DiM^3: Bridging Multilingual and Multimodal Models via Direction- and Magnitude-Aware Merging

Zijing Wang[1] **Mingyang Wang**[2,3] **Ercong Nie**[4] **Yongkang Liu**[1*] **Shi Feng**[1]
Mengjie Zhao[1] **Daling Wang**[1*] **Xiaocui Yang**[1] **and Hinrich Schütze**[2,3]
[1]Northeastern University, China; [2]CIS, LMU Munich, Germany
[3]Munich Center for Machine Learning (MCML), Germany
[4]Shanghai Jiao Tong University, China;
wzj1718@gmail.com

Abstract

Towards more general and human-like intelligence, large language models should seamlessly integrate both multilingual and multimodal capabilities; however, extending an existing multimodal model to many languages typically requires expensive multilingual multimodal data construction and repeated end-to-end retraining. We study a training-free alternative: injecting multilingual capability into an existing multimodal model by composing residual updates in the shared language model backbone. The key challenge is that multilingual and multimodal updates are heterogeneous, reflecting different functional roles in the shared model. To address this, we propose **Di**rection- and **M**agnitude-aware **M**ultilingual **M**ultimodal merging (**DiM^3**), which selectively composes the two updates at each parameter dimension while preserving the original vision encoder and multimodal projector. Experiments on multilingual benchmarks in both text-only and vision-language settings, covering 57 languages across LLaVA- and Qwen-based backbones, show that DiM^3 consistently outperforms existing merging baselines, substantially improves multilingual performance over the original multimodal model, and remains competitive with dedicated multilingual multimodal fine-tuning while largely retaining general multimodal ability. We further show that DiM^3 can be directly applied to already trained multilingual multimodal models and still yield additional gains. Further interpretability analysis shows that DiM^3 primarily reshapes intermediate-layer semantic representations, strengthening cross-lingual alignment under both text-only and multimodal inputs while preserving higher-layer task-sensitive structure. Our repository is on https://github.com/wzj1718/DiM3.

1 Introduction

Multimodal large language models (MLLMs) [1, 2, 3, 4] have demonstrated remarkable capabilities across a wide range of multimodal tasks, particularly in high-resource languages (HRLs). However, their performance degrades substantially in low-resource languages (LRLs) [5, 6, 7, 8], where linguistic resources and multimodal training data are scarce. A key reason is that existing MLLMs remain highly English-centric, with pretraining dominated by English data and only limited multilingual exposure [9, 10, 11, 12, 13]. Such limited multilingual exposure is insufficient to support reliable reasoning and knowledge generalization in low-resource settings, leading to persistent performance gaps.

To improve MLLMs performance in LRLs, prior work typically relies on supervised fine-tuning (SFT) or continual pre-training (CPT) [14]. SFT adapts models to downstream tasks using large amounts of

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

labeled data, while CPT leverages large-scale target-language corpora to enhance language modeling and fluency [15]. Despite their effectiveness, both approaches require substantial amounts of data, either labeled or unlabeled, and extensive computational resources, making them costly and difficult to scale. These limitations motivate more efficient alternatives that can enhance multilingual capability without data-intensive retraining.

Model merging [16, 17, 18, 19, 20] has emerged as a promising direction, enabling the integration of complementary capabilities from separately trained models without costly large-scale training. Existing methods typically represent the parameter differences between a fine-tuned model and a shared base model as task vectors [21] and combine them through importance-weighted aggregation [22, 23, 24]. While effective for relatively homogeneous model merging settings, these approaches struggle when task vectors are heterogeneous, differing in modality, geometric structure, and training origin, as well as functional roles. In particular, multilingual updates primarily enhance cross-lingual semantic alignment [25, 26, 27], whereas multimodal updates adapt the backbone for visual grounding [28, 29]. Naively combining such mismatched updates can therefore lead to interference and degraded performance.

To address this limitation, we propose **DiM^3**, a direction- and magnitude-aware merging strategy tailored to heterogeneous multilingual multimodal composition. Our key insight is that task vectors encode two complementary signals [30, 31]: **magnitude**, reflecting the strength of parameter updates, and **direction**, capturing their functional reorientation relative to the base model. This decomposition is crucial in heterogeneous settings, where multilingual and multimodal updates may differ not only in update scale, but also in their functional directions. Instead of assuming global compatibility, **DiM^3** performs *selective composition* rather than uniform aggregation, adaptively assigning their contributions based on local geometric importance in parameter space. By jointly leveraging magnitude and direction, the method strengthens cross-lingual semantic alignment where multilingual signals are salient, while preserving multimodal grounding where multimodal updates dominate, enabling training-free multilingual transfer without compromising visual understanding.

We evaluate **DiM^3** on multilingual text-only and vision-language benchmarks covering 57 languages and multiple backbone families. Our method consistently outperforms existing merging baselines and substantially improves over the original multimodal model, while remaining competitive with dedicated multilingual multimodal fine-tuning despite requiring no additional training. Furthermore, DiM^3 can be applied to already trained multilingual multimodal models to yield additional gains, while maintains strong performance on general multimodal benchmarks. Analysis of internal representations shows that these improvements arise from reshaping intermediate-layer semantics, leading to stronger cross-lingual alignment under both text-only and multimodal inputs.

Our contributions are threefold: **(1)** We formulate multilingual multimodal transfer as a training-free residual composition problem on a shared backbone, and demonstrate that multilingual and multimodal task vectors exhibit fundamental heterogeneity; to the best of our knowledge, this setting has not been previously explored. **(2)** We propose a direction- and magnitude-aware merging rule that selectively composes multilingual and multimodal updates at the parameter level, going beyond naive addition to better capture their distinct functional roles. **(3)** We demonstrate that this approach substantially improves multilingual multimodal performance while largely preserving general multimodal ability, can be directly applied to already trained multilingual multimodal models for further gains, and achieves these improvements by strengthening cross-lingual alignment in intermediate-layer representations.

2 Related Work

Multilingual Representation in LLMs. Recent work suggests that multilingual capability in LLMs is supported not only by token coverage, but also by cross-lingual alignment in internal representations [32, 33, 34, 35]. Such alignment has been linked to stronger cross-lingual transfer and multilingual generalization [36, 37]. Layer-wise analyses further indicate that middle layers tend to encode more transferable, language-agnostic information, whereas upper layers are more closely tied to language-specific generation [38, 39]. These results imply that multilingual fine-tuning produces non-uniform, functionally structured residual changes in the shared backbone.

Multimodal LLMs and Multilingual Multimodal Models. MLLMs are typically built from a vision encoder, a multimodal projector, and a pretrained language model, with much of multimodal reasoning mediated by the shared language backbone [40, 41, 42]. Existing efforts to improve multilingual capability in MLLMs therefore mainly rely on additional training, including multilingual instruction tuning, large-scale multilingual multimodal data construction, and language-aware visual instruction learning [43, 9, 10, 44]. Recent studies also show that multimodal training can degrade multilingual fidelity, for example by inducing English-biased outputs under visual inputs, and address this issue through multilingual regularization or layer-specific enhancement [45, 46]. While effective, these approaches remain training- and data-intensive, motivating our study of whether training-free parameter composition can inject multilingual capability into an existing multimodal model.

Model Merging. Model merging aims to compose capabilities from independently adapted checkpoints directly in parameter space [47, 48, 49, 50, 51, 52]. Early work shows that fine-tuning residuals can be treated as task vectors and combined through simple arithmetic [21]. Subsequent methods improve this paradigm by reducing redundancy and mitigating interference, for example through sparsification and rescaling [22], sign-aware conflict resolution [23], and sparse update selection [53, 54]. More recent work explores finer-grained compatibility modeling: PCB-Merging balances parameter importance and inter-task competition during composition [55], STF formulates merging through the superposition of task-specific features [56] in representation space, and NeuroMerging further studies interference disentanglement at the neuron level [57]. However, most existing methods are designed for general multi-task composition and do not explicitly address the heterogeneous updates considered here. In our setting, multilingual adaptation and multimodal adaptation play different functional roles in a shared backbone, making uniform residual addition insufficient. Our method therefore focuses on a direction-magnitude aware composition rule tailored to multilingual multimodal transfer.

3 Methodology

3.1 Problem Setup

In our setting, we consider three models that share the same language model backbone: a base model N, a multilingual model M, and a multimodal model V. The multilingual model extends N through continual pre-training on large-scale multilingual data to improve cross-lingual understanding and generation, while the multimodal model augments the same backbone with a vision encoder and a multimodal projector for grounded reasoning, i.e., $V = V^{\mathrm{lm}} \otimes V^{\mathrm{vis}} \otimes V^{\mathrm{proj}}$. As the backbone is the only component shared across all three models, transfer is performed in this parameter space.

Let θ_N and θ_M denote the parameters of the base model N and the multilingual model M, and let $\theta_{V^{\mathrm{lm}}}$ denote the language model parameters inside the multimodal model V. We define two source residuals relative to the same base point:

$$\Delta_{\mathrm{ml}} = \theta_M - \theta_N, \qquad \Delta_{\mathrm{mm}} = \theta_{V^{\mathrm{lm}}} - \theta_N. \tag{1}$$

Here Δ_{ml} captures multilingual adaptation and Δ_{mm} captures the language-side adaptation induced by multimodal training.

Our goal is to inject multilingual capability into V while preserving the visual grounding and multimodal coordination already established by its visual modules. We therefore keep the vision encoder and multimodal projector fixed, and only replace the shared language model backbone with a merged backbone $\widetilde{\theta}_{V^{\mathrm{lm}}}$. Under the standard residual view of fine-tuning, this reduces the problem to composing two source residuals around a common reference point, rather than retraining a new multilingual multimodal model from scratch.

3.2 Residual Heterogeneity and Geometric Motivation

A natural baseline is direct task arithmetic,

$$\theta_{\mathrm{merge}} \approx \theta_N + \Delta_{\mathrm{ml}} + \Delta_{\mathrm{mm}}, \tag{2}$$

but this approximation is reliable only when the two residuals are sufficiently compatible [58]. In our setting, multilingual and multimodal fine-tuning are driven by different supervision signals: the former primarily improves cross-lingual semantic alignment and generation, whereas the latter reshapes the

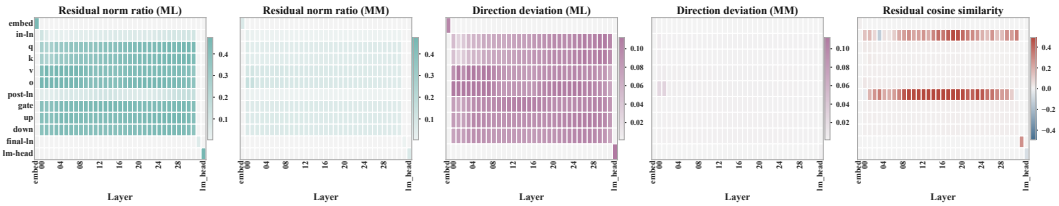

Figure 1: **Residual heterogeneity in the shared language model backbone.** The panels show residual norm, base-relative reorientation, and cross-residual alignment for Δ_{ml} and Δ_{mm} across layers and modules. Together, these diagnostics reveal that multilingual and multimodal adaptations differ in both update magnitude and geometry, motivating selective rather than uniform composition.

same backbone to cooperate with projected visual features. This motivates a *heterogeneity hypothesis*: multilingual and multimodal residuals should not be treated as interchangeable parameter offsets, because they support different functional roles in the shared backbone.

To characterize the heterogeneity between the multilingual and multimodal updates, we study three complementary diagnostics: (1) *residual norm*, which measures update magnitude; (2) *base-relative reorientation*, which captures directional change relative to the base model; and (3) *cross-residual alignment*, which measures how well the multilingual and multimodal updates align with each other.

For each layer weight matrix $W \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$, we adopt a DoRA-style column-wise decomposition [31], representing each column as a magnitude-direction pair:

$$m(W)_j = \|W_{:,j}\|_2, \qquad D(W)_{:,j} = \frac{W_{:,j}}{m(W)_j + \epsilon}, \tag{3}$$

where j indexes the input column and ϵ is a small numerical stabilizer. Relative to the base model N, we define magnitude deviation and directional deviation for each source $k \in \{\mathrm{ml}, \mathrm{mm}\}$ as

$$\delta^{\mathrm{mag}}_{k,j} = \left| m_k(j) - m_N(j) \right|, \qquad \delta^{\mathrm{dir}}_{k,j} = 1 - \cos\big(D_k(:,j), D_N(:,j)\big). \tag{4}$$

Additionally, we quantify cross-residual alignment through the cosine similarity between the two source residuals at the same column.

