VFA: Empowering Multilingual MLLMs via Vision-Free Adaptation

Yixia Li[1*†], **Yaqing Shi**[3†], **Zhiwen Ruan**[1], **Dongdong Zhang**[2], **Lingjie Jiang**[4], **Shaohan Huang**[2], **Yun Chen**[3,5], **Guanhua Chen**[1‡], **Furu Wei**[2]
[1]Southern University of Science and Technology, [2]Microsoft Research Asia
[3]Shanghai University of Finance and Economics, [4]Peking University
[5]MoE Key Laboratory of Interdisciplinary Research of Computation and Economics

Abstract

Multimodal large language models have advanced rapidly, yet most remain English-centric, as scaling multilingual multimodal instruction tuning is limited by the scarcity and high cost of high-quality non-English image-text supervision. Although multilingual text data is abundant, naive textual fine-tuning can disrupt vision-language alignment and induce catastrophic forgetting. We propose Vision-Free Adaptation (VFA), a framework that decouples multilingual language enhancement from visual alignment by composing complementary task vectors over a shared LLM backbone. Specifically, we fine-tune a base LLM on multilingual text data to derive a multilingual task vector, which is then merged with the vision-aligned task vector of an MLLM. Experiments on five MLLMs across six multilingual multimodal benchmarks show consistent improvements while preserving both general multimodal and text-only capabilities. Moreover, using less than 2% of the text data, VFA narrows the gap to the fully multimodal-trained model, demonstrating its data efficiency.

1 Introduction

Multimodal large language models (MLLMs) have rapidly advanced the ability to reason over images and text in a unified interface, enabling visual question answering, grounded dialogue, and multimodal reasoning (Tong et al., 2025). As MLLMs move toward real-world deployment, multilingual and cross-cultural (Song et al., 2026) competence is becoming increasingly important: models should not only parse and follow non-English instructions, but also ground culturally specific concepts in visual contexts (Bai et al., 2025a; Nyandwi et al., 2025). Yet most open-source MLLMs remain English-centric, and their performance often degrades substantially when prompts, outputs, or benchmarks shift beyond English, limiting their applicability across global users and scenarios (Yue et al., 2024).

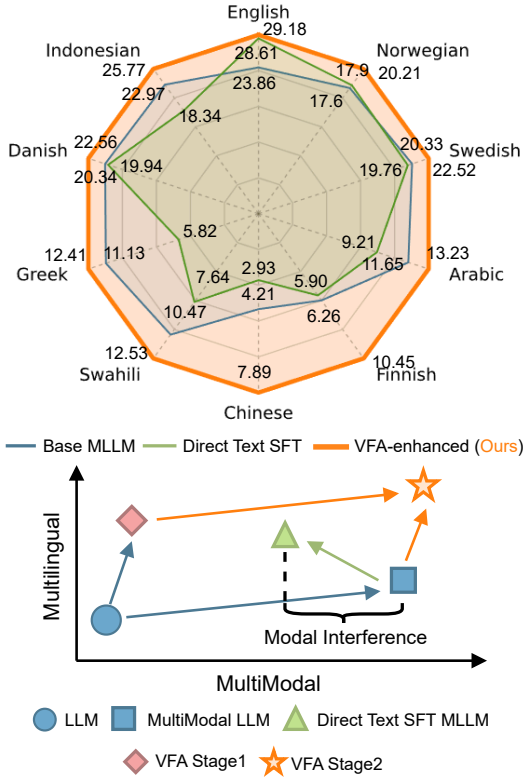

Figure 1: **Overview of VFA. Top:** Multilingual visual performance of Idefics3-8B on XM100, where VFA improves multilinguality while direct text-only tuning degrades performance. **Bottom:** VFA decouples multimodal adaptation via a two-stage vision-free strategy: Stage 1 fine-tunes the LLM on text data, and Stage 2 merges multilingual knowledge into the MLLM.

A common approach to building multilingual MLLMs is to scale multilingual multimodal instruction tuning using imagetext pairs (Yue et al., 2024; Dash et al., 2025). While effective, this paradigm is fundamentally constrained by data availability: high-quality non-English multimodal supervision remains scarce, expensive to curate, and unevenly distributed across languages and cultures, making comprehensive multilingual cover-

*Work done during internship at Microsoft Research Asia.
†Equal contribution.
‡Corresponding Author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8998–9015
July 2-7, 2026

age difficult to achieve in practice (Zhang et al., 2025b). Moreover, multilingual multimodal training incurs substantial computational overhead, as visual tokens lengthen input sequences and significantly increase attention and memory costs (Kuo et al., 2025). Taken together, these limitations impede the scalability of large-scale multilingual imagetext training as a general solution for multilingual adaptation.

Multilingual text data offers a compelling alternative: it is rich, cheaper to obtain, and far more scalable for multilingual learning. However, directly fine-tuning an MLLM on text can substantially interfere with its established cross-modal alignment, potentially leading to catastrophic forgetting of multimodal capabilities, as shown in Figure 1. This failure mode suggests that effective multilingual adaptation should strengthen language competence while minimizing disruption to the vision–language alignment learned during multimodal pre-training (Sanyal et al., 2025).

To address this challenge, we propose *Vision-Free Adaptation* (VFA), which reframes multilingual and multimodal adaptation as the composition of task vectors defined over a shared LLM backbone. Concretely, we view a trained MLLM as the original LLM augmented by a vision-aligned task vector, representing the parameter changes that encode cross-modal grounding. Separately, we derive a multilingual task vector by fine-tuning the same original LLM on multilingual text data. VFA then fuses these task vectors to inject multilingual competence into the MLLM while preserving its existing vision language alignment. This formulation mitigates the catastrophic forgetting commonly observed in direct text fine-tuning and reduces reliance on scarce multilingual image-text pairs.

We evaluate VFA on five MLLMs across different scales and model families on six multilingual multimodal benchmarks. VFA consistently enhances multilingual performance, yielding improvements of +2.99 and +0.57 on LLaVA-OneVision-1.5 at 8B and 4B scales, respectively. Importantly, these gains are obtained with general multimodal and text-only capabilities largely retained. Moreover, despite being trained on only 100K text samples, VFA substantially narrows the gap to multilingual models that rely on millions of imagetext pairs for full multimodal training, demonstrating the effectiveness and data efficiency of our approach. Overall, VFA provides a practical and resource-efficient pathway toward multilingual MLLMs, paving the way for future research on domain-specific multimodal models.[1]

2 Related Work

Multimodal LLMs Currently, MLLMs have achieved breakthrough progress in architectural paradigms and general capabilities (Wang et al., 2026, 2025b). Closed-source models like GPT-5 series (OpenAI, 2025) and Gemini 2.5 Pro (Comanici et al., 2025) have demonstrated remarkable capabilities, while in the open-source domain, the LLaVA series (Liu et al., 2023) established a mainstream paradigm by aligning vision encoders with LLMs via a projection layer. Subsequent works, such as LLaVA-OneVision-1.5 (An et al., 2025) and Qwen3-VL (Bai et al., 2025b), have further optimized training strategies. However, most open-source MLLMs heavily rely on large-scale text-image pair datasets for instruction tuning. Consequently, although the language backbones of these MLLMs possess inherent multilingual capabilities, the models often suffer from performance degradation or hallucinations in non-English tasks due to the scarcity of visually aligned multilingual data.

Multilingual MLLMs Although recent research tries to improve MLLMs' multilingual language ability by expanding data scale, the lack of high-quality image-text data and high training costs pose a significant challenge (Chen et al., 2023a; Ruan et al., 2025). For instance, Yue et al. (2024) introduces Pangea, a series of fully open-source multilingual multimodal models. By releasing the 39-language PangeaInstruct dataset, they demonstrate that broader language coverage is key to boosting cross-cultural capabilities in MLLMs. Aya Vision (Dash et al., 2025) utilizes high-quality synthetic data to tackle the non-English instruction following gap by capitalizing on massive text generation. However, these resource-intensive, data-driven approaches are hindered by the scarcity of high-quality non-English image-text pairs. In contrast, we address these challenges by utilizing pure text data, thereby significantly improving the multilingual performance of MLLMs.

[1]Code is available at https://github.com/sustech-nlp/VFA.

