R2-MultiOmnia: Leading Multilingual Multimodal Reasoning via Self-Training

Leonardo Ranaldi **Federico Ranaldi** **Giulia Pucci**
School of Informatics, University of Edinburgh, UK
Human-centric ART, University of Rome Tor Vergata, IT
University of Aberdeen, UK
{first_name.last_name}@ed.ac.uk

Abstract

Reasoning is an intricate process that transcends both language and vision; because of its inherently modality-agnostic nature, developing effective multilingual and multimodal reasoning capabilities is a substantial challenge for Multimodal Large Language Models (MLLMs). They struggle to activate complex reasoning behaviours, delivering step-wise explanation, questioning and reflection, particularly in multilingual settings where high-quality supervision across languages is lacking. Recent works have introduced eclectic strategies to enhance MLLMs' reasoning; however, they remain related to a single language.

To make MLLMs' reasoning capabilities aligned among languages and improve modality performances, we propose **R2-MultiOmnia**, a modular approach that instructs the models to abstract key elements of the reasoning process and then refine reasoning trajectories via self-correction. Specifically, we instruct the models producing multimodal synthetic demonstrations by bridging modalities and then self-improving their capabilities. To stabilise learning and the reasoning processes structure, we propose Curriculum Learning Reasoning Stabilisation with structured output rewards to gradually refine the models' capabilities to learn and deliver robust reasoning processes. Experiments show that **R2-MultiOmnia** improves multimodal reasoning, gets aligned performances among the languages approaching strong models.

1 Introduction

Reasoning is a fundamental cognitive ability that allows humans to tackle complex problems, make critical decisions, and adapt to their environment. It transcends language and vision, functioning independently through visual intuition, symbolic manipulation, or spatial representation, before finding expression in written or spoken form (Altmann, 2001; Johnson-Laird, 2010; Cuskley et al., 2024).

In the context of Multimodal Large Language Models (MLLMs), reasoning remains notably challenging. Recent approaches, such as Multimodal Chain-of-Thought *et inter alia* (Zhang et al., 2024), attempt to instil structured multimodal reasoning within these models by operating on manually crafted resources and using supervised fine-tuning (SFT). Although they exhibit performance gains, these approaches often lead to manufactured rationales—superficial approximations lacking key cognitive processes such as questioning and reflection—limiting the models' efficacy on complex multimodal reasoning tasks (Yin et al., 2024). To overcome these limitations, Huang et al. (2025); Yang et al. (2025) leverage Reinforcement Learning (RL) as a powerful post-SFT strategy, following the R1 paradigm (DeepSeek-AI et al., 2025), aiming to boost the self-emergence of complex reasoning capacities. Yet, developing genuine human-like reasoning remains pivotal to advancing MLLMs' multimodal reasoning capabilities. In parallel to the modality-specific limitations, a layer of complexity is introduced by the language. Indeed, even though reasoning is fundamentally independent of language, significant performance gaps arise due to the imbalance of multilingual training data. LLMs trained predominantly on English-centric data exhibit significantly stronger performance in English language reasoning (Ranaldi et al., 2024a). These imbalances limit the global applicability of MLLMs, underlining the pressing need for culturally inclusive strategies.

To improve MLLMs' reasoning capabilities and align them among the languages, we propose **R2-MultiOmnia**, a modular approach that instructs the models to structure the reasoning abstractly and self-refine their capabilities likewise for all languages. Complementing the foundations works proposed by Yang et al. (2025); Huang et al. (2025), we extend multimodal capabilities beyond English and, starting from the concept that reasoning is

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8220–8234

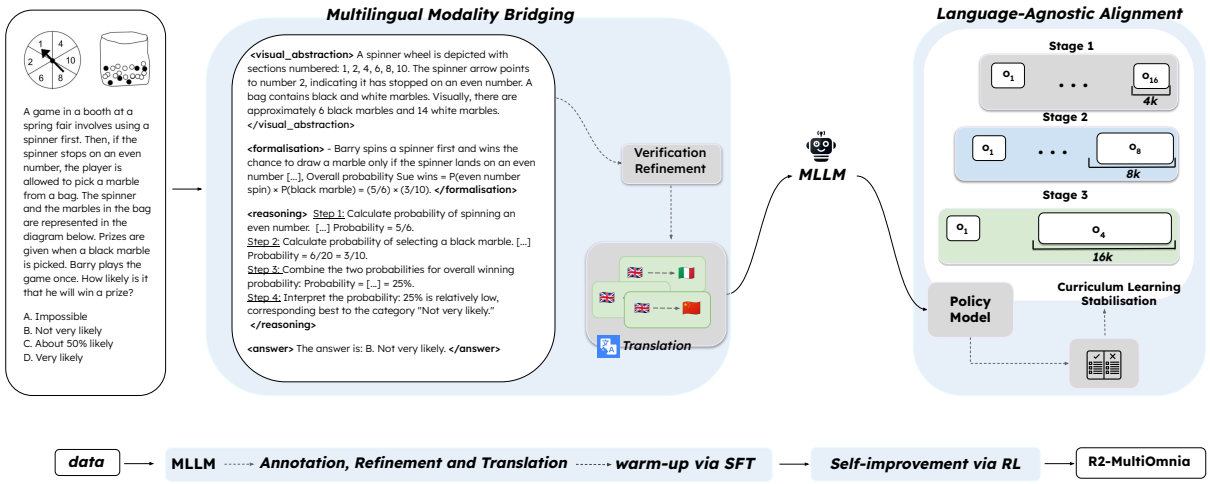

Figure 1: Overview of **R2-MultiOmnia** framework. The modular architecture consists of two stages: *Multilingual Modality Bridging*, which instructs the MLLM to generate structured multimodal reasoning demonstrations across languages by abstracting visual and textual cues into step-wise rationales; and *Language-Agnostic Reasoning Alignment*, which refines these capabilities via RL based on Curriculum Learning Stabilisation.

language-agnostic (Ranaldi and Pucci, 2025), we propose a framework that disentangles logical reasoning from content and then reaches the final solution respecting input languages.