Proposition 3.1 (Residual geometry)**.** *Ignoring ϵ for clarity, the squared residual norm of column j decomposes as*

$$\|W_{k,:,j} - W_{N,:,j}\|_2^2 = \big(\delta^{\mathrm{mag}}_{k,j}\big)^2 + 2\, m_k(j) m_N(j)\, \delta^{\mathrm{dir}}_{k,j}. \tag{5}$$

Proof. Since $W_{k,:,j} = m_k(j) D_k(:,j)$, $W_{N,:,j} = m_N(j) D_N(:,j)$, and $\|D_k(:,j)\|_2 = \|D_N(:,j)\|_2 = 1$,

$$\|W_{k,:,j} - W_{N,:,j}\|_2^2 = m_k(j)^2 + m_N(j)^2 - 2 m_k(j) m_N(j) \cos\big(D_k(:,j), D_N(:,j)\big). \tag{6}$$

Rearranging gives

$$\big(m_k(j) - m_N(j)\big)^2 + 2 m_k(j) m_N(j) \big(1 - \cos\big(D_k(:,j), D_N(:,j)\big)\big) \tag{7}$$

$$= \big(\delta^{\mathrm{mag}}_{k,j}\big)^2 + 2\, m_k(j) m_N(j)\, \delta^{\mathrm{dir}}_{k,j}, \tag{8}$$

which is exactly the claimed decomposition. □

This decomposition shows that base-relative residual departure contains both a radial term and an angular term. In our setting, it provides a geometric basis for treating magnitude deviation and direction deviation as complementary signals: differences in norm structure favor magnitude-aware selection, while differences in reorientation favor direction-aware selection. Figure 1 confirms that Δ_{ml} and Δ_{mm} differ across layers and modules not only in residual norm, but also in base-relative reorientation and cross-residual alignment. This pattern indicates that the compatibility between multilingual and multimodal updates is local rather than global: some parts of backbone should remain closer to multimodal structure, whereas others can absorb stronger multilingual transfer. We therefore seek a selective merge strategy that uses both magnitude and direction signals.

3.3 Direction- and Magnitude-Aware Multilingual Multimodal Merging

Following the analysis above, we merge the two source residuals in a column-wise manner within the shared language model backbone. We use column-wise aggregation because it is a natural unit for both the geometry and the function of the linear map. The magnitude-direction decomposition is defined on column vectors, and each column $W_{:,j}$ determines how input dimension j contributes to all output dimensions. For each source $k \in \{\mathrm{mm}, \mathrm{ml}\}$ and each column j, we use both $\delta^{\mathrm{mag}}_{k,j}$ and $\delta^{\mathrm{dir}}_{k,j}$ to estimate source salience. The former measures how strongly a source rewrites a column, while the latter measures how differently that column is reoriented relative to the base model.

For either branch $* \in \{\mathrm{mag}, \mathrm{dir}\}$, let $\delta^{*}_{k,j}$ denote the corresponding deviation value. We first normalize deviations within each matrix by their relative ranks across columns:

$$r^{*}_{k,j} = \frac{\mathrm{rank}(\delta^{*}_{k,j})}{d_{\mathrm{in}}}. \tag{9}$$

We then convert these normalized ranks into cross-source salience scores:

$$s^{*}_{k,j} = \frac{\exp(r^{*}_{k,j})}{\sum_{k'} \exp(r^{*}_{k',j})}. \tag{10}$$

This yields a relative importance score based on within-source column salience rather than absolute deviation scale. As suggested by Figure 1, the multilingual and multimodal residuals exhibit substantial cross-source heterogeneity across layers and modules, both in residual norm and direction deviation. In this setting, raw deviations are not directly comparable across the two sources: a large raw value may reflect a broader deviation range for one source in a given matrix, rather than stronger relative salience at the current column. We therefore compare the two sources through within-matrix ranks, which preserve relative ordering inside each source while reducing sensitivity to source- and module-specific scale variation. Rank normalization also makes the comparison less sensitive to a small number of extreme entries. We validate this design through salience-estimator ablations in Section 5.4.

In the two-source case, the softmax can be written as

$$s^{*}_{\mathrm{ml},j} = \sigma\big(r^{*}_{\mathrm{ml},j} - r^{*}_{\mathrm{mm},j}\big), \qquad s^{*}_{\mathrm{mm},j} = 1 - s^{*}_{\mathrm{ml},j}, \tag{11}$$

where $\sigma(x) = 1/(1+e^{-x})$. Hence the gate is driven by the salience gap between the multilingual and multimodal residuals, rather than by absolute update scale.

Since rank-softmax enforces a relative allocation across sources, it may attenuate residual contributions even when the corresponding dimensions exhibit high within-source salience. To preserve such contributions, we apply a salience calibration step before merging. For each source k and branch $* \in \{\mathrm{mag}, \mathrm{dir}\}$, we calibrate the corresponding salience score as

$$\widetilde{s}^{*}_{k,j} = \begin{cases} 1.0, & r^{*}_{k,j} > \rho\left(\frac{1}{d_{\mathrm{in}}}\sum_{j'=1}^{d_{\mathrm{in}}} r^{*}_{k,j'}\right), \\ s^{*}_{k,j}, & \text{otherwise.} \end{cases} \tag{12}$$

We set $\rho = 0.8$ throughout all experiments. Thus, the rank-softmax score serves as the default cross-source allocation, while dimensions whose within-source salience exceeds the calibration threshold retain the full corresponding residual contribution.

We aggregate the two calibrated salience branches by simple averaging:

$$\omega_{k,j} = \frac{1}{2}\left(\widetilde{s}^{\mathrm{mag}}_{k,j} + \widetilde{s}^{\mathrm{dir}}_{k,j}\right). \tag{13}$$

This symmetric design treats magnitude- and direction-based evidence equally by default, and is later validated in the ablation study. It can also be interpreted as the closed-form solution of the following consensus objective:

$$\omega_j = \arg\min_{\omega \in \mathbb{R}^2} \|\omega - \widetilde{s}^{\mathrm{mag}}_{j}\|_2^2 + \|\omega - \widetilde{s}^{\mathrm{dir}}_{j}\|_2^2. \tag{14}$$

Thus, a column receives a larger multilingual weight when multilingual salience is high in either branch, and remains closer to the multimodal update when multimodal salience dominates. The role of this symmetric aggregation is examined in Section 5.4.

The merged parameter is then computed column-wise as

$$\widetilde{W}_{:,j} = W_{N,:,j} + \sum_{k\in\{\mathrm{mm},\mathrm{ml}\}} \omega_{k,j}\, \Delta_k^{(W)}{}_{:,j}. \tag{15}$$

For one-dimensional parameters such as normalization weights or bias terms, the direction-magnitude decomposition is not available. We therefore use absolute base-relative deviation,

$$\delta_{k,i}^{\mathrm{1d}} = |w_{k,i} - w_{N,i}|, \tag{16}$$

derive element-wise salience scores in the same rank-based manner, and merge by

$$\widetilde{w}_i = w_{N,i} + \sum_{k\in\{\mathrm{mm},\mathrm{ml}\}} \gamma_{k,i}\, \Delta_k^{(w)}{}_i. \tag{17}$$

After merging all shared backbone parameters, we write $\widetilde{\theta}_{V^{\mathrm{lm}}}$ back into the multimodal anchor while leaving the vision encoder and projector unchanged. The final merged multimodal model is

$$\widetilde{V} = \big(\theta_{V^{\mathrm{vis}}}, \theta_{V^{\mathrm{proj}}}, \widetilde{\theta}_{V^{\mathrm{lm}}}\big). \tag{18}$$

Table 1: Role assignment of models in each shared-backbone setting.

Family	Backbone	MM anchor	Multilingual expert	Extra references
LLaMA	LLaMA-2-7B [59]	LLaVA-v1.5-7B [29]	Emma-500 [60]	PLAST-7B [46]; Palo-7B/13B [43]
Qwen2	Qwen2-7B-Instruct [61]	Pangea-7B [9]	Qwen2-7B-Multilingual-RP	Pangea-7B
Qwen3	Qwen3-8B [4]	Qwen3-VL-8B-Instruct [62]	AfriqueQwen-8B [63]	Qwen3-VL-8B-Instruct

4 Experimental Setup

Baselines We conduct experiments under three shared-backbone settings, summarized in Table 1. In all cases, merging is performed only on the shared language model parameters, while the visual components of the multimodal model are kept fixed.

For **merging method baselines**, we compare DiM^3 with representative training-free merging approaches, including Task Arithmetic [21], DARE [22], TIES-Merging [23], Breadcrumbs [53], PCB-Merging [55], STF [56], and NeuroMerging [57]. These methods span several major directions in model merging, including direct task vector composition, sparsity-based selection, conflict resolution, and more structured compatibility modeling. This comparison allows us to test whether DiM^3 provides advantages over existing merging strategies in multilingual and multimodal settings. Details on these baselines are provided in Appendix A.

For **model baselines**, we report the performance of the backbone-compatible source models and multilingual multimodal reference systems used in each setting. In the LLaMA-based setup, these include the base model LLaMA-2-7B, the multilingual model Emma-500, the multimodal anchor LLaVA-v1.5-7B, PLAST-7B [46], and Palo-7B/13B [43], where PLAST-7B is a parameter-efficient multilingual enhancement of LLaVA-v1.5-7B and Palo is a multilingual multimodal model fine-tuned from the same backbone. In the Qwen3-based setting, we further report Qwen3-VL-8B-Instruct [62] and AfriqueQwen-8B [63] as the multimodal and multilingual source models, respectively. To examine whether DiM^3 can also be directly applied to an existing multilingual multimodal model, we additionally conduct merging experiments based on Pangea-7B [9], using it as a pretrained multilingual multimodal checkpoint in the Qwen2 family.

Benchmarks We evaluate DiM^3 on three categories of benchmarks.

For **multilingual text-only tasks**, we consider XCOPA [64], XStoryCloze [65], and XNLI [66]. On these benchmarks, we evaluate the shared language model backbones of LLaMA-2-7B, Emma-500, LLaVA-v1.5-7B, and the models produced by different merging methods. This setting isolates multilingual semantic transfer in pure text scenarios and tests whether merging preserves the multilingual gains of the source model. Detailed benchmark settings are provided in Appendix A. We also list all languages used in our experiments and their corresponding ISO 639-1/2 codes in Appendix Table 7.

For **multilingual multimodal tasks**, we use xMMMU [9], MaXM [6], MaRVL [67], xGQA [5], CVQA [8], and Afri-MCQA [68]. Together, these benchmarks are used to evaluate multilingual

visual understanding and cross-lingual multimodal reasoning. In the LLaMA-based setting, we report results on all six benchmarks. We further include Afri-MCQA in the Qwen2- and Qwen3-based shared-backbone settings to examine DiM^3 under additional backbone families.

For **general multimodal tasks**, we report results on MMStar [69], MMMU [70], and SEED-Bench-2-Plus [71]. We use these benchmarks in the main LLaMA-based setting to assess whether multilingual transfer degrades the general multimodal competence of the anchor model.

All evaluation results are obtained using the **lm-evaluation-harness** [72] and **lmms-eval** [73] libraries under the same settings to ensure fair comparison. For the merging baselines, hyperparameters are chosen by grid search on validation sets, and we report the best configuration for each method.

5 Main Results

5.1 Multilingual Capability Transfer in Text-Only Benchmarks

Comparison with Existing Merging Methods on Multilingual Text Benchmarks. We first evaluate whether DiM^3 can transfer multilingual capability to the shared language backbone under text-only inputs. Following Section 3, we compose the multilingual residual from Emma-500 with the language-backbone residual of LLaVA-v1.5-7B around the common base model LLaMA-2, while keeping the visual modules fixed. For text-only evaluation, we extract the language model component from each merged model and evaluate it independently. Table 2 summarizes the average multilingual score $\mathrm{Avg}_{\mathrm{mul}}$ on XCOPA, XStoryCloze, and XNLI, while Appendix Tables 8, 9, 10, and 11 report the full language-wise results. Our method achieves the best average multilingual score on all three benchmarks, consistently outperforming existing merging baselines. This shows that direction-magnitude-aware merging is better suited to composing multilingual and multimodal residuals than prior parameter-merging rules.

