NEURON-ALIGNED MODEL MERGING FOR CROSS-MODAL REASONING TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing neuron-level model merging methods transfer reasoning capabilities from LLMs to VLMs by selectively merging reasoning-relevant neurons at the same neuron indices. However, reasoning training and multimodal adaptation can reshape neuron responses, breaking the *same-index correspondence assumption*: same-index neurons may become functionally mismatched, while cross-index correspondences may be overlooked. To address this limitation, we propose Targeted Alignment and Neuron Grafting via Optimal Transport (TANGO), a training-free framework for cross-modal reasoning transfer through neuron-aligned model merging. TANGO uses activations on shared inputs and optimal transport to establish functional correspondences between source and target neuron spaces, rather than relying on fixed index-wise matches. The resulting correspondences guide both the selection of reasoning-relevant transfer locations and the alignment of source parameters for selective merging. Across four VLMs from the Qwen and InternVL families, TANGO consistently improves multimodal mathematical reasoning and general multimodal understanding, achieving the best or tied-best result in 27 of 28 mathematical reasoning settings and ranking first in 18 of 20 general multimodal and text-only reasoning comparisons.

1 INTRODUCTION

Recent advances in large language models (LLMs) have produced strong mathematical and general reasoning capabilities (Guo et al., 2025; Yang et al., 2024), while vision-language models (VLMs) still face challenges in complex multimodal reasoning (Ghosh et al., 2025; Hou et al., 2025; Tian et al., 2026). This has motivated transferring reasoning capabilities already learned by LLMs to VLMs (Luo et al., 2025; Xu et al., 2026). Model merging provides a training-free approach to such transfer by directly integrating model parameters (Yadav et al., 2023; Yu et al., 2024; Chen et al., 2025), with recent methods progressively refining the transfer granularity from parameters and subspaces to individual neurons (Hu et al., 2025; Huang et al., 2026; Cui et al., 2026b).

At the neuron level, existing methods typically select reasoning-relevant neurons that share the same indices across the source LLM and target VLM (Cui et al., 2026b), implicitly relying on a *same-index correspondence assumption*. However, this assumption need not hold after reasoning training and multimodal adaptation, which can reshape neuron responses and decouple neuron indices from functional correspondence. As illustrated in Figure 1, same-index matching in (a) can pair a target VLM neuron with a functionally mismatched source LLM neuron while overlooking a more compatible neuron at a different index, which TANGO recovers in (b). Comparing the activation responses in (a) and (b), the different-index source neuron identified by TANGO follows the target neuron substantially more closely than its same-index counterpart (Pearson correlation: 0.851 vs. 0.640). This observation motivates establishing cross-model neuron correspondence directly from functional responses rather than fixed neuron indices.

To address this problem, we propose Targeted Alignment and Neuron Grafting via Optimal Transport (TANGO), a training-free framework for cross-modal reasoning transfer through neuron-aligned model merging. Rather than assuming same-index correspondence, TANGO formulates neuron alignment as an **optimal transport** problem over activation responses to shared inputs, yielding a weighted correspondence between the source and target neuron spaces. It first identifies task-relevant neurons in both models using their respective task data, and then combines target task

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

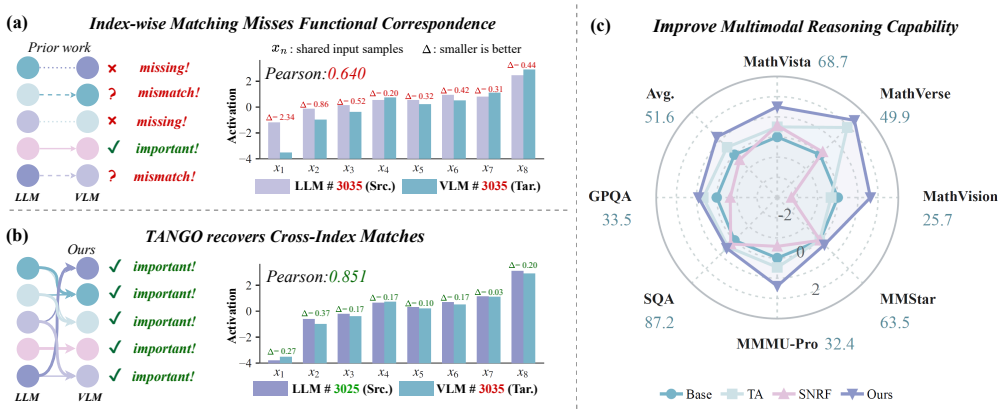

Figure 1: **Neuron alignment for reasoning transfer.** (a) Same-index matching can produce functionally mismatched pairs while overlooking compatible neurons at different indices. (b) TANGO aligns neurons by their functional responses, recovering a cross-index source neuron that more closely matches the target neuron. Each bar pair compares the source and target neuron responses to the same input sample, and Pearson correlation summarizes their similarity across all shared samples. (c) Performance gains over the base VLM across multimodal reasoning benchmarks.

relevance with source reasoning relevance transported through the learned correspondence to select transfer locations. The same transport plan is then reused to align source parameters with the selected target neurons before selective merging, ensuring that where reasoning is transferred and how source parameters are aligned follow the same functional correspondence.

We compare TANGO with five training-free reasoning-transfer baselines across four VLMs from the Qwen (Bai et al., 2025b; Bai et al., 2025a) and InternVL (Chen et al., 2024b; Wang et al., 2025) families, on benchmarks spanning multimodal mathematical reasoning, general multimodal understanding, and text-only scientific reasoning. As summarized in Figure 1(c), TANGO consistently improves reasoning performance across diverse benchmarks. Across 28 comparable combinations of models and mathematical reasoning settings, TANGO achieves the best result in 27 combinations, and ranks first in 18 of 20 comparisons on general multimodal and text-only reasoning. Ablation studies further show that the identified shared neurons are functionally important, while both functional alignment and soft OT-based transport are essential for effective capability transfer.

Our contributions are three-fold: (1) **Functional Neuron Correspondence:** We reveal the limitation of same-index neuron matching for cross-modal reasoning transfer and establish functional neuron correspondences from shared-input activations via optimal transport. (2) **Methodology:** We propose TANGO, a training-free neuron-aligned model merging framework that uses the learned transport plan for both transfer-location selection and source-parameter alignment. (3) **Performance:** Across four VLMs from the Qwen and InternVL families, TANGO achieves the best result in 27 of 28 mathematical reasoning settings and ranks first in 18 of 20 general reasoning comparisons.

2 Related Work

Vision language models. Vision-language models (VLMs) integrate visual perception with large language models (LLMs) to support unified multimodal understanding and generation (Liu et al., 2023; Dai et al., 2023). Recent advances have further extended VLMs beyond conventional perception-oriented tasks toward complex multimodal reasoning, including mathematical reasoning (Lu et al., 2024) and knowledge-intensive reasoning (Yue et al., 2024). To enhance these capabilities, prior work has primarily relied on training-based adaptation, ranging from full supervised or instruction tuning with curated multimodal data (Liu et al., 2023; Dai et al., 2023) to parameter-efficient methods that update only a small subset of model parameters to reduce training costs (Luo et al., 2023; Sung et al., 2022). More recent reasoning-oriented post-training leverages explicit reasoning traces and synthetic supervision (Xu et al., 2025; Dong et al., 2025), together with pref-

