you can.

[Spanish]
User: Read the text that appears in the image and translate it into Spanish.
VLM: Este es un ejemplo de escritura a mano. Escribe tan bien como puedas.
[English] *VLM:* This is a handwritten example. Write as good as you can.

[Catalan]
User: Read the text that appears in the image and translate it into Catalan.
VLM: Aquest és un exemple de text escrit a mà. Escriu tan bé com puguis.
[English] *VLM:* This is a handwritten example. Write as good as you can.

[French]
User: Read the text that appears in the image and translate it into French.
VLM: C'est un exemple de lettre écrite à la main. Ecrivez comme vous pouvez.
[English] *VLM:* This is an example of a handwritten letter. Write as you can.

[German]
User: Read the text that appears in the image and translate it into German.
VLM: Dies ist ein Beispiel für einen handgeschriebenen Text. Schreiben Sie so gut wie Sie können.
[English] *VLM:* This is a handwritten example. Write as good as you can.

[Italian]
User: Read the text that appears in the image and translate it into Italian.
VLM: Questo è un esempio di scrittura a mano. Scrivete bene come voi potete.
[English] *VLM:* This is a handwritten example. Write as good as you can.

[Dutch]
User: Read the text that appears in the image and translate it into Dutch.
VLM: Dit is een handgeschreven voorbeeld. Schrijf zo goed als je kunt.
[English] *VLM:* This is a handwritten example. Write as good as you can.

[Portuguese]
User: Read the text that appears in the image and translate it into Portuguese.
VLM: Este é um exemplo de um texto escrito à mão. Escreva como bom como você pode.
[English] *VLM:* This is a handwritten example. Write as good as you can.

[Galician]
User: Read the text that appears in the image and translate it into Galician.
VLM: Isto é un exemplo de escrita a man. Escribe ben como se pode.
[English] *VLM:* This is a handwritten example. Write as good as you can.

(Continues on next page...)

[Russian]
User: Read the text that appears in the image and translate it into Russian.
VLM: Это пример написанного от руки текста. Напишите как можно лучше, чем вы можете.
[English] *VLM:* This is an example of a handwritten text. Write as best as you can.

[Polish]
User: Read the text that appears in the image and translate it into Polish.
VLM: Oto tekst w języku polskim: "To jest przykładowy handwriting example. Napisz tak, jak potrafisz"
[English] *VLM:* Here is the text in Polish: "This is a handwriting example. Write as much as you can"

[Swedish]
User: Read the text that appears in the image and translate it into Swedish.
VLM: <u>This is a handwritten example. Write as good as you can.</u>
[English] *VLM:* This is a handwritten example. Write as good as you can.

[Norwegian]
User: Read the text that appears in the image and translate it into Norwegian.
VLM: <u>This is a handwritten example. Write as good as you can.</u>
[English] *VLM:* This is a handwritten example. Write as good as you can.

[Hungarian]
User: Read the text that appears in the image and translate it into Hungarian.
VLM: Ez egy kézzel írott példa. Írjál úgy, mint te tudsz.
[English] *VLM:* This is a handwritten example. Write the way you know how.

Figure 8: Multilingual generation examples with the model trained with Text Regularization and merged with the original backbone LLM. Words in English whenever it is not the target are underlined.

Multi-Image Reasoning

[English]

User: Which dish from the ones shown in these images should I eat if I am lactose intolerant?
VLM: If you are lactose intolerant, you should eat the dish on the left, which is a stir fry of chicken and vegetables. This dish does not contain any dairy products, making it a good option for those who are lactose intolerant.

[Spanish]

User: ¿Qué plato de los que aparecen en estas imágenes debería de comer si soy intolerante a la lactosa?
VLM: Si eres intolerante a la lactosa, no deberías comer el plato de pizza.
[English] *VLM:* If you are lactose intolerant, you should not eat the pizza dish.

[Catalan]

User: Quin plat de les imatges donades hauria de menjar si sóc intolerant a la lactosa?
VLM: Si ets intolerant a la lactosa, hauries de menjar el plat de la primera imatge, ja que no conté productes lactis.
[English] *VLM:* If you are lactose intolerant, you should eat the dish in the first image, as it does not contain dairy products.

[French]

User: Quel plat parmi ceux qui apparaissent dans ces images devrais-je manger si je suis intolérant au lactose ?
VLM: Si vous êtes intolérant au lactose, je vous conseillerais de manger le plat de poulet grillé et de légumes.
[English] *VLM:* If you're lactose intolerant, I'd recommend eating the grilled chicken and vegetable dish.