Table 2: Results on multilingual text-only tasks. Best and second-best results within each dataset are highlighted.

Models	XCOPA	XStoryCloze	XNLI
language models			
Llama_2_7B	56.27	55.74	39.07
EMMA-500	62.47	64.77	**43.64**
LLaVA-v1.5-7B	51.98	49.48	35.55
merging methods			
Task Arithmetic	56.78	58.41	40.07
DARE	56.55	56.81	39.14
TIES-Merging	57.44	58.57	40.47
Breadcrumbs	56.98	57.49	40.42
PCB-Merging	55.55	54.18	37.14
STF	55.40	56.25	39.14
NeuroMerging	56.09	56.54	40.01
DiM^3	**63.38**	**64.82**	42.93

Comparison with Base Models on Multilingual Text Benchmarks. The same pattern holds when compared with backbone-compatible source models. Relative to LLaVA-v1.5-7B, DiM^3 yields large gains on all three text benchmarks, confirming that multilingual capability is effectively transferred into the shared language backbone. At the same time, it remains close to Emma-500 overall, and even surpasses it in $\mathrm{Avg}_{\mathrm{mul}}$ on XCOPA and XStoryCloze. Overall, these results show that our method can effectively inherit multilingual capability from the multilingual source model while maintaining compatibility with the original LLaVA backbone.

5.2 Multilingual Transfer in Multimodal Benchmarks

We next evaluate multilingual capability transfer under multimodal inputs. Table 3 summarizes the average multilingual score $\mathrm{Avg}_{\mathrm{mul}}$ on six benchmarks: xMMMU, MaXM, MaRVL, xGQA, CVQA, and Afri-MCQA, while Appendix Tables 12, 13, 14, and 15 report the full language-wise results. Across all six benchmarks, our method achieves the best average multilingual performance, showing that DiM^3 transfers multilingual capability effectively in multimodal settings.

A notable phenomenon is that some existing merging baselines become unstable in this heterogeneous multimodal setting. This is especially evident on generative multilingual tasks such as MaXM, where several methods produce repetitive, weakly relevant, or degenerate outputs; the problem is particularly severe for PCB-Merging, which we omit from later multimodal evaluations. This failure mode suggests that successful multimodal merging must preserve the coordination between language generation and visual grounding, not merely interpolate parameters. In this regard, DiM^3 is substantially more robust than prior merging strategies.

The appendix further shows that the gains are broadly distributed across languages and are particularly pronounced in lower-resource settings. Table 4 further validates this trend under the additional shared-backbone settings in Table 1, while Appendix Table 14 reports the corresponding results of different

merging methods. In the Qwen3 setting, DiM^3 improves Qwen3-VL-8B-Instruct from 0.383 to 0.452 average accuracy, confirming that the method remains effective across a different backbone family. In the Pangea-based Qwen2 setting, merging Pangea-7B with Qwen2-7B-Multilingual-RP raises the average score from 0.283 to 0.370, showing that DiM^3 can further improve an already fine-tuned multilingual multimodal model. Compared with dedicated multilingual multimodal fine-tuning baselines such as Palo, our method remains highly competitive while requiring no end-to-end multilingual multimodal retraining.

Table 3: Results on multilingual multimodal tasks.

Models	xMMMU	MaXM	MaRVL	xGQA	CVQA	Afri-MCQA
LLaVA-v1.5-7B	31.48	20.10	49.78	30.62	39.32	24.61
Task Arithmetic	32.67	22.27	49.81	36.54	40.47	24.39
DARE	31.38	19.54	49.76	31.51	38.42	24.32
TIES-Merging	32.27	24.23	49.71	36.64	39.33	23.84
Breadcrumbs	29.92	21.92	49.64	36.03	34.24	21.00
PCB-Merging	27.95	0.06	–	–	7.80	0.00
STF	32.05	11.09	50.87	10.87	34.59	23.86
NeuroMerging	33.47	20.53	49.84	20.98	39.38	23.67
Palo-7B	31.30	16.45	50.73	36.93	37.00	24.47
Palo-13B	31.30	28.68	49.70	30.60	39.59	23.95
DiM^3-7B	**33.85**	**33.44**	**51.89**	**48.04**	**43.78**	**26.31**

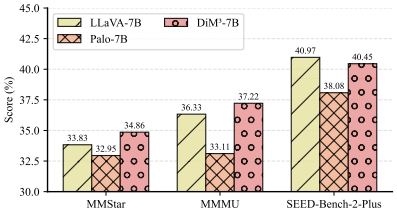

Figure 2: Results on three general multimodal benchmarks.

Table 4: Afri-MCQA results under the Qwen2- and Qwen3-based shared-backbone settings.

Backbone	Model	Am	Ha	Ig	Lug	Om	Rw	Ki	So	Ti	Ak	Yo	Tn	Ny	Zu	Sot	Ln	Avg
Qwen2	Pangea-7B	**0.357**	0.271	0.352	0.320	0.130	0.370	0.253	0.310	0.287	0.332	0.380	0.185	0.265	0.198	0.131	0.389	0.283
	DiM^3	0.317	**0.342**	**0.407**	**0.425**	**0.320**	**0.480**	**0.376**	**0.345**	**0.347**	**0.388**	**0.390**	**0.315**	**0.390**	**0.333**	**0.331**	**0.419**	**0.370**
Qwen3	Qwen3-VL-8B-Instruct	0.377	0.342	0.347	0.415	0.335	0.460	0.392	**0.405**	0.357	**0.430**	0.420	0.350	0.380	0.302	**0.323**	**0.490**	0.383
	DiM^3	**0.493**	**0.498**	**0.362**	**0.450**	**0.445**	**0.645**	**0.407**	**0.555**	**0.505**	0.388	**0.445**	**0.420**	**0.480**	**0.406**	0.277	0.460	**0.452**

5.3 Retaining General Multimodal Ability

We further examine whether multilingual transfer compromises general multimodal ability. As shown in Figure 2, DiM^3-7B remains comparable to the LLaVA-7B anchor, improving on MMStar and MMMU and staying close on SEED-Bench-2-Plus. Meanwhile, it consistently outperforms the directly fine-tuned multilingual multimodal baseline Palo-7B, with scores of 34.86 vs. 32.95, 37.22 vs. 33.11, and 40.45 vs. 38.08, respectively. These results indicate that DiM^3 better preserves the anchor model's general multimodal competence than the fine-tuned multilingual multimodal baseline. Overall, the trade-off is mild: the merged model remains comparable to LLaVA on general multimodal tasks while delivering substantially stronger multilingual gains.

5.4 Ablations

We ablate two components of the merge rule: the salience estimator in Eq. 9 and the branch aggregation rule in Eq. 13. Table 5 reports results on five representative multimodal benchmarks. For salience estimation, we compare the proposed rank-based softmax with four alternatives: Raw, which applies softmax directly to the original deviations; ZScore, which standardizes deviations within each tensor before softmax; MinMax, which replaces rank normalization with min-max normalization; and Ratio, which assigns source weights in proportion to their raw deviations without the softmax comparison. For branch aggregation, we compare the symmetric average in Eq. 13 with four alternatives: Dir-Weighted (Dir-W) and Mag-Weighted (Mag-W), which use asymmetric combinations of the two branches, and Magnitude-Only (Mag-Only) and Direction-Only (Dir-Only), which keep only one branch.

Table 5: Ablations on salience estimation and branch aggregation.

Variant	MMStar	MMMU	SEED	MaXM	MaRVL
DiM^3	**34.86**	**37.22**	40.45	**33.44**	51.89
MinMax	33.74	29.00	39.48	32.57	52.28
Ratio	2.56	25.33	10.01	3.58	6.94
Raw	33.64	31.00	39.48	33.00	52.45
ZScore	34.50	36.67	**40.80**	33.32	49.92
Dir-W	34.27	36.67	40.71	31.71	53.79
Mag-W	34.26	36.33	40.40	33.19	50.62
Mag-Only	33.89	34.78	40.10	33.34	51.91
Dir-Only	34.01	30.33	40.01	29.82	**55.86**

Two conclusions emerge. First, rank-based salience is the most robust estimator: Ratio degrades sharply on all five benchmarks, while Raw and MinMax are consistently weaker; ZScore is more competitive, but still underperforms our method on four of the five tasks. Second, the joint two-branch design matters: both Magnitude-Only and Direction-Only fall short of the full model, and both asymmetric variants are slightly worse than the symmetric average.

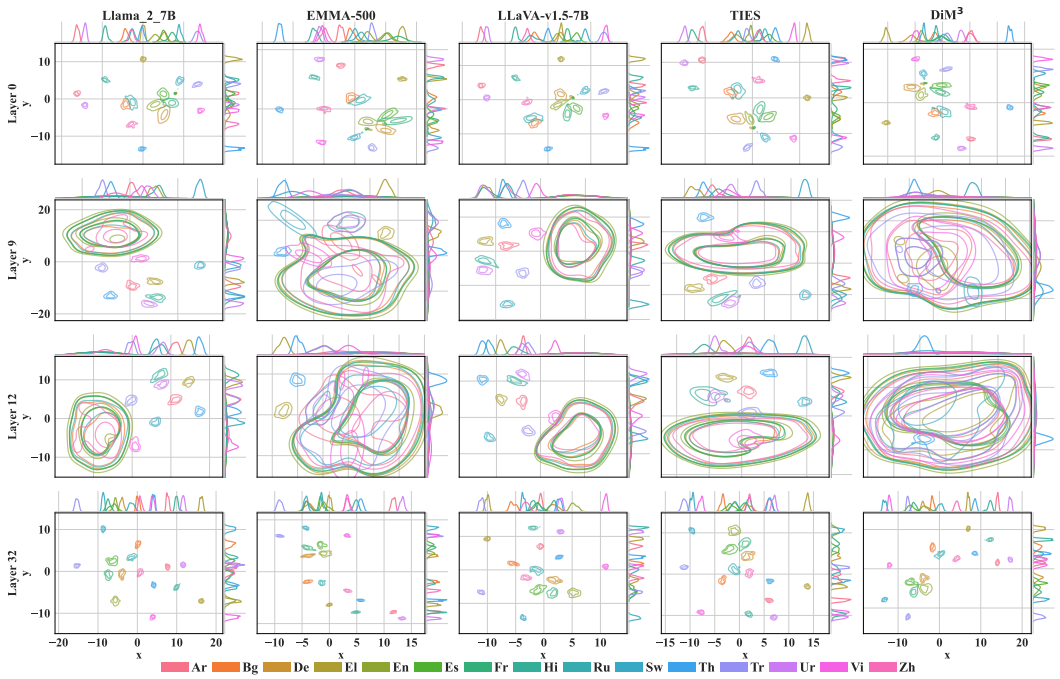

Figure 3: t-SNE visualizations of average-pooled hidden states on multilingual text inputs from XNLI across selected layers for LLaMA-2-7B, Emma-500, LLaVA-v1.5-7B, TIES-Merging, and DiM^3. Colors denote languages. DiM^3 shows more overlapping language clusters at intermediate layers, consistent with stronger cross-lingual alignment.

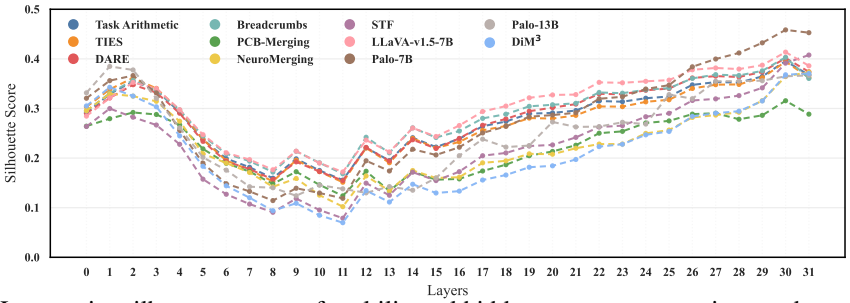

Figure 4: Layer-wise silhouette scores of multilingual hidden-state representations under multilingual multimodal inputs. Lower values indicate stronger cross-lingual overlap in the representation space.