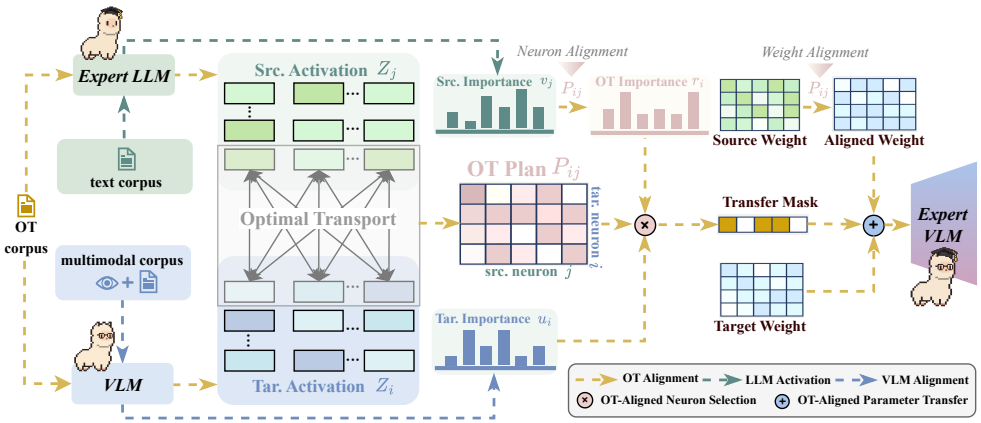

Figure 2: **Overview of** TANGO. Task-specific text and multimodal corpora provide neuron importance scores for the source expert LLM and target VLM, respectively. Activations elicited by a shared calibration corpus are used to estimate an optimal transport (OT) plan between source and target neurons. The same plan supports two operations: transporting source importance scores into the target neuron space for combination with target importance scores to construct a transfer mask, and aligning source weights with target neurons. The mask then guides selective merging of the aligned source weights with the target weights to obtain the enhanced VLM.

erence optimization and reinforcement learning (Dong et al., 2025; Tan et al., 2025), to improve multi-step reasoning. Despite their effectiveness, these approaches require additional supervision and gradient-based optimization, resulting in non-trivial data and computational overhead (Hu et al., 2022). These limitations motivate alternative approaches that reuse capabilities already encoded in existing models without extensive retraining.

Reasoning transfer to VLMs. A growing line of work explores transferring reasoning capabilities from language models to VLMs, mainly through representation-level or parameter-level approaches. Representation-based methods transfer internal reasoning representations from language models to VLMs through inference-time intervention (Zhan et al., 2026; Gan et al., 2026). In parallel, parameter-space approaches transfer reasoning capabilities by directly manipulating or integrating model parameters. General model merging methods provide training-free mechanisms for combining capabilities in parameter space (Jin et al., 2023; Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024). Building on these techniques, BR2V demonstrates direct reasoning transfer from LLMs to VLMs through model merging (Chen et al., 2025). Subsequent methods progressively refine reasoning transfer through parameter and layer selection with subspace alignment (Hu et al., 2025), subspace-level modulation (Huang et al., 2026), and neuron-level fusion (Cui et al., 2026b). Yet existing selective transfer methods primarily focus on identifying where or what to transfer. Even neuron-level transfer typically relies on predefined one-to-one correspondence within a shared language backbone, rather than explicitly inferring functional correspondence between source and target neurons (Cui et al., 2026b). Activation-based optimal transport has recently been explored to establish neuron correspondence for cross-architecture LLM merging (Cui et al., 2026a). Its role in selective reasoning transfer from LLMs to VLMs, however, remains underexplored.

3 METHOD

We introduce TANGO, a training-free framework for transferring reasoning capabilities from a source LLM to a target VLM. As illustrated in Figure 2, TANGO identifies task-relevant neurons in the source and target models, establishes their functional correspondence through activation-based optimal transport, and selectively transfers the aligned source parameters to the target VLM.

3.1 PRELIMINARIES

Model merging. Model merging combines capabilities from independently trained models in a single parameter set without additional training (Matena & Raffel, 2022). Given weight matrices $\mathbf{W}_A$ and $\mathbf{W}_B$ from the same layer and projection type of two models, standard linear merging requires identical input and output dimensions and interpolates the matrices element-wise:

$$\widehat{\mathbf{W}} = (1-\alpha)\mathbf{W}_A + \alpha\mathbf{W}_B, \qquad \alpha \in [0,1]. \tag{1}$$

The coefficient α balances the contributions of the two parameter sets.

Optimal transport. Optimal transport (OT) compares two probability distributions by minimizing the cost of transporting mass from one to the other. For discrete distributions supported on n target elements and m source elements, the pairwise costs form $\mathbf{C} \in \mathbb{R}^{n\times m}$, where C_{ij} is the cost of matching target element i with source element j. Entropy-regularized OT solves:

$$\begin{aligned} \mathbf{P}^\star &= \underset{\mathbf{P}\in\Pi(\mathbf{a},\mathbf{b})}{\arg\min} \langle \mathbf{P}, \mathbf{C}\rangle - \varepsilon H(\mathbf{P}), \\ \Pi(\mathbf{a},\mathbf{b}) &= \left\{\mathbf{P} \in \mathbb{R}_+^{n\times m} \mid \mathbf{P}\mathbf{1}_m = \mathbf{a},\ \mathbf{P}^\top \mathbf{1}_n = \mathbf{b}\right\}. \end{aligned} \tag{2}$$

Here, $\mathbf{P}$ is a nonnegative transport plan, $\mathbf{1}_k$ denotes the k-dimensional all-ones vector, and $H(\mathbf{P}) = -\sum_{i,j} P_{ij} \log P_{ij}$ is the entropy of the plan. The marginals are uniform, with $\mathbf{a} = \frac{1}{n}\mathbf{1}_n$ and $\mathbf{b} = \frac{1}{m}\mathbf{1}_m$. The coefficient $\varepsilon > 0$ controls the strength of entropy regularization.

3.2 NEURON PROFILING AND ACTIVATION EXTRACTION

Setup. The target VLM and source LLM are denoted by $\mathcal{M}_t$ and $\mathcal{M}_s$, respectively. `TANGO` updates only the language-model parameters of $\mathcal{M}_t$, while keeping its vision encoder and multimodal projector fixed. Within each Transformer layer, transfer covers `q_proj`, `k_proj`, `v_proj`, `o_proj`, `up_proj`, and `down_proj`. For a linear projection: $\mathbf{y} = \mathbf{W}\mathbf{x}$. A neuron corresponds to one input or output coordinate of $\mathbf{W}$. Thus, neurons associated with `q_proj`, `k_proj`, `v_proj`, and `up_proj` index rows, whereas those associated with `o_proj` and `down_proj` index columns. Query and key/value neurons are handled separately under grouped-query attention. Omitting the layer and projection indices, d_t and d_s denote the numbers of target and source neurons, respectively.

Task-specific neuron profiling. Effective transfer requires identifying target neurons involved in multimodal reasoning and source neurons that encode transferable reasoning capabilities. We therefore profile the target VLM on a multimodal reasoning corpus and the source model on a text-only reasoning corpus. In our experiments, the multimodal corpus focuses on mathematical reasoning. The resulting task-relevance indicators form binary vectors:

$$\mathbf{u} \in \{0,1\}^{d_t}, \qquad \mathbf{v} \in \{0,1\}^{d_s}.$$

Here, $u_i = 1$ and $v_j = 1$ indicate that target neuron i and source neuron j are relevant to their respective reasoning tasks. These indicators determine the candidate target and source neurons considered for subsequent transfer.

Cross-model activation extraction. Task-specific profiling identifies relevant neurons within each model, but does not determine which source neurons functionally correspond to the target neurons. To establish such correspondence, we evaluate both models on the same set of N text-only calibration inputs and collect their activation responses:

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{d_t}] \in \mathbb{R}^{N\times d_t}, \qquad \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{d_s}] \in \mathbb{R}^{N\times d_s}. \tag{3}$$

Each row corresponds to one shared input, while each column records the responses of one neuron across all N inputs. Thus, $\mathbf{x}_i$ and $\mathbf{y}_j$ provide comparable functional profiles of target neuron i and source neuron j on a common input basis.

For a gated feed-forward block, the neuron response used for profiling is measured after the gating operation:

$$\mathbf{h} = \mathrm{SiLU}(\mathbf{W}_{\mathrm{gate}}\mathbf{x}) \odot (\mathbf{W}_{\mathrm{up}}\mathbf{x}). \tag{4}$$

For cross-model alignment, activations are instead extracted on the projection side corresponding to the neuron dimension being aligned: output-side responses are used for `q_proj`, `k_proj`, `v_proj`, and `down_proj`, whereas input-side responses are used for `o_proj` and `up_proj`. We do not directly transfer `gate_proj` parameters.