[Italian]

User: Quale piatto tra quelli mostrati in queste immagini dovrei mangiare se sono intollerante al lattosio?
VLM: Se sei intollerante al lattosio, dovresti mangiare il primo piatto, il stir fry di pollo e verdura.
[English] *VLM:* If you are lactose intolerant, you should eat the first course, chicken and vegetable stir fry.

(Continues on next page...)

[German]
User: Welches Gericht aus den in diesen Bildern gezeigten sollte ich essen, wenn ich laktoseintolerant bin?
VLM: Wenn Sie laktoseintolerant sind, sollten Sie das erste Bild, das Chicken and Vegetables, essen.
[English] *VLM:* If you are lactose intolerant, you should eat the first picture, the Chicken and Vegetables.

[Dutch]
User: Welk gerecht van de getoonde gerechten in deze afbeeldingen zou ik moeten eten als ik lactose-intolerant ben?
VLM: Als je lactose-intolerant bent, zou je het gerecht in de eerste afbeelding moeten eten. Het is een salade met een dressing die geen lactose bevat. De salade is gemaakt van grilled chicken, rode en groene peppers, en is niet gebakken of gefrituurd, wat betekent dat het geen lactose bevat. De salade is dus een lactose-vrije optie.
[English] *VLM:* If you are lactose intolerant, you should eat the dish in the first picture. It is a salad with a dressing that does not contain lactose. The salad is made with grilled chicken, red and green peppers, and is not baked or fried, which means that it does not contain lactose. The salad is therefore a lactose-free option.

[Portuguese]
User: Qual prato entre os mostrados nestas imagens eu deveria comer se for intolerante à lactose?
VLM: Se você é intolerante à lactose, o prato que você deve evitar é o pizza. A pizza é feita com queijo, que é uma fonte de lactose, e pode causar sintomas como dor abdominal, diarreia, náuseas e vômitos. Se você tem intolerância à lactose, é importante evitar alimentos que contenham lactose para evitar esses sintomas.
[English] *VLM:* If you are lactose intolerant, the food you should avoid is pizza. Pizza is made with cheese, which is a source of lactose, and can cause symptoms such as abdominal pain, diarrhea, nausea, and vomiting. If you are lactose intolerant, it is important to avoid foods that contain lactose to avoid these symptoms.

[Galician]
User: Que prato das imaxes dadas debería comer se son intolerante á lactosa?
VLM: Se é intolerante á lactosa, debería comer o prato da primeira imaxe, que é unha ensalada de pollo e verduras.
[English] *VLM:* If you are lactose intolerant, you should eat the dish in the first picture, which is a chicken and salad.

[Polish]
User: Jakie danie z pokazanych na tych obrazkach powinienem zjeść, jeśli jestem nietolerancyjny na laktozę?
VLM: Jeżeli jesteś nietolerancki na laktozę, to nie powinieneś jeść pizzy. Pizza jest gotowa z mozzarella, a mozzarella jest produkowana z mleka. Jeżeli wolisz jeść potrawy bez laktozy, to możesz zjeść kurczaka z warzywami. Kurczak jest bezlaktozny, a warzywa są bezlaktozne.
[English] *VLM:* If you are lactose intolerant, you should not eat pizza. Pizza is made with mozzarella, and mozzarella is made from milk. If you prefer to eat lactose-free food, you can eat chicken with vegetables. Chicken is lactose-free, and vegetables are lactose-free.

(Continues on next page...)

[Russian]

User: Какое блюдо из показанных на этих изображениях мне следует съесть, если я непереносим к лактозе?

VLM: Если вы не переносите лактозу, то вам следует выбрать блюдо, которое не содержит молочных продуктов. На изображении 1 показаны кусочки курицы с овощами, которые не содержат лактозы. На изображении 2 показана пицца с сыром и другими toppings, которые могут содержать лактозу. Однако, если вы хотите попробовать что-то, что не содержит лактозы, то вам следует обратиться к другим блюдам, которые не содержат молочные продукты.

[English] *VLM:* If you are lactose intolerant, you should choose a dish that does not contain dairy products. Picture 1 shows chicken pieces with vegetables, which do not contain lactose. Picture 2 shows a pizza with cheese and other toppings, which may contain lactose. However, if you want to try something that does not contain lactose, you should turn to other dishes that do not contain dairy products.

[Swedish]

User: Vilken rätt av de som visas i dessa bilder bör jag äta om jag är laktosintolerant?