6 Analysis: What Does the Merge Change?

Hidden-State Geometry on Multilingual Text To understand why merging improves multilingual transfer, we analyze the hidden-state geometry of multilingual text inputs. Figure 3 presents t-SNE visualizations of average-pooled hidden states on XNLI for LLaMA-2-7B, EMMA-500, LLaVA-v1.5-7B, TIES-Merging, and DiM^3; full-layer results and additional baselines are provided in Appendix Figures 6, 7, 8, and 9. Across models, we observe a consistent layer-wise pattern: lower layers remain largely language-dependent, reflecting lexical and surface-form information. In middle layers, however, DiM^3 exhibits substantially stronger cross-lingual overlap than LLaMA-2-7B, LLaVA-v1.5-7B, and existing merging baselines. In upper layers, representations become more dispersed again across languages, indicating the reintroduction of language-sensitive structure for generation. Overall, this analysis suggests that merging primarily reshapes the intermediate semantic space, strengthening cross-lingual alignment without disrupting higher-layer task-sensitive representations.

Hidden-State Geometry in Multimodal Multilingual Inputs We next examine whether this representation-level effect persists under multimodal conditioning. Appendix Figure 5 shows t-SNE

visualizations of average-pooled hidden states from xMMMU for LLaVA-v1.5-7B and our merged model, and Figure 4 reports the corresponding layer-wise silhouette scores for all baselines, where lower values indicate stronger cross-lingual overlap. The multimodal results closely mirror the text-only setting. Lower layers retain clear language separation, whereas in the middle layers DiM^3 yields substantially stronger cross-lingual overlap than LLaVA, indicating a more shared semantic space under visual inputs. This trend is supported quantitatively by the silhouette scores, where DiM^3 consistently achieves lower values across most layers compared to LLaVA and other merging baselines. The increase in higher layers further suggests that improved mid-layer alignment does not collapse representations into a single undifferentiated cluster. Overall, these findings confirm that DiM^3 primarily reshapes intermediate semantic representations, now under visual conditioning, providing a consistent explanation for the observed multilingual multimodal gains in Section 5.

7 Conclusion

In this work, we formulate multilingual multimodal composition as a training-free residual composition problem on a shared vision-language backbone, and show that multilingual and multimodal adaptations should be treated as heterogeneous residual updates rather than homogeneous parameter offsets. We propose a direction- and magnitude-aware merging strategy that selectively composes these updates while preserving the visual modules of the multimodal anchor. Experiments on multilingual text and multilingual multimodal benchmarks show that DiM^3 consistently improves multilingual performance, outperforms representative merging baselines, remains competitive with dedicated multilingual multimodal fine-tuning, and largely preserves general multimodal ability. Overall, our results highlight that effective capability composition depends not only on the magnitude of parameter changes, but also on their directional alignment, with the primary impact occurring in intermediate semantic representations.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (Grant Nos. 62272092, 62676083 and 62606078), the National Science Foundation for Young Scientists of China (Grant No. 62502081), and the Fundamental Research Funds for the Central Universities (Grant No. N2523011).

References

[1] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[2] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[3] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[4] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[5] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xgqa: Cross-lingual visual question answering. In *Findings of the association for computational linguistics: ACL 2022*, pages 2497–2511, 2022.

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

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

[8] David Orlando Romero Mogrovejo, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Villa Cueva, Jinheon Baek, Soyeong Jeong, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

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

[10] Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. Parrot: Multilingual visual instruction tuning. In *Forty-second International Conference on Machine Learning*.

[11] Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. mblip: Efficient bootstrapping of multilingual vision-llms. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 7–25, 2024.

[12] Yongkang Liu, Shi Feng, Daling Wang, and Yifei Zhang. Mulzdg: Multilingual code-switching framework for zero-shot dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 648–659, 2022.

[13] Yongkang Liu, Shi Feng, Daling Wang, Yifei Zhang, and Hinrich Schütze. Pvgru: Generating diverse and relevant dialogue responses via pseudo-variational mechanism. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3295–3310, 2023.

[14] Mingxu Tao, Chen Zhang, Quzhe Huang, Tianyao Ma, Songfang Huang, Dongyan Zhao, and Yansong Feng. Unlocking the potential of model merging for low-resource languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8705–8720, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[15] Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia, May 2024. ELRA and ICCL.

[16] 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. 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, 2022.

[17] Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, and Kun Kuang. Model tailor: mitigating catastrophic forgetting in multi-modal large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62581–62598, 2024.

[18] Shuqi Liu, Han Wu, Bowei He, Xiongwei Han, Mingxuan Yuan, and Linqi Song. Sens-merging: Sensitivity-guided parameter balancing for merging large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19243–19255, 2025.

[19] Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. Twin-merging: Dynamic integration of modular expertise in model merging. *Advances in Neural Information Processing Systems*, 37:78905–78935, 2024.

[20] Zijing Wang, Yongkang Liu, Mingyang Wang, Ercong Nie, Mengjie Zhao, Yunpu Ma, Kang Liu, Zihan Wang, Shi Feng, Daling Wang, and Hinrich Schütze. Orthogonal yet coupled: Decoupling geometric components for model merging, 2026.

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

[22] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*, 2024.

[23] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.

[24] Yingfeng Luo, Dingyang Lin, Junxin Wang, Ziqiang Xu, Kaiyan Chang, Tong Zheng, Bei Li, Anxiang Ma, Tong Xiao, Zhengtao Yu, and Jingbo Zhu. One size does not fit all: A distribution-aware sparsification for more precise model merging, 2025.

[25] Katharina Hämmerl, Jindřich Libovickỳ, and Alexander Fraser. Understanding cross-lingual alignment—a survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, 2024.

[26] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498, 2021.

[27] Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6101–6117, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

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

[30] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[31] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

[32] Tyler Chang, Zhuowen Tu, and Benjamin Bergen. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, 2022.

[33] Joanito Agili Lopo, Muhammad Ravi Shulthan Habibi, Tack Hwa Wong, Muhammad Ilham Ghozali, Fajri Koto, Genta Indra Winata, Peerat Limkonchotiwat, Alham Fikri Aji, and Samuel Cahyawijaya. Language surgery in multilingual large language models. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 438–467, 2025.

[34] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, 2024.

[35] Yuxin Chen, Zhengzhou Cai, Xiangtian Ji, Weixiang Zhao, An Zhang, Xiang Wang, and Tat-Seng Chua. Understanding multilingualism in mixture-of-experts llms: Routing mechanism, expert specialization, and layerwise steering. *arXiv preprint arXiv:2601.14050*, 2026.

[36] Chongxuan Huang, Yongshi Ye, Biao Fu, Qifeng Su, and Xiaodong Shi. From neurons to semantics: Evaluating cross-linguistic alignment capabilities of large language models via neurons alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28956–28974, 2025.

[37] Mengyu Bu and Yang Feng. Language on demand, knowledge at core: Composing llms with encoder-decoder translation models for extensible multilinguality. *arXiv preprint arXiv:2603.17512*, 2026.

[38] Mengyu Bu, Shaolei Zhang, Zhongjun He, Hua Wu, and Yang Feng. Alignx: Advancing multilingual large language models with multilingual representation alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6471–6500, 2025.

[39] Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319, 2024.

[40] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[41] Zijing Wang, Yongkang Liu, Mingyang Wang, Ercong Nie, Deyuan Chen, Zhengjie Zhao, Shi Feng, Daling Wang, Xiaocui Yang, Yifei Zhang, et al. Plam: Training-free plateau-guided model merging for better visual grounding in mllms. *arXiv preprint arXiv:2601.07645*, 2026.

[42] Xinhan Zheng, Huyu Wu, Xueting Wang, and Haiyun Jiang. Unveiling intrinsic text bias in multimodal large language models through attention key-space analysis. *arXiv preprint arXiv:2510.26721*, 2025.

[43] Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. Palo: A polyglot large multimodal model for 5b people. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1745–1754, 2025.

[44] Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. Centurio: On drivers of multilingual ability of large vision-language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2831–2881, 2025.

[45] Iñigo Pikabea, Iñaki Lacunza, Oriol Pareras Velasco, Carlos Escolano, Aitor Gonzalez-Agirre, Javier Hernando, and Marta Villegas. Breaking language barriers in visual language models via multilingual textual regularization. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 299–337, 2025.

[46] Yuchun Fan, Yilin Wang, Yongyu Mu, Lei Huang, Bei Li, Xiaocheng Feng, Tong Xiao, and Jingbo Zhu. Language-specific layer matters: Efficient multilingual enhancement for large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12473–12500, 2025.

[47] Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in llms, mllms, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8):1–41, 2026.

[48] Zijing Wang, Xingle Xu, Yongkang Liu, Yiqun Zhang, Peiqin Lin, Shi Feng, Xiaocui Yang, Daling Wang, and Hinrich Schütze. Why do more experts fail? a theoretical analysis of model merging. *arXiv preprint arXiv:2505.21226*, 2025.

[49] Zijing Wang, Yongkang Liu, Yingfeng Luo, Ming Wang, Zhen Song, Shi Feng, Xiaocui Yang, Dingyang Lin, Daling Wang, Yifei Zhang, et al. Scaling intelligence through model merging: A comprehensive survey. *Authorea Preprints*, 2025.

[50] Yifei He, Yuzheng Hu, Yong Lin, Tong Zhang, and Han Zhao. Localize-and-stitch: Efficient model merging via sparse task arithmetic. *arXiv preprint arXiv:2408.13656*, 2024.

[51] Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *Forty-second International Conference on Machine Learning*.

[52] Yiyang Du, Xiaochen Wang, Chi Chen, Jiabo Ye, Yiru Wang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Zhifang Sui, et al. 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, 2025.

[53] MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*, pages 270–287. Springer, 2024.

[54] Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. Magmax: Leveraging model merging for seamless continual learning. In *European Conference on Computer Vision*, pages 379–395. Springer, 2024.

[55] Guodong Du, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim K Goh, Ho-Kin Tang, Daojing He, et al. Parameter competition balancing for model merging. *Advances in Neural Information Processing Systems*, 37:84746–84776, 2024.

[56] Haiquan Qiu, You Wu, Dong Li, Jianmin Guo, and Quanming Yao. Superpose task-specific features for model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4200–4214, 2025.

[57] Zitao Fang, Guodong Du, Shuyang Yu, Yifei Guo, Yiwei Zhang, Yiyao Cao, Jing Li, Ho-Kin Tang, and Sim Kuan Goh. To see a world in a spark of neuron: Disentangling multi-task interference for training-free model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15731–15751, 2025.

[58] Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, Lu Hou, Sinan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. Optmerge: Unifying multimodal llm capabilities and modalities via model merging. *arXiv preprint arXiv:2505.19892*, 2025.

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[60] Shaoxiong Ji, Zihao Li, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. Emma-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*, 2024.

[61] Qwen2 technical report. 2024.

[62] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[63] Hao Yu, Tianyi Xu, Michael A Hedderich, Wassim Hamidouche, Syed Waqas Zamir, and David Ifeoluwa Adelani. Afriquellm: How data mixing and model architecture impact continued pre-training for african languages. *arXiv preprint arXiv:2601.06395*, 2026.

[64] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, 2020.

[65] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9019–9052, 2022.

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

[67] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, 2021.

[68] Atnafu Lambebo Tonja, Srija Anand, Emilio Villa-Cueva, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Muhidin A Mohamed, Debela Desalegn Yadeta, Negasi Haile Abadi, Abigail Oppong, Nnaemeka Casmir Obiefuna, et al. Afri-mcqa: Multimodal cultural question answering for african languages. *arXiv preprint arXiv:2601.05699*, 2026.

[69] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

[70] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024.

[71] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

[72] Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

[73] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025.

A Baselines and Benchmarks

All experiments were conducted on NVIDIA RTX A6000 GPUs with 48GB memory per GPU.

A.1 Merging Methods

Task Arithmetic. Task Arithmetic [21] represents a task as the parameter difference between a fine-tuned model and its base model, and composes capabilities through linear operations on these task vectors. It provides the basic formulation of training-free model editing and serves as the foundation for many subsequent merging methods.

DARE. DARE [22] sparsifies fine-tuning deltas by randomly dropping a large fraction of parameters and rescaling the remainder. This procedure is motivated by the redundancy of task vectors and is intended to reduce interference while largely preserving the effect of the original update.

TIES-Merging. TIES-Merging [23] trims small-magnitude updates, resolves sign disagreement across task vectors, and merges only the coordinates that are consistent with the elected sign. The method is designed to explicitly reduce destructive interference caused by conflicting residual directions.