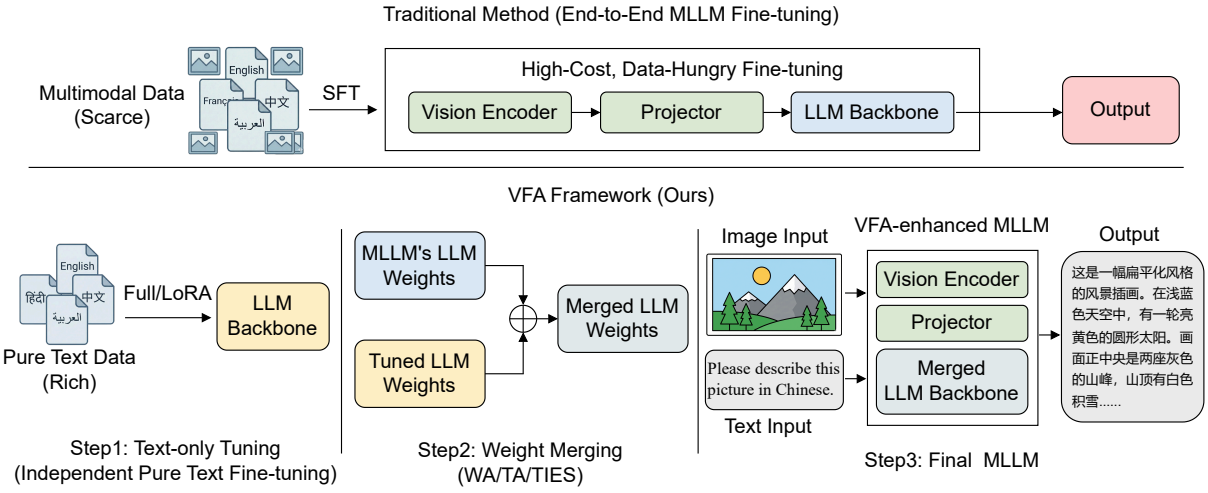

Figure 2: **Overall Architecture of VFA.** Traditional multilingual MLLM tuning depends on scarce, costly non-English image–text supervision and expensive end-to-end training. VFA leverages rich multilingual *text-only* data, fine-tunes the base LLM to obtain a multilingual task vector, and merges it into a vision-aligned MLLM to enhance multilinguality while preserving vision–language alignment.

Model Merging for MLLM Model merging offers a cost-efficient way to expand model capabilities without increasing inference overhead (Yang et al., 2024; Zhang et al., 2025a; Du et al., 2025). However, its potential for multilingual enhancement in MLLMs remains underexplored. At present, although several studies have explored enhancing MLLMs with mathematical (Chen et al., 2025) or coding (Jiang et al., 2025) capabilities via model merging, there is limited research on its application to multilingual adaptation. To address this gap, we propose a method that combines fine-tuning with model merging, enabling the efficient construction of multilingual MLLMs without relying on large-scale imagetext datasets.

3 Methodology

3.1 Vision-Free Adaptation (VFA)

Directly fine-tuning an MLLM on multilingual text often induces modal interference, as language-only updates can disturb the vision–language alignment learned during multimodal training and weaken cross-lingual multimodal integration. As a result, improvements in multilingual text understanding do not reliably translate into stronger multilingual multimodal performance. To address this issue, VFA decouples multilingual learning from visual grounding through a task-vector formulation. Under this view, fine-tuning is treated as a task-specific parameter update that can be composed with the vision-aligned parameters of an existing MLLM.

Formally, VFA first fine-tunes the base LLM initialization θ_{base} on multilingual text data $\mathcal{D}_{\text{multi}}$ to obtain θ_{FT}:

$$\theta_{\text{FT}} \leftarrow \text{Tune}(\theta_{\text{base}}, \mathcal{D}_{\text{multi}}), \tag{1}$$

where Tune denotes either full-parameter fine-tuning or parameter-efficient adaptation (Wang et al., 2025a). The parameter difference $(\theta_{\text{FT}} - \theta_{\text{base}})$ is interpreted as a multilingual task vector and is injected into the MLLM backbone θ_{MLLM} via a scaling factor α:

$$\theta_{\text{merged}} = \theta_{\text{MLLM}} + \alpha(\theta_{\text{FT}} - \theta_{\text{base}}), \tag{2}$$

where θ_{merged} is the resulting model parameterized for downstream tasks.

In comparison, conventional multimodal adaptation directly updates the MLLM parameters θ_{MLLM} using image–text supervision data $\mathcal{D}_{\text{visual}}$:

$$\theta_{\text{tuned}} = \text{Tune}(\theta_{\text{MLLM}}, \mathcal{D}_{\text{visual}}), \tag{3}$$

where θ_{tuned} represents the parameters of the conventionally fine-tuned model.

As illustrated in Figure 2, VFA enhances multilingual capability by leveraging large-scale text-only data, while preserving visual grounding by keeping the vision encoder and projection modules frozen. Multilingual knowledge is acquired independently of visual supervision and subsequently composed with the vision-aligned MLLM, thereby mitigating interference between linguistic and visual adaptation. This design substantially reduces reliance on scarce multilingual image–text

supervision, offering a more resource-efficient alternative to joint multimodal fine-tuning. Moreover, when multiple MLLMs share the same underlying LLM, the multilingual task vector can be *trained once and reused* across models via lightweight merging, whereas end-to-end multimodal adaptation must be performed separately for each MLLM, incurring significantly higher training costs.

3.2 Merging Methods

We consider three representative merging operators for composing the multilingual update with the vision-aligned MLLM, each offering a distinct trade-off between multilingual capability injection and preservation of vision–language alignment.

Weight Averaging (WA). Weight averaging (Wortsman et al., 2022) linearly interpolates between the pre-trained backbone and the fine-tuned parameters:

$$\begin{aligned}\Delta_{\text{FT}} &= \theta_{\text{FT}} - \theta_{\text{base}},\\ \Delta_{\text{MLLM}} &= \theta_{\text{MLLM}} - \theta_{\text{base}},\\ \theta_{\text{merged}} &= \theta_{\text{base}} + \alpha \cdot \Delta_{\text{MLLM}} + (1-\alpha)\cdot \Delta_{\text{FT}},\end{aligned} \tag{4}$$

where θ_{FT} and θ_{base} denote the fine-tuned LLM and the original LLM backbone within the VLM, respectively. The mixing coefficient $\alpha \in [0, 1]$ balances the trade-off between general multimodal capabilities and task-specific expertise.

Task Arithmetic (TA). Task arithmetic (Ilharco et al., 2023) constructs a task vector τ that captures the learned update and adds it to the backbone:

$$\theta_{\text{merged}} = \theta_{\text{base}} + \alpha \cdot (\Delta_{\text{FT}} + \Delta_{\text{MLLM}}), \tag{5}$$

where α is a scaling factor used to control the strength of the injected capabilities.

TIES-Merging. To mitigate parameter interference, TIES (Yadav et al., 2023) applies Trimming, Electing, and Sign-merging on the task vector. The merged parameters are derived as:

$$\theta_{\text{merged}} = \theta_{\text{base}} + \alpha \cdot \text{TIES}(\Delta_{\text{FT}} + \Delta_{\text{MLLM}}), \tag{6}$$

where the function $\text{TIES}(\cdot)$ represents the sequential process of: (i) Trimming to retain only the top-$k\%$ parameters by magnitude, (ii) Electing to determine the dominant sign of each parameter, and (iii) Sign-merging to aggregate values that align with the elected sign. The scaling factor α controls the intensity of the combined multimodal and task-specific updates.

Dimension	Benchmarks
MULTILINGUAL MULTIMODAL	MaXM, xGQA, xMMMU, XM100, MaRVL, M3Exam
GENERAL MULTIMODAL	OCRBench, MMBench, MMMU, MathVista
TEXT-ONLY MULTILINGUAL	TyDiQA, MMMLU, XNLI, HellaSwag, MLogiQA, M-IFEval, FLORES

Table 1: Benchmark suite grouped by evaluation axis.

3.3 Merged MLLM

To ensure consistent token representations and avoid disrupting vision–language alignment during model merging, we follow the standard practice (Chen et al., 2025) and exclude the visual input and output embeddings from the merging process. The resulting VFA model is composed by integrating the merged language backbone with frozen visual modules:

$$\theta_{\text{VFA}} = \{\underbrace{\theta_{\text{v-enc}}, \theta_{\text{proj}}}_{\text{Visual Backbone}}, \underbrace{\theta^{\text{m}}_{\text{emb}}, \theta^{\text{m}}_{\text{trans}}, \theta^{\text{m}}_{\text{head}}}_{\text{Merged Language Backbone}}\}, \tag{7}$$

where the superscript m denotes merged parameters. Specifically, $\theta_{\text{v-enc}}$ and θ_{proj} represent the visual encoder and vision language projector inherited from the original MLLM, respectively. For the language backbone, $\theta^{\text{m}}_{\text{emb}}$, $\theta^{\text{m}}_{\text{trans}}$, and $\theta^{\text{m}}_{\text{head}}$ correspond to the merged token embeddings, transformer blocks, and output head. By keeping the visual stream fixed, this design retains the models vision–language grounding while enabling the seamless integration of multilingual knowledge without additional training. Importantly, VFA produces a single unified model with no additional inference latency or memory overhead compared to the original MLLM.