R2-MultiOmnia is composed by *Multilingual Modality Bridging* followed by *Language-Agnostic Reasoning Alignment*. In the *Multilingual Modality Bridging* phase, we instruct an MLLM to deliver rationales from multimodal image-text pairs, which explicitly abstract key elements to answer the question through vision descriptions and a structured step-wise reasoning process, exposing clear vision information in a language format. We then feed these enriched reasoning texts back into the MLLM to double-check and get the answer. We refine these annotations through rejection sampling based on rule-based criteria. The resulting dataset contains 10K multimodal reasoning samples, which, to make the training fair and multilingual, are in 8 different languages. In the *Language-Agnostic Reasoning Alignment* phase, we operate via Group Relative Policy Optimisation (GRPO) (Shao et al., 2024) to enhance the reasoning capability of the warm-up model. Following Huang et al. (2025), we avoid the overthinking phenomenon, and propose a Curriculum Learning heuristic, incorporating a formatting result reward function. This approach enables **R2-MultiOmnia** to compress reasoning steps early in the RL, internalising correct reasoning methods while progressively extending its reasoning span over time to tackle complex problems properly.

We conducted an extensive empirical evaluation and ablation studies to demonstrate the robustness of the proposed approach in achieving improvements in multimodal reasoning tasks, resulting in consistent performance gains across languages on the evaluated benchmarks. These results support the following key findings and conclusions:

- Structuring multilingual reasoning in MLLMs through a cognitive-inspired approach and self-refinement enhances reasoning capabilities. **R2-MultiOmnia** combines strategic SFT demonstrations with incremental RL, enabling a systematic strategy to furnish models with the ability to abstract logical reasoning from linguistic and visual content.

- The approach demonstrates increased robustness and consistency of performance in multilingual settings. We propose an instruction strategy inspired by human reasoning, which, unlike Multimodal CoT-based methods, is not exposed to content bias and is inherently language-agnostic. We employ structured demonstrations to conduct an initial warm-up phase, followed by incremental self-refinement, enabling the model to acquire increasingly complex reasoning behaviours.

To the best of our knowledge, our work is the first to apply heuristics to instruct MLLMs to solve a multilingual task using SFT empowered via RL-based reasoning approaches.

2 Method

Reasoning across languages and modalities presents distinct challenges for MLLMs, as it requires high-level abstraction and the integration of both visual and linguistic information across diverse languages and modalities. To address these limitations, we propose **R2-MultiOmnia**, a modular framework that disentangles content from reasoning and delivers aligned reasoning trajectories. Our method consists of two main stages: *Multilingual Modality Bridging* (§2.1) and *Language-Agnostic Reasoning Alignment* (§2.2).

2.1 Multilingual Modality Bridging

To enable robust multimodal reasoning aligned among the languages, we construct a high-quality multimodal dataset, developed through a human-inspired methodology grounded in semi-structured dynamic representations that furnish interpretable constructs to distil the core logical elements of reasoning, integrating complex cognitive processes and visual information within demonstrations through a Modality Bridging process.

Thus, starting from a multimodal dataset defined by $\mathcal{L} = \{(x_i, a_i)\}_{i=1}^{N}$ each tuple comprises: x_i (an image-text input) and a_i (the final answer). We instruct a MLLM to abstract the key visual element to answer the question and deliver a caption and a rationale, i.e., ${c_i}^N$ and ${y_i}^N$. Hence, to improve the quality of data, we follow Huang et al. (2025) and feed these texts back into the MLLM to get robust demonstrations. We instruct models structuring the reasoning process around strategic logical components, enabling more generalisable and interpretable inference across languages and modalities (Ranaldi and Pucci, 2025). These demonstrations, ${r_i}^N$, are then filtered via rejection sampling by rule-based criteria:

$$\mathcal{L}^* = \{(x_i, r_i, a_i) \in \mathcal{L} \mid F(x_i, r_i, a_i) = 1\} \quad (1)$$

where $F(\cdot)$ is a binary filter accepting only structurally and logically coherent samples as reported in Appendix K.

Supervised Fine-Tuning Objective. The filtered corpus $\mathcal{L}^*$ is employed for supervised fine-tuning (SFT), by minimising the objective:

$$\mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}_{(x,r)\sim\mathcal{L}^*} \left[\log f_\theta(r \mid x)\right] \quad (2)$$

where f_θ denotes the model's output distribution, x is the input, and r is the reasoning demonstration.

2.2 Language-Agnostic Reasoning Alignment

Following the initial warm-up phase (§2.1), we introduce a dedicated alignment procedure to reinforce reasoning capabilities. Unlike approaches that rely on language-specific cues or strict structural constraints, we employ Group Relative Policy Optimisation (GRPO) (Shao et al., 2024) to enable consistent reasoning behaviour across both languages and modalities.

Composite Reward Formulation. To direct the alignment of multilingual and multimodal reasoning, we define a composite reward function that aggregates multiple constraints. Each sampled response o_i receives a total reward:

$$R_i = \sum_{k=1}^{K} w_k \, r_k(o_i) \quad (3)$$

where $r_k(o_i)$ denotes the k-th constraint-specific reward component and w_k is its corresponding weight. This formulation flexibly combines factual accuracy, format compliance, structural integrity, and robustness to minor deviations. Implementation details of all constraint-based reward components are provided in Appendix K.

Group-Normalised Advantage. To encourage diversity and robust policy learning, rewards are normalised within each group of n sampled outputs as follows:

$$A_i = \frac{R_i - \text{mean}(\mathbf{R})}{\text{std}(\mathbf{R})} \quad (4)$$

where $\mathbf{R} = \{R_1, \ldots, R_n\}$ is the vector of group rewards. This group-relative advantage amplifies intra-group differences and stabilises training dynamics.

GRPO Policy Objective. The model parameters θ are optimised to maximise the expected log-likelihood of generated responses, weighted by their group-normalised advantages and regularised via a Kullback–Leibler divergence penalty:

$$J(\theta) = \mathbb{E}\left[\frac{1}{n}\sum_i \log \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i - \beta \, D_{\text{KL}}\right] \quad (5)$$

where the expectation $\mathbb{E}[\cdot]$ is taken over questions q and groups of n responses $\{o_i\}$ sampled from the previous policy $\pi_{\theta_{\text{old}}}$ given q. A_i denotes the group-normalised advantage for response o_i, D_{KL} denotes the Kullback–Leibler divergence between

the current policy π_θ and the reference policy π_{ref}, and β controls the strength of the regularisation.

KL Annealing. To promote training stability and facilitate the emergence of deeper reasoning patterns, the KL penalty coefficient β is dynamically adjusted using a cosine annealing schedule:

$$\hat{\beta} = \frac{\beta}{2}\left(1 + \cos\left(\pi \frac{T_{\text{cur}}}{T_{\max}}\right)\right) \tag{6}$$

where T_{cur} and $T_{\max}$ denote the current and maximum training iteration, respectively. This phase ensures that the model learns robust reasoning strategies that generalise across languages, without overfitting to language-specific artifacts or shallow format cues.