Breadcrumbs. Breadcrumbs [53] constructs sparse masks over task vectors to retain informative updates while filtering out both negligible changes and large disruptive outliers.

PCB-Merging. PCB-Merging [55] formulates merging as a parameter competition balancing problem. It estimates parameter importance within each task model and competition across task models, then selectively retains and rescales updates to preserve task-relevant information while reducing harmful conflicts during composition.

STF. STF [56] approaches model merging from a representation perspective by superposing task-specific features rather than relying solely on direct parameter arithmetic. It emphasizes that compatibility across tasks can be modeled through structured interactions in feature space, providing an alternative to purely parameter-level heuristics.

NeuroMerging. NeuroMerging [57] studies merging through neuronal structure by disentangling task-specific information into complementary neuronal subspaces. It aims to mitigate multi-task interference at the neuron level and thereby better preserve complementary knowledge across source models.

A.2 Models

Emma-500. Emma-500 [60] is a massively multilingual language model obtained by continual pre-training of LLaMA-2-7B on the MaLA corpus, with coverage extending to 546 languages. The model is designed to improve multilingual performance, especially for underrepresented languages, while preserving compatibility with the original LLaMA-2-7B backbone. In our experiments, Emma-500 serves as the multilingual source model that provides the multilingual residual to be merged into the multimodal anchor.

Palo. Palo [43] is a multilingual multimodal model developed to support visual reasoning in 10 major languages covering roughly 5 billion people. It is built through multilingual multimodal instruction tuning, using a semi-automated translation pipeline to extend English multimodal supervision to multiple target languages, and is reported at several model scales including 7B. In our experiments, Palo is used as a trained multilingual multimodal reference model rather than a merging source, providing a comparison point for performance achievable through additional multilingual multimodal training.

PLAST-7B. PLAST [46] is an efficient multilingual enhancement method for large vision-language models. Starting from LLaVA-v1.5-7B, it identifies shallow language-specific layers through neuron-activation analysis and then fine-tunes only these layers using question-translation pairs, with the goal of improving multilingual understanding while updating only a small fraction of parameters.

AfriqueQwen-8B. AfriqueQwen-8B [63] is a multilingual language model obtained by continued pre-training Qwen3-8B on a 26B token mixture covering 20 African languages. It is part of the AfriqueLLM suite, which systematically studies how data composition and model architecture affect adaptation to African languages. In our experiments, AfriqueQwen-8B serves as the multilingual source model in the Qwen3-based shared-backbone setting.

Pangea-7B. Pangea-7B [9] is a multilingual multimodal model trained through multilingual multimodal instruction tuning on PangeaIns, a 6M-sample dataset spanning 39 languages. It is designed to support multilingual and cross-cultural multimodal reasoning, and is evaluated in the original paper on PangeaBench, a benchmark suite covering 14 datasets in 47 languages. In our experiments, Pangea-7B is used as an already trained multilingual multimodal checkpoint in the Qwen2-based shared-backbone setting to test whether the proposed method can further improve an existing multilingual multimodal model.

A.3 Benchmarks

XCOPA. XCOPA [64] is a multilingual benchmark for causal commonsense reasoning in 11 languages. Each example asks the model to choose the more plausible cause or effect for a given event, making it a compact test of cross-lingual commonsense understanding.

XStoryCloze. XStoryCloze [65] is a multilingual story completion benchmark spanning 11 languages. The task requires selecting the more coherent ending for a short narrative context, probing discourse-level commonsense and narrative understanding across languages.

XNLI. XNLI [66] is a standard cross-lingual natural language inference benchmark covering 15 languages. It extends MultiNLI with translated development and test sets and is widely used to evaluate multilingual sentence understanding.

xMMMU. xMMMU [9] is the multilingual extension of MMMU introduced in PangeaBench, designed to evaluate multilingual multimodal understanding and reasoning across a broad range of subjects and languages.

MaXM. MaXM [6] is a multilingual visual question answering benchmark covering seven languages. It is constructed through a scalable translation-based pipeline and serves as a test-only benchmark for multilingual VQA.

MaRVL. MaRVL [67] is a multilingual and multicultural benchmark for visually grounded reasoning. It pairs images with statements written by native speakers and asks the model to judge whether each statement is true or false, with data designed to reflect linguistic and cultural diversity.

xGQA. xGQA [5] extends the GQA benchmark to seven typologically diverse languages for cross-lingual visual question answering. It is designed to reveal multilingual multimodal misalignment by measuring how well models transfer visual reasoning ability beyond English. This makes it particularly informative for testing whether multilingual transfer remains compatible with visual grounding.

CVQA. CVQA [8] is a culturally diverse multilingual visual question answering benchmark that goes beyond translation-based multilingual evaluation. It includes culturally grounded images and questions collected across diverse countries, languages, and scripts, making it a useful testbed for both linguistic diversity and cultural coverage.

Afri-MCQA. Afri-MCQA [68] is a multilingual cultural multimodal question-answering benchmark covering about 7.5k image-grounded QA pairs across 16 African languages from 12 countries. The benchmark provides parallel English and native-language question-answer pairs in both text and speech, and is constructed by native speakers.

MMStar. MMStar [69] is a multimodal evaluation benchmark designed to test whether large vision-language models genuinely use visual information rather than exploiting shortcuts or benchmark artifacts. It emphasizes visually indispensable questions and more reliable evaluation design.

MMMU. MMMU [70] is a large-scale benchmark for multimodal understanding and reasoning at a college-exam level of difficulty. It covers a wide range of disciplines, subjects, and image types, and is intended to measure advanced perception and domain-specific reasoning.

SEED-Bench-2-Plus. SEED-Bench-2-Plus [71] is a benchmark for text-rich visual comprehension in multimodal large language models. It focuses on scenarios where substantial textual content appears inside images, requiring models to jointly process visual layout and embedded text.

B Other Ablations

We further ablate the merge scope by applying DiM^3 to different modules of the shared language model. 'Embed_Only' merges only the token embedding layer, 'LLM_Only' merges only the transformer backbone, and 'Lmhead_Only' merges only the output language modeling head. We also evaluate partial-layer variants that merge 'embed_tokens', a contiguous block of transformer layers, and 'lm_head', namely '0_7layers', '8_15layers', '16_23layers', and '24_31layers'. As shown in Table 6, merging only a single module is insufficient, and partial-layer merging is consistently weaker than applying DiM^3 to the full shared backbone. Among the partial variants, '24_31layers' is the strongest and remains relatively close to the original LLaVA on general multimodal performance, suggesting that upper layers preserve a substantial portion of task-sensitive multimodal structure. Nevertheless, the full DiM^3 merge yields the best overall results, which is consistent with the analysis in Section 6: the gain is not attributable to a single isolated module, but is associated with improved shared semantic alignment across the backbone, with the most pronounced changes occurring in intermediate layers while higher-layer task-sensitive structure is still preserved.

Table 6: Ablation on merge scope: results of applying DiM^3 to different modules and layer ranges of the shared language model.

Variant	MMStar	MMMU	SEED	Afri-MCQA	MaXM	MaRVL
DiM^3	**34.86**	**37.22**	40.45	**26.31**	**33.44**	51.89
Embed_Only	33.33	35.67	41.81	24.67	19.79	49.99
LLM_Only	32.70	28.33	35.92	19.01	16.91	**54.17**
Lmhead_Only	20.31	25.67	24.68	0.00	2.68	49.72
0_7layers	10.40	27.44	12.08	1.01	2.66	49.72
8_15layers	16.51	25.67	14.32	0.00	2.01	50.28
16_23layers	28.11	30.33	35.70	0.00	1.82	50.40
24_31layers	34.36	35.22	**42.03**	24.74	27.13	49.85

C Limitations

Our method currently relies on a reasonably compatible shared backbone between the multilingual source model and the multimodal anchor. When the source checkpoints differ substantially in architecture or tokenizer design, parameter-space composition may become less stable because the two residual updates no longer admit a clean common reference point. Extending training-free merging beyond this shared-backbone setting, especially to models with tokenizer or architectural mismatches, remains an important direction for future work.

Table 7: Languages used in our experiments and their corresponding ISO 639-1/2 codes (57 in total).

ISO 639	Language	ISO 639	Language	ISO 639	Language
Ak	Akan/Twi	**Id**	Indonesian	**Ro**	Romanian
Am	Amharic	**Ig**	Igbo	**Ru**	Russian
Ar	Arabic	**It**	Italian	**Rw**	Kinyarwanda
Bg	Bulgarian	**Ja**	Japanese	**Si**	Sinhala
Bn	Bengali	**Jv**	Javanese	**So**	Somali
Br	Breton	**Ki**	Kikuyu	**Sot**	Sesotho
De	German	**Ko**	Korean	**Su**	Sundanese
El	Greek	**Ln**	Lingala	**Sw**	Swahili
En	English	**Lug**	Luganda	**Ta**	Tamil
Es	Spanish	**Min**	Minangkabau	**Te**	Telugu
Et	Estonian	**Mn**	Mongolian	**Th**	Thai
Eu	Basque	**Mr**	Marathi	**Ti**	Tigrinya
Fil	Filipino	**Ms**	Malay	**Tn**	Setswana
Fr	French	**My**	Burmese	**Tr**	Turkish
Ga	Irish	**No**	Norwegian	**Ur**	Urdu
Ha	Hausa	**Ny**	Chichewa	**Vi**	Vietnamese
He	Hebrew	**Om**	Oromo	**Yo**	Yoruba
Hi	Hindi	**Pt**	Portuguese	**Zh**	Chinese
Ht	Haitian	**Qu**	Quechua	**Zu**	Zulu

Table 8: Results on XCOPA and XStoryCloze across multilingual settings. $\mathrm{Avg}_{\mathrm{mul}}$ denotes the average performance over all languages except English.

XCOPA

Methods	Et	Ht	Id	It	Qu	Sw	Ta	Th	Tr	Vi	Zh	$\mathbf{Avg}_{\mathbf{mul}}$
Task Arithmetic	48.20	50.60	65.00	68.00	49.80	49.80	52.00	55.40	55.60	63.80	66.40	56.78
DARE	48.60	50.60	63.60	67.20	50.80	50.60	52.80	**57.40**	56.00	60.20	64.20	56.55
TIES-Merging	48.00	49.80	67.00	68.40	49.40	50.80	52.20	56.40	56.20	66.20	67.40	57.44
Breadcrumbs	48.40	50.60	65.80	66.20	51.00	51.00	51.40	57.00	55.80	65.60	64.00	56.98
PCB-Merging	48.00	52.80	60.40	60.40	51.20	56.60	55.80	56.20	55.00	56.60	58.00	55.55
STF	53.80	55.00	59.80	53.40	49.80	52.80	56.60	53.00	57.40	59.80	58.00	55.40
NeuroMerging	52.40	55.20	58.20	55.60	47.40	51.40	56.40	54.60	58.20	64.00	63.60	56.09
DiM[3]	**60.60**	**55.80**	**73.40**	**70.80**	**55.00**	**61.40**	**60.80**	56.60	**62.00**	**70.00**	**70.80**	**63.38**

XStoryCloze

Methods	Ar	Es	Eu	Hi	Id	My	Ru	Sw	Te	Zh	En	$\mathbf{Avg}_{\mathbf{mul}}$
Task Arithmetic	53.67	70.22	51.82	56.52	64.06	48.91	67.44	51.75	55.13	64.53	80.41	58.41
DARE	51.62	67.70	52.22	53.34	61.42	48.91	64.92	50.50	54.20	63.27	78.95	56.81
TIES-Merging	54.07	**71.08**	51.56	56.32	64.86	49.11	67.97	51.03	55.06	64.59	**80.94**	58.57
Breadcrumbs	52.42	69.82	51.03	55.26	63.20	48.78	65.72	50.56	55.66	62.41	80.87	57.49
PCB-Merging	53.21	62.41	52.08	53.34	53.34	43.94	60.42	49.50	52.08	61.48	67.31	54.18
STF	60.23	55.06	53.54	58.90	56.12	50.03	57.31	53.54	58.44	59.36	61.15	56.25
NeuroMerging	58.31	57.78	53.21	58.37	58.24	48.25	60.89	52.68	55.86	61.81	65.72	56.54
DiM[3]	**64.06**	69.82	**63.80**	**64.92**	**68.96**	**55.20**	**70.75**	**62.14**	**62.67**	**65.92**	79.02	**64.82**

Table 9: Results on XNLI across 15 languages.

