LAYER-SPECIALIZED MODEL MERGING FOR TRAINING-FREE CROSS-MODAL REASONING TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-only reasoning LLMs are strong multi-step problem solvers, but transferring this behavior to multimodal LLMs (MLLMs) usually requires costly multimodal post-training and scarce verifiable reasoning supervision. We study a training-free alternative: fusing a text-only reasoning decoder with a homologous MLLM decoder while keeping the MLLM visual encoder and projector unchanged. This setting is more delicate than standard model merging because the fused decoder must integrate reasoning behavior from a text-only model while continuing to interpret visual inputs. We introduce ANO (**A**ttention-calibrated **N**orm-**O**rthogonal merging), a layer-specialized fusion method for cross-modal reasoning transfer. ANO represents the two decoders as task vectors relative to a shared base model and computes closed-form, label-free fusion weights from task-vector salience. To account for depth-dependent visual-token use, it calibrates these weights with the MLLM's layer-wise visual-attention decay, using a continuous prior that favors the vision branch in shallow layers and the reasoning branch in deeper layers. Across homologous MLLM–reasoner pairs, ANO improves reasoning-heavy multimodal benchmarks without gradient updates or labeled multimodal reasoning data. We also evaluate the boundary of this transfer: MME-RW perception results and MathVista performance reveal task-dependent tradeoffs between reasoning gains and visual capabilities. Together, the results suggest that layer-specialized merging is a practical route for training-free cross-modal reasoning transfer within the studied homologous model families.

1 INTRODUCTION

Text-only reasoning models have become increasingly capable of solving problems that require multiple steps of inference (Guo et al., 2025; Lightman et al., 2024; Guha et al., 2025). Bringing these advances to multimodal large language models (MLLMs) could help them reason about the images, diagrams, and visual evidence they encounter. A common route is additional multimodal post-training, which requires suitable supervision and optimization (Xu et al., 2024; Dong et al., 2025; Huang et al., 2025; Zhang et al., 2025a). Yet reasoning capabilities are also available in existing text-only checkpoints. For an MLLM with a compatible language decoder, this creates an opportunity to reuse an independently trained reasoner rather than repeat a multimodal training stage. We study how to make this form of training-free transfer effective.

Model merging provides a parameter-space route for combining compatible checkpoints or their task vectors (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024). In our setting, the MLLM and the text-only reasoner share a homologous decoder, so their parameters can be combined while retaining the MLLM's visual encoder and projector. Their contributions, however, are not interchangeable. The MLLM decoder has been adapted to interpret visual inputs, whereas the reasoning donor has been adapted through text. Updating their shared decoder therefore changes not only reasoning behavior, but also how the multimodal model uses visual information. Parameter compatibility makes merging possible; choosing how the two adaptations should contribute remains the central design question.

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.