2.3 Curriculum Learning Stabilisation

Given the instability and shortcut risks of RL-based training, we integrate a structured Curriculum Learning strategy inspired by the paradigm introduced in (Shao et al., 2024).

Progressive Length Constraint. We incrementally increase the allowable reasoning length ℓ_k at each stage k:

$$\ell_k = \ell_0 \cdot 2^{k-1} \tag{7}$$

and enforce $\text{len}(\text{r}) \leq \ell_k$. where ℓ_0 is the initial length. This constraint is enforced such that $\text{len}(\text{r}) \leq \ell_k$ for generated completions at stage k.

Structured Output Rewards. In addition to the rewards defined in §2.1, we introduce explicit constraints at each curriculum stage: a reward is assigned only if the output matches the required configuration (defined in Appendix K), while excessive reasoning length is penalised as

$$r_{\text{length}} = -\lambda \cdot \max(0, \text{len}(\text{r}) - \ell_k). \tag{8}$$

where λ controls the penalty strength. These additional constraints ensure that outputs remain concise and well-structured throughout the training process.

Stage-wise Training Objective. Let S denote the total number of Curriculum stages. At each stage $s \in \{1, 2, \ldots, S\}$, completions are restricted to the output space $O^{(s)} = \{o : |o| \leq \ell_s\}$. The training objective for stage s is:

$$J^{(s)}(\theta) = \mathbb{E}\left[\frac{1}{n_s}\sum_i \log \frac{\pi_\theta(o_i^{(s)}|q)}{\pi_{\theta_{\text{old}}}(o_i^{(s)}|q)} A_i^{(s)} - \beta\, D_{\text{KL}}\right] \tag{9}$$

where n_s is the number of sampled outputs at stage s, and $A_i^{(s)}$ is the group-normalised advantage as defined previously, reflecting both format and length rewards.

Iterative Training. These optimisation steps are iteratively repeated, with reward structure and length constraints relaxed across subsequent curriculum stages as the model demonstrates improved stability and performance. As represented in Figure 2, we conduct three phases, changing the parameters to ensure both conciseness and the eventual emergence of more complex reasoning strategies. Hence, we set the parameters as $S = 3$, $L_s \in \{4K, 8K, 16K\}$, and $G_s \in \{16, 8, 4\}$. This curriculum-driven stabilisation is essential for preventing degenerate solutions and promoting the emergence of grounded, generalisable reasoning trajectories across both modalities and languages.

3 Experiments

We aim to propose a method for improving the reasoning abilities of MLLMs beyond language boundaries. To this end, we evaluate our approach on multilingual multimodal reasoning tasks but on monolingual as well. Moreover, we include language-based tasks to systematically assess the extent to which our method enhances language-agnostic reasoning and mitigates content bias. We use models reported in §3.1, trained as described in §3.2 and evaluated on tasks (§3.3), using the configurations described in §3.4.

3.1 Models

We conduct our study using two different models to facilitate comparison. We use Qwen2.5-VL-Instruct (Bai et al., 2025) and DeepSeek-VL2 (Wu et al., 2024) precisely 3B for the first and 2.8B version for the second. Furthermore, to demonstrate the scalability and effectiveness of our approach on additional models, we present further evaluations on larger versions, specifically the 7B and 4.5B versions.

3.2 Training Methods

As introduced in §2, we conduct two stages: *Multilingual Modality Bridging* (§2.1) and *Language-Agnostic Reasoning Alignment* (§2.2). Specifically, we follow standard practice and perform a warm-up phase based on an SFT step using demonstrations constructed as discussed in §2.1 on data reported in §3.3.1. Then, we conduct the self-refinment by

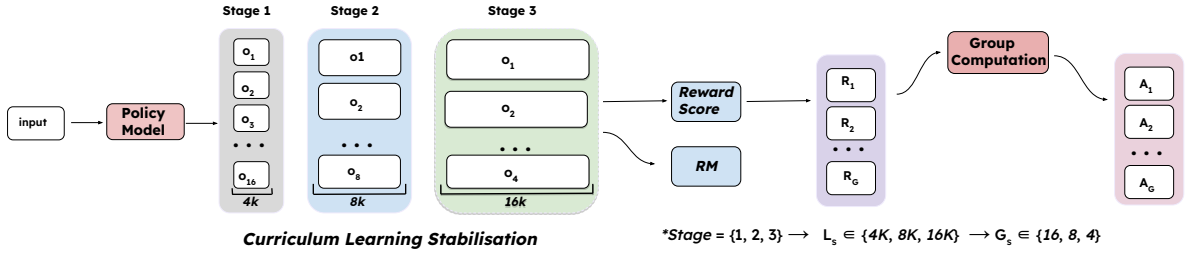

Figure 2: Curriculum Learning Stabilisation used during §2.2 and introduced in §2.3. At each stage, context length is progressively increased (4K, 8K, 16K tokens), with corresponding group sizes of 16, 8, and 4. The GRPO reward is based on the formatting result function detailed in Appendix K.

applying RL optimisation algorithm (GRPO, as presented in §2.2).

Supervised Fine-tuning Regarding the SFT phase, we tune the model for one epoch (warm-up), using learning rates specified according to the model configuration, as detailed in Appendix H.

Preference Optimisation RL We employ the HuggingFace trainer ($GRPO_{trainer}$) to ensure reproducibility. We set the learning rate to *5e-7* and β to 0.04. The sampling temperature is set to 0.9 following the recommended practice and the generation configuration for each stage as described in §2.3. The optimisation process is set at a maximum of 1000 for each stage. Details in Appendix H.

3.3 Data

3.3.1 Training Set

We employ synthetic demonstrations to train models to solve tasks following the two phases in Figure 1, complementing (Huang et al., 2025). To conduct comparative experiments, we utilise the training introduced by Xu et al. (2025). In contrast to this latter, our strategy employs the instruction strategy reported in Appendix A.

Multilingual Demonstrations To produce consistent results, we make gold parallel annotations in 8 different languages. In particular, behind the annotation process, as outlined in Appendix A, we assess the quality of the demonstrations using rule-based heuristics (details are provided in Appendix D). Then, on the filtered annotations, we perform a translation and extract a total of 10k balanced demonstrations for all languages (full details in Appendix G). The remaining demonstrations are used for the RL phase. Since this process may be influenced by biases introduced by either the translation system or the annotation procedure within the experimental setting, we discuss the different dynamics between translation and annotation.

3.3.2 Evaluation Set

To study the multimodal reasoning performances, we operate via xMMMLU, M3Exam, MAXM, MathVision, MathVista and introduce Multi-MathVista. Then we use two language-based reasoning datasets, i.e., mGSM and mGSM-Symbolic.