Methods	Ar	Bg	De	El	Es	Fr	Hi	Ru	Sw	Th	Tr	Ur	Vi	Zh	En	$\mathbf{Avg}_{\mathbf{mul}}$
Task Arithmetic	33.53	42.01	46.10	38.76	43.78	47.71	39.56	40.32	36.31	36.02	41.00	34.02	43.41	**38.51**	56.91	40.07
DARE	33.69	43.29	42.49	37.91	45.22	45.42	38.55	40.76	34.82	34.46	38.07	33.57	41.73	37.91	56.63	39.14
TIES-Merging	33.49	42.49	47.63	39.72	**45.82**	48.88	41.33	41.24	36.47	35.82	41.29	34.50	41.16	36.79	58.51	40.47
Breadcrumbs	33.78	43.61	**47.75**	39.28	45.22	49.08	40.28	43.09	36.22	35.74	40.80	33.53	39.20	38.27	**58.67**	40.42
PCB-Merging	33.13	41.73	44.02	33.41	41.12	45.10	33.33	40.44	37.35	33.49	34.38	33.45	35.90	33.17	40.24	37.14
STF	33.49	**46.10**	42.77	40.24	40.84	**51.41**	38.07	36.71	33.33	33.65	**43.57**	38.31	37.51	31.93	38.55	39.14
NeuroMerging	33.61	41.29	44.10	41.04	43.57	47.71	42.81	39.60	34.02	35.50	40.84	43.29	38.27	34.46	41.04	40.01
DiM[3]	**35.10**	43.25	47.71	**43.17**	44.18	45.82	**45.94**	**44.02**	**43.57**	**39.24**	42.37	**43.90**	**44.62**	38.11	50.68	**42.93**

Table 10: Results on XCOPA and XStoryCloze across multilingual settings.

XCOPA												
Models	**Et**	**Ht**	**Id**	**It**	**Qu**	**Sw**	**Ta**	**Th**	**Tr**	**Vi**	**Zh**	**$\mathbf{Avg_{mul}}$**
Llama_2_7B	49.00	50.40	62.00	62.80	50.80	52.20	53.60	56.40	54.80	63.40	63.60	56.27
EMMA-500	**61.40**	**57.00**	**73.60**	65.40	50.00	**64.80**	60.60	56.20	61.00	**70.20**	67.00	62.47
LLaVA-v1.5-7B	49.40	51.60	50.80	50.00	51.20	52.00	53.80	54.60	54.20	51.00	53.20	51.98
DiM^3	60.60	55.80	73.40	**70.80**	**55.00**	61.40	**60.80**	**56.60**	**62.00**	70.00	**70.80**	**63.38**
XStoryCloze												
Models	**Ar**	**Es**	**Eu**	**Hi**	**Id**	**My**	**Ru**	**Sw**	**Te**	**Zh**	**En**	**$\mathbf{Avg_{mul}}$**
Llama_2_7B	50.17	67.70	50.56	53.54	59.36	48.51	63.34	50.43	54.33	59.50	77.10	55.74
EMMA-500	**65.25**	**69.89**	**64.39**	63.80	**69.03**	**58.44**	67.84	**63.34**	**64.00**	61.75	75.65	64.77
LLaVA-v1.5-7B	47.98	51.89	50.23	47.58	48.64	47.32	50.17	48.58	53.67	48.78	66.71	49.48
DiM^3	64.06	69.82	63.80	**64.92**	68.96	55.20	**70.75**	62.14	62.67	**65.92**	**79.02**	**64.82**

Table 11: Results on XNLI across multilingual settings.

Models	**Ar**	**Bg**	**De**	**El**	**Es**	**Fr**	**Hi**	**Ru**	**Sw**	**Th**	**Tr**	**Ur**	**Vi**	**Zh**	**En**	**$\mathbf{Avg_{mul}}$**
Llama-2-7B	34.22	42.13	**48.11**	36.59	42.45	**50.28**	38.71	39.88	35.02	36.27	36.87	33.90	38.07	34.46	**55.66**	39.07
EMMA-500	34.78	**46.27**	47.07	**45.86**	**46.75**	34.78	**47.59**	**46.55**	**46.18**	**41.97**	**44.86**	**45.98**	**47.03**	35.22	53.78	**43.64**
LLaVA-v1.5-7B	33.57	33.45	38.71	36.22	37.55	37.87	34.98	35.30	33.65	34.06	35.82	34.58	36.06	35.86	52.81	35.55
DiM^3	**35.10**	43.25	47.71	43.17	44.18	45.82	45.94	44.02	43.57	39.24	42.37	43.90	44.62	**38.11**	50.68	42.93

Table 12: Results on xMMMU and MaXM across multilingual settings.

	xMMMU								**MaXM**							
Models	Ar	Fr	Hi	Id	Ja	Pt	En	$\mathrm{Avg_{mul}}$	Fr	Hi	He	Ro	Th	Zh	En	$\mathrm{Avg_{mul}}$
LLaVA-v1.5-7B	28.90	34.90	27.80	32.00	31.60	33.70	35.70	31.48	32.20	16.92	12.50	15.14	16.04	27.80	49.81	20.10
Task Arithmetic	30.20	36.60	27.10	33.00	32.70	36.40	**35.90**	32.67	33.33	21.15	12.50	16.90	18.66	31.05	49.42	22.27
DARE	28.90	34.90	27.50	31.60	33.10	32.30	35.60	31.38	34.47	17.69	12.50	11.27	16.04	25.27	49.81	19.54
TIES-Merging	28.50	32.20	26.80	33.00	**35.70**	**37.40**	34.70	32.27	33.71	26.92	21.79	16.90	17.54	28.52	48.25	24.23
Breadcrumbs	27.20	32.60	25.80	30.30	30.90	32.70	31.90	29.92	31.44	24.62	19.64	19.01	10.82	25.99	48.64	21.92
PCB-Merging	21.50	31.50	30.20	28.30	27.90	28.30	24.70	27.95	0.38	0.00	0.00	0.00	0.00	0.00	12.45	0.06
STF	31.50	34.20	30.90	34.00	29.00	32.70	31.00	32.05	10.61	10.00	11.79	7.75	11.94	14.44	18.29	11.09
NeuroMerging	28.50	**37.20**	30.60	**36.70**	33.10	34.70	33.80	33.47	25.00	14.23	26.79	9.15	22.76	25.27	30.74	20.53
Palo-7B	30.90	33.20	28.90	34.00	27.10	33.70	31.70	31.30	33.71	16.54	12.14	11.27	14.55	10.47	**51.36**	16.45
Palo-13B	29.50	33.90	**32.00**	32.30	26.80	33.30	33.80	31.30	**40.91**	**39.62**	12.86	33.45	14.55	30.69	**51.36**	28.68
PLAST-7B	-	-	-	-	-	-	-	-	-	10.70	10.20	**36.70**	20.8	-	-	-
DiM^3-7B	**34.60**	36.60	29.20	32.30	**35.70**	34.70	32.60	**33.85**	35.98	35.77	**40.71**	14.79	**36.19**	**37.18**	50.58	**33.44**

Table 13: Results on MaRVL and xGQA across multilingual settings.

	MaRVL							**xGQA**								
Models	Id	Sw	Ta	Tr	Zh	En	$\mathrm{Avg_{mul}}$	Bn	De	Id	Ko	Pt	Ru	Zh	En	$\mathrm{Avg_{mul}}$
LLaVA-v1.5-7B	50.09	49.37	49.36	50.00	50.10	51.22	49.78	15.48	28.50	33.17	38.00	27.53	32.97	38.67	**61.89**	30.62
Task Arithmetic	50.00	49.46	49.60	50.08	49.90	51.12	49.81	16.88	38.03	37.33	41.93	34.42	42.01	45.20	61.50	36.54
DARE	50.00	49.37	49.60	49.83	50.00	50.92	49.76	14.78	32.96	34.23	34.07	30.85	36.47	37.19	60.16	31.51
TIES-Merging	50.00	49.37	49.28	50.00	49.90	50.99	49.71	18.54	38.21	38.24	40.59	35.71	41.85	43.32	58.37	36.64
Breadcrumbs	50.00	49.37	49.52	50.00	49.31	51.13	49.64	15.66	41.30	38.26	37.01	38.39	40.80	40.78	53.27	36.03
STF	51.33	**50.27**	**51.05**	50.59	51.09	52.49	50.87	0.91	16.96	11.11	9.42	10.40	11.24	16.06	20.56	10.87
NeuroMerging	49.91	49.28	49.84	50.17	50.00	50.93	49.84	2.44	28.56	23.34	17.60	24.56	26.62	23.74	36.19	20.98
Palo-7B	51.95	49.55	49.44	52.29	50.40	**55.39**	50.73	**41.92**	38.97	36.67	41.28	31.86	21.84	45.94	60.25	36.93
Palo-13B	50.00	49.37	49.60	49.75	49.80	51.10	49.70	31.91	21.22	28.28	29.39	26.75	32.85	43.78	59.08	30.60
DiM^3-7B	**52.13**	48.92	50.56	**54.58**	**53.26**	54.14	**51.89**	38.96	**51.98**	**49.21**	**48.00**	**49.80**	**47.93**	**50.42**	59.27	**48.04**

Table 14: Results on Afri-MCQA across multilingual settings.

Method	Am	Ha	Ig	Lug	Om	Rw	Ki	So	Ti	Ak	Yo	Tn	Ny	Zu	Sot	Ln	Avg
Llama-2-7B																	
LLaVA-v1.5-7B	0.216	**0.256**	0.256	0.265	0.225	0.295	0.237	0.270	0.199	0.276	**0.235**	0.255	0.200	**0.260**	0.254	0.237	0.246
Task Arithmetic	0.216	**0.256**	0.251	0.265	0.225	0.290	0.237	0.270	0.199	0.271	**0.235**	0.245	0.200	0.250	0.254	0.237	0.244
DARE	0.216	0.251	0.246	0.260	0.225	0.295	0.237	0.270	0.199	0.276	**0.235**	**0.260**	0.195	0.240	0.254	0.232	0.243
TIES-Merging	0.216	0.251	0.241	0.265	0.215	0.275	0.227	0.260	0.199	0.271	**0.235**	0.245	0.195	0.240	0.246	0.232	0.238
Breadcrumbs	0.060	0.246	0.236	0.250	0.210	0.265	0.232	0.260	0.046	0.229	0.215	0.240	0.180	0.229	0.223	0.237	0.210
PCB-Merging	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
STF	**0.251**	0.246	0.221	0.280	0.235	0.300	0.242	0.235	0.199	0.271	0.170	**0.260**	0.205	0.208	0.246	0.248	0.239
NeuroMerging	0.166	0.246	0.266	0.295	0.215	0.300	**0.253**	0.270	0.167	0.257	0.210	0.235	0.200	0.208	0.246	0.253	0.237
Palo-7B	0.216	0.251	0.261	0.255	0.225	0.290	0.247	0.270	0.199	0.280	0.225	0.250	0.205	**0.260**	0.246	0.232	0.245
Palo-13B	0.216	**0.256**	0.246	0.260	0.220	0.270	0.242	0.260	0.199	0.262	0.230	0.250	0.195	0.240	0.254	0.232	0.240
DiM^3-7B	0.241	**0.256**	**0.281**	**0.320**	**0.245**	**0.335**	0.232	**0.320**	**0.204**	**0.304**	**0.235**	0.250	**0.215**	0.240	**0.269**	**0.263**	**0.263**
Qwen2-7B-Instruct																	
Pangea-7B	0.357	0.271	0.352	0.320	0.130	0.370	0.253	0.310	0.287	0.332	0.380	0.185	0.265	0.198	0.131	0.389	0.283
Task Arithmetic	0.332	0.312	0.357	**0.425**	0.240	0.475	0.345	0.340	0.306	0.360	0.390	0.265	0.345	**0.333**	0.285	0.424	0.346
DARE	**0.367**	0.292	0.322	0.350	0.225	0.430	0.335	0.320	0.301	0.322	0.320	0.265	0.340	0.271	0.231	0.404	0.318
TIES-Merging	0.337	0.327	0.357	0.415	0.275	**0.480**	0.366	0.335	0.301	0.351	0.395	0.315	0.385	**0.333**	0.292	**0.455**	0.357
Breadcrumbs	0.327	0.276	0.332	0.395	0.295	0.455	0.351	0.345	0.324	0.346	**0.440**	0.335	0.395	0.302	0.315	0.414	0.353
PCB-Merging	0.337	0.271	0.327	0.375	**0.325**	0.455	0.371	0.345	0.333	0.346	0.390	0.315	0.385	0.323	0.277	0.429	0.350
STF	**0.367**	0.312	0.387	0.380	0.260	0.400	0.304	0.325	0.333	0.318	0.350	0.265	0.275	0.292	0.208	0.394	0.323
NeuroMerging	0.307	0.281	**0.407**	0.415	0.305	0.455	0.340	**0.380**	0.343	0.346	0.380	**0.355**	**0.400**	**0.333**	0.308	0.424	0.361
DiM^3	0.317	**0.342**	**0.407**	**0.425**	0.320	**0.480**	**0.376**	0.345	**0.347**	**0.388**	0.390	0.315	0.390	**0.333**	**0.331**	0.419	**0.370**
Qwen3-8B																	
Qwen3-VL-8B-Instruct	0.377	0.342	0.347	0.415	0.335	0.460	0.392	0.405	0.357	**0.430**	0.420	0.350	0.380	0.302	0.323	0.490	0.383
Task Arithmetic	0.402	0.337	0.352	0.420	0.340	0.495	0.381	0.435	0.357	0.383	0.415	0.335	0.385	0.281	0.308	**0.505**	0.383
DARE	0.231	0.181	0.236	0.220	0.245	0.190	0.289	0.250	0.273	0.224	0.265	0.225	0.280	0.250	0.300	0.248	0.244
TIES-Merging	0.382	0.342	**0.372**	0.410	0.390	0.490	**0.407**	0.425	0.366	0.383	0.415	0.335	0.395	0.292	0.269	**0.505**	0.386
Breadcrumbs	0.010	0.015	0.025	0.000	0.000	0.015	0.000	0.000	0.028	0.000	0.000	0.005	0.025	0.000	0.000	0.020	0.009
PCB-Merging	0.402	0.342	0.307	0.395	0.370	0.475	0.366	0.430	0.426	0.374	0.425	0.320	0.370	0.313	**0.346**	0.404	0.379
STF	0.427	0.442	0.367	0.400	0.415	**0.645**	**0.407**	0.530	0.458	0.336	**0.475**	0.340	0.440	0.354	0.262	0.404	0.419
NeuroMerging	0.211	0.292	0.226	0.240	0.275	0.380	0.191	0.385	0.273	0.215	0.335	0.235	0.235	0.281	0.200	0.187	0.260
DiM^3	**0.493**	**0.498**	0.362	**0.450**	**0.445**	**0.645**	**0.407**	**0.555**	**0.505**	0.388	0.445	**0.420**	**0.480**	**0.406**	0.277	0.460	**0.452**