VLM: Om du är laktosintolerant, skulle jag rekommendera att du äter den första bilden, som är en stir-fry med kyckling, redbetor, och pepper. Det har ingen laktos i sig och passar bra för dem som är laktosintoleranta.

[English] *VLM:* If you are lactose intolerant, I would recommend eating the first picture, which is a stir-fry with chicken, beetroot, and pepper. It has no lactose in it and is suitable for those who are lactose intolerant.

[Norwegian]

User: Hvilket måltid av de som vises på disse bildene bør jeg spise hvis jeg er laktoseintolerant?

VLM: For en person som er laktoseintolerant, er det best å spise en av de to alternativene i disse bildene:

1. **Chicken and Vegetable Stir-Fry:** - Chicken er en bra kilde til protein, men det er ikke laktose.
- Vegetable Stir-Fry er full av fiber, vitaminer, og mineraler, og har ingen laktose.
2. **Pepperoni and Sausage Pizza:** - Pepperoni og sausage er ikke laktoseintolerant, men de har ikke så store mengder laktose som fullpizza.

Det er viktig å merke seg at disse alternativene ikke er full substitut for en fullpizza, men de er bedre enn fullpizza for en laktoseintolerant person.

[English] *VLM:* or a person who is lactose intolerant, it is best to eat one of the two options in these pictures:

1. **Chicken and Vegetable Stir-Fry:** - Chicken is a good source of protein, but it's not lactose. - Vegetable Stir-Fry is full of fiber, vitamins, and minerals, and has no lactose.
2. **Pepperoni and Sausage Pizza:** - Pepperoni and Sausage are not lactose intolerant, but they don't have as high amounts of lactose as full pizzas.

It's important to note that these options are not a full substitute for a full pizza, but they are better than full pizza for a lactose intolerant person.

[Hungarian]

User: Melyik ételt kellene enned a képeken látható ételek közül, ha laktózérzékeny vagyok?

VLM: Ha laktózérzékeny vagy, akkor az ételt, amely nem tartalmaz laktózt, kellene enned. A laktózmentes alternatívaként a csirkét és a zöldségeket ajánlanám. A csirkét és a zöldségeket laktózmentesen is elkészítheted, és nem tartalmaznak laktózt.

[English] *VLM:* If you are lactose intolerant, you should eat the food that does not contain lactose. As a lactose-free alternative, I would recommend chicken and vegetables. You can also make chicken and vegetables lactose-free and they do not contain lactose. You can combine chicken and vegetables to get a delicious lactose-free meal.

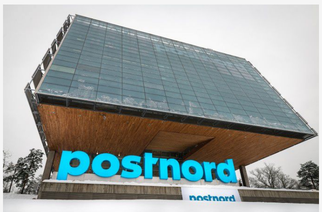
Figure 9: Multilingual generation examples with the model trained with Text Regularization and merged with the original backbone LLM.

Generation Codeswitching Examples

Target Language: Spanish
Dos tacos con salsa y cilantro sobre una tortilla, **served in a paper wrapper and a lime wedge, with a white plate with a blue floral pattern underneath**.

Target Language: Dutch
Het is een scenische boulevard met hotels aan de waterkant, een restaurant met **outdoor seating**, en een **sailboat** aan de promenade..

Target Language: Russian
«Postnord» — это название компании, которая занимается логистикой и дистрибуцией товаров. **Company logo features bold, blue text on a wooden background, which gives a rustic and modern feel**.



Target Language: German
In einem urbanen **Setting** steht ein orange VW Golf auf einem **cobblestone**. Die license plate **reads** ḦH 7293, **and the car has a small antenna on the roof. A person stands behind the car, with a backpack on their back.**

Figure 10: Code switching examples observed when generating the captions of the Crossmodal-3600 dataset images using the Tr-3S-0.05x model. These examples showcase the need of adding a complementary evaluation apart from GlotLid's language detection in order to check language consistency. The words in English are shown in bold letters.