Multimodal Reasoning We use four multitask multilingual benchmark: xMMMLU (Yue et al., 2025), M3Exam (Zhang et al., 2023) and MAXM (Changpinyo et al., 2023). Then, to asses the logical reasoning abilities we employ two monolingual mathematical task: MathVista (Lu et al., 2024) and MathVision (Wang et al., 2024). Finally, we produce an extended version of MathVista, namely xMathVista, in 7 different languages to challenge models in multilingual scenarios (Data are available on GitHub at the following link)

Language-based Reasoning We use the extension of GSM8K, i.e., Multilingual Grade School Math (mGSM). In original cases, the authors proposed a benchmark of English mathematical problems with the following structure: a word problem in natural language and a corresponding numerical answer. For both versions, a subset of instances from the official list of examples was translated into 11 different languages, maintaining the structure of the input and output.

mGSM-Symbolic Mirzadeh et al. (2024) improved GSM8k (MGSM ancestor) by proposing GSM-Symbolic. This introduces symbolic patterns in GSM8k that complicate the task and disadvantage the LLMs' capabilities. Ranaldi and Pucci (2025) propose mGSM-Symbolic, the multilingual GSM-Symbolic extension.

Evaluation Metrics To evaluate the performance, we use the accuracy of the final answer, assessed through a flexible match between the generated response and the ground truth.

3.4 Experimental Setup

In the main discussion evaluate the models introduced in §3.1 using Qwen2.5-VL-3B- Instruct (Qwen2.5-VL) and DeepSeek-VL2-Small (DeepSeek-VL2) as backbone models. We then report the performances of different closed-/open-source models using the following configurations:

Baseline We prompt without any tuning.

Training We assess the impact of the proposed approaches by conducting different configurations:

- **SFT, RL and SFT+RL** We tune the models using the demonstrations released by Xu et al. (2025) as detailed §3.3.1. Hence, we conduct SFT, RL, and SFT (as a warm-up) and RL as detailed in §3.2. Regarding the SFT phase, we propose using the released demonstrations, specifically those from LLaVA-CoT. The SFT+RL configuration, on the other hand, aims to expand the method proposed by Yang et al. (2025) for multiple languages. Hence, we adapted the resource in accordance with the practices for R2-MultiOmnia.
- **R2-MultiOmnia** We warm-up the models using the synthetic demonstrations generated via our prompting strategy and conduct the self-training strategies using both policies as introduced in §2. Generally, this is similar to the SFT+RL configuration, however: *(i)* the used demonstrations have different structure and *(ii)* it employs an incremental stabilisation RL strategy (§2.3).

4 Results & Discussion

Reasoning is not constrained to a specific modality (textual, visual) not to a single natural language, yet it is a matter of fact and operates through abstract structures, the highest form of act common to all modalities. **R2-MultiOmnia** transferred this concept to the reasoning refinement of MLLMs. Specifically, it leads to a modality-agnostic reasoned solution, directing MLLMs in delivering robust reasoned pathways to reach the final solution. The proposed training approach enables the models to achieve structured reasoning trajectories by obtaining results approaching those of the strongest models (§4.1). Self-training is definitely more performant than single SFT or RL and allows the models to achieve better results with significantly fewer training data (§4.2). The emerging dynamics between languages demonstrate the scalability of the proposed method by tailoring and obtaining parallel gains between languages (§4.3). In in-depth studies, we perform ablation studies to prove the effectiveness, scalability and robustness of the proposed approach (§5).

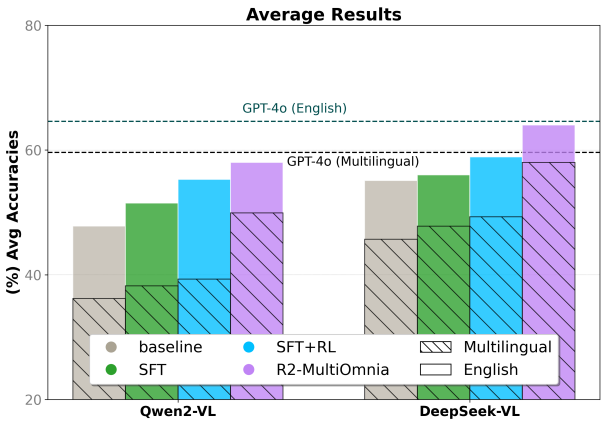

Figure 3: Average results on English and Multilingual using **SFT**, **SFT+RL** and **R2-MultiOmnia** (§3.4).

4.1 Reasoning in Multidimensional Spaces

Figure 3 shows that models instructed via the **R2-MultiOmnia** framework achieve strong results in all proposed multimodal tasks, outperforming definitely alternative tuning approaches such as SFT and SFT+RL and approaching the results obtained by state-of-the-art models such as GPT-4o. Table 1 shows in detail that the two models instructed via **R2-MultiOmnia** strategy perform consistently better in multimodal tasks when compared with other models. In particular, when compared with stronger models, they get promising results (see the deltas in Table 1). When compared with the respective baselines, they outperform by 38.2% Qwen2.5-VL and by 26.9% DeepSeek-VL2. The framework achieves strong results in monolingual-multimodal tasks as well, confirming the actual benefits of the modular structure of the proposed framework based on *Multilingual Modality Bridging* and *Language-Agnostic Alignment* (as presented in §2). However, to gain a clearer understanding of the respective contributions of the two components to the final results and to analyse the dynamics and advantages that arise during the tuning process, in §4.2 we examine these elements in greater detail.