3.3 NEURON ALIGNMENT AND SELECTION

Activation-based neuron alignment. The activation matrices in equation 3 make target and source neurons comparable through their responses to the same inputs. For target neuron i and source neuron j, the Pearson distance defines the alignment cost:

$$C_{ij} = 1 - \rho(\mathbf{x}_i, \mathbf{y}_j), \tag{5}$$

where $\rho(\mathbf{x}_i, \mathbf{y}_j) \in [-1, 1]$ is the Pearson correlation computed over the N shared calibration inputs. A lower C_{ij} indicates more similar response patterns. Solving equation 2 with this cost matrix and $n = d_t$, $m = d_s$ yields a transport plan $\mathbf{P}^\star$. Normalizing each row to sum to one gives $\mathbf{P}[P_{ij}] \in \mathbb{R}_+^{d_t \times d_s}$, where P_{ij} is the transport weight from source neuron j to target neuron i. Thus, row i defines a soft correspondence distribution over source neurons for target neuron i.

Task-aware neuron selection. The transport plan assigns source contributions to every target neuron without accounting for their task relevance. The relevance indicators $\mathbf{u}$ and $\mathbf{v}$ are therefore combined with the transport plan to rank target neurons for transfer. For target neuron i, the transported source relevance, alignment concentration, and transfer score are defined as:

$$r_i = \sum_{j=1}^{d_s} P_{ij} v_j, \qquad c_i = \max_{1 \le j \le d_s} P_{ij}, \qquad s_i = u_i r_i c_i. \tag{6}$$

Here, r_i measures the source reasoning relevance transported to target neuron i, c_i captures the concentration of its alignment, and u_i enforces target-side task relevance. Their product s_i is the transfer score for target neuron i.

To impose a fixed transfer budget for each projection type, scores are pooled across all layers. For projection group g, the B_g highest-scoring eligible layer–neuron pairs are selected:

$$\mathcal{S}_g = \mathrm{Top}_{B_g} \left\{ (\ell, i; s_{\ell i}) \,|\, u_{\ell i} = 1 \right\}. \tag{7}$$

The selected pairs define a binary mask $\mathbf{M}$ over the corresponding weight coordinates, which determines the parameters updated in Section 3.4.

3.4 SELECTIVE PARAMETER TRANSFER

After selecting the transfer locations, we first align the source weights with the target parameter coordinates. For a projection of group g, let $\mathbf{W}_t$ and $\mathbf{W}_s$ denote the target and source weight matrices. The transport plan is applied along the aligned neuron dimension:

$$\widetilde{\mathbf{W}}_s = \begin{cases} \mathbf{P}^{\mathrm{out}} \mathbf{W}_s, & g \in \{\texttt{q_proj}, \texttt{k_proj}, \texttt{v_proj}, \texttt{down_proj}\}, \\ \mathbf{W}_s (\mathbf{P}^{\mathrm{in}})^\top, & g \in \{\texttt{o_proj}, \texttt{up_proj}\}. \end{cases} \tag{8}$$

Here, $\mathbf{P}^{\mathrm{out}}$ and $\mathbf{P}^{\mathrm{in}}$ denote the transport plans defined on the output and input neuron dimensions.

Finally, TANGO interpolates only the selected coordinates:

$$\mathbf{W}_{\mathrm{new}} = \mathbf{W}_t + \alpha \mathbf{M} \odot \left(\widetilde{\mathbf{W}}_s - \mathbf{W}_t \right), \tag{9}$$

where α is the transfer strength for the current projection group. All unselected language parameters retain their target values, and all vision-specific parameters remain unchanged.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and transfer settings. We evaluate TANGO on four target VLMs from the Qwen and InternVL families. For Qwen, we transfer from Qwen2.5-Math-7B to Qwen2.5-VL-7B and from DeepSeek-R1-0528-Qwen3-8B to Qwen3-VL-8B (Bai et al., 2025b;a; Yang et al., 2024; Guo et al., 2025). For InternVL, we transfer from Qwen2.5-3B-Instruct-Math-GRPO to InternVL2.5-4B and from DeepSeek-R1-0528-Qwen3-8B to InternVL3.5-8B (Chen et al., 2024b; Wang et al., 2025;

Table 1: **Main results on multimodal mathematical reasoning benchmarks.** Avg. denotes the mean over all seven evaluation settings. For each model, the best and second-best results are shown in **bold** and underlined, respectively.

Model	Method	MathVista			MathVerse		Math Vision	MMStar-Math	Avg.
		CoT	Format	Solution	T.D.	V.I.			
Qwen2.5-VL-7B	*Base*	69.1	69.0	68.1	54.57	45.30	25.66	63.48	56.46
	TA	70.2	68.7	68.8	57.74	45.94	25.33	63.46	57.17
	DARE	69.9	68.4	68.5	57.99	43.36	25.99	63.36	56.79
	BR2V	68.4	68.4	67.9	49.24	45.62	25.33	58.87	54.82
	L2V-CoT	71.3	69.0	68.1	57.36	45.50	26.45	61.83	57.08
	SNRF	69.9	69.4	68.6	54.44	45.81	25.66	63.48	56.43
	TANGO	**71.9**	**69.6**	**69.2**	**58.50**	**46.19**	**27.28**	**63.79**	**58.07**
Qwen3-VL-8B	*Base*	71.0	65.6	71.3	57.99	49.37	18.75	48.49	54.64
	TA	71.3	65.8	73.4	58.50	49.73	19.41	45.08	54.75
	DARE	69.7	66.2	72.6	58.12	**49.96**	19.08	44.89	54.36
	BR2V	71.0	63.0	71.6	57.11	49.22	19.74	44.96	53.80
	L2V-CoT	71.2	65.2	69.8	58.63	49.75	19.41	47.80	54.54
	SNRF	71.0	65.8	72.9	58.37	49.86	20.07	47.41	55.06
	TANGO	**72.3**	**66.7**	**73.6**	**58.76**	**49.96**	**20.72**	**49.57**	**55.94**
InternVL2.5-4B	*Base*	60.9	65.8	62.4	40.61	35.28	25.33	58.55	49.84
	TA	61.0	66.4	62.5	41.88	36.06	25.66	60.42	50.56
	DARE	60.4	65.9	62.3	41.88	35.41	25.00	58.55	49.92
	BR2V	57.5	66.3	62.5	42.65	35.67	26.64	58.65	49.99
	L2V-CoT	59.7	65.9	62.1	40.99	36.13	25.33	60.59	50.11
	SNRF	60.4	66.5	**63.1**	42.51	35.36	25.33	59.05	50.32
	TANGO	**61.3**	**66.7**	62.9	**43.15**	**36.29**	**27.96**	**61.22**	**51.36**
InternVL3.5-8B	*Base*	70.5	71.4	69.9	47.21	38.58	24.34	69.00	55.85
	TA	70.8	71.2	69.7	46.83	39.21	24.34	69.18	55.89
	DARE	70.9	71.6	70.3	47.21	38.83	24.01	69.88	56.10
	BR2V	69.3	71.6	70.2	**47.72**	38.96	25.00	68.51	55.90
	L2V-CoT	70.7	71.7	70.1	47.21	39.09	**25.66**	69.29	56.25
	SNRF	70.3	71.3	**70.5**	47.21	38.96	25.32	69.98	56.22
	TANGO	**71.1**	**71.9**	**70.5**	**47.72**	**39.34**	**25.66**	**70.49**	**56.67**

Williammsq, 2025). For each pair, neuron correspondences are estimated using shared calibration inputs, and only selected language-model parameters are merged, while the vision encoder and multimodal projector remain unchanged.

Evaluation benchmarks. We evaluate multimodal mathematical reasoning on MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024), and MMStar-Math, the mathematics subset of MMStar (Chen et al., 2024a). We further assess general multimodal capabilities on MME (Fu et al., 2025), MMStar, MMMU (Yue et al., 2024), MMMU-Pro (Yue et al., 2025), and ScienceQA (Lu et al., 2022), and evaluate cross-domain reasoning on GPQA (Rein et al., 2024). Benchmarks used for ablation and sensitivity analyses are specified in their respective subsections. We report the official aggregate score for MME and accuracy (%) for all other benchmarks. In Table 1, Avg. denotes the mean across the seven reported mathematical reasoning evaluation settings.