4 Experiments

4.1 Experimental Setup

Multilingual Adaptation. **(1) Data.** For multilingual adaptation, we fine-tune the base LLMs on a 100K-example subset of the Multilingual-SFT dataset,[2] which combines xP3mt (Muennighoff et al., 2022), Bactrian-X (Li et al., 2023), the text subset of Aya Vision (Singh et al., 2024), and LLM-generated instruction data. **(2) Models.** We evaluate VFA on five MLLMs spanning multiple model families and scales, including Qwen2.5 (Team, 2024), Qwen3 (Team, 2025), Llama3 (AI@Meta, 2024), and Llama3.1 (Meta

[2] https://huggingface.co/datasets/agentlans/multilingual-sft

Model	Method	MaXM	xGQA	xMMMU	XM100	MaRVL	M3Exam	Avg.
Qwen2.5-VL-7B ↪Qwen2.5-7B-base	Base	50.37	47.81	48.34	15.74	53.50	61.97	46.29
	Direct Text SFT	49.68	31.20	49.69	13.80	0.00	61.86	34.37
	Δ *Gains*	–0.69	–16.61	+1.35	–1.94	–53.50	–0.11	–11.92
	VFA (ours)	51.69	48.45	48.15	15.43	65.83	62.55	48.68
	Δ *Gains*	+1.32	+0.64	–0.19	–0.31	+12.33	+0.58	+2.39
Idefics3-8B ↪Llama3.1-8B-inst	Base	46.40	48.73	42.83	13.80	26.50	27.93	34.37
	Direct Text SFT	46.56	44.95	36.45	14.06	38.00	44.79	37.47
	Δ *Gains*	+0.16	–3.78	–6.38	+0.26	+11.50	+16.86	+3.10
	VFA (ours)	50.11	48.10	43.27	16.20	62.67	49.73	45.01
	Δ *Gains*	+3.71	–0.63	+0.44	+2.40	+36.17	+21.80	+10.64
LLaVA-Next-8B ↪Llama3-8B-inst	Base	30.26	43.60	37.73	1.63	46.33	45.48	34.17
	Direct Text SFT	42.43	44.01	35.50	13.92	6.50	42.24	30.77
	Δ *Gains*	+12.17	+0.41	–2.23	+12.29	–39.83	–3.24	–3.40
	VFA (ours)	42.80	45.39	38.62	14.43	55.33	46.57	40.52
	Δ *Gains*	+12.54	+1.79	+0.89	+12.80	+9.00	+1.09	+6.35
LLaVA-OV-1.5-8B ↪Qwen3-8B-base	Base	53.97	29.85	54.71	16.51	62.17	63.86	46.85
	Direct Text SFT	52.92	28.73	55.80	15.07	61.83	64.08	46.41
	Δ *Gains*	–1.05	–1.12	+1.09	–1.44	–0.34	+0.22	–0.44
	VFA (ours)	56.14	43.97	53.67	16.02	65.50	63.75	49.84
	Δ *Gains*	+2.17	+14.12	–1.04	–0.49	+3.33	–0.11	+2.99
LLaVA-OV-1.5-4B ↪Qwen3-4B-base	Base	49.52	36.49	53.44	14.86	60.50	59.79	45.77
	Direct Text SFT	49.47	32.17	53.05	14.32	60.33	59.75	44.85
	Δ *Gains*	–0.05	–4.32	–0.39	–0.54	–0.17	–0.04	–0.92
	VFA (ours)	51.64	37.71	53.42	14.34	60.67	60.26	46.34
	Δ *Gains*	+2.12	+1.22	–0.02	–0.52	+0.17	+0.47	+0.57

Table 2: **Multilingual Multimodal Results of VFA across Various MLLMs.** Green and Red values in Δ Gains rows denote relative changes compared to the base MLLM. ↪indicates the backbone of the MLLM.

AI, 2024) series (4B–8B). **(3) Training.** All experiments are run on 4×A100 (80GB) GPUs using LlamaFactory (Zheng et al., 2024). Detailed dataset descriptions, models, and hyperparameters are provided in Appendix A.

Multimodal Model Merging. To incorporate multilingual capability into MLLMs, we compare three merging methods (WA, TA, and TIES). For each model pair, we conduct experiments with mixing coefficients $\alpha \in \{0.5, 0.7, 0.9, 1.0\}$ and utilize CVQA (Mogrovejo et al., 2024) as a validation set to select the best combination.

Evaluation. For a comprehensive assessment, we evaluate models along three axes: (i) multilingual multimodal capability, (ii) general multimodal capability, and (iii) text-only multilingual capability. Table 1 summarizes the benchmark suite for each axis. We utilize OpenCompass (Contributors, 2023) framework for text-only benchmarks and lmms-eval (Zhang et al., 2024) framework for multimodal benchmarks, using vLLM for inference. We set the maximum sequence length to 2048 and the batch size to 512. Additional details are provided in Appendix B.

4.2 Multilingual Multimodal Results

As shown in Table 2, we compare VFA with the original MLLM (*Base*) and direct text-only fine-tuning (*Direct Text SFT*). Across all evaluated models, VFA consistently improves multilingual multimodal performance under text-only supervision, leading to higher overall averages without degrading existing capabilities. The most pronounced gains are observed on culturally grounded visual reasoning tasks (MaRVL), with improvements of +12.33 on Qwen2.5-VL-7B and +36.17 on Idefics3-8B. This pattern indicates that VFA is particularly effective for language-intensive cross-cultural grounding, where accurate linguistic interpretation plays a central role in visual reasoning. Similar trends hold across model families and scales, yielding average gains of +2.39 (Qwen2.5-VL-7B), +10.64 (Idefics3-8B), and +6.35 (LLaVA-Next-8B), as well as substantial improvements on Qwen3-based LLaVA-OV models (+2.99 for 8B and +0.57 for 4B), with notable gains on xGQA. Per-language breakdowns for all benchmarks and qualitative examples are provided in Appendix C and D, respectively.

In contrast, direct text-only fine-tuning frequently leads to degraded multilingual multimodal performance, reflecting interference between linguistic adaptation and pre-existing vision–language alignment. For Qwen2.5-VL-7B, Direct Text SFT reduces the overall average by 11.92 and causes a near-complete collapse on MaRVL (from 53.50 to 0.00), despite remain-

Model	Method	OCRBench	MMBench	MMMU	MathVista	Avg.
Qwen2.5-VL-7B ↪Qwen2.5-7B-base	Base	14.36	87.80	51.11	61.90	53.79
	Direct Text SFT	14.56	87.20	51.33	65.90	54.75
	Δ *Gains*	+0.20	–0.60	+0.22	+4.00	+0.96
	VFA	14.47	87.20	51.78	65.70	54.79
	Δ *Gains*	+0.11	–0.60	+0.67	+3.80	+1.00
Idefics3-8B ↪Llama3.1-8B-inst	Base	5.96	84.50	38.78	24.00	38.31
	Direct Text SFT	6.90	83.70	38.11	33.40	40.53
	Δ *Gains*	+0.94	–0.80	–0.67	+9.40	+2.22
	VFA	4.50	82.50	39.89	31.80	39.67
	Δ *Gains*	–1.46	–2.00	+1.11	+7.80	+1.36
LLaVA-Next-8B ↪Llama3-8B-inst	Base	6.44	79.60	38.00	20.60	36.16
	Direct Text SFT	6.13	78.90	38.67	24.20	36.98
	Δ *Gains*	–0.31	–0.70	+0.67	+3.60	+0.82
	VFA	6.40	80.10	40.11	20.80	36.85
	Δ *Gains*	–0.04	+0.50	+2.11	+0.20	+0.69
LLaVA-OV-1.5-8B ↪Qwen3-8B-base	Base	12.50	88.60	55.33	32.80	47.31
	Direct Text SFT	5.26	98.00	55.00	38.33	49.15
	Δ *Gains*	–7.24	+9.40	–0.33	+5.53	+1.84
	VFA	12.34	88.10	56.22	31.80	47.12
	Δ *Gains*	–0.16	–0.50	+0.89	–1.00	–0.19
LLaVA-OV-1.5-4B ↪Qwen3-4B-base	Base	11.44	89.90	53.22	36.50	47.77
	Direct Text SFT	5.14	99.00	53.00	32.33	47.37
	Δ *Gains*	–6.30	+9.10	–0.22	–4.17	–0.40
	VFA	11.44	90.10	54.22	37.50	48.32
	Δ *Gains*	+0.00	+0.20	+1.00	+1.00	+0.55

Table 3: **General Multimodal Results of VFA across Various MLLMs.** Green and Red values in Δ Gains rows denote relative changes compared to the base MLLM. ↪indicates the backbone of the MLLM.

ing competitive on several other benchmarks. A similar trend is observed for LLaVA-Next-8B: although Direct Text SFT improves MaXM and XM100, it substantially degrades performance on MaRVL (-39.83), resulting in an overall average drop of 3.40. Taken together, these results align with the motivation of VFA: composing a multilingual task vector with a vision-aligned MLLM enables broad multilingual gains while mitigating the alignment degradation that can arise when the MLLM backbone is directly adapted using text-only supervision.