Models	Multilingual						Monolingual			Average	
	xMMMU		M3Exam		MaXM		xMathVista		MathVision		
	en	mul	en	mul	en	mul	mul	en	en	en	mul
GPT4-o	**69.1**	**58.3**	**68.0**	**61.0**	60.7	**65.4**	**58.9**	63.8	**30.6**	**65.4**	**60.5**
Gemini-1.5-Pro	36.2	31.5	32.3	29.0	56.4	63.5	54.3	63.9	19.2	47.2	44.6
Pangea-7b	45.7	43.7	61.4	42.1	58.0	45.5	46.3	56.2	16.6	55.3	44.4
Qwen2.5-VL	34.2	33.0	46.0	37.5	52.0	24.8	49.4	59.2	20.0	37.8	36.2
DeepSeek-VL2	43.7	40.7	60.4	41.1	54.6	52.3	52.6	61.9	22.2	55.1	46.7
(**SFT**) Qwen2.5-VL	39.8	37.2	50.2	47.8	56.1	37.9	51.7	60.4	21.8	51.6	41.5
(**SFT+RL**) Qwen2.5-VL	45.6	41.0	52.6	51.6	59.2	44.0	54.2	62.6	24.4	55.0	45.3
(**SFT**) DeepSeek-VL2	47.3	43.2	59.7	44.3	54.6	54.8	53.0	62.5	25.0	56.0	48.8
(**SFT+RL**) DeepSeek-VL2	49.0	49.6	64.2	47.8	59.0	56.7	54.9	63.5	26.8	58.9	52.3
R2-Qwen2.5-VL	51.2	45.3	56.8	58.6	60.0	54.4	56.3	64.0	26.5	58.0	49.9
Δ over GPT-4o	-17.9	-13.0	-9.4	-2.4	-2.6	-11.0	-2.6	**-0.2**	-4.1	+0.4	-10.6
R2- DeepSeek-VL2	54.9	53.0	65.7	60.8	**61.7**	59.9	58.2	**65.2**	30.0	61.8	57.8
Δ over GPT-4o	-14.2	-5.3	-2.3	**-0.2**	**+1.0**	-5.5	**-0.7**	**+1.4**	**-0.6**	-3.6	-2.7

Table 1: Overall performance on proposed multimodal benchmarks (§3.1). The best-performing open model on each task is in **bold** and in Δ the differences with GPT-4o, which for most tasks represents the SOTA model.

4.2 R2 Training Strategy

The training processes, based on a modular strategy founded by *Multilingual Modality Bridging* (MMB) and *Language-Agnostic Alignment* (LAA), deliver robust models by consistently increasing performance and employing less training data than other approaches. Table 2 shows the improvement gained from MMB and LAA (which are the founding parts of MultiOmnia) over the single phases and the first phase (i.e., MMB) and RL based on standard GRPO. To better interpret the effect of the proposed approaches, we now analyse these components in detail.

Model	Strategy	xMMMLU	xMathVista
Qwen2.5-VL	MMB	40.2	51.4
	RL	35.0	47.3
	LAA	36.8	49.6
	MMB+*RL*	41.5	52.3
	MultiOmnia	**45.3**	**56.3**
DeepSeek-VL2	MMB	44.7	54.8
	RL	43.5	52.9
	LAA	45.8	53.7
	MMB+*RL*	47.5	55.0
	MultiOmnia	**53.0**	**58.2**

Table 2: Average accuracies achieved performing *Multilingual Modality Bridging* (MMB), *Language-Agnostic Alignment* (LAA), standard reinforcement learning using GRPO (RL).

The role of RL The results in Table 2 demonstrate that the *Language-Agnostic Alignment* (LLA) is effective and yields gains for both models. For instance, on a multilingual multimodal mathematical task (xMathVista) LLA gain +4.0 points and +3.2 points for Qwen2.5-VL and DeepSeek-VL2 when compared with standard RL based on GRPO. To get a better understanding, we compare different settings of LAA with Curriculum Learning CL and without, leaving the length of the generations fixed (4K, 8K or 16K). Figure 4 demonstrates the impact of CL on model performance, both when output length is held constant and when it is incrementally increased; yet, when there is the incremental setting, it is possible to observe significantly better performances. These results confirm that the proposed strategy is optimal.

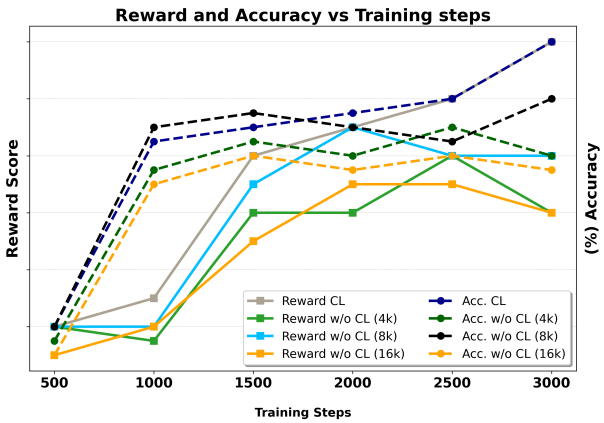

Figure 4: Reward Score and average Accuracy on DeepSeek-VL2 over xMathVista after MMB (the first warm-up phase) using incremental output length rewards and fixed (w/o CL).

The impact of Training Demonstrations Current alignment strategies typically rely on demonstrations produced by an expert model from the same architectural family, highlighting the greater influence of in-family learning on student model performance (Ranaldi and Freitas, 2024a,b). In the MMB (i.e., warm-up phase), we employ self-generated demonstrations. To evaluate the robustness of our approach, we perform a comparative analysis using demonstrations generated by alternative models. As shown in Figure 5, GPT -4 generations lead to more performant models; however, these require an annotation budget, i.e., costs associated with the use of APIs. On the other hand, self-generated annotations have a significantly reduced cost and very good performance (note the differences in Figure 5 around two percentage points).

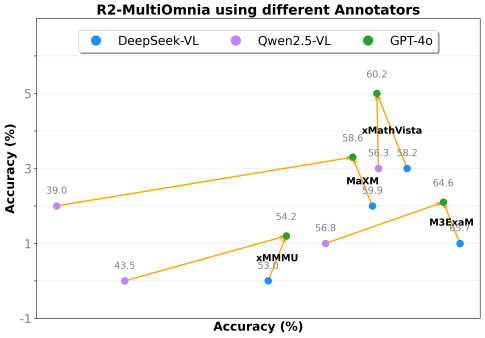

Figure 5: Average results on multimodal multilingual tasks (§3.3) using demonstrations self-generated (as original **R2-MultiOmnia**) and generated from GPT-4o.

4.3 Language Improvements

Reasoning is modality-agnostic; nevertheless, natural language is used to deliver and externalise reasoning processes. In the previous sections, we observed the performance achieved by the proposed framework in multimodal tasks. To show R2-MultiOmnia's multidimensional functionality, we now evaluate this framework on text-based multilingual mathematical reasoning tasks introduced in §3.3.2. Table 3 shows the improvement over baseline models and over other stringer models (GPT-4o and Gemini-1.5-Pro). The models tuned via the proposed framework achieve higher average accuracies. Although they do not outperform in all cases, GPT-4o, they definitely get improvement when compared to baseline models and when compared with Gemini-1.5-Pro. This demonstrates that R2-MultiOmnia: *(i)* separating reasoning and content both in vision-based and text-based reasoning tasks and obtains substantial benefits both in the more robust models capable of performing this abstraction step and *(ii)* in the smaller models (as the proposed ones) consistently improves results by providing performance alignment between languages.

