

DiM³ : Bridging Multilingual and Multimodal Models via Direction- and Magnitude-Aware Merging

Zijing Wang¹, Mingyang Wang^{2,3}, Ercong Nie⁴, Yongkang Liu¹, Shi Feng¹,
Mengjie Zhao¹, Daling Wang¹, Xiaocui Yang¹, and Hinrich Schütze^{2,3}

1 Northeastern University, China; 2 CIS, LMU Munich, Germany;
3 Munich Center for Machine Learning, Germany; 4 Shanghai Jiao Tong University, China
Contact: wzj1718@gmail.com

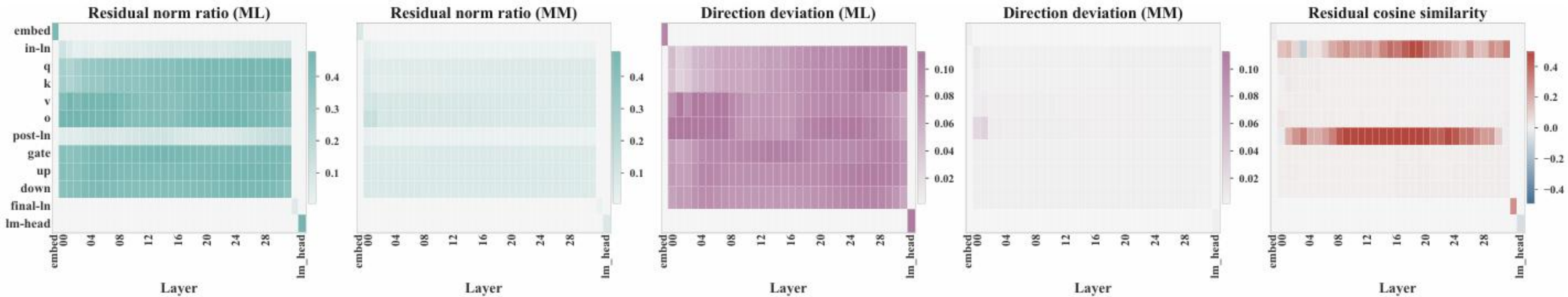


Motivation: Toward Multilingual Multimodal Intelligence

- ❑ **The capability gap:** MLLMs work well in high-resource languages, but performance can drop sharply in low-resource languages.
- ❑ **The conventional route:** Multilingual SFT or continued pre-training requires large datasets, compute, and repeated adaptation.
- ❑ **The opportunity:** Model merging can combine capabilities from compatible checkpoints without end-to-end retraining.

Can we inject multilingual capability into an existing MLLM without additional training—while preserving its visual understanding?

Observation: Residual Heterogeneity in Parameter Space



- ❑ **Magnitude and Direction Disparity:** Multilingual updates generally exhibit larger relative residual norms and stronger base-relative directional deviations than multimodal updates.
- ❑ **Non-uniform Residual Alignment:** Cross-residual cosine similarities vary across layers and modules, revealing locally heterogeneous compatibility between multilingual and multimodal updates.
- ❑ **Key Insight:** Uniform residual aggregation overlooks these geometric differences, motivating adaptive merging based on both update **magnitude and direction**.

Methodology

Given a base model N , multilingual model M , and multimodal model V sharing the same language backbone, we define:

$$\Delta_{\text{ml}} = \theta_M - \theta_N, \quad \Delta_{\text{mm}} = \theta_V^{\text{lm}} - \theta_N.$$

Merge only the shared language backbone, keeping the vision encoder and projector fixed.

1. Measure Magnitude and Direction Deviations

$$\begin{aligned} \delta_{k,j}^{\text{mag}} &= |||W_{k,:j}|_2 - |W_{N,:j}|_2|, \\ \delta_{k,j}^{\text{dir}} &= 1 - \cos(W_{k,:j}, W_{N,:j}). \end{aligned}$$

2. Estimate and Calibrate Rank-Based Saliency

$$\begin{aligned} r_{k,j}^* &= \frac{\text{rank}(\delta_{k,j}^*)}{d_{\text{in}}}, & s_{k,j}^* &= \frac{e^{r_{k,j}^*}}{\sum_{k'} e^{r_{k',j}^*}}. \\ \tilde{s}_{k,j}^* &= \begin{cases} 1, & r_{k,j}^* > \rho \bar{r}_k^*, \\ s_{k,j}^*, & \text{otherwise.} \end{cases} \end{aligned}$$

Rank normalization reduces scale sensitivity; calibration preserves highly salient updates.

3. Fuse Saliency and Merge Residuals

$$\begin{aligned} \omega_{k,j} &= \frac{1}{2} (\tilde{s}_{k,j}^{\text{mag}} + \tilde{s}_{k,j}^{\text{dir}}), \\ \tilde{W}_{:,j} &= W_{N,:j} + \sum_{k \in \{\text{ml}, \text{mm}\}} \omega_{k,j} \Delta_{k,:j} \end{aligned}$$

Notation: $k \in \{\text{ml}, \text{mm}\}$ denotes the source, j the weight column, and $*$ $\in \{\text{mag}, \text{dir}\}$.

Experiments

□ Multilingual Text Transfer

Models	XCOPA	XStoryCloze	XNLI
<i>language models</i>			
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
<i>merging methods</i>			
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

- DiM³ leads the merging baselines on all three text benchmarks.
- Large gains on XCOPA, XStoryCloze, and XNLI indicate successful multilingual transfer.

Experiments

□ Multilingual Multimodal Transfer

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

- DiM³ achieves **the best average multilingual performance** across all six benchmarks, demonstrating effective multilingual capability transfer in multimodal settings.