4.3 General Multimodal Results

As shown in Table 3, VFA maintains or improves *general-purpose vision–language performance*. Overall, we observe no systematic degradation: VFA remains close to neutral on average for most models and can be beneficial in some cases, such as +1.36 on Idefics3-8B and +1.00 on Qwen2.5-VL-7B, while staying nearly unchanged on LLaVA-OV-1.5-8B (–0.19). This pattern indicates that introducing multilingual capability through VFA does not inherently interfere with the visual grounding learned during multimodal training.

The observed gains are not uniform across benchmarks. Improvements are primarily concentrated on reasoning-centric evaluations such as MMMU and MathVista, while MMBench and OCRBench exhibit only modest, backbone-dependent variations. For instance, the average improvement on Idefics3-8B is driven largely by a marked gain on MathVista (+7.80), despite minor declines on OCRBench and MMBench. This pattern suggests that enhancing the language backbone predominantly benefits language-intensive multimodal reasoning, including problem interpretation and multi-step inference, whereas perception-heavy skills such as OCR are less directly affected. Incorporating additional multimodal supervision may therefore be necessary to further improve performance on perception-focused benchmarks.

4.4 Text-only Multilingual Results

Table 4 presents text-only multilingual performance after injecting multilingual knowledge into MLLMs via VFA. Across both Qwen and Llama model families, VFA preserves or improves per-

Model	Method	TyDIQA	MMMLU	XNLI	HellaSwag	MLogiQA	M-IFEval	FLORES	Avg.
Qwen2.5-VL-7B ↪Qwen2.5-7B-base	Base	24.89	44.70	67.67	59.02	46.62	66.56	37.04	49.50
	VFA	31.05	48.85	68.59	60.31	48.88	66.67	33.90	51.18
	Δ *Gains*	+6.16	+4.15	+0.92	+1.29	+2.26	+0.11	–3.14	+1.68
Idefics3-8B ↪Llama3.1-8B-inst	Base	31.60	45.35	63.83	52.73	29.62	53.86	33.66	44.38
	VFA	42.95	45.98	60.75	53.22	41.00	57.91	19.72	45.93
	Δ *Gains*	+11.35	+0.63	–3.08	+0.49	+11.38	+4.05	–13.94	+1.55
LLaVA-Next-8B ↪Llama3-8B-inst	Base	25.47	41.95	53.75	45.89	36.00	52.50	30.57	40.88
	VFA	45.55	42.55	60.50	49.73	36.62	61.77	33.33	47.15
	Δ *Gains*	+20.08	+0.60	+6.75	+3.84	+0.62	+9.27	+2.76	+6.27
LLaVA-OV-1.5-8B ↪Qwen3-8B-base	Base	28.85	50.38	67.84	68.04	48.25	70.21	38.37	53.13
	VFA	30.43	51.52	67.25	67.94	50.00	67.81	38.00	53.28
	Δ *Gains*	+1.58	+1.14	–0.59	–0.10	+1.75	–2.40	–0.37	+0.15
LLaVA-OV-1.5-4B ↪Qwen3-4B-base	Base	30.68	49.75	65.42	63.10	45.88	73.23	36.88	52.13
	VFA	31.38	49.85	66.33	64.55	46.88	70.83	36.81	52.38
	Δ *Gains*	+0.70	+0.10	+0.91	+1.45	+1.00	–2.40	–0.07	+0.25

Table 4: **Multilingual Text-task Results of VFA across Various MLLMs.** Green and Red values in Δ Gains rows denote relative changes compared to the base MLLM. ↪indicates the backbone of the MLLM.

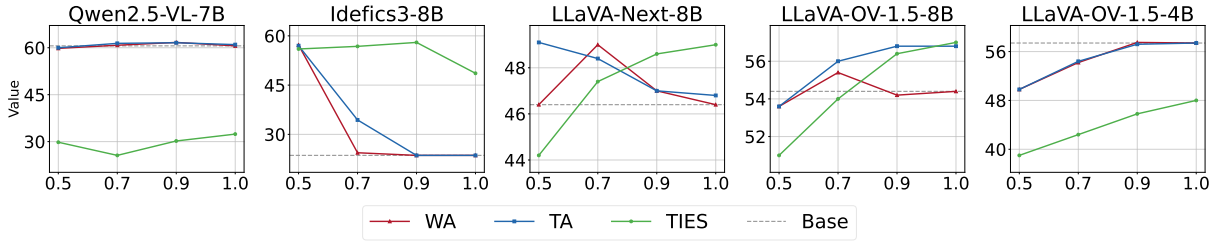

Figure 3: **Effect of Merging Operators and Mixing Coefficient.** Validation performance on CVQA under different merging operators (WA, TA, and TIES) and mixing coefficients α.

formance on the vast majority of benchmarks. In particular, it brings modest gains to Qwen-based models and delivers substantial improvements for LLaVA-Next-8B (+6.27 on average). The main outlier is Idefics3-8B, which achieves strong gains on reasoning and knowledge-intensive tasks but exhibits a pronounced decline on FLORES (−13.94) and a smaller drop on XNLI (−3.08).

Inspecting outputs on the FLORES translation task reveals that this regression is primarily an artifact of generation behavior rather than a fundamental loss of translation ability. The merged Idefics3 model typically translates the first sentence correctly but fails to emit the EOS token, over-generating hallucinated text until the length limit. Since FLORES uses single-sentence references, BLEU heavily penalizes this via precision dilution, drastically suppressing the score despite the accurate initial translation. This issue is most pronounced in languages with concise references (e.g., Chinese) and is absent in LLaVA-Next-8B, suggesting that the EOS-control mechanisms of underlying instruction-tuned LLMs react differently to weight merging across MLLM families. Representative examples of this EOS-failure pattern are in Appendix E.

5 Analysis

5.1 Merging Strategies

We analyze how merging choices affect the merged MLLM. Figure 3 compares the validation performance (CVQA) of three operators (WA, TA, and TIES) across different mixing coefficients α. The results indicate that the performance trends are highly dependent on architecture. Specifically, for Qwen2.5-VL-7B and LLaVA-OV-1.5-4B, WA and TA perform similarly and significantly outperform TIES. In contrast, for Idefics3-8B, TIES demonstrates stronger overall performance, while WA and TA experience substantial degradation at higher coefficients. For LLaVA-Next-8B and LLaVA-OV-1.5-8B, the optimal operator varies with α, showing no single dominant method across the range. Consequently, rather than applying a

Model	Method	Training Dataset	MaXM	xGQA	xMMMU	XM100	MaRVL	M3Exam	Avg.
Pangea-7B	–	6M Multimodal Samples	51.27	62.58	44.00	16.72	78.33	49.15	50.34
Qwen2-VL-7B	Base	–	45.03	42.64	47.97	2.05	57.83	57.43	42.16
	VFA	100K Text-Only Samples	48.68	51.64	48.45	13.86	66.67	57.76	47.84
	Δ *Gains*	–	+3.65	+9.00	+0.48	+11.81	+8.84	+0.33	+5.68

Table 5: **Comparison with Multimodal Fine-tuning on Multilingual Multimodal Benchmarks.** We compare VFA (trained on limited text-only data) against a baseline MLLM trained with large-scale image-text supervision, highlighting the trade-off between performance and training efficiency.

Model	Method	CodeVision	HumanEval-V	Avg.
InternVL3-8B	Base	39.96	18.28	29.12
	VFA	47.49	19.93	33.71
	Δ *Gains*	+7.53	+1.65	+4.59
Llama3-LLaVA-8B	Base	8.35	6.04	7.20
	VFA	8.43	6.61	7.52
	Δ *Gains*	+0.08	+0.57	+0.32
Idefics3-8B	Base	3.67	9.69	6.68
	VFA	15.58	11.01	13.30
	Δ *Gains*	+11.91	+1.32	+6.62

Table 6: **VFA on Multimodal Visual Coding Tasks**.

universal setting, our final configuration for each backbone is determined by selecting the specific combination of operator and coefficient α that yields the highest absolute validation score. Detailed parameter choices and recommended defaults are summarized in Appendix F.

5.2 Comparison with Multimodal Fine-tuning

We compare VFA to conventional multimodal training that relies on large-scale paired image–text supervision. As shown in Table 5, Pangea-7B is trained with 6M multimodal samples on Qwen2-7B-Instruct and achieves 50.34 on average, whereas VFA starts from Qwen2-VL-7B and uses only 100K text-only samples (about 2% of Pangea's training size) to reach 47.84, leaving a 2.50-point gap. From an efficiency-frontier perspective, VFA shifts the performance vs. data-cost trade-off leftward by replacing expensive multilingual image–text curation with rich multilingual text. The remaining gap is plausibly attributable to capabilities that benefit from paired supervision, such as stronger visual grounding and broader multilingual vision–language alignment coverage; closing it may require higher-quality multilingual text, stronger language backbones, or lightweight hybrid objectives that reintroduce limited multimodal signals.