Models	MGSM	MGSM-SYMBOLIC
GPT-4o	70.9(86.8)	67.3(83.2)
Gemini-1.5-Pro	69.5(77.5)	60.2(75.3)
Qwen-2.5-VL	55.7(64.7)	52.9(59.8)
+SFT+RL	60.2(67.3)	56.6(66.5)
+R2-MultiOmnia	69.8 (83.5)	66.6 (76.8)
DeepSeek-VL2	60.5(66.2)	55.8(61.0)
+SFT+RL	66.4(70.5)	61.2(69.3)
+R2-MultiOmnia	**70.9** (85.0)	**69.8** (75.0)

Table 3: Average performances on MGSM and MGSM-SYMBOLIC. In brackets are the performances for the English subset.

5 Additional Studies

Scaling models and data To demonstrate the impact of the proposed framework on scalability and operability for larger models, we operated with models of the same family, scaling the number of parameters. Specifically, we operated via Qwen-7B-2.5-VL and YDeepSeek-VL2-Small (named as DeepSeek-VL2-S), adopting the same tuning approach proposed for Qwen-3B-2.5-VL and DeepSeek-VL2-Tiny (see §3.4). Table 4 shows that the models evaluated without tuning (baseline) outperform their smaller-parameter counterparts (see values in brackets). However, the R2-MultiOmnia framework also proves effective on these models, enhancing their base performance. Moreover, the bracketed values, which indicate the performance gap with the smaller models trained through R2-MultiOmnia, reveal that such differences are significantly smaller compared to the baseline models.

Models	XMMMLU	XMATHV	M3EXAM
GPT-4o	58.3	58.9	61.0
Gemini-1.5-Pro	31.5	54.3	29.0
Qwen-7B-2.5-VL	40.9 (-6.7)	54.5 (-4.6)	44.8 (-7.3)
+R2-MultiOmnia	49.1 (-3.8)	51.8 (-1.2)	60.3 (-1.7)
DeepSeek-VL2-S	44.3 (-3.6)	56.9 (-4.3)	45.0 (-3.9)
+R2-MultiOmnia	55.1 (-2.1)	59.2 (-1.0)	61.6 (-0.8)

Table 4: Performances of bigger versions of the models used in the previous experiments.

6 Related Work

The performance of large language models (LLMs) on reasoning tasks has been shown to improve significantly when they are guided to simulate human-like cognitive processes and follow stepwise reasoning strategies. In response, a growing body of research has focused on developing methods to structure and enhance LLM reasoning. These approaches often involve human-designed formats that scaffold outputs into interpretable steps. Some examples include CoT, plan-based methods such as Tree-of-Thought (Lightman et al., 2023; Ranaldi et al., 2024b), and the construction of complex Supervised-Fine-Tuning (SFT) datasets (Ranaldi et al., 2025d,e).

Recent advances have shown that reinforcement learning (RL) with structured rewards can foster sophisticated, human-like reasoning in LLMs (DeepSeek-AI et al., 2025), enhancing their performance on complex tasks. However, applying these methods to Multilingual LLMs (MLLMs) remains largely unexplored and raises several open challenges. MLLMs process inputs from various modalities by translating them into textual representations, which are subsequently analysed by LLM architectures. This technique has consistently delivered strong performance across multiple vision-related understanding tasks, as evidenced by numerous recent studies (Liu et al., 2024). Motivated by these successes and parallel advancements in reasoning for MLLMs, considerable effort has been directed towards enhancing the reasoning capabilities. Notably, approaches employing Multimodal CoT prompting (Zhang et al., 2024) and the creation of data that explicitly incorporates structured reasoning (Yao et al., 2024) have gained traction. Recent approaches that integrate tuning initialisation with targeted RL training have shown promise in developing richer reasoning and enhancing performances (Zhang et al., 2024). While useful, these methods often fail to capture key cognitive aspects of human reasoning, such as questioning and reflection, limiting their effectiveness on complex tasks. In response, we propose a strategy to instruct models to abstract reasoning logic from image and text content. Building on our prior work (Ranaldi et al., 2023, 2025b,c), we extend these insights to multilingual and multimodal contexts.

Despite utility, the reasoning generated through these methods frequently displays deficiencies, particularly in reflecting natural cognitive processes integral to human reasoning, such as critical questioning, reflective analysis, and iterative inspection. Hence, their overall effectiveness in complex problem-solving scenarios remains limited. To this end, we propose a strategy to instruct models to abstract reasoning logic from image and text content. Building on our prior work (Ranaldi and Pucci, 2025), we go beyond language by transferring the previous findings into multimodal contexts.

7 Conclusion

Although reasoning inherently transcends modality and language, MLLMs typically exhibit inconsistent performance due to modality-specific and linguistic biases in training data. To enhance multilingual and multimodal reasoning capabilities equitably, we introduced **R2-MultiOmnia**, a modular approach instructing models to abstract reasoning processes independent of specific languages or modalities, followed by structured self-correction. Leveraging Multilingual Modality Bridging, we synthesised modality-neutral reasoning demonstrations, enabling fair proficiency across languages. Our Language-Agnostic Reasoning Alignment, enhanced via Curriculum Learning, significantly improved reasoning consistency and depth. Empirical results confirmed that R2-MultiOmnia reduces cross-lingual performance gaps, offering robust, precise multilingual reasoning. These outcomes underscore the extent to which structured abstraction and incremental RL refinements advance multilingual, modality-agnostic reasoning capabilities.

Limitations & Future Work

In future developments, we plan to extend the analysis to other models and take care of the efficiency of the methodologies used. Concerning tasks, we will expand the analysis to tasks of cultural commonsense reasoning. Instead, we explore annotation strategies that require fewer computational and data resources. Considering the limitations posed by existing evaluation tools and the financial costs associated with external model APIs, our experimentation necessarily focused on a limited subset of tasks and linguistic contexts, thereby addressing only a fraction of the world's language diversity. Moreover, future work should explore models tailored to specific languages. While such resources are currently limited, their expected growth will enable deeper multilingual research, facilitating more comprehensive research in multilingual modelling.

References

Gerry T. M. Altmann. 2001. The language machine: Psycholinguistics in review. *British Journal of Psychology*, 92(1):129–170.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. MaXM: Towards multilingual visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682, Singapore. Association for Computational Linguistics.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models.

Philip N. Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models.