Table 15: Multilingual performance comparison of different merging methods on CVQA.

Language	LLaVA-v1.5-7B	Task Arithmetic	DARE	TIES-Merging	Breadcrumbs	PCB-Merging	STF	NeuroMerging	Palo-7B	Palo-13B	DiM^3
Am	26.50	29.06	29.91	27.78	25.21	0.43	**35.90**	30.34	19.23	26.07	33.76
Bn	29.37	31.47	31.47	32.87	29.02	1.05	23.43	25.52	**37.76**	36.01	32.52
Br	29.63	31.36	31.11	29.38	27.16	20.74	23.95	**33.58**	29.14	33.33	30.86
Bg	39.35	39.08	40.43	37.74	34.23	11.05	34.77	**42.05**	36.66	40.97	39.89
Zh	48.57	**48.95**	45.89	47.80	41.87	0.19	44.17	47.23	45.51	46.65	47.80
Ar	32.51	35.96	34.98	33.00	30.05	2.46	32.51	33.99	29.06	27.59	**40.39**
Fil	46.31	46.80	45.32	45.32	37.44	5.91	37.44	**47.29**	38.92	46.80	**47.29**
Hi	39.80	37.31	38.81	39.80	31.34	1.49	45.27	**51.74**	42.79	45.27	51.24
Ig	32.50	**35.50**	34.00	33.00	28.00	8.00	12.00	28.00	29.50	33.00	**35.50**
Id	42.23	46.12	42.23	46.60	39.56	5.83	36.17	46.60	41.26	44.90	**49.03**
Ga	43.87	47.24	44.48	**48.16**	38.04	7.06	34.97	45.09	43.25	37.12	47.24
Ja	34.98	36.45	32.02	33.00	28.08	0.00	31.03	31.53	30.54	**37.93**	32.51
Jv	36.36	39.06	32.66	36.36	29.97	11.78	29.29	39.39	32.32	38.72	**42.42**
Rw	34.47	35.74	32.34	34.04	29.79	20.00	32.34	33.62	28.94	28.94	**43.40**
Ko	**51.03**	**51.03**	50.69	**51.03**	46.55	0.00	47.24	43.79	44.48	46.55	47.93
Ms	47.94	51.11	47.30	**51.75**	40.32	9.52	40.95	44.76	42.86	46.35	48.57
Mr	36.14	36.14	35.64	34.16	26.24	0.99	42.08	36.63	39.11	39.11	**45.05**
Min	36.65	37.05	33.07	34.66	33.86	9.56	29.48	41.04	31.87	39.04	**42.63**
Mn	29.49	29.17	28.53	32.37	27.24	19.55	30.13	30.13	28.53	26.92	**36.22**
No	52.51	**53.51**	51.17	50.17	42.14	7.36	43.81	51.17	48.83	53.18	50.50
Om	**37.38**	35.05	34.58	32.24	29.44	18.22	17.29	33.18	32.71	31.78	35.98
Pt	57.39	**58.80**	56.69	55.28	52.46	17.61	47.54	57.39	56.69	56.69	53.52
Ro	50.99	53.64	51.32	**55.96**	46.69	3.31	44.70	50.99	46.03	55.63	53.64
Ru	57.50	57.50	52.50	55.50	46.50	8.00	56.50	60.00	47.00	52.50	**61.00**
Si	24.44	25.33	27.11	28.00	26.22	2.22	9.33	11.56	27.56	24.89	**39.11**
Es	54.57	**54.71**	52.43	52.19	47.72	18.90	44.22	52.19	49.66	51.51	51.60
Su	38.00	38.50	34.50	38.00	33.00	8.50	27.00	41.00	32.00	40.50	**44.50**
Sw	36.26	39.56	37.36	38.10	33.33	18.68	35.90	47.25	36.26	40.66	**49.45**
Ta	31.78	30.84	27.57	30.84	28.50	0.93	34.11	28.50	27.57	32.71	**40.65**
Te	34.50	32.50	30.00	25.50	25.00	0.00	26.50	16.00	28.50	26.50	**36.50**
Ur	25.92	30.05	25.00	28.67	26.38	2.52	42.20	39.22	42.43	39.45	**46.33**
Avg_{mul}	39.32	40.47	38.42	39.33	34.24	7.80	34.59	39.38	37.00	39.59	**43.78**

Table 16: Results on three general multimodal benchmarks: MMStar, MMMU, and SEED-Bench-2-Plus. This table reports the exact scores corresponding to the comparison summarized in Figure 2.

Models	MMStar	MMMU	SEED-Bench-2-Plus
LLaVA-v1.5-7B	33.83	36.33	40.97
Task Arithmetic	33.53	36.89	42.07
DARE	**36.11**	35.78	41.90
TIES-Merging	34.27	36.33	**43.13**
Breadcrumbs	32.01	33.33	40.32
STF	31.51	30.67	34.43
NeuroMerging	34.58	36.89	40.84
Palo-7B	32.95	33.11	38.08
DiM^3-7B	34.86	**37.22**	40.45

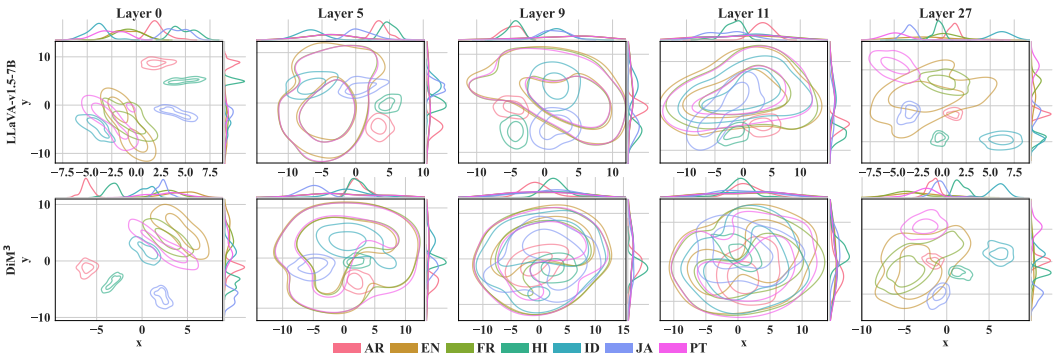

Figure 5: t-SNE visualizations of average-pooled hidden states for the question spans in multilingual multimodal inputs from xMMMU across selected layers. We focus on question tokens to avoid confounding effects from English instruction templates and image tokens.

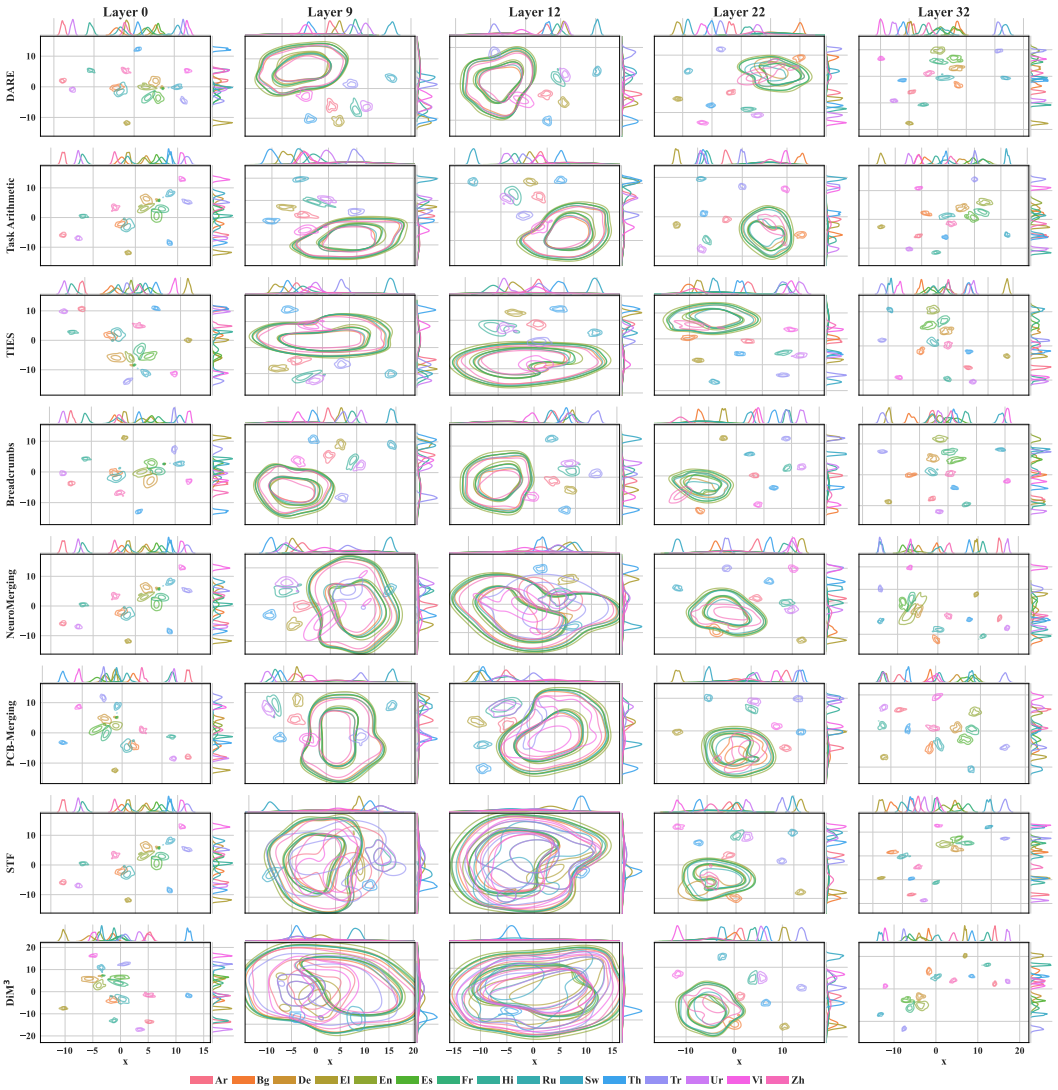

Figure 6: Additional t-SNE visualizations of average-pooled hidden states on multilingual text inputs from XNLI across selected layers. The figure compares a broader set of merging baselines to illustrate differences in cross-lingual hidden-state geometry.