Baselines and evaluation protocol. We compare TANGO with the unmodified target VLM (Base) and five training-free baselines: Task Arithmetic (TA) (Ilharco et al., 2023), DARE (Yu et al., 2024), BR2V (Chen et al., 2025), L2V-CoT (Zhan et al., 2026), and SNRF (Cui et al., 2026b). These baselines cover several representative approaches to training-free capability transfer. TA and DARE construct and merge task vectors in parameter space; BR2V and SNRF transfer reasoning capabilities by modifying selected model parameters; and L2V-CoT performs latent-representation intervention during inference. To ensure a fair comparison, we use the same source and target checkpoint across methods whenever permitted by their original formulations.

All methods are evaluated zero-shot with deterministic decoding, following the benchmark-specific protocols for answer extraction and scoring. Within each transfer setting, the same prompts, image preprocessing, decoding configurations, and evaluators are used across methods.

Table 2: **General multimodal and text-only reasoning performance.** Results are reported for the Qwen and InternVL model families. MMStar and GPQA use their overall scores; SQA denotes ScienceQA. The best and second-best results within each target model are shown in **bold** and underlined, respectively.

(a) **Qwen series**

Method	Reasoning			Cross-domain	
	MME	MMMU-Pro	MMStar	SQA	GPQA
Qwen2.5-VL-7B					
Base	1680.70	32.43	62.63	87.20	33.48
TA	1679.69	32.89	62.80	87.66	34.15
DARE	1685.97	33.30	62.56	87.61	34.15
BR2V	1666.29	32.66	60.37	86.07	31.92
L2V-CoT	1681.77	32.37	62.58	87.06	33.93
SNRF	1711.17	31.85	59.86	87.46	32.81
TANGO	**1722.21**	**33.82**	**63.73**	**87.75**	**34.38**
Qwen3-VL-8B-Instruct					
Base	1727.06	35.43	62.11	94.30	39.51
TA	1735.64	35.43	61.11	94.15	40.18
DARE	1737.29	35.49	60.99	94.15	39.29
BR2V	1695.02	34.22	61.10	94.15	39.06
L2V-CoT	1739.19	35.43	62.27	94.35	40.18
SNRF	1731.22	35.49	61.77	94.40	39.73
TANGO	**1744.43**	**35.90**	**62.61**	**94.70**	**40.63**

(b) **InternVL series**

Method	Reasoning			Cross-domain	
	MME	MMMU-Pro	MMStar	SQA	GPQA
InternVL2.5-4B					
Base	1671.39	23.29	59.12	97.37	29.46
TA	1680.28	23.29	59.43	**97.42**	29.02
DARE	1672.50	23.24	58.94	97.32	29.69
BR2V	1670.14	23.24	58.37	97.03	29.46
L2V-CoT	1679.01	22.95	59.49	97.32	29.69
SNRF	1678.11	22.72	59.29	**97.42**	29.91
TANGO	**1685.50**	**23.41**	**59.55**	97.37	**30.58**
InternVL3.5-8B-Instruct					
Base	1696.63	29.94	64.12	97.12	41.07
TA	1696.24	30.17	64.08	96.98	41.74
DARE	1694.99	**30.23**	64.12	97.22	39.96
BR2V	1703.83	29.02	63.23	96.98	41.07
L2V-CoT	1695.13	29.94	64.19	97.22	41.52
SNRF	1700.35	29.94	64.17	97.12	41.52
TANGO	**1705.66**	30.17	**64.52**	**97.37**	**42.63**

4.2 Main Results

Multimodal mathematical reasoning. Table 1 summarizes the results across four target VLMs. TANGO consistently outperforms the training-free baselines, achieving the best or tied-best result in 27 of the 28 model–evaluation-setting comparisons and the highest average score for all four target models. Representative gains over the strongest baselines include Qwen2.5-VL-7B (+0.90 pts in Avg.), Qwen3-VL-8B-Instruct (+0.88 pts in Avg.), and InternVL2.5-4B (+1.32 pts on MathVision). On InternVL3.5-8B-Instruct, TANGO matches or exceeds the strongest baseline on every evaluation setting while attaining the highest overall average. The consistent gains across both Qwen and InternVL demonstrate that the transferred reasoning capability remains effective across different VLM families and model scales.

General multimodal and cross-domain reasoning. Table 2 further examines whether these reasoning gains generalize beyond the mathematical benchmarks. TANGO ranks first in 18 of the 20 model–benchmark comparisons and achieves the second-best result in the remaining two. It obtains the highest MME score for all four target models, with representative gains over the strongest baselines on Qwen2.5-VL-7B (+11.04), Qwen3-VL-8B-Instruct (+5.24), and InternVL2.5-4B (+5.22). The advantage also extends to cross-domain reasoning: TANGO achieves the highest GPQA score for every target model, including a gain of +1.11 on InternVL3.5-8B-Instruct. Together, these results show that TANGO improves multimodal mathematical reasoning without sacrificing general multimodal capability, while the gains extend beyond the source reasoning domain.

4.3 Ablation Studies

Functional importance of the identified shared neurons. We examine whether the shared neurons identified by TANGO are functionally important by perturbing their associated parameters while leaving the remaining parameters unchanged. Specifically, we either zero these parameters or replace them with random values, and evaluate the resulting models across mathematical, multimodal, and cross-domain reasoning benchmarks. As shown in Table 3, zeroing the identified parameters collapses performance on all five benchmarks for every target model, while random replacement consistently causes substantial degradation. For example, random replacement reduces GPQA on Qwen3-VL-8B-Instruct to 23.95, compared with 39.51 for the original model. These results indicate that the identified shared neurons correspond to functionally important parameter regions, and that their learned values are essential for preserving the associated capabilities.

Table 3: **Functional importance of identified shared neurons.** We intervene on the identified shared neurons by either replacing their parameters with zeros or random values, and compare the resulting performance with the original model. Higher is better for all metrics.

Model	Intervention	MathVista-CoT	MathVerse-TD	MME	MMMU-Pro	GPQA
Qwen3-VL-8B	None(TANGO)	72.3	57.99	1727.06	35.43	39.51
	Zero	0.0	0.00	0.00	0.00	0.00
	Random	60.9	40.61	1671.39	30.21	23.95
Qwen2.5-VL-7B	None(TANGO)	71.9	54.57	1680.70	32.43	33.48
	Zero	0.0	0.00	0.00	0.00	0.00
	Random	59.3	46.45	1623.3	25.43	22.54
InternVL2.5-4B	None(TANGO)	61.3	40.61	1671.39	23.29	29.46
	Zero	0.0	0.00	0.00	0.00	0.00
	Random	52.6	33.50	1618.82	18.15	22.54
InternVL3.5-8B	None(TANGO)	71.1	47.21	1696.63	29.94	41.07
	Zero	0.0	0.00	0.00	0.00	0.00
	Random	64.7	38.96	1639.24	26.13	33.93

Table 4: **Ablation study of key components in** TANGO. We examine the effects of functional neuron alignment and soft OT-based parameter transport. Overall denotes the average over perception, reasoning, and mathematics. Higher is better for all metrics.

Method	OT Alignment	OT Transfer	Perception	Reasoning	Math	Overall
Base	✗	✗	76.59	73.78	48.49	66.29
w/o OT alignment	✗	✗	77.06	74.28	47.80	66.38
Shuffled alignment	✗	✓	78.56	73.02	49.16	66.91
Hard transport	✓	✗	76.11	74.65	48.08	66.28
TANGO (full)	✓	✓	**78.75**	**75.43**	**49.57**	**67.92**

Effects of alignment and soft transport. We conduct this ablation on Qwen2.5-VL-7B using MMStar. Perception, Reasoning, and Math correspond to the coarse perception, logical reasoning, and mathematics subskills, respectively, while Overall denotes the average across all six MMStar subskills. We evaluate the two key design choices in TANGO by removing OT alignment, shuffling the learned source–target correspondence, or replacing soft OT transport with hard matching. As shown in Table 4, the full method achieves the strongest overall performance, while each ablated variant leads to a decline. Shuffling the correspondence weakens the gain despite retaining the same transport mechanism, indicating that the learned functional correspondence provides more effective source–target matching than arbitrary alignment. Hard transport nearly removes the improvement over the base model (66.28 vs. 66.29), showing that preserving the soft correspondence structure is important for effective parameter transfer. Together, these results support both functional alignment and soft transport as key components of TANGO.