Leonardo Ranaldi and Andre Freitas. 2024a. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Leonardo Ranaldi and Andre Freitas. 2024b. Self-refine instruction-tuning for aligning reasoning in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2347, Miami, Florida, USA. Association for Computational Linguistics.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025a. Multilingual retrieval-augmented generation for knowledge-intensive task.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2025b. When natural language is not enough: The limits of in-context learning demonstrations in multilingual reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7369–7396, Albuquerque, New Mexico. Association for Computational Linguistics.

Leonardo Ranaldi and Giulia Pucci. 2023. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

Leonardo Ranaldi and Giulia Pucci. 2025. Multilingual reasoning via self-training. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11566–11582, Albuquerque, New Mexico. Association for Computational Linguistics.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2024a. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7961–7973, Bangkok, Thailand. Association for Computational Linguistics.

Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2024b. A tree-of-thoughts to broaden multi-step reasoning across languages. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1229–1241, Mexico City, Mexico. Association for Computational Linguistics.

Leonardo Ranaldi, Giulia Pucci, and Fabio Massimo Zanzotto. 2023. Modeling easiness for training transformers with curriculum learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 937–948, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Leonardo Ranaldi, Federico Ranaldi, Fabio Massimo Zanzotto, Barry Haddow, and Alexandra Birch. 2025c. Improving multilingual retrieval-augmented language models through dialectic reasoning argumentations.

Leonardo Ranaldi, Marco Valentino, and Andre Freitas. 2025d. Eliciting critical reasoning in retrieval-augmented generation via contrastive explanations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11168–11183, Albuquerque, New Mexico. Association for Computational Linguistics.

Leonardo Ranaldi, Marco Valentino, Alexander Polonsky, and Andrè Freitas. 2025e. Improving chain-of-thought reasoning via quasi-symbolic abstractions.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. 2024. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

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.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models.

A Instruction Template for R2-MultiOmnia

#Role You are an expert in visual reasoning, skilled at abstracting and integrating information from images.

#Task Given a problem that includes both an image and a text question, follow the steps below to extract, abstract and structure relevant information, formalise the problem, and solve it rigorously.

#Instructions

1) Visual Abstraction: Analyse the image and identify all visual elements, patterns, or relationships that are important for solving the problem. Clearly describe these elements in a structured, concise way. *Label this step as* *<visual_abstraction>...</visual_abstraction>*

2) Formalisation: Transform the abstracted visual information into key logical components of the problem, including the relevant visual information, variables, operations, and constraints. Structure these elements to clearly formulate the problem. *Label this step as* *<formalisation>...</formalisation>*

3) Reasoning: Solve the problem by breaking it into clear, logically coherent steps, integrating both the structured visual abstraction and the formalised problem statement. Clearly explain your reasoning and justify each step. *Label this step as* *<reasoning>...</reasoning>*

Final Answer: State the answer clearly as "**The answer is:**". *Label this step as* *<answer>...</answer>*

#Question {`image, question`}

Table 5: The template instructs the model to abstract key visual components, formalise the problem, reason stepwise, and present a clear answer.

B Instruction Template for Step-wise Verification and Refinement

#Role You are an expert reviewer in visual reasoning, specialised in critically evaluating, correcting, and refining multimodal problem-solving steps.

#Task Given a solution structured following the template (visual abstraction, formalisation, reasoning, answer), carefully review each step to detect any misleading passages. Clearly identify any issues and provide a corrected, refined version, ensuring clarity, rigour, and logical soundness.

#Instructions

1) Step-by-Step Verification: For each section, evaluate the content for correctness, clarity, and logical consistency. Indicate any factual mistakes, omitted reasoning, misleading steps, or unclear explanations.
2) Correction and Refinement: For every identified issue, provide a corrected and improved version of the relevant steps. Ensure the refined solution: abstracts and structures the visual information; formalises logical components; presents coherent reasoning steps; states the answer clearly and unambiguously. **3)** Present the Refined Solution: Complete the revised steps into a complete solution, using the same template (`<visual_abstraction>`, `<formalisation>`, `<reasoning>`, `<answer>`).

#Input

Table 6: The template instructs the model to rigorously verify, correct, and refine each step of a multimodal reasoning solution, ensuring logical soundness and clarity.

C Annotations Pipeline

We construct our tuning set starting from LLaVA-CoT (Xu et al., 2025). We generate synthetic demonstrations for the warm-up phase §2.1. We use both the self-generation strategy and GPT-4o as an annotator. We then conduct the self-training phase §2.2. As a common practice, we name the annotations demonstrations. As described in the main paper, they are generated by prompting the models using instructions detailed in Appendix A. After generating these, they go back to the model and are reviewed i.e. checked and corrected with the instructions in Appendix B. Although the generations are basically good after this double-check, there are still some misleading cases. To handle this, we evaluated the quality of the generated demonstrations by filtering out inaccurate examples to get a gold instruction set. In particular, we removed all inaccurate answers (outputs that do not match the exact target string metric). Then, we verify that the demonstrations follow the steps indicated in our prompt (see Table 5) using GPT-4o-mini and the prompt in Appendix D.

D Evaluation Metrics

We used a double-check to assess the accuracy of the responses delivered in the different experiments. In the first step, we used an exact-match heuristic. However, since some experiments required a more accurate response check, we used GPT-4o-mini as a judge. Hence, we prompt the model as follows:

Evaluation Prompt

#Role:
You are an experienced expert skilled in answering complex problems through logical reasoning and structured analysis.
#Instructions:
Given the following "#Input", you are a decider that decides whether the "Generated Answer" follows the "Required Format" and the final answer is the same as the "Target Answer". If the output doesn't align with the required format and target answer, respond with '0', whereas if it's correct, then respond with '1'. *Please, do not provide any other answer beyond '0' or '1'.*
#Senteces:
Generated Answer: {model_result}
Required Format: {format}
Target Answer: {correct_answer}.

E Data Composition

Therefore, to produce the training set, we start from LLaVA-CoT. We take a random sample of 60k out of the total 100k. We then annotate these and discard misleading and poorly formatted outputs. The results are about 20k. Table 7 shows the instances in the final testing. Then, we conduct the translation phase by making the data available for eight languages (including the original one, i.e. English). To ensure the languages are perfectly balanced, we translated 10k samples from English to (it, zh, fr, pt, ja, es, de). We conduct this phase using the nllb-200-distilled-600M as the translation system. However, to understand the quality, we back-translated the outputs and performed sanity checks.

Resource	Selected	Filtred*	Train. Set
R2-MultiOmnia (warm-up)	$60k$	$20k$	$10k$
SFT	$60k$	$20k$	$10k$
R2-MultiOmnia (RL)	$60k$	$20k$	$2k$
RL	$60k$	$20k$	$2k$

Table 7: Initial training data. *(in SFT and RL we used the demonstrations released by (Xu et al., 2025) that we translated into the different languages).