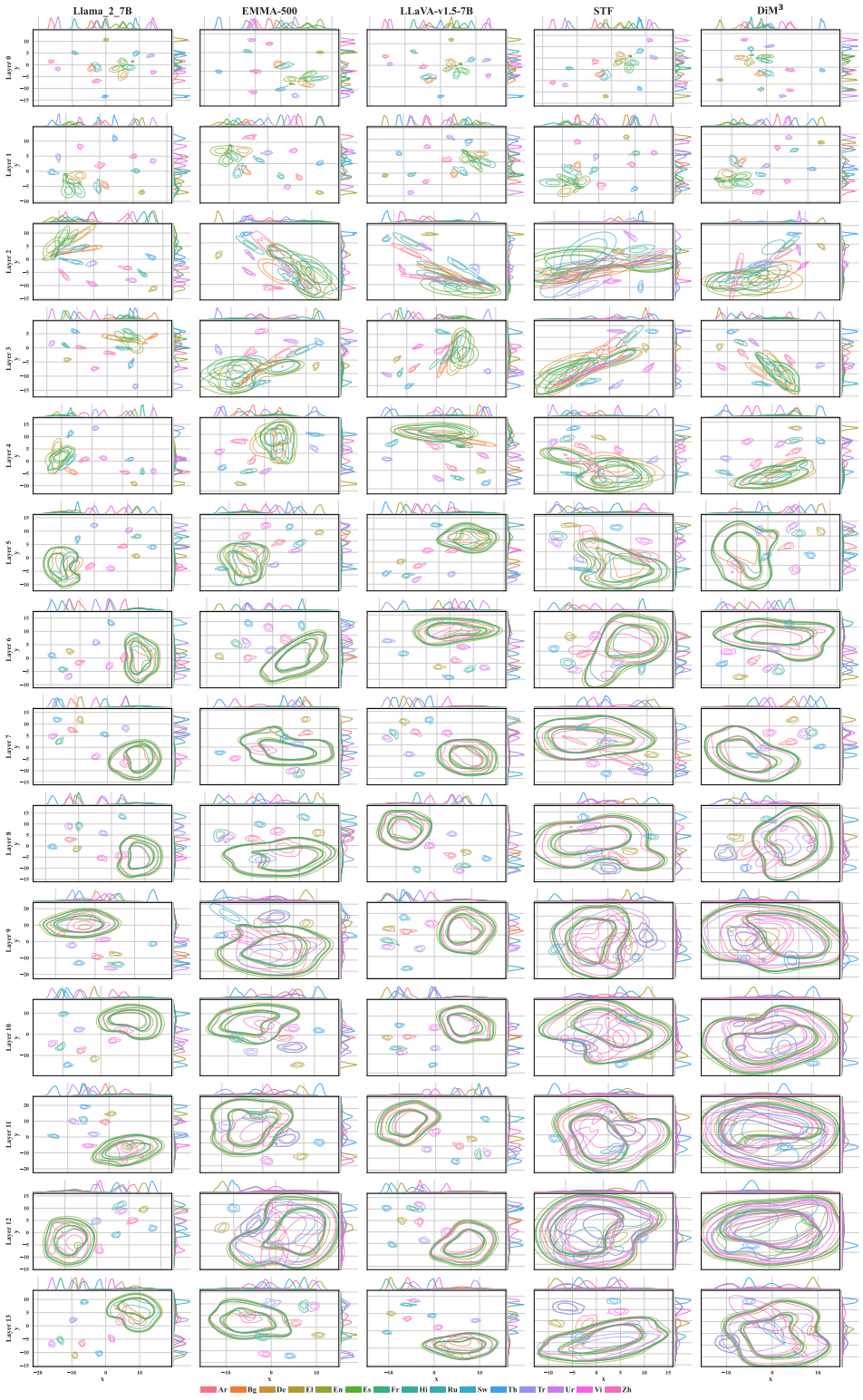

Figure 7: Full layer-wise t-SNE visualizations of average-pooled hidden states on multilingual text inputs from XNLI for layers 0–13.

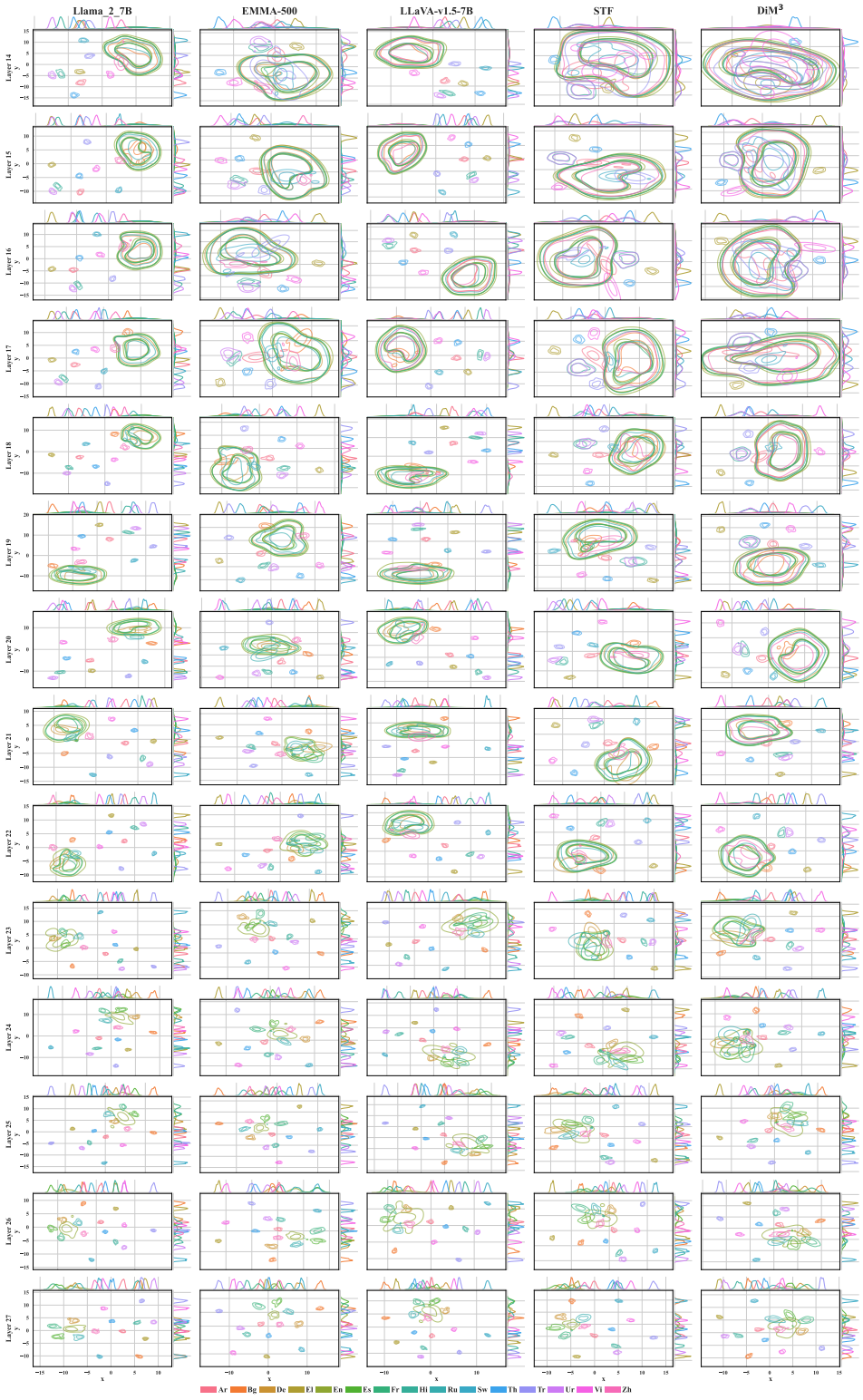

Figure 8: Full layer-wise t-SNE visualizations of average-pooled hidden states on multilingual text inputs from XNLI for layers 14–27.

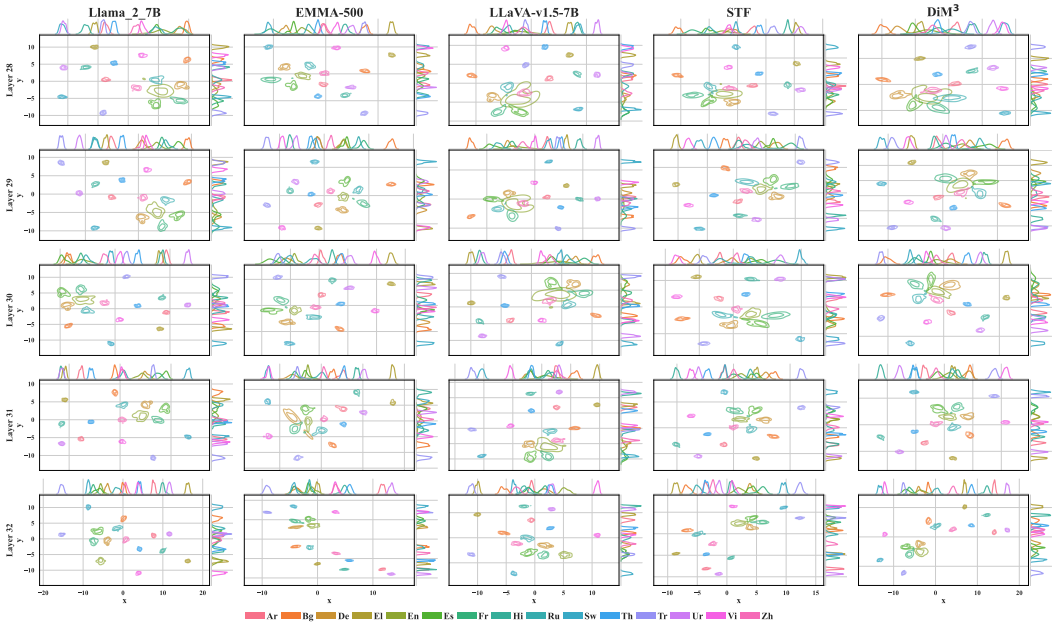

Figure 9: Full layer-wise t-SNE visualizations of average-pooled hidden states on multilingual text inputs from XNLI for layers 28–32.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 3 formalizes the training-free shared-backbone merging setting, Section 5 shows gains on multilingual text-only and vision-language benchmarks together with retention of general multimodal ability, and the additional shared-backbone experiments further verify applicability to already trained multilingual multimodal models. The representation analysis is also consistent with the interpretability claims made in the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the "Limitations" section (Appendix Section C), we discuss that our method assumes a compatible shared backbone between the multilingual and multimodal source models. We further note that tokenizer or architectural mismatches may reduce the stability of parameter-space composition; extending the method beyond the shared-backbone setting is therefore left for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The proposition in Section 3.2 states the necessary assumptions and notation and is followed by a complete proof deriving the residual-norm decomposition from the column-wise parameter decomposition.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 describes the experimental settings, including the source-model combinations, benchmarks, baselines, and evaluation procedure. We also provide an anonymous code repository to support reproduction of the main empirical findings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide an anonymous code repository containing the implementation, scripts, and usage instructions for reproducing the experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide an anonymous code repository containing the information.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the GPU type used for all experiments in the appendix. Because the method is training-free, the main cost comes from model loading, merging, and evaluation; however, we do not provide detailed wall-clock runtimes for every experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to all the ethics guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: We propose a training-free method for constructing multilingual multimodal models. This work does not have direct ties to broader societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Existing datasets and code are properly cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Paper does not exist new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This paper does not involve LLMs in any meaningful capacity.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.