Experiments

□ Multilingual Multimodal Transfer

Models	xMMMU								MaXM							
	Ar	Fr	Hi	Id	Ja	Pt	En	Avg _{mul}	Fr	Hi	He	Ro	Th	Zh	En	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

Models	MaRVL							xGQA								
	Id	Sw	Ta	Tr	Zh	En	Avg _{mul}	Bn	De	Id	Ko	Pt	Ru	Zh	En	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

- Improvements are **not confined to a small subset of languages**; the appendix shows gains across languages, with particularly pronounced improvements in lower-resource settings.

Experiments

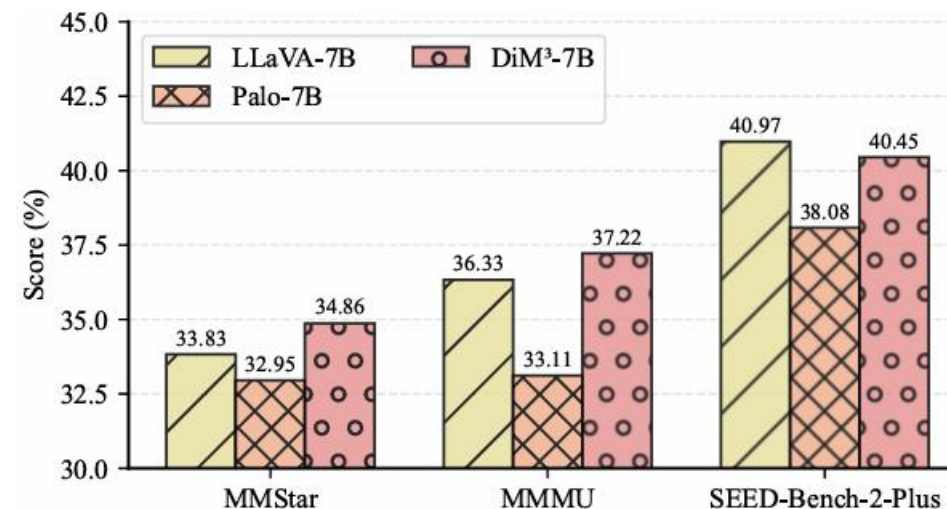
□ Multilingual Multimodal Transfer

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

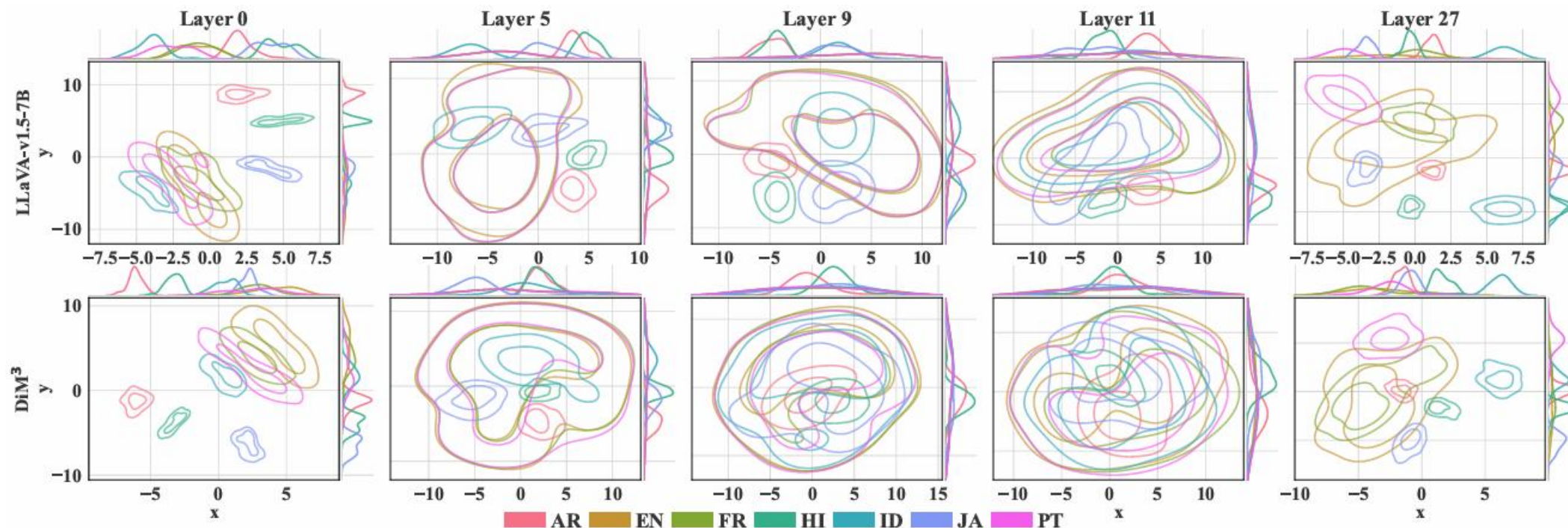
- DiM³ delivers consistent gains **across different backbone families** and remains competitive with dedicated multilingual multimodal fine-tuning baselines, demonstrating its flexibility beyond a single model architecture.

□ General multimodal ability

- DiM³ remains competitive with the original multimodal anchor and exceeds the Palo-7B fine-tuned baseline on all three reported benchmarks.



Analysis



- Enhanced mid-layer alignment: DiM³ improves cross-lingual overlap under multimodal inputs.
- Preserved language structure: Lower- and upper-layer representations remain language-sensitive, avoiding representation collapse.