5.3 Beyond Multilingual: VFA on Visual Coding Tasks

While we primarily focus on multilingual adaptation, VFA is inherently a versatile vision-free framework. To illustrate its broader applicability, we present an additional case study applying the vision-free injection paradigm to multimodal visual coding. As shown in Table 6, VFA achieves consistent performance improvements across all three evaluated backbones, yielding the most significant gains on InternVL3-8B (+4.59) and Idefics3-8B (+6.62). These findings underscore VFA's potential as a lightweight, reusable mechanism for expanding MLLM capabilities beyond multilingual contexts.

6 Conclusion

In this paper, we introduce VFA, an efficient framework designed to enhance the multilingual capabilities of MLLMs without relying on additional text-image paired datasets. Additionally, VFA decouples linguistic knowledge injection from visual representation learning, effectively circumventing the catastrophic forgetting of visual alignment that often plagues direct text fine-tuning methods. Extensive evaluations across five diverse architectures and six multilingual multimodal benchmarks show that VFA not only preserves generic multimodal robustness but also surpasses models trained on massive multimodal datasets in some tasks. Overall, this work provides a practical and efficient pathway to broaden the linguistic capabilities of MLLMs.

Limitations

The multilingual data used for adaptation does not yet incorporate explicit quality filtering, and its scale is intentionally kept modest to highlight the efficiency of VFA. In addition, as a text-only approach, VFA primarily enhances language-intensive reasoning, while perception-heavy tasks

such as OCR may still benefit from supplementary imagetext supervision. Our empirical evaluation mainly focuses on models in the 4B–8B parameter range. Exploring larger model and data scales, adopting stricter data curation, and developing more advanced merging algorithms to more consistently mitigate this interference remain important directions for future work.

Acknowledgements

This project was supported by National Key R&D Program of China (No. 2025YFB4007600), National Natural Science Foundation of China (No. 62306132), Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011564), Natural Science Foundation of Shanghai (No. 25ZR1402136). We thank the anonymous reviewers for their insightful feedback on this work.

References

AI@Meta. 2024. Llama 3 model card.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025a. The power of many: Multi-agent multimodal models for cultural image captioning. 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 2970–2993.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025b. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023a. mclip: Multilingual clip via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. 2025. Bring reason to vision: Understanding perception and reasoning through model merging. In *Forty-second International Conference on Machine Learning*.

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, Wan Xiang, and Benyou Wang. 2023b. MultilingualSIFT: Multilingual Supervised Instruction Fine-tuning.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2475–2485.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, et al. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*.

Peter Devine. 2024. Tagengo: A multilingual chat dataset. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 106–113.

Yiyang Du, Xiaochen Wang, Chi Chen, Jiabo Ye, Yiru Wang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Zhifang Sui, et al. 2025. Adamms: Model merging for heterogeneous multimodal large language models with unsupervised coefficient optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9413–9422.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Lingjie Jiang, Shaohan Huang, Xun Wu, Yixia Li, Dongdong Zhang, and Furu Wei. 2025. Viscodex: Unified multimodal code generation via merging vision and coding models. *arXiv preprint arXiv:2508.09945*.

Chia-Wen Kuo, Sijie Zhu, Fan Chen, Xiaohui Shen, and Longyin Wen. 2025. D-attn: Decomposed attention for large vision-and-language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23935–23944.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuạt Nguyễn, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation. *Preprint*, arXiv:2305.15011.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021b. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3622–3628.

Meta AI. 2024. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/. Accessed: 2024-07 (or the date you accessed it).

David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, and Others. 2024. CVQA: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, et al. 2022. No language left behind: Scaling human-centered machine translation.

Jean De Dieu Nyandwi, Yueqi Song, Simran Khanuja, and Graham Neubig. 2025. Grounding multilingual multimodal llms with cultural knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24198–24242.

OpenAI. 2025. Gpt-5 system card. Accessed: January 6, 2026.

Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. Humaneval-xl: A multilingual code generation benchmark for cross-lingual natural language generalization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8383–8394.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. 2025. Palo: A large multilingual multimodal language model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025)*.

Zhiwen Ruan, Yixia Li, He Zhu, Longyue Wang, Weihua Luo, Kaifu Zhang, Yun Chen, and Guanhua Chen. 2025. LayAlign: Enhancing Multilingual Reasoning in Large Language Models via Layer-Wise Adaptive Fusion and Alignment Strategy. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1481–1495.

Sunny Sanyal, Hayden Prairie, Rudrajit Das, Ali Kavis, and Sujay Sanghavi. 2025. Upweighting easy samples in fine-tuning mitigates forgetting. In *Forty-second International Conference on Machine Learning*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 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.

Yuchen Song, Andong Chen, Wenxin Zhu, Kehai Chen, Xuefeng Bai, Muyun Yang, and Tiejun Zhao. 2026. Culture in a frame: C^{3}b as a comic-based benchmark for multimodal culturally awareness. In *The Fourteenth International Conference on Learning Representations*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. 2025. Metamorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17001–17012.

Hanqing Wang, Yixia Li, Shuo Wang, Guanhua Chen, and Yun Chen. 2025a. MiLoRA: Harnessing minor singular components for parameter-efficient LLM finetuning. 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 4823–4836, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaolong Wang, Zhaolu Kang, Wangyuxuan Zhai, Xinyue Lou, Yunghwei Lai, Ziyue Wang, Yawen Wang, Kaiyu Huang, Yile Wang, Peng Li, and Yang Liu. 2025b. MUCAR: Benchmarking multilingual cross-modal ambiguity resolution for multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15026–15048, Suzhou, China. Association for Computational Linguistics.

Zexuan Wang, Chenghao Yang, Yingqi Que, Zhenzhu Yang, Huaqing Yuan, Yiwen Wang, Zhengxuan Jiang, Shengjie Fang, Zhenhe Wu, Zhaohui Wang, et al. 2026. Worldtravel: A realistic multimodal travel-planning benchmark with tightly coupled constraints. *arXiv preprint arXiv:2602.08367*.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*.

Dingkun Zhang, Shuhan Qi, Xinyu Xiao, Kehai Chen, and Xuan Wang. 2025a. Merge then realign: Simple and effective modality-incremental continual learning for multimodal LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13148–13164, Suzhou, China. Association for Computational Linguistics.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024. Lmms-eval: Reality check on the evaluation of large multimodal models. *Preprint*, arXiv:2407.12772.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025b. Less, but better: Efficient multilingual expansion for llms via layer-wise mixture-of-experts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17948–17963.

Yidan Zhang, Yu Wan, Boyi Deng, Baosong Yang, Hao-Ran Wei, Fei Huang, Bowen Yu, Dayiheng Liu, Junyang Lin, and Jingren Zhou. 2025c. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4809–4836.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *CoRR*.

A Training Settings

A.1 Datasets

Data Construction. The 100K multilingual text subset is constructed deterministically from the `train` split of the HuggingFace dataset `agentlans/multilingual-sft` (configuration: `clustered_k100000`), traversed in its default order without random sampling. We apply the following fixed filtering rules: (1) Field validation: samples are excluded if the input or source fields are not valid strings; (2) Multimodal token filtering: instances containing <audio>, <image>, or <video> tags are removed to ensure pure text-only supervision. The resulting dataset is summarized in Table 7, and Table 8 further lists the ISO 639-1 language codes represented.

xP3mt. xP3mt (Muennighoff et al., 2022) is a large-scale multilingual instruction tuning dataset. Derived from the xP3 dataset, xP3mt aims to address the scarcity of non-English prompts by utilizing machine translation to translate English prompts into 20 diverse languages. The dataset includes 46 languages and 13 training tasks, including QA, Summarization, and Program Synthesis.

Bactrian-X. Bactrian-X (Li et al., 2023) is a comprehensive multilingual instruction dataset designed to democratize instruction following capabilities across 52 languages. It includes approximately 3.4 million instruction-response pairs. The dataset was constructed by translating English instructions from Alpaca and Dolly-15k into 51 target languages, followed by feeding these translated prompts into gpt-3.5-turbo to generate authentic, native-like responses.

Aya Dataset. The Aya Dataset (Singh et al., 2024) is a large-scale human-curated multilingual instruction tuning dataset, comprising approximately 204K instruction-response pairs covering 65 languages. Its human-centric curation helps mitigate the biases and artifacts commonly found in machine-translated corpora, particularly for underrepresented low-resource languages.