Visual Data			
Dataset	**Field**	**Stage**	**Citation**
LLaVA Pretrain LCS-558K	Image Captions	1	Liu et al. (2023b)
BLIP558K	Detailed Description	1.5	Liu et al. (2024b)
CC3M	Detailed Description	1.5	Liu et al. (2024b)
COCO118K	Detailed Description	1.5	Liu et al. (2024b)
Evol Instruct	Math/Reasoning	1.5	Chen et al. (2024a)
UReader	OCR	1.5	Ye et al. (2023)
SynthDOG	Language	1.5	Kim et al. (2022)
AI2D	Infographics	2/2.5	Kembhavi et al. (2016)
Allava Instruct	General	2/2.5	Chen et al. (2024a)
AOKVQA	General	2/2.5	Schwenk et al. (2022)
Cambrian (filtered)	General	2/2.5	Tong et al. (2024)
Chart2Text	Infographics	2/2.5	Obeid and Hoque (2020)
ChartQA	Infographics	2/2.5	Masry et al. (2022)
ChromeWriting	OCR	2/2.5	-
CLEVR	General	2/2.5	Johnson et al. (2017)
CLEVR-Math	Math/Reasoning	2/2.5	Johnson et al. (2017)
COCO Caption	General	2/2.5	Lin et al. (2014)
Diagram Image2Text	Infographics	2/2.5	-
DocVQA	Infographics	2/2.5	Mathew et al. (2021)
DVQA	Infographics	2/2.5	Kafle et al. (2018)
FigureQA	Infographics	2/2.5	Kahou et al. (2017)
GQA	Math/Reasoning	2/2.5	Hudson and Manning (2019)
Geo170K Align	Math/Reasoning	2/2.5	Gao et al. (2023)
Geo170K QA	Math/Reasoning	2/2.5	Gao et al. (2023)
Geo3K	Math/Reasoning	2/2.5	-
Geometry3K	Math/Reasoning	2/2.5	Lu et al. (2021a)
GeoMVerse	Math/Reasoning	2/2.5	Kazemi et al. (2024)
GeoQA+	Math/Reasoning	2/2.5	Chen et al. (2021)
GEOS	Math/Reasoning	2/2.5	Seo et al. (2015)
Hateful Memes	General	2/2.5	Kiela et al. (2020)
HiTab	Infographics	2/2.5	Cheng et al. (2022)
HME100K	OCR	2/2.5	Yuan et al. (2022)
IAM	OCR	2/2.5	Marti and Bunke (2002)
IconQA	General	2/2.5	Lu et al. (2021b)
IIIT5K	OCR	2/2.5	Mishra et al. (2012)
Infographic VQA	Infographics	2/2.5	Mathew et al. (2022)
InterGPS	General	2/2.5	Lu et al. (2021a)
Image Textualization	General	2/2.5	Pi et al. (2024)
K12 Printing	OCR	2/2.5	-
LLaVA-158K	General	2/2.5	Liu et al. (2023b)
LLaVA-Wild (train)	General	2/2.5	Liu et al. (2023b)
LLaVAR	General	2/2.5	Zhang et al. (2023b)
LRV-Chart	Infographics	2/2.5	Liu et al. (2023a)
LRV-Normal	Math/Reasoning	2/2.5	Liu et al. (2023a)
Magpie Pro	Language	2/2.5	Xu et al. (2024a)

(Continues on next page...)

Dataset	Field	Stage	Citation
MapQA	Math/Reasoning	2/2.5	Chang et al. (2022a)
MathQA	Math/Reasoning	2/2.5	Amini et al. (2019)
MAVIS	Math/Reasoning	2/2.5	Zhang et al. (2024)
OKVQA	General	2/2.5	Marino et al. (2019)
OCR-VQA	OCR	2/2.5	Mishra et al. (2019)
RAVEN	Math/Reasoning	2/2.5	Zhang et al. (2019)
RefCOCO	General	2/2.5	Yu et al. (2016)
Rendered Text	OCR	2/2.5	-
RoBUT	Infographics	2/2.5	Zhao et al. (2023)
ScienceQA	General	2/2.5	Lu et al. (2022)
Screen2Words	Infographics	2/2.5	Wang et al. (2021)
ShareGPT4O	General	2/2.5	Cui et al. (2024)
ShareGPT4V	General	2/2.5	Chen et al. (2025)
ST-VQA	General	2/2.5	Biten et al. (2019)
Super-CLEVR	Math/Reasoning	2/2.5	Li et al. (2023c)
TabMWP	Math/Reasoning	2/2.5	Lu et al. (2023)
TallyQA	General	2/2.5	Acharya et al. (2019)
TextCaps	OCR	2/2.5	Sidorov et al. (2020)
TextOCR-GPT4	OCR	2/2.5	Carter (2024)
TQA	Infographics	2/2.5	Kembhavi et al. (2017)
UniGeo	Math/Reasoning	2/2.5	Chen et al. (2022)
Ureader	Infographics	2/2.5	Ye et al. (2023)
Vision FLAN	General	2/2.5	Xu et al. (2024b)
Visual7W	General	2/2.5	Zhu et al. (2016)
Visual Genome	Math/Reasoning	2/2.5	Krishna et al. (2017)
VisText	General	2/2.5	Tang et al. (2023)
VisualMRC	Infographics	2/2.5	Tanaka et al. (2021)
VizWiz	General	2/2.5	Gurari et al. (2018)
VQARAD	General	2/2.5	Lau et al. (2018)
VQAv2	General	2/2.5	Antol et al. (2015)
VSR	General	2/2.5	Liu et al. (2023b)
WebSight	General	2/2.5	Laurençon et al. (2024)
Spot-the-Diff	Multi-Image	2.5	Jhamtani and Berg-Kirkpatrick (2018)
Birds-to-Words	Multi-Image	2.5	Forbes et al. (2019)
CLEVR-Change	Multi-Image	2.5	Park et al. (2019)
HQ-Edit-Diff	Multi-Image	2.5	Hui et al. (2024)
MagicBrush-Diff	Multi-Image	2.5	Zhang et al. (2023a)
IEdit	Multi-Image	2.5	Tan et al. (2019)
AESOP	Multi-Image	2.5	Ravi et al. (2021)
FlintstonesSV	Multi-Image	2.5	Gupta et al. (2018)
PororoSV	Multi-Image	2.5	Li et al. (2019)
VIST	Multi-Image	2.5	Huang et al. (2016)
WebQA	Multi-Image	2.5	Chang et al. (2022b)
TQA (MI)	Multi-Image	2.5	Kembhavi et al. (2017)
OCR-VQA (MI)	Multi-Image	2.5	Mishra et al. (2019)
DocVQA (MI)	Multi-Image	2.5	Mathew et al. (2021)