4.4 Hyperparameter sensitivity

We analyze three hyperparameters in TANGO: the OT entropy regularization coefficient ϵ, the number N of shared calibration samples used for OT estimation, and the parameter injection strength β. As shown in Figure 3, we study ϵ and N on Qwen2.5-VL-7B, and β on Qwen3-VL-8B. For each sweep, all other settings are kept fixed. In particular, when varying β, we keep the learned OT correspondence and neuron-selection mask unchanged so that only the amount of transferred parameters is varied. Dashed lines indicate the performance of the corresponding unmodified target models.

OT entropy regularization. We vary ϵ to examine how the sharpness of the OT correspondence affects transfer. Performance improves as ϵ increases from a small value, remains strong around $\epsilon = 0.03$–0.05, and decreases when the transport plan becomes overly smoothed. This trend suggests that effective transfer benefits from a correspondence that is neither excessively concentrated nor overly diffuse. Accordingly, we fix $\epsilon = 0.03$ for all model pairs in the main experiments.

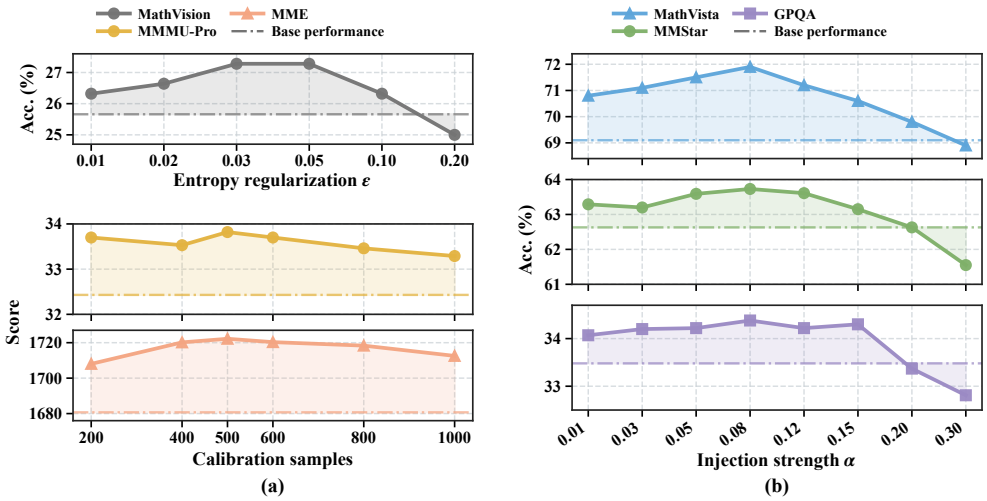

Figure 3: **Hyperparameter sensitivity of TANGO.** (a) Sensitivity to the OT entropy regularization coefficient ϵ and the number of shared calibration samples on Qwen2.5-VL-7B. (b) Sensitivity to the parameter injection strength α on Qwen3-VL-8B.

Shared calibration set size. We vary the number of shared calibration samples used to estimate functional correspondence. Both MME and MMMU-Pro remain above the corresponding base-model performance throughout the evaluated range and reach their strongest results around $N = 500$. Increasing the calibration set further provides no consistent benefit, indicating that reliable cross-model correspondence can be estimated from a moderate number of shared inputs. Based on this saturation behavior, $N = 500$ is used uniformly across all models.

Parameter injection strength. We vary α while keeping the learned correspondence and selected transfer locations fixed. Moderate injection consistently improves MathVista, MMStar, and GPQA, with $\alpha = 0.08$ providing the strongest overall operating point in this setting. As α becomes larger, performance gradually deteriorates and eventually falls below the base model, showing that excessive source injection can disrupt capabilities already present in the target VLM. These results complement the alignment ablations in Table 4: effective transfer depends not only on identifying where to transfer parameters, but also on controlling how much source information is introduced. The preferred injection strength can vary across model pairs; additional model-specific sensitivity results are provided in the appendix.

5 Conclusion & Limitations

Conclusion. We presented TANGO, a training-free framework for transferring mathematical reasoning from language models to vision-language models through functional neuron correspondence. By combining task-specific neuron profiling with activation-based optimal transport, TANGO identifies target neurons for transfer and aligns source parameters for selective merging without additional model training. Experiments across four target VLMs from the Qwen and InternVL families show consistent gains in mathematical reasoning while largely preserving general multimodal and text-only reasoning capabilities.

Limitations. TANGO depends on the data used for task-specific neuron profiling and shared activation calibration, since distribution mismatch may affect the estimated neuron relevance and source–target correspondences. Its effectiveness on more heterogeneous model architectures and specialized capabilities beyond mathematical reasoning remains to be further studied. From a broader-impact perspective, TANGO may reduce the training costs of enhancing VLM reasoning, but transferred models may also inherit biases or hallucination tendencies from their source models. These effects require evaluation in downstream applications.

AI USE STATEMENT

In this work, we used generative AI tools, primarily OpenAI's GPT-5.6 Sol and GPT-5.5, for language polishing, translation, formatting, and supporting programming tasks such as data processing, debugging, and auxiliary script generation. All AI-assisted content was manually reviewed and verified by the authors, who take full responsibility for the final manuscript and research artifacts.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. No human subjects or animal experiments were involved. All datasets used in this study are publicly available and were used in accordance with their respective licenses and usage policies. No personally identifiable information was used, and no additional privacy or security concerns were introduced. We also took care to avoid bias and maintain research integrity throughout the study.

REPRODUCIBILITY STATEMENT

We have taken steps to make the results in this paper reproducible. We provide the implementation of TANGO together with instructions for environment setup, checkpoint preparation, functional-profile collection, OT alignment, model merging, and evaluation. The paper and accompanying documentation specify the calibration data, neuron-selection budgets, OT settings, injection strengths, model configurations, and benchmark protocols. We evaluate on MathVista, MathVerse, MathVision, MM-Star, MME, MMMU, MMMU-Pro, ScienceQA, and GPQA using Qwen2.5-VL-7B, Qwen3-VL-8B, InternVL2.5-4B, and InternVL3.5-8B as target models. We identify the exact source and target checkpoints used in each transfer setting and apply the same evaluation protocol to methods compared within a setting. These materials enable independent verification of our results, subject to the access requirements of individual benchmarks and checkpoints.

REFERENCES

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

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? In *NeurIPS*, 2024a.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring reason to vision: Understanding perception and reasoning through model merging. In *ICML*, 2025.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Chenhang Cui, Binyun Yang, Fei Shen, Yuxin Chen, Jingnan Zheng, Xiang Wang, An Zhang, and Tat-Seng Chua. Transport and merge: Cross-architecture merging for large language models. In *ICML*, 2026a.

Chenhang Cui, An Zhang, Yuxin Chen, Gelei Deng, Jingnan Zheng, Zhenkai Liang, Xiang Wang, and Tat-Seng Chua. Do llms and vlms share reasoning neurons? evidence and mechanisms of cross-modal transfer. In *CVPR Findings*, 2026b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony M. H. Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *CVPR*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *NeurIPS*, 2025.

Woody Haosheng Gan, Deqing Fu, Julian Asilis, Ollie Liu, Vatsal Sharan, Robin Jia, and Willie Neiswanger. Textual steering vectors can improve visual understanding in multimodal large language models. In *ACL*, 2026.

Sreyan Ghosh, Chandra Kiran Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning in lvlms. In *ICLR*, volume 2025, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 2025.

Yifan Hou, Buse Giledereli, Yilei Tu, and Mrinmaya Sachan. Do vision-language models really understand visual language? In *ICML*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Yijie Hu, Zihao Zhou, Kaizhu Huang, Xiaowei Huang, and Qiufeng Wang. Can mllms absorb math reasoning abilities from llms as free lunch? In *NeurIPS*, 2025.

Chenyu Huang, Peng Ye, Xudong Tan, Jinhan Mu, Shenghe Zheng, Li Shen, and Tao Chen. Frism: Fine-grained reasoning injection via subspace-level model merging for vision-language models. In *ICML*, 2026.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *ICLR*, 2023.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *ICLR*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *ACL*, 2021.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

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

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. In *NeurIPS*, 2023.