Tagengo-GPT4. Tagengo-GPT4 (Devine, 2024) is a high-quality multilingual instruction tuning dataset developed by Lightblue, explicitly designed to bridge the gap between English-centric LLMs and global language accessibility. The dataset includes approximately 76,000 diverse prompt-response pairs covering 74 languages.

Multilingual-SIFT. Multilingual-SIFT (Chen et al., 2023b) is a comprehensive dataset collection specifically constructed to enhance the multilingual instruction-following capabilities of LLMs. It is derived from high-quality English instruction corpora, including Alpaca-GPT4, Evol-Instruct, and ShareGPT. To ensure linguistic diversity, the authors employed GPT-3.5 Turbo to translate these datasets into multiple target languages.

A.2 Models

Table 12 presents the MLLM models used in this paper and their corresponding LLMs.

A.3 Hyperparameters

For the fine-tuning experiments, we set the total batch size to 128, with a learning rate of 1e-5, and train for 1 epoch.

A.4 Training Efficiency

Table 9 compares the training efficiency between standard multimodal SFT and VFA. By relying solely on text-only data and avoiding visual token

Category	Source	#Sub	Count	Prop.
Primary large-scale	Aya, Bactrian-X, xP3mt, Tagengo	4	92,907	92.91%
Multi. Alpaca	FreedomIntel. (Alpaca-GPT4)	11	4,965	4.97%
Multi. Evol	FreedomIntel. (Evol-Instruct)	11	2,128	2.13%
Total		**26**	**100,000**	**100%**

Table 7: Composition of the 100K multilingual text subset.

Code	Language	Code	Language	Code	Language
af	Afrikaans	it	Italian	su	Sundanese
ar	Arabic	iw (he)	Hebrew	sw	Swahili
bn	Bengali	ja	Japanese	ta	Tamil
de	German	jv	Javanese	th	Thai
en	English	ko	Korean	tr	Turkish
es	Spanish	mn	Mongolian	ur	Urdu
fr	French	pt	Portuguese	vi	Vietnamese
hi	Hindi	ro	Romanian	zh	Chinese
id	Indonesian	ru	Russian	min	Minangkabau
rw	Kinyarwanda	–	–	–	–

Table 8: Mapping between ISO 639-1 language codes and full names.

Method	Training Data	GPU Hours (8B)	Throughput
Multimodal SFT	760K image-text	∼320 A100-h	∼2.4K/GPU-h
VFA	100K text-only	∼10 A100-h	∼10K/GPU-h

Table 9: Comparison of training efficiency between standard multimodal SFT and VFA.

Task	Benchmarks	Metric
Machine Translation	FLORES-200 (NLLB Team et al., 2022)	BLEU
Natural Language Understanding	XNLI (Conneau et al., 2018), M-HellaSwag (Lai et al., 2023)	Accuracy
Code Generation	HumanEval-XL (Peng et al., 2024)	Pass@1
Mathematical Reasoning	MGSM (Shi et al., 2023)	Accuracy
Logical Reasoning	M-LogiQA (Liu et al., 2021b)	Accuracy
General Knowledge	M-MMLU (Hendrycks et al., 2021)	Accuracy
Instruction Following	M-IFEval (Zhou et al., 2023)	Accuracy

Table 10: **Overview of the P-MMEval Benchmark.** Summary of evaluation tasks, multilingual benchmarks, and their corresponding metrics. This suite evaluates the core linguistic and reasoning capabilities of models across diverse languages.

processing, VFA reduces training time, GPU memory consumption, and data requirements, making it a practical and scalable alternative to conventional multilingual multimodal fine-tuning.

B Evaluation

We compare VFA-enhanced MLLMs with baselines on the following benchmarks:

TyDiQA. TyDiQA (Clark et al., 2020) is a benchmark for evaluating QA systems across 11 typologically diverse languages, comprising approximately 200K human-annotated QA pairs.

P-MMEval. P-MMEval (Zhang et al., 2025c) is a large-scale parallel multilingual multitask benchmark specifically designed to enable consistent and fair evaluation of LLMs across diverse linguistic landscapes. It provides strictly parallelized evaluation samples across 10 typologically distinct languages. As shown in Table 10, the benchmark integrates fundamental NLP tasks with capability-specialized challenges, serving as a rigorous testbed for assessing a model's cross-lingual transferability and core reasoning proficiency without the noise of dataset inconsistency.

Tasks	Datasets	Forms	Size	Languages	Metric
Multimodal Chat	xChatBench (Yue et al., 2024)	Long	400	zh, en, hi, id, ja, rw, ko, es	LLM-as-Judge
	M-LlavaBench (Rasheed et al., 2025)	Long	600	ar, bn, zh, fr, hi, ja, ru, es, ur, en	LLM-as-Judge
Captioning	XM100 (Yue et al., 2024)	Long	3.6K	36 languages	ROUGE-L
Cultural Understanding	CVQA (Mogrovejo et al., 2024)	MC	21K	en, zh, ko, mn, ja, id, jv, min, su	Accuracy
	MaRVL (Liu et al., 2021a)	Short	6K	id, sw, ta, tr, zh	Accuracy
Multilingual VQA	xGQA (Pfeiffer et al., 2022)	Short	77K	en, de, pt, ru, id, bn, ko, zh	Accuracy
	MaXM (Changpinyo et al., 2023)	MC	2K	hi, th, zh, fr, en, iw, ro	Accuracy
Reasoning (Multi-subject)	xMMMU (Yue et al., 2024)	Short/MC	3K	en, ar, fr, hi, id, ja, pt	Accuracy
	M3Exam (Zhang et al., 2023)	MC	3K	en, zh, it, pt, vi, th, af	Accuracy

Table 11: **Summary of Evaluation Datasets.** We select CVQA for model validation. MaRVL is utilized for multilingual visual reasoning, serving as an extension of the NLVR2 task. The table lists task categories, dataset forms, language coverage, and metrics.

Qualitative Case Study: Multilingual Image Captioning (VFA vs. Base)

Prompt	Provide a one-sentence caption for the provided image in English.
GT	Two cameras on a table.
Base	A black and white photo of a camera with the word Rolleiflex on it. (*Score: 0.3099*)
VFA	Two cameras sit on a table, one of which is a Rolleiflex. (*Score: 0.5806*)

Prompt	Provide a one-sentence caption for the provided image in Chinese.
GT	一辆停在路上的灰色美式典型老式汽车近景。
Base	一辆银色的奔驰轿车停在路边。(*Score: 0.3214*)
VFA	一辆复古风格的灰色美式老爷车停在公路旁，展现出经典的设计线条。(*Score: 0.8901*)

Prompt	Provide a one-sentence caption for the provided image in Swahili.
GT	Mwanaume mzee kidogo anaucheza piano na kuna mwanaume mzee kidogo nyuma yake ameshikilia chombo cha ngoma.
Base	A man in a white shirt is playing a piano. (*Score: 0.2295*)
VFA	Mwanamume linaweza kusaidia kufanya kazi zao kama vile kuchagua chombo cha piano. (*Score: 0.478*)

Figure 4: Qualitative comparison between VFA and base MLLM on XM100.

Qualitative Evidence: EOS-Failure on FLORES (Idefics3-8B, Base vs. VFA)	
Case 1. EOS failure: faithful first sentence followed by hallucinated chronology.	
Source (en)	Dr. Ehud Ur, professor of medicine at Dalhousie University in Halifax, Nova Scotia and chair of the clinical and scientific division of the Canadian Diabetes Association cautioned that the research is still in its early days.
Reference	埃胡德 · 乌尔博士（新斯科舍省哈利法克斯市达尔豪西大学医学教授，加拿大糖尿病协会临床与科学部门教授）提醒，这项研究仍处在早期阶段。
Base	达尔豪西大学医学教授、加拿大糖尿病协会临床和科学部主任 Ehud Ur 博士在哈利法克斯，诺瓦斯科舍省表示，研究仍在初期阶段。*(stops at EOS, 61 chars)*
VFA	达尔豪西大学（Dalhousie University）哈利法克斯分校（Halifax）医学教授、加拿大糖尿病协会临床和科学部门主席 Ehud Ur 博士表示，研究仍处于早期阶段。*(faithful first sentence, 78 chars)* *2011 年 12 月，Ehud Ur 博士在加拿大糖尿病协会的会议上发表了他的研究成果。2012 年 1 月，Ehud Ur 博士在加拿大糖尿病协会的会议上发表了他的研究成果。2012 年 2 月，… 2012 年 12 月，… 2013 年 1 月，… (hallucinated continuation until max-tokens; total 2722 chars)*
Case 2. Counter-example: when VFA does emit EOS, its translation matches the baseline.	
Source (en)	Although an experimental vaccine appears able to reduce Ebola mortality, no drugs have yet been clearly shown suitable for treating existing infection.
Reference	虽然有一种实验性疫苗看似能够降低埃博拉病毒的死亡率，但迄今为止，还没明确证明任何药物适合治疗现有的感染。
Base	虽然有一种实验性疫苗似乎能够降低埃博拉病毒的死亡率，但到目前为止，尚未有任何药物被明确证明适合治疗现有的感染。*(stops at EOS, 56 chars)*
VFA	尽管有一种实验性疫苗似乎能够降低埃博拉病毒的死亡率，但直到现在，尚未有任何药物被明确证明适合治疗现有的感染。*(stops at EOS, 54 chars)*

Figure 5: **Comparison of Outputs from the Idefics3-8B Base MLLM and VFA on FLORES Chinese. Case 1** illustrates the EOS failure driving the BLEU regression: VFA translates the first sentence accurately but hallucinates thereafter until the length limit. **Case 2** shows proper EOS emission, yielding a translation comparable to the base MLLM. Hallucinations are in red.