(Continues on next page...)

Dataset	Field	Stage	Citation
MIT-StateCoherence	Multi-Image	2.5	Isola et al. (2015)
MIT-PropertyCoherence	Multi-Image	2.5	Isola et al. (2015)
RecipeQA ImageCoherence	Multi-Image	2.5	Yagcioglu et al. (2018)
VISION	Multi-Image	2.5	Bai et al. (2023)
Multi-VQA	Multi-Image	2.5	Li et al. (2023a)
IconQA	Multi-Image	2.5	Lu et al. (2021b)
Co-Instruct	Multi-Image	2.5	Wu et al. (2024)
DreamSim	Multi-Image	2.5	Fu et al. (2023)
ImageCoDe	Multi-Image	2.5	Krojer et al. (2022)
nuScenes	Multi-Image	2.5	Caesar et al. (2020)
ScanQA	Multi-Image	2.5	Azuma et al. (2022)
ALFRED	Multi-Image	2.5	Shridhar et al. (2020)
ContrastCaption	Multi-Image	2.5	Jiang et al. (2024)
VizWiz (MI)	Multi-Image	2.5	Gurari et al. (2018)
ScanNet	Multi-Image	2.5	Dai et al. (2017)
COMICS Dialogue	Multi-Image	2.5	Iyyer et al. (2017)
NLVR2	Multi-Image	2.5	Suhr et al. (2019)
NExT-QA	Video	2.5	Xiao et al. (2021)
Ego-4D	Video	2.5	Grauman et al. (2024)
YouCook2	Video	2.5	Zhou et al. (2018)
ActivityNet	Video	2.5	Yu et al. (2019)
Charades	Video	2.5	Sigurdsson et al. (2016)
ShareGPT4Video	Video	2.5	Chen et al. (2024c)

Table 10: English only visual datasets used throughout this work. The same data as proposed in LLaVA-OneVision has been used.

Text-Only Data			
Dataset	**Field**	**Stage**	**Citation**
Aya Dataset	General	1.5/2/2.5	Singh et al. (2024)
CoqCat	Conversation QA	1.5/2/2.5	Gonzalez-Agirre et al. (2024)
Databricks Dolly 15k	General	1.5/2/2.5	Conover et al. (2023)
Databricks Dolly 3k CA	General	1.5/2/2.5	-
FLORES-200 (Instructions)	Translations	1.5/2/2.5	Costa-jussà et al. (2024)
MentorCA	General	1.5/2/2.5	-
No Robots	General	1.5/2/2.5	Rajani et al. (2023)
OASST	General	1.5/2/2.5	Köpf et al. (2023)
OASST-CA	General	1.5/2/2.5	-
RAG Multilingual	General	1.5/2/2.5	-
Tower-Blocks-v0.1	Text-Insight	1.5/2/2.5	Alves et al. (2024)

Table 11: Multilingual text-only datasets added throughout the visual instruction process.