Grace Luo, Trevor Darrell, and Amir Bar. Vision-language models create cross-modal task representations. In *ICML*, 2025.

Michael S Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *NeurIPS*, 2022.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *COLM*, 2024.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. PaliGemma 2: A family of versatile VLMs for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*, 2022.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. In *NeurIPS*, 2025.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, Mengqi He, Fabian Waschkowski, Lukas Wesemann, Peter Tu, and Jing Zhang. More thought, less accuracy? on the dual nature of reasoning in vision-language models. In *ICLR*, 2026.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *NeurIPS*, 2024.

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.

Williammsq. Qwen2.5-3b-instruct-math-grpo. https://huggingface.co/Williammsq/Qwen2.5-3B-Instruct-Math-GRPO, 2025. Accessed: 2025-09-08.

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

Zhiyu Xu, Lean Wang, Yuanxin Liu, Lei Li, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Investigating cross-modal skill injection: Scenarios, methods, and hyperparameters. In *ACL*, 2026.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *NeurIPS*, 2023.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

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 *ICML*, 2024.

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 *CVPR*, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *ACL*, 2025.

Yu-Liang Zhan, Xinyu Tang, Han Wan, Jian Li, Jirong Wen, and Hao Sun. L2v-cot: Cross-modal transfer of chain-of-thought reasoning via latent intervention. In *AAAI*, 2026.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality check on the evaluation of large multimodal models. In *Findings of NAACL*, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *ECCV*, 2024.

A Experimental Details

A.1 From Functional Correspondence to Selective Transfer

The main paper presents the formulation of TANGO; here, we clarify how its components work together as an end-to-end transfer procedure. The central difficulty is that neuron indices are architectural coordinates rather than functional identities. Consequently, intersecting independently selected neurons at identical indices can both miss useful cross-index correspondences and retain neurons whose apparent overlap does not reflect similar behavior. TANGO therefore separates the transfer process into two questions: *which source and target neurons exhibit corresponding behavior*, and *which of these correspondences should be used for parameter transfer*. This separation is important because a strong activation in both models does not by itself imply that the two same-index neurons implement the same function.

TANGO first records neuron-response profiles from the source LLM and target VLM on a shared calibration corpus. These profiles provide a common behavioral coordinate system in which neurons can be compared without assuming index alignment. The resulting pairwise similarities are converted into an entropy-regularized optimal-transport plan. Unlike a hard permutation, this soft plan can distribute the relevance of one neuron across several functionally related neurons, thereby representing ambiguous or distributed correspondence between independently trained models. Source-side relevance is then projected through the transport plan into the target neuron space and combined with the target model's own relevance. The selective mask retains target neurons that are supported by both transferred source evidence and target-side evidence, while suppressing correspondences that are weak, inconsistent, or irrelevant to the target model.

Finally, the mask is applied to parameter transfer rather than to model outputs or hidden states. Only the selected target coordinates receive the aligned source update; all remaining coordinates preserve the original VLM parameters. This design turns the transport plan into a conservative editing mechanism: functional knowledge is transferred where cross-model evidence is sufficient, while the target model's existing multimodal capability is left unchanged elsewhere. In this way, the response profiles, transport plan, selective mask, and parameter fusion are not independent modules, but successive stages that translate behavioral correspondence into localized model editing. The controlled permutation experiment in Fig. 4 verifies that this correspondence can be recovered when the ground truth is known, while Fig. 5 shows that the learned correspondence in real models contains substantial cross-index structure.

A.2 Implementation Details

Model configurations. Table 5 summarizes the model pairs used in our experiments. `TANGO` transfers directly from the reasoning source to the target VLM and does not require a reference model. Reference checkpoints are used only by task-vector baselines such as TA, DARE, and BR2V.

Task-specific activation profiling. We collect target-VLM activations from 200 multimodal mathematical examples sampled from Geometry3K (Lu et al., 2021) and source-LLM activations from 200 GSM-style reasoning examples (Cobbe et al., 2021). We use the same question–answer-with-rationale prompt format for both sides and deterministic generation with `do_sample=False`. Activation statistics are collected only from the language tower.

Specifically, we profile the output coordinates of the q, k, and v projections, the input coordinates of the attention output projection, and the shared FFN middle space at the input of `down_proj`. The SwiGLU gate is used when measuring FFN neuron responses, but `gate_proj` parameters are not directly transferred.

Functional alignment and transfer selection. We estimate source–target correspondence using 500 held-out GSM examples (Cobbe et al., 2021). These examples are distinct from the task-specific examples used to construct the relevance signals. For every layer and projection group, neuron activation profiles are standardized across calibration examples. We define the transport cost as

$$C_{ij} = 1 - \rho(\mathbf{z}_i^v, \mathbf{z}_j^r),$$

and solve the entropy-regularized transport problem using Sinkhorn iterations with $\epsilon = 0.03$.

We match the total selection budget of the corresponding hard-intersection baseline, apply the budget separately to each projection group, and multiply the FFN budget by 0.5. The same learned transport plans are reused during parameter transfer; no additional alignment is estimated at merge time.

Aligned parameter transfer. `TANGO` updates only the selected language-model projection parameters. The vision encoder, multimodal projector, token embeddings, language-model head, and all unselected parameters are copied unchanged from the target VLM. Query and output projections are transported in their corresponding hidden spaces; key and value projections use their grouped-query coordinates; and both FFN directions use the plan estimated in the shared intermediate space. Merged checkpoints are stored in BF16. The procedure is entirely training-free and does not perform gradient-based optimization.

A.3 Baseline Configurations

For Qwen2.5, TA uses $\alpha = 0.02$, and DARE uses $\alpha = 0.02$ with a drop rate of 0.2. For Qwen3-8B and InternVL3.5-8B, the corresponding values are $\alpha = 0.05$ and a DARE drop rate of 0.5. InternVL2.5 uses $\alpha = 0.10$ and a drop rate of 0.2. DARE uses seed 0 and rescales the retained task-vector entries.

BR2V uses target and reasoning coefficients of 0.9 and 0.1, respectively. SNRF uses $(r, \beta) = (4, 0.5)$ for Qwen2.5, $(8, 0.15)$ for Qwen3-8B and InternVL3.5-8B, and $(32, 0.5)$ for InternVL2.5. L2V-CoT uses an intervention strength of 0.3 and the 400-example direction-estimation protocol. All parameter-merging baselines modify only compatible language-tower parameters.

A.4 Evaluation Protocol

We use the local `lmms-eval` (Zhang et al., 2025) implementation and evaluate all models zero-shot with batch size 1 and deterministic decoding. Prompts, image preprocessing, decoding parameters, answer extraction, and evaluators are held fixed across methods within each model family.

MathVista reports the CoT, format, and solution settings; MathVerse reports the vision-intensive and text-dominant settings; and MathVision uses its testmini split. API-judged MathVista and MathVerse predictions are scored with GPT-4o using temperature 0. We report the aggregate MME score and accuracy percentages for the remaining benchmarks.

Qwen models use SDPA attention. InternVL3.5 is loaded with its official remote code under Transformers 4.52.1, without overriding its default system message, and uses `do_sample=False`.

B Additional Experimental Results

B.1 Controlled Functional-Correspondence Recovery

Because the ground-truth neuron correspondence between independently obtained models is unknown, we design a controlled experiment that isolates correspondence recovery from differences in model function. Specifically, on Layer 12 of DeepSeek-R1-0528-Qwen3-8B, we construct a functionally identical model by permuting its FFN intermediate neurons.

Table 5: **Model configurations and transfer strengths.** β_{attn} denotes the interpolation strength applied to the selected attention parameters.