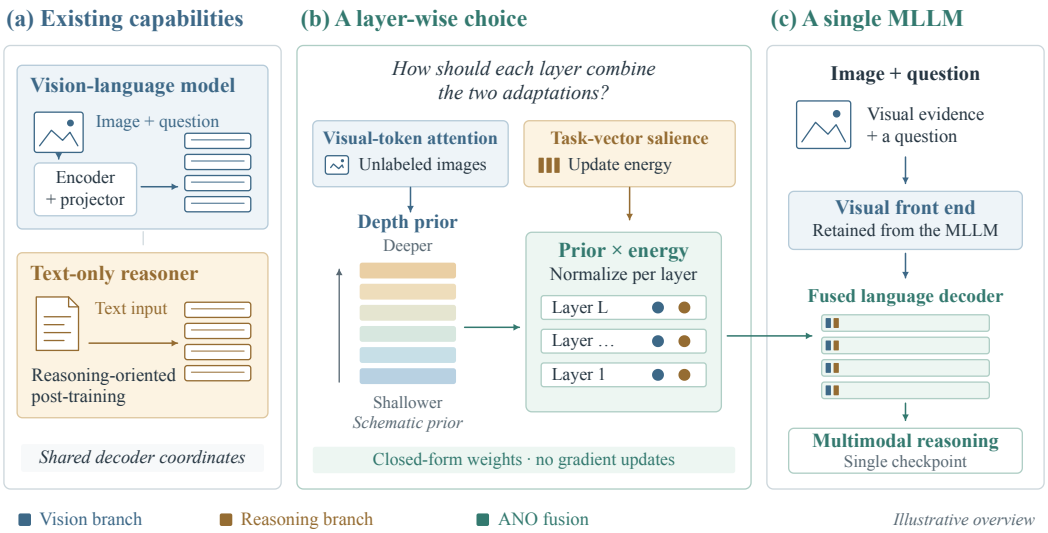


Figure 1: **Compatible parameters, different modality roles.** **(a)** A vision-language model and a text-only reasoner provide compatible decoder adaptations with different input histories. **(b)** ANO combines a visual-attention-derived depth prior with task-vector salience to determine layer-wise fusion weights. **(c)** The fused decoder retains the MLLM's visual encoder and projector in a single multimodal model. Layer blocks and prior colors are schematic: they are neither measured fusion weights nor an exclusive division of visual and reasoning functions.

Figure 1 illustrates this distinction. The two branches offer compatible parameters but bring different input histories and adaptation goals. A global interpolation coefficient specifies their overall balance, while task-vector magnitudes describe the scale of their parameter changes. Neither signal explicitly describes how the target MLLM uses visual tokens across decoder depth. Layer-wise visual attention provides a complementary, model-derived signal: in the models examined, direct attention to visual tokens is concentrated toward shallow layers, consistent with earlier observations of depth-dependent visual-token use (Chen et al., 2024a). This motivates a continuous modality prior for fusion. The prior guides the relative emphasis on the two branches without assigning each layer an exclusively visual or reasoning function.

We introduce ANO, a layer-specialized merging method that combines this modality prior with task-vector salience. ANO represents the two adaptations relative to a shared base decoder and uses their squared norms together with an attention-calibrated prior to compute closed-form fusion weights for each layer. The prior is fitted using unlabeled images; no gradient updates or labeled downstream examples are needed for fusion. As shown in Figure 1b,c, model-derived calibration determines the prior, while the MLLM's visual front end is carried into a single fused multimodal model. This design makes modality information part of the weight construction rather than relying on parameter magnitudes alone.

We evaluate ANO on six homologous MLLM–reasoner pairs across five multimodal reasoning benchmarks. The average improves over the corresponding source MLLM for every pair in this study. Matched-checkpoint comparisons assess the fusion rule against alternative merging configurations, while no-prior and schedule ablations examine the contribution of the modality prior. Perception, grounding, and captioning evaluations further characterize the cost of transfer: the aim is to improve multimodal reasoning while mitigating visual degradation, not to guarantee that every capability remains unchanged. The observed MathVista regressions on two pairs make this boundary explicit. Together, these results evaluate both the practical benefit of layer-specialized fusion and the conditions under which it should be used.

Our contributions are:

- We introduce **ANO**, a training-free, layer-wise fusion rule that combines task-vector salience with a visual-attention-calibrated modality prior.

- We evaluate **cross-modal reasoning transfer** across six homologous model pairs, with matched merging comparisons and ablations of the weight construction.
- We analyze **the behavior and boundaries of transfer** through calibration and geometry diagnostics, direct visual-capability measurements, and task-level failures.

2 Related Work

Multimodal reasoning and post-training. Text-only reasoning LLMs have improved through process supervision, reinforcement-learning-style post-training, and reasoning-data recipes (Lightman et al., 2024; OpenAI, 2024; Guo et al., 2025; Team et al., 2025b; Ma et al., 2025; Guha et al., 2025). MLLMs connect visual encoders to language decoders for image understanding and instruction following (Wang et al., 2023; Liu et al., 2023; Sun et al., 2024; Bai et al., 2023; 2025; Li et al., 2024; Wang et al., 2025; Liu et al., 2024a;b; Lin et al., 2024; Chen et al., 2024b; Wang et al., 2024b; Team et al., 2025a). Multimodal reasoning is often elicited by prompting (Wei et al., 2022; Mitra et al., 2024; Luan et al., 2024; Zhang et al., 2024) or improved through supervised traces, search, reflection, and reinforcement learning (Xu et al., 2024; Dong et al., 2025; Yao et al., 2024; Liu et al., 2025; Huang et al., 2025; Peng et al., 2025; Zhang et al., 2025a; Meng et al., 2025; Wei et al., 2025a). These methods either add inference-time scaffolding or require multimodal reasoning data and optimization. ANO instead transfers reasoning behavior from an existing text-only reasoner into a homologous MLLM without labeled multimodal reasoning supervision or gradient updates.