Family	MLLM	Base LLM
Qwen	Qwen2-VL-7B-Instruct	Qwen2-7B
	Qwen2.5-VL-7B-Instruct	Qwen2.5-7B
	LLaVA-OV-1.5-4B-Inst	Qwen3-4B-Base
	LLaVA-OV-1.5-8B-Inst	Qwen3-8B-Base
Llama	llama3-llava-next-8b-hf	Meta-Llama-3-8B-Inst
	Idefics3-8B-Llama3	Llama-3.1-8B-Inst

Table 12: Multimodal models and their corresponding LLM backbones used in this paper.

Model	Merge Operator	α
Qwen2.5-VL-7B	WA	0.9
Idefics3-8B	TIES	0.9
LLaVA-Next-8B	TA	0.5
LLaVA-OV-1.5-8B	TIES	1.0
LLaVA-OV-1.5-4B	WA	0.9

Table 13: Selected merge operator and mixing coefficient (α) per model based on CVQA validation.

PangeaBench. PangeaBench (Yue et al., 2024) is a comprehensive evaluation designed to assess the multilingual and multicultural capabilities of MLLMs. Including 47 languages across 14 diverse datasets, PangeaBench extends beyond standard translation-based metrics by incorporating culturally specific tasks. As shown in Table 11, the benchmark spans five primary categories, ranging from multimodal chat to multi-subject reasoning.

C PangeaBench Results by Language

We present the performance of the VFA-enhanced MLLM and the base MLLM on the MaXM, xGQA, xMMMU, MaRVL, M3Exam, and XM100 benchmarks in Tables 14–19.

D Qualitative Analysis

Furthermore, to offer a more intuitive perspective, we present qualitative comparisons between VFA-equipped MLLMs and standard base models, as illustrated in Figure 4. Through various multilingual visual question answering tasks, these examples vividly demonstrate how VFA practically enhances the precision and overall quality of non-English generations. Notably, these gains avoid the typical vision-language alignment tax.

E Generation Behavior on FLORES

As discussed in Section 4.4, Figure 5 provides raw output examples illustrating the EOS-failure pattern in Idefics3-8B. While both the base model and VFA model successfully generate accurate first-sentence translations, the VFA model frequently fails to emit the EOS token. Consequently, it over-generates hallucinated text until reaching the

Model	Method	English	Thai	Chinese	French	Hindi	Hebrew	Romanian
Qwen2.5-VL-7B	Base	57.26	64.55	46.57	46.97	51.15	49.64	37.68
	VFA	56.42	65.30	49.82	50.00	51.15	49.29	40.85
	Δ *Gains*	–0.84	+0.75	+3.25	+3.03	+0.00	–0.35	+3.17
Idefics3-8B	Base	55.25	44.78	36.46	43.56	61.54	46.43	38.38
	VFA	52.14	56.72	39.71	44.32	63.08	54.29	41.55
	Δ *Gains*	–3.11	+11.94	+3.25	+0.76	+1.54	+7.86	+3.17
LLaVA-Next-8B	Base	49.81	33.58	29.60	33.33	19.62	20.71	26.41
	VFA	48.64	53.36	41.88	40.53	41.54	37.14	37.32
	Δ *Gains*	–1.17	+19.78	+12.28	+7.20	+21.92	+16.43	+10.91
LLaVA-OV-1.5-8B	Base	55.25	66.04	49.10	61.74	56.54	48.57	41.90
	VFA	60.70	68.28	50.90	62.88	58.85	46.79	46.13
	Δ *Gains*	+5.45	+2.24	+1.80	+1.14	+2.31	–1.78	+4.23
LLaVA-OV-1.5-4B	Base	57.59	61.94	41.52	56.82	49.23	40.36	40.85
	VFA	62.65	63.06	48.38	56.44	51.92	41.07	39.79
	Δ *Gains*	+5.06	+1.12	+6.86	–0.38	+2.69	+0.71	–1.06

Table 14: Performance on MAXM across multiple languages.

Model	Method	Bengali	German	English	Indonesian	Korean	Portuguese	Russian	Chinese
Qwen2.5-VL-7B	Base	44.60	47.70	64.70	43.20	45.60	48.30	49.70	38.70
	VFA	45.10	49.50	63.00	44.30	46.40	48.60	50.80	39.90
	Δ *Gains*	+0.50	+1.80	–1.70	+1.10	+0.80	+0.30	+1.10	+1.20
Idefics3-8B	Base	39.80	50.20	57.90	46.00	48.10	48.20	48.80	50.80
	VFA	44.20	50.00	56.30	46.20	46.90	47.10	46.60	47.50
	Δ *Gains*	+4.40	–0.20	–1.60	+0.20	–1.20	–1.10	–2.20	–3.30
LLaVA-Next-8B	Base	12.20	43.30	68.20	40.00	43.20	47.00	45.20	49.70
	VFA	17.20	49.20	63.80	41.40	44.70	50.20	48.00	48.60
	Δ *Gains*	+5.00	+5.90	–4.40	+1.40	+1.50	+3.20	+2.80	–1.10
LLaVA-OV-1.5-8B	Base	40.60	30.70	64.50	28.10	17.60	27.80	29.50	0.00
	VFA	45.50	52.40	62.20	49.40	42.20	52.20	47.10	0.80
	Δ *Gains*	+4.90	+21.70	–2.30	+21.30	+24.60	+24.40	+17.60	+0.80
LLaVA-OV-1.5-4B	Base	37.40	51.00	64.70	40.00	18.60	44.50	35.60	0.10
	VFA	38.40	51.30	64.30	40.70	23.70	44.60	38.70	0.00
	Δ *Gains*	+1.00	+0.30	–0.40	+0.70	+5.10	+0.10	+3.10	–0.10

Table 15: Performance on XGQA across multiple languages.

length limit (as seen in Example 1). This qualitative evidence confirms that the underlying translation capability is preserved, and the observed BLEU regression is primarily a generation-length artifact.

F Recommended Merge Defaults

As shown in Table 13, we establish practical guidelines for model merging based on our empirical findings.

G Monolingual Adaptation

To demonstrate the versatility of our method, beyond the multilingual setting, we also evaluate VFA under monolingual adaptation by individually fine-tuning on Hindi and Romaniantwo languages not included in our main multilingual suite. Results are reported in Table 20, showing that VFA remains effective for targeted, language-specific adaptation.

Model	Method	Arabic	English	French	Hindi	Indonesian	Japanese	Portuguese
Qwen2.5-VL-7B	Base	42.60	50.20	50.70	44.00	49.20	46.10	51.50
	VFA	43.30	51.10	48.70	42.60	48.10	46.80	50.20
	Δ *Gains*	+0.70	+0.90	–2.00	–1.40	–1.10	+0.70	–1.30
Idefics3-8B	Base	37.20	42.70	44.60	42.60	43.40	46.10	43.80
	VFA	38.60	44.40	44.00	43.60	39.70	43.50	46.80
	Δ *Gains*	+1.40	+1.70	–0.60	+1.00	–3.70	–2.60	+3.00
LLaVA-Next-8B	Base	38.90	38.40	41.30	34.70	35.40	36.10	37.70
	VFA	35.90	41.70	39.60	34.00	36.40	35.70	40.40
	Δ *Gains*	–3.00	+3.30	–1.70	–0.70	+1.00	–0.40	+2.70
LLaVA-OV-1.5-8B	Base	51.30	56.10	55.40	49.80	58.60	52.40	56.20
	VFA	53.00	55.60	53.70	47.40	57.60	50.90	53.20
	Δ *Gains*	+1.70	–0.50	–1.70	–2.40	–1.00	–1.50	–3.00
LLaVA-OV-1.5-4B	Base	52.30	53.80	56.70	46.40	53.50	53.90	56.60
	VFA	51.30	54.20	55.00	47.10	54.20	54.30	56.20
	Δ *Gains*	–1.00	+0.40	–1.70	+0.70	+0.70	+0.40	–0.40

Table 16: Performance on xMMMU (validation set split) across multiple languages.