Setting	Target VLM	Reasoning Source	Reference LLM	β_{attn}
Qwen2.5-7B	Qwen2.5-VL-7B	Qwen2.5-Math-7B-Instruct	Qwen2.5-7B	0.08
Qwen3-8B	Qwen3-VL-8B-Instruct	DeepSeek-R1-0528-Qwen3-8B	Qwen3-8B-Base	0.12
InternVL2.5-4B	InternVL2.5-4B	Qwen2.5-3B-Instruct-Math-GRPO	Qwen2.5-3B-Instruct	0.08
InternVL3.5-8B	InternVL3.5-8B-Instruct	DeepSeek-R1-0528-Qwen3-8B	Qwen3-8B-Base	0.13
Qwen3-4B	Qwen3-VL-4B-Instruct	General-Reasoner-Qwen3-4B	Qwen3-4B-Base	0.08

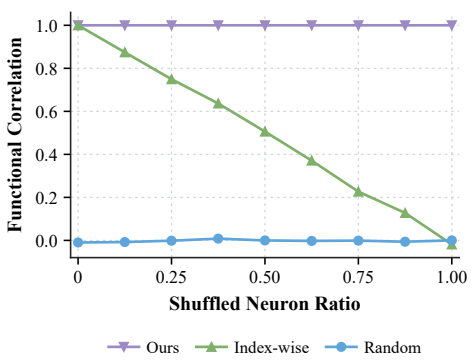

Figure 4: **Functional-correspondence recovery under function-preserving neuron permutation.** We permute different fractions of selected SwiGLU FFN neurons while applying the corresponding inverse permutation to the down projection, leaving the model function unchanged. TANGO recovers the latent correspondence and maintains near-perfect held-out functional correlation, whereas index-wise matching degrades with increasing permutation ratio and random matching remains near zero.

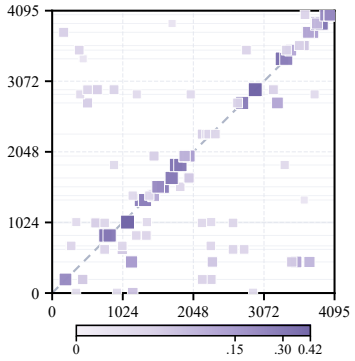

Figure 5: **Visualization of cross-index functional correspondence.** Representative rows of the row-normalized attention-Q transport plan from Layer 13 are shown. The dashed diagonal denotes same-index correspondence, while marker size and color encode off-diagonal transport mass, which shows that functional correspondence extends beyond neurons sharing the same index. This confirms that TANGO captures distributed cross-index alignments overlooked by index-wise matching.

For a SwiGLU FFN with $\mathbf{z} = \text{SiLU}(W_g\mathbf{h}) \odot (W_u\mathbf{h})$, we apply a permutation matrix P such that $W'_g = PW_g$, $W'_u = PW_u$, and $W'_d = W_dP^\top$. This gives $\mathbf{z}' = P\mathbf{z}$ while preserving the FFN output:

$$W'_d\mathbf{z}' = W_dP^\top P\mathbf{z} = W_d\mathbf{z}.$$

The applied permutation therefore provides an exact ground-truth correspondence without changing the model's computation.

We permute varying fractions of the 512 most active FFN neurons, estimate correspondence using 256 calibration examples, and evaluate it on 128 disjoint held-out examples over three random seeds. TANGO constructs the Pearson cost $C_{ij} = 1 - \rho(\mathbf{z}^t_i, \mathbf{z}^s_j)$ and solves the resulting transport problem using Sinkhorn OT with $\epsilon = 0.03$.

The controlled recovery experiment and the visualization of the transport plan learned on an actual LLM–VLM pair are presented side by side in Figs. 4 and 5.

As shown in Fig. 4, TANGO maintains held-out functional correlation close to 1.0 across all permutation ratios. In contrast, index-wise matching falls to 0.506 at a 50% permutation ratio, 0.227 at 75%, and -0.019 under complete permutation, while random matching remains near zero.

These results demonstrate that neuron indices are not stable functional identifiers and that activation-based OT can recover a known latent correspondence after function-preserving reindexing. This controlled result validates the correspondence mechanism itself; it does not assume that independently trained LLMs and VLMs admit a perfect one-to-one alignment.

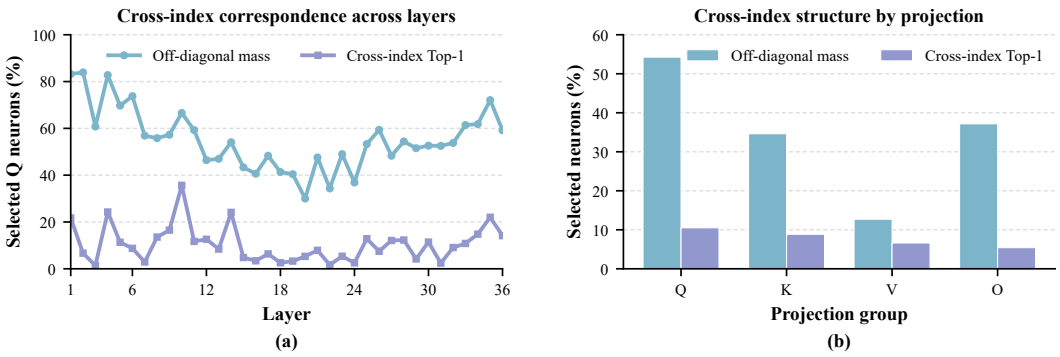

Figure 6: **Cross-index functional correspondence across layers and projections.** (a) Off-diagonal OT mass and cross-index Top-1 ratio for selected Q-projection neurons across all 36 layers. Off-diagonal mass measures the total soft correspondence assigned to differently indexed source neurons, whereas cross-index Top-1 records whether the dominant source match has a different index. (b) The same statistics aggregated over the Q, K, V, and O projection groups. Cross-index correspondence is widespread across layers and is especially pronounced in the Q projection.

B.2 Structure of the Learned Transport Plan

To examine the correspondence learned on an actual LLM–VLM pair, Fig. 5 visualizes the row-normalized attention-Q transport plan from Layer 13. Because the full transport matrix is visually dense, we deterministically select one cross-index-dominant target row from each 128-index interval and retain its five largest transport entries. Same-index entries are omitted from the scatter plot, while the dashed diagonal is retained as a reference.

Among all 112 selected target neurons in this layer, 54.1% of the transport mass lies off the diagonal, 59.8% assign a majority of their mass to cross-index source neurons, and 24.1% have a cross-index neuron as their largest-mass correspondence.

This visualization complements the controlled permutation study: the latter verifies that OT can recover a known correspondence, whereas this analysis shows that the plan learned between the actual source LLM and target VLM contains substantial non-index-wise and distributed correspondence.

B.3 Analysis of Functional Correspondence

Cross-index correspondence is widespread and structured. Figure 6(a) examines the selected Q-projection neurons across all 36 layers. Averaged over the selected neurons, 54.26% of the OT mass is assigned to source neurons with different indices, while 10.50% of the dominant Top-1 matches are cross-index. The gap between these quantities shows that cross-index correspondence is often distributed over multiple source neurons even when the same-index neuron remains the largest individual match. Moreover, the off-diagonal mass remains substantial throughout the network rather than being confined to a particular layer.

Figure 6(b) shows that this structure also depends on projection type. The Q projection exhibits the largest off-diagonal mass and cross-index Top-1 ratio, reaching 54.26% and 10.50%, respectively. The K and O projections also contain substantial off-diagonal mass, whereas V exhibits a more index-aligned structure. Thus, TANGO does not indiscriminately rearrange neurons; the degree of cross-index correspondence varies systematically across projection groups.

Functional correspondence is non-local and soft. Figure 7(a) measures the index distance between each selected target neuron and its dominant cross-index source match. At Layer 13, the median absolute index displacement is 62, and 14.8% of the matches span more than 256 indices. At Layer 34, the median displacement increases to 324, with 50.0% of the matches spanning more than 256 indices. Cross-index correspondence therefore cannot be explained solely by small local shifts between neighboring neurons.