Model merging and task arithmetic. Model soups and task arithmetic combine separately adapted models or task vectors without increasing inference-time cost (Wortsman et al., 2022; Ilharco et al., 2023; Zhang et al., 2023). Later methods reduce interference with Fisher weighting, sign agreement, sparsification, rescaling, dataless fusion, or task-specific arithmetic (Matena & Raffel, 2022; Yadav et al., 2023; Yu et al., 2024; Jin et al., 2023; Zhou et al., 2024). Most of these methods merge comparable task adaptations in a shared parameter space. Our setting is asymmetric: the MLLM branch must keep visual grounding, while the text-only branch contributes reasoning behavior. This asymmetry motivates depth-dependent fusion rather than a single global coefficient.

MLLM merging and local-surrogate methods. Recent MLLM merging work studies perception and reasoning behavior through parameter fusion, optimizes merge coefficients for heterogeneous multimodal models, or merges multimodal capabilities directly (Chen et al., 2025; Du et al., 2025; Wei et al., 2025b). Visual-token use in MLLM decoders also changes sharply with depth (Chen et al., 2024a), which motivates our attention-derived layer prior. Data-free merging methods such as DOGE and WUDI use task-vector geometry or local surrogate objectives to reduce interference (Wei et al., 2025c; Cheng et al., 2025), echoing local-linearization intuitions from NTK-style analyses (Jacot et al., 2018; Zhou et al., 2024; Touvron et al., 2023; Zhong et al., 2024). We use this view only as a local approximation and check it with orthogonality, interpolation, and prior-sensitivity diagnostics.

3 Method

ANO combines compatible vision and reasoning adaptations using layer-wise weights informed by task-vector energy and visual-token use. It first calibrates a modality prior on the source MLLM using unlabeled images, then computes fusion coefficients and assembles a single multimodal checkpoint. Figure 2 summarizes the construction. We introduce the fusion setup, the prior, and the normalized coefficient rule in that order.

3.1 Overview and Fusion Setup

Let θ_0 denote the shared base decoder, θ_V the MLLM decoder, and θ_R the text-only reasoning decoder. Merged parameters require compatible layer structure, tensor correspondence, and parameter semantics. We retain the MLLM's visual encoder, projector, tokenizer, and image preprocessing. Embeddings and output-head parameters are fused only when their parameter and tokenizer mappings are compatible; detailed checks appear in Appendix A.

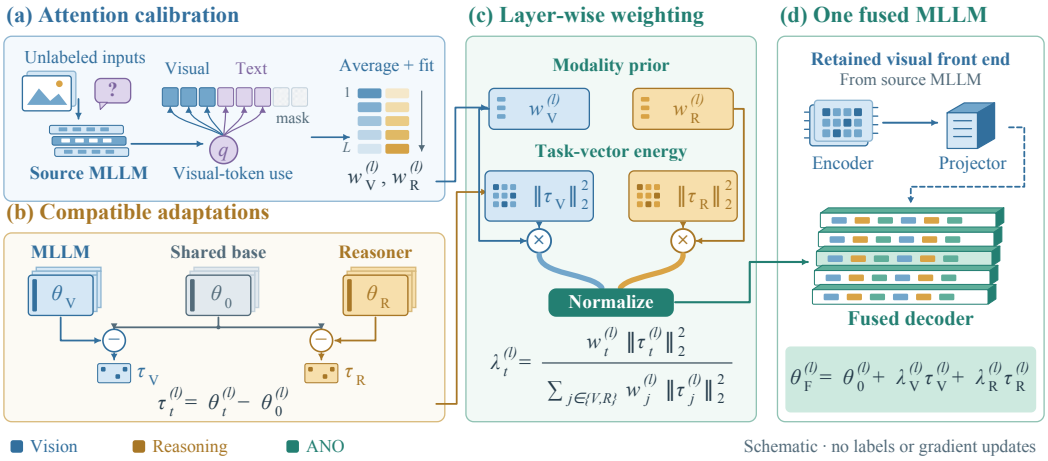


Figure 2: **ANO overview. (a)** Unlabeled images and prompts are processed by the source MLLM. Generated-query attention is averaged over heads, queries, and examples; fitting and mapping across depth yields the modality priors. **(b)** Compatible vision and reasoning task vectors use the same base. **(c)** At each layer, prior-weighted task-vector energies are normalized across $t \in \{V, R\}$. **(d)** The weighted adaptations construct the fused decoder, which is combined with the retained visual encoder and projector. Token links, prior colors, parameter tiles, and layer counts are schematic, not empirical values; compact energy icons omit the layer superscript. The equations summarize the nondegenerate construction in Section 3.

For decoder block $l \in \{1, \dots, L\}$, $\theta_t^{(l)}$ denotes the selected block tensors flattened into one vector. The branch adaptations and fused block are

$$\tau_t^{(l)} = \theta_t^{(l)} - \theta_0^{(l)}, \quad t \in \{V, R\}, \qquad \theta_F^{(l)} = \theta_0^{(l)} + \lambda_V^{(l)} \tau_V^{(l)} + \lambda_R^{(l)} \tau_R^{(l)}, \tag{1}$$

where $\lambda_V^{(l)}, \lambda_R^{(l)} \geq 0$ and $\lambda_V^{(l)} + \lambda_R^{(l)} = 1$. Thus the central choice is a pair of fusion coefficients for each layer. ANO determines them from two complementary quantities: a modality prior describing the intended depth-dependent branch preference, and the squared norm of each branch's task vector describing its adaptation scale.

3.2 Calibrating a Modality Prior

We obtain the modality signal from the *source MLLM*, before fusion. For a calibration input, let $\mathcal{V}$ index visual keys and $\mathcal{T}_q$ index nonvisual keys visible to generated query q. With attention probability $A_{l,h,qk}$ at layer l and head h, the visual-attention ratio and its average are

$$r_{l,h,q} = \frac{\sum_{k \in \mathcal{V}} A_{l,h,qk}}{\sum_{k \in \mathcal{V} \cup \mathcal{T}_q} A_{l,h,qk}}, \qquad a_l = \mathbb{E}_{h,q,x}[r_{l,h,q}]. \tag{2}$$

The average is taken over heads, generated query positions, and calibration examples. It measures direct attention to visual tokens, rather than all possible uses of visual information within the decoder.

Using unlabeled MSCOCO images (Lin et al., 2014), we fit $a_l \approx C \exp(-\hat{\alpha} l)$ by least squares on $\{(l, \log(a_l + \epsilon))\}_{l=1}^{L}$, including an intercept and a small fixed ϵ for numerical stability. The exponential summarizes the shallow-heavy trend observed in the source models; it is used as a smooth prior, not as a hard division of visual and reasoning layers. We use the fitted slope rather than the amplitude, which can depend on tokenization, image-token count, and prompt format.

For a positive fitted decay and $L > 1$, the normalized profile and complementary branch priors are

$$\tilde{a}_l = \exp[-\hat{\alpha}(l-1)], \qquad w_V^{(l)} = \epsilon_w + (1 - 2\epsilon_w) \frac{\tilde{a}_l - \tilde{a}_L}{\tilde{a}_1 - \tilde{a}_L}, \qquad w_R^{(l)} = 1 - w_V^{(l)}, \tag{3}$$

where $0 < \epsilon_w < 1/2$. The visual prior decreases from $w_V^{(1)} = 1 - \epsilon_w$ to $w_V^{(L)} = \epsilon_w$, while the reasoning prior increases. Both remain positive, so the construction expresses a continuous preference rather than excluding either branch at a layer boundary. The measured attention statistics and fit diagnostics are reported separately in Section 4.4.

3.3 MODALITY-CALIBRATED LAYER-WISE FUSION

Task-vector magnitude and the modality prior play different roles. Magnitude supplies a label-free proxy for the extent of adaptation within a block; the prior incorporates the target MLLM's visual-attention depth pattern. We combine them as a branch salience and normalize within each layer:

$$s_t^{(l)} = w_t^{(l)} \|\tau_t^{(l)}\|_2^2, \qquad \lambda_t^{(l)} = \frac{w_t^{(l)} \|\tau_t^{(l)}\|_2^2}{w_V^{(l)} \|\tau_V^{(l)}\|_2^2 + w_R^{(l)} \|\tau_R^{(l)}\|_2^2}. \tag{4}$$

For positive total salience, these coefficients are nonnegative and sum to one. Although the prior is monotone across depth, the final weights also depend on the two task-vector energies and need not be monotone. The prior therefore guides, rather than completely determines, the allocation.

Coefficient-space interpretation. For fixed saliences, consider the objective

$$J^{(l)}(\lambda_V, \lambda_R) = s_V^{(l)}(1-\lambda_V)^2 + s_R^{(l)}(1-\lambda_R)^2, \quad \lambda_V + \lambda_R = 1. \tag{5}$$

Each branch favors retaining coefficient one on its own adaptation, while the sum-to-one constraint balances the two preferences. The salience weights the corresponding penalty. For $s_V^{(l)} + s_R^{(l)} > 0$, the unique minimizer is

$$\lambda_t^{(l)} = \frac{s_t^{(l)}}{s_V^{(l)} + s_R^{(l)}}. \tag{6}$$

This gives Eq. 4 after substituting the calibrated saliences. The closed form is exact for this specified coefficient objective; choosing task-vector energy and the modality prior is the empirical design evaluated by the ablations. Appendix B gives the proof and distinguishes this objective from the original model losses. The orthogonality and interpolation measurements characterize the evaluated checkpoints; they are not additional algebraic requirements for solving this objective.

Magnitude-only variant. Removing the modality prior gives

$$s_t^{(l)} = \|\tau_t^{(l)}\|_2^2, \qquad \lambda_t^{(l)} = \frac{\|\tau_t^{(l)}\|_2^2}{\|\tau_V^{(l)}\|_2^2 + \|\tau_R^{(l)}\|_2^2}. \tag{7}$$

This is the no-prior comparison used to examine the contribution of modality information beyond adaptation magnitude. With the normalized coefficients, the fused block can equivalently be written as $\theta_F^{(l)} = \lambda_V^{(l)}\theta_V^{(l)} + \lambda_R^{(l)}\theta_R^{(l)}$. The base cancels from this interpolation once the weights are fixed, but still affects the task-vector energies used to calculate them.

3.4 CONSTRUCTING THE FUSED MODEL

The construction has two stages: calibration on the source MLLM, followed by parameter arithmetic on the compatible decoder blocks. After estimating the prior, we compute each block's task vectors, energies, and normalized coefficients, apply Eq. 1, and combine the fused decoder with the original visual front end. No source-model loss, labeled downstream validation set, or gradient update is used to choose the coefficients.

Algorithm 1 ANO construction for the nondegenerate setting

Require: Compatible decoders $\theta_0, \theta_V, \theta_R$; source visual modules; unlabeled calibration inputs

1: Measure source-MLLM visual-attention ratios and average them using Eq. 2.
2: Fit $\hat{\alpha}$ and construct the modality priors using Eq. 3.
3: **for** each compatible decoder block l **do**
4: Form $\tau_V^{(l)}$ and $\tau_R^{(l)}$ relative to $\theta_0^{(l)}$.
5: Compute calibrated saliences and normalized weights using Eq. 4.
6: Construct $\theta_F^{(l)}$ using Eq. 1.
7: **end for**
8: Assemble the fused decoder with the retained MLLM visual encoder and projector.
9: **return** One fused multimodal checkpoint.

Algorithm 1 summarizes the nondegenerate procedure. Zero-energy blocks, flat or nondecaying fits, and embedding/head treatment are discussed in Appendices B and A.

Apart from attention calibration, fusion arithmetic is linear in the number of merged decoder parameters. It produces one checkpoint with the source MLLM's architecture; generation length can still affect end-to-end latency. Visual retention is evaluated empirically rather than guaranteed by the coefficients.

4 EXPERIMENTS

We evaluate checkpoint reuse through four questions: does it improve multimodal reasoning, what does ANO add beyond generic merging, how much does attention calibration contribute, and what visual-capability trade-offs accompany the gains? The evaluation follows this order, separating the breadth of transfer from controlled comparisons of the fusion rule.

4.1 EXPERIMENTAL SETUP

Models and benchmarks. We study six compatible MLLM–reasoner pairs spanning Llama- and Qwen-derived decoders, with MLLM sizes from 4B to 38B (Table 4). Each comparison uses a fixed source MLLM, reasoning donor, and shared base; the unfused source MLLM is the reference. The reasoning suite comprises MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b), MathVista (Lu et al., 2024), MathVision (Wang et al., 2024a), and WeMath (Qiao et al., 2024). MME-RW (Zhang et al., 2025b) evaluates visual perception; RefCOCO (Yu et al., 2016) and COCO Caption (Lin et al., 2014) add direct grounding and description measurements, with CIDEr (Vedantam et al., 2015) for captioning. These visual scores are reported separately rather than averaged with reasoning accuracy.

Fusion and comparison protocol. ANO uses unlabeled calibration images and parameter arithmetic, without gradient updates or labeled downstream examples for coefficient selection. The controlled 4B setting