Model	Method	English	Indonesian	Swahili	Tamil	Turkish	Chinese
Qwen2.5-VL-7B	Base	62.00	67.00	44.00	67.00	39.00	42.00
	VFA	75.00	71.00	54.00	67.00	61.00	67.00
	Δ *Gains*	+13.00	+4.00	+10.00	+0.00	+22.00	+25.00
Idefics3-8B	Base	49.00	18.00	25.00	19.00	23.00	25.00
	VFA	71.00	59.00	58.00	61.00	70.00	57.00
	Δ *Gains*	+22.00	+41.00	+33.00	+42.00	+47.00	+32.00
LLaVA-Next-8B	Base	31.00	54.00	52.00	45.00	42.00	54.00
	VFA	54.00	49.00	56.00	54.00	56.00	63.00
	Δ *Gains*	+23.00	–5.00	+4.00	+9.00	+14.00	+9.00
LLaVA-OV-1.5-8B	Base	67.00	54.00	51.00	60.00	74.00	67.00
	VFA	69.00	65.00	52.00	63.00	73.00	71.00
	Δ *Gains*	+2.00	+11.00	+1.00	+3.00	–1.00	+4.00
LLaVA-OV-1.5-4B	Base	65.00	59.00	50.00	61.00	66.00	62.00
	VFA	64.00	60.00	50.00	61.00	67.00	62.00
	Δ *Gains*	–1.00	+1.00	+0.00	+0.00	+1.00	+0.00

Table 17: Performance on MaRVL across multiple languages.

Model	Method	Afrikaans	Chinese	English	Italian	Portuguese	Thai	Vietnamese
Qwen2.5-VL-7B	Base	60.74	86.37	67.09	67.00	50.44	40.40	39.66
	VFA	61.96	86.14	67.47	67.00	52.67	40.40	41.38
	Δ *Gains*	+1.22	–0.23	+0.38	+0.00	+2.23	+0.00	+1.72
Idefics3-8B	Base	34.97	22.40	23.33	48.87	24.00	23.94	27.59
	VFA	56.44	54.27	55.61	54.66	46.00	34.66	32.76
	Δ *Gains*	+21.47	+31.87	+32.28	+5.79	+22.00	+10.72	+5.17
LLaVA-Next-8B	Base	42.94	48.04	53.72	51.13	39.33	31.92	34.48
	VFA	41.10	48.27	55.61	52.64	42.00	31.17	36.21
	Δ *Gains*	–1.84	+0.23	+1.89	+1.51	+2.67	–0.75	+1.73
LLaVA-OV-1.5-8B-Inst	Base	69.94	73.67	69.86	70.78	54.22	45.39	55.17
	VFA	69.33	74.83	69.99	70.53	54.67	44.14	51.72
	Δ *Gains*	–0.61	+1.16	+0.13	–0.25	+0.45	–1.25	–3.45
LLaVA-OV-1.5-4B-Inst	Base	63.80	66.28	66.71	67.00	51.56	43.14	47.41
	VFA	66.26	66.51	67.34	65.24	52.44	45.14	45.69
	Δ *Gains*	+2.46	+0.23	+0.63	–1.76	+0.88	+2.00	–1.72

Table 18: Performance of different MLLMs on M3Exam across multiple languages.

Model	Method	Arabic	Bengali	Czech	Danish	German	Greek	English	Spanish	Persian
Qwen2.5-VL-7B	Base	11.04	12.74	16.65	22.24	17.69	12.07	25.30	23.94	20.57
	VFA	11.38	10.49	15.43	20.82	16.61	12.67	28.34	24.02	22.02
	Δ *Gains*	+0.34	–2.25	–1.22	–1.42	–1.08	+0.60	+3.04	+0.08	+1.45
Idefics3-8B	Base	11.65	15.59	13.15	20.34	14.35	11.13	23.86	21.10	16.09
	VFA	12.30	16.07	12.94	22.99	17.87	13.66	30.37	24.31	20.95
	Δ *Gains*	+0.65	+0.48	–0.21	+2.65	+3.52	+2.53	+6.51	+3.21	+4.86
LLaVA-Next-8B	Base	0.00	0.00	1.98	1.04	2.05	0.00	30.38	1.48	0.00
	VFA	8.28	8.79	13.58	23.23	17.65	7.30	31.04	26.18	21.79
	Δ *Gains*	+8.28	+8.79	+11.60	+22.19	+15.60	+7.30	+0.66	+24.70	+21.79
Model	**Method**	**Finnish**	**Filipino**	**French**	**Hebrew**	**Hindi**	**Croatian**	**Hungarian**	**Indonesian**	**Italian**
Qwen2.5-VL-7B	Base	9.29	21.13	23.78	10.66	15.82	13.41	7.16	24.45	20.86
	VFA	9.04	20.60	21.82	12.09	14.99	12.59	16.09	22.06	20.36
	Δ *Gains*	–0.25	–0.53	–1.96	+1.43	–0.83	–0.82	+8.93	–2.39	–0.50
Idefics3-8B	Base	6.26	20.11	19.03	8.76	12.94	9.93	13.02	22.97	17.28
	VFA	8.84	23.71	23.45	13.97	15.72	11.64	15.38	25.84	22.11
	Δ *Gains*	+2.58	+3.60	+4.42	+5.21	+2.78	+1.71	+2.36	+2.87	+4.83
LLaVA-Next-8B	Base	0.00	0.39	0.91	1.71	0.00	0.00	0.07	3.25	0.67
	VFA	9.56	18.49	25.72	11.42	13.96	8.86	14.36	19.12	22.32
	Δ *Gains*	+9.56	+18.10	+24.81	+9.71	+13.96	+8.86	+14.29	+15.87	+21.65
Model	**Method**	**Japanese**	**Korean**	**Maori**	**Dutch**	**Norwegian**	**Polish**	**Portuguese**	**Quechua**	**Romanian**
Qwen2.5-VL-7B	Base	1.29	5.65	12.71	24.38	21.24	16.52	23.76	0.63	16.11
	VFA	1.39	6.52	12.79	25.27	19.76	14.69	22.81	0.23	15.94
	Δ *Gains*	+0.10	+0.87	+0.08	+0.89	–1.48	–1.83	–0.95	–0.40	–0.17
Idefics3-8B	Base	6.43	4.64	21.88	22.23	17.55	11.15	18.70	1.49	13.14
	VFA	6.11	6.61	16.76	25.51	20.29	16.30	24.02	1.56	17.18
	Δ *Gains*	–0.32	+1.97	–5.12	+3.28	+2.74	+5.15	+5.32	+0.07	+4.04
LLaVA-Next-8B	Base	3.91	0.00	0.00	0.14	3.36	0.62	0.39	1.90	0.47
	VFA	4.63	5.79	8.33	27.99	19.68	14.07	24.55	0.31	16.83
	Δ *Gains*	+0.72	+5.79	+8.33	+27.85	+16.32	+13.45	+24.16	–1.59	+16.36
Model	**Method**	**Russian**	**Swedish**	**Swahili**	**Telugu**	**Thai**	**Turkish**	**Ukrainian**	**Vietnamese**	**Chinese**
Qwen2.5-VL-7B	Base	23.06	23.87	9.70	4.74	0.54	13.34	14.72	28.61	6.97
	VFA	23.42	23.20	8.61	4.97	0.54	12.20	14.45	29.36	8.06
	Δ *Gains*	+0.36	–0.67	–1.09	+0.23	+0.00	–1.14	–0.27	+0.75	+1.09
Idefics3-8B	Base	14.62	20.33	10.47	6.08	1.19	13.65	9.31	22.23	4.21
	VFA	16.54	23.04	11.28	6.07	0.31	14.28	11.89	26.49	6.99
	Δ *Gains*	+1.92	+2.71	+0.81	–0.01	–0.88	+0.63	+2.58	+4.26	+2.78
LLaVA-Next-8B	Base	1.51	0.34	1.43	0.08	0.00	0.19	0.20	0.23	0.00
	VFA	15.42	21.89	9.56	2.76	0.00	10.32	12.44	18.99	4.43
	Δ *Gains*	+13.91	+21.55	+8.13	+2.68	+0.00	+10.13	+12.24	+18.76	+4.43

Table 19: Performance of different MLLMs on XM100 across multiple languages.

Model	Method	MaXM (hi)	XMMU (hi)	MaXM (ro)
LLaVA-Next-8B	Base	19.23	34.70	26.41
	VFA	44.23	35.70	33.45
	Δ *Gains*	+25.00	+1.00	+7.04
Idefics3-8B	Base	62.31	42.30	37.68
	VFA	65.77	46.00	44.01
	Δ *Gains*	+3.46	+3.70	+6.33

Table 20: Multilingual performance of VFA on Hindi (hi) and Romanian (ro) benchmarks.