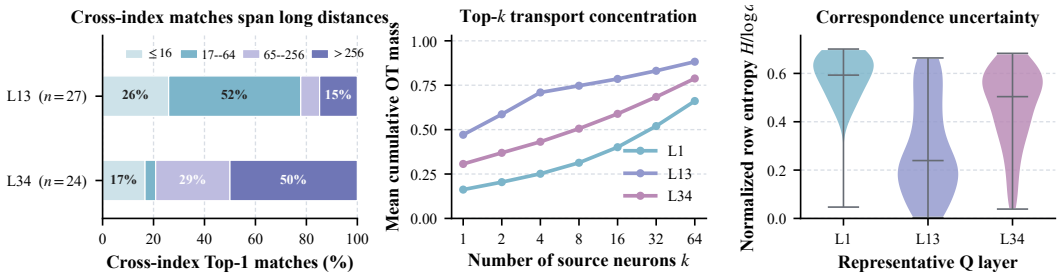

Figure 7: **Distance, concentration, and uncertainty of functional correspondence.** (a) Absolute index-distance distribution among cross-index Top-1 matches at representative Layers 13 and 34. A substantial fraction of the matches span hundreds of neuron indices, showing that functional correspondence is not restricted to neighboring neurons. Layer 1 is omitted from this panel because it contains only three cross-index Top-1 matches. (b) Mean cumulative row-normalized OT mass captured by the top-k source neurons at Layers 1, 13, and 34. (c) Distribution of normalized OT-row entropy at the same layers. Together, these results show that correspondence structure varies with depth and often requires soft transport over multiple source neurons.

Figure 7(b) and (c) characterize the softness of the transport plan. At Layer 13, the highest-mass source neuron accounts for 47.10% of the OT mass on average, while the top four source neurons jointly account for 70.93%. The concentration and entropy distributions also vary substantially with depth. This distributed correspondence supports soft OT-based transport: restricting each target neuron to a single source neuron would discard a non-negligible portion of its functional correspondence.

B.4 Qualitative Response Comparison

Aggregate benchmark scores do not reveal whether an improvement arises from better visual grounding or from more reliable downstream reasoning. We therefore examine representative examples from the MathVista testmini split using Qwen2.5-VL-7B. We select cases in which TANGO predicts the correct answer while both the unmodified base model and SNRF fail, covering visual counting, geometry, and table reasoning. The response excerpts are shortened only for readability; their reasoning conclusions and final answers remain unchanged.

The examples expose three characteristic failure modes. In the visual-counting case, the base model returns an unsupported answer, while SNRF undercounts the objects. TANGO enumerates all nine objects before performing the requested subtraction. In the geometry case, both baselines incorrectly treat the labeled angles as supplementary and solve $x + 2x = 180°$, whereas TANGO recognizes that the two labeled angles and the square's right angle together form a straight angle. In the table-reasoning case, the baselines include values above the specified upper bound of 81, while TANGO correctly enforces $60 \leq n < 81$.

These examples qualitatively indicate that the gains from TANGO are reflected not only in final-answer accuracy, but also in more complete visual evidence and reasoning that is consistent with the problem constraints. This analysis complements, rather than replaces, the aggregate evaluation in the main paper.

B.5 Cross-Model Results on MMMU

To examine whether the effectiveness of TANGO generalizes across model families and scales, we compare all methods on MMMU under four target-model settings. As shown in Table 6, TANGO consistently achieves the best performance, improving over both the corresponding base models and the strongest competing merging methods. These results indicate that the benefit of functional neuron alignment is not specific to a particular architecture or model scale.

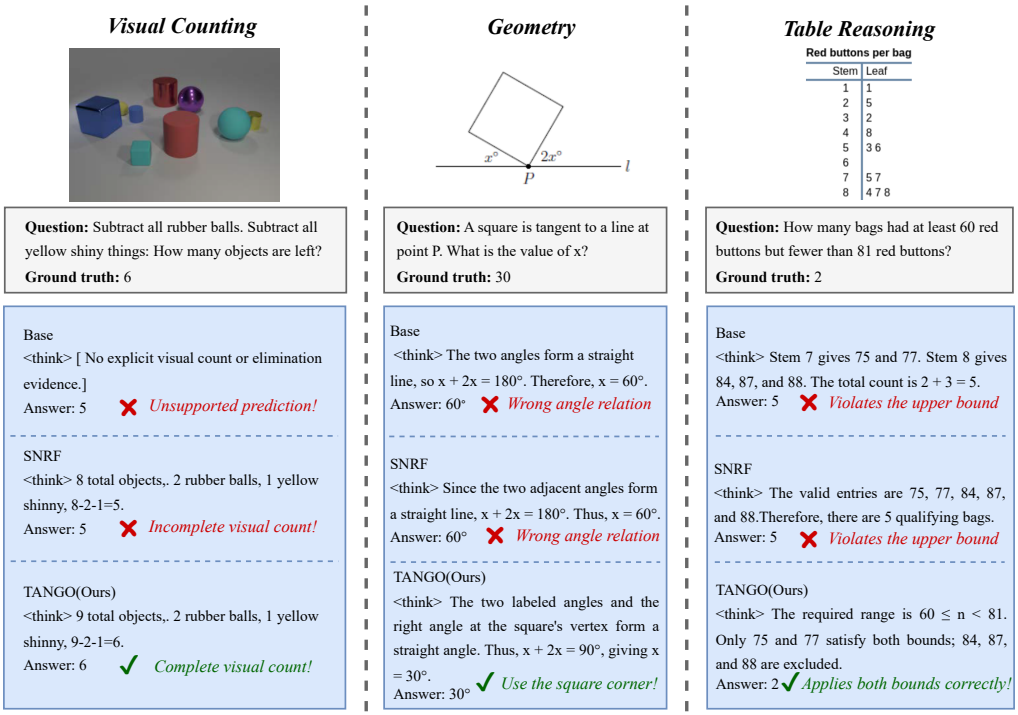

Figure 8: **Qualitative comparison on MathVista.** We compare the responses of the Qwen2.5-VL-7B base model, SNRF, and TANGO on three representative problems involving visual counting, geometric reasoning, and table interpretation. Each case presents the input image, question, ground-truth answer, abbreviated reasoning response, and final prediction. The base model and SNRF fail because of an incomplete visual count, an incorrect geometric relation, or a violation of the specified numerical range. In contrast, TANGO identifies the complete visual evidence and applies the appropriate reasoning constraints, producing the correct answer in all three cases. Reasoning responses are shortened only for readability, while the reasoning conclusions and final answers remain unchanged. Red crosses and green check marks denote incorrect and correct predictions, respectively.

Table 6: **Cross-model comparison on MMMU.** We report accuracy (%) across four target-model settings. Higher is better. The best result in each column is shown in **bold**, and the second-best result is underlined.

Method	Qwen2.5-VL-7B	InternVL2.5-4B	Qwen3-VL-8B	InternVL3.5-8B
Base	50.56	49.67	53.89	56.33
TA	50.22	50.00	54.00	57.22
DARE	50.33	49.33	54.00	57.11
BR2V	48.78	50.26	53.44	57.89
L2V-CoT	51.11	49.44	53.89	56.33
SNRF	50.02	50.11	54.00	57.33
TANGO	**51.78**	**50.44**	**54.67**	**58.22**

B.6 Additional Results on PaliGemma 2-10B

We further evaluate the model-merging methods using PaliGemma 2-10B (Steiner et al., 2024). As shown in Table 7, TANGO achieves the best overall performance across the evaluated visual and reasoning benchmarks.

Table 7: **Additional results on PaliGemma 2-10B.** We compare TANGO with representative model-merging baselines across multimodal perception and reasoning benchmarks. Higher is better for all metrics. The best result is shown in **bold**, and the second-best result is underlined.

Method	MathVista-CoT	MathVerse-TD	MME	MathVision	MMStar	ScienceQA	MMMU	MMMU-Pro
Base	33.3	16.50	1312.37	13.82	55.79	94.60	36.89	7.75
TA	32.9	17.39	1327.43	14.80	**56.25**	94.60	37.33	8.38
DARE	32.8	17.01	<u>1330.55</u>	14.80	56.00	94.70	37.00	8.04
BR2V	<u>34.0</u>	<u>18.78</u>	1313.29	14.47	<u>56.14</u>	<u>95.04</u>	<u>37.56</u>	<u>8.44</u>
L2V-CoT	33.4	16.62	1321.43	14.47	55.68	93.85	37.00	7.57
SNRF	<u>34.0</u>	17.51	1289.91	<u>14.99</u>	54.83	93.16	37.33	7.80
TANGO	**34.6**	**19.02**	**1347.81**	**15.13**	**56.25**	**95.14**	**38.56**	**8.50**