As described above, the demonstrations used in the SFT phase are those released in LLaVA-CoT. In order to have consistent and comparable experiments, we selected the same set that we used for R2-MultiOmnia. However, as described in Appendix C, the qualities of the annotations are not perfect. Hence, after filtering the annotations, we obtained a gold dataset, respectively 20k. Again, to have balanced data, we use 10k for SFT and 2k for RL phases (standard and the one proposed in the R2-MultiOmnia framework). The numbers discussed are native, i.e., in English and then transferred into different languages.

F Details Evaluation Data

Dataset	Lan	Size
XMMMU	en, ar, fr, hi, id, ja, pt	3K
M3EXAM	en, zh, it, pt, vi, th, ar	3K
MAXM	en, hi, th, zh, fr, iw, ro	2K
XMATHVISTA	en, it, zh, pt, jp, es, de	2K
MGSM	en, de, es, fr, ja, th, zh	1.7K
MGSM-SYMBOLIC	en, de, es, fr, ja, th, zh, it	2K

Table 8: Overview of evaluation datasets. Five multimodal and two text-only multilingual datasets are included.

G Translation Process

We translated the demonstrations from English into seven target languages (it, zh, fr, pt, ja, es, de) using nllb-200-distilled-600M (NLLB-200-600M). The purpose of the translation process is to translate the tense parts of the demonstrations to align proficiency and warm-up models in different languages. For this purpose, we only translate the text of the demonstrations and not the markers (i.e., <visual_abstraction>, <formalisation>, <reasoning>, <answer> left original). Similar pipelines used language augmentation strategies (Ranaldi and Pucci, 2023; Ranaldi et al., 2024a) and systematic translations (Ranaldi et al., 2025a), and evaluating the quality of the translations.
To evaluate translation quality, we adopted a multi-step strategy. First, we performed back-translation into English and computed BLEU-4 scores, a standard metric that measures n-gram overlap up to four words between a candidate and a reference text. BLEU-4 allowed us to capture both lexical accuracy and local coherence by comparing the original demonstrations (English) with their back-translated versions. Additionally, we assessed semantic similarity by computing cosine similarity scores: we used the all-MiniLM-L6-v2 sentence-transformer to compare original and back-translated demonstrations, and the paraphrase-multilingual-MiniLM-L12-v2 model to compare the original demonstrations with their direct translations.

Lang	Back-translation	Translation	BLEU-4
it	0.95	0.88	51
de	0.93	0.82	48
pt	0.94	0.84	47
fr	0.90	0.83	46
zh	0.96	0.87	42
ja	0.89	0.76	40

Table 9: Cosine similarity between original and translated demonstrations across languages. The third column are the BLEU-4 scores using NLLB-200-600M.

Lang	Back-translation	Translation	BLEU-4
it	0.98	0.90	54
de	0.95	0.86	50
pt	0.97	0.87	50
fr	0.93	0.85	49
zh	0.98	0.88	47
ja	0.92	0.82	45

Table 10: Cosine similarity between original and translated demonstrations across languages. The third column are the BLEU-4 scores using NLLB-200-1.3B.

H Models and Hyperparameters

Models In our experimental setting, as introduced in §3.1, we propose different models (detailed in Table 11). We choose the generation temperature for (mostly) deterministic outputs, with a maximum token length related to our CL strategy. The other parameters are left unchanged as recommended by the official resources. We use four 48GB NVIDIA RTXA600 GPUs for all experiments.

Hyperparameters In §3.2, we described the standard RL setting. We have proposed different experimental settings. In the self-training experimental setting, we conducted three iterations. In the SFT-only settings, we warm-up for one epoch. We conducted this setting after the pilot experiments shown in the previous sections.

I Models Vesions

Model	Version
Qwen2.5-VL-3B-Instruct	Qwen/Qwen2.5-VL-3B-Instruct
Qwen2.5-VL-7B-Instruct	Qwen/Qwen2.5-VL-3B-Instruct
DeepSeek-vl2-tiny	deepseek-ai/deepseek-vl2-tiny
DeepSeek-vl2-small	deepseek-ai/deepseek-vl2-small
GPT-4o	gpt-4o-2024-08-06
GPT-4o-mini	gpt-4o-mini-2024-07-18

Table 11: List the versions of the models proposed in this work, which can be found on huggingface.co. We used all the default configurations proposed in the repositories for each model.

J Dataset

Dataset	Version
MAXM	neulab/PangeaBench-maxm
XMMMLU	neulab/PangeaBench-xmmmu
M3EXAM	neulab/PangeaBench-m3exam

Table 12: List the versions of the models proposed in this work, which can be found on huggingface.co.

K RL Tuning

This procedure evaluates model-generated completions using **key constraints**, each contributing to the composite reward function used in the Language-Agnostic Reasoning Alignment phase (§2.2).

1. Strict Reward ($r_{\mathbf{s}}$) Assesses factual correctness by verifying whether the extracted answer matches the ground truth. The reward is assigned as follows:

$$r_{\mathrm{s}}(y) = \begin{cases} 2 & \text{if } \texttt{extract\_match}(y) = \hat{y} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where `extract_match(y)` match the target answer and generated response (both the **score** and the **query language**).

2. Format Reward ($r_{\mathbf{num/choice}}$) Promotes adherence to a structured reasoning format via regular expression matching:

$$r_{\mathrm{ans}}(y) = \begin{cases} 0.5 & \text{if } \texttt{answer\_format}(y) \text{ is valid} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

follow `The answer is: []` (in specific language).

3. Format Reward ($r_{\mathbf{f}}$) Enforces compliance with a rigid reasoning structure using regex:

$$r_{\mathrm{f}}(y) = \begin{cases} 0.5 & \text{if response matches } s_1 \wedge s_n \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

4. Structural Integrity Reward ($r_{\mathbf{SI}}$) Assign incremental rewards on correct placement and penalising excessive content:

$$r_{\mathrm{SI}}(y) = \sum_{i=1}^{4} w_i \cdot \mathbb{1}(s_i \in y) - \lambda \cdot \textit{extra\_content} \tag{13}$$

where: $w_i = 0.125$ for placing s_1 and s_2 and $\lambda = 0.001$ additional content.

5. CL Stabilisation Rewards ($r_{\mathbf{CL}}$) In order to avoid overthinking (excessive text generation) and obtain robust and consistent output, we propose dynamic tuning by progressively increasing the generation token limit while decreasing the generated output. We then use the configurations introduced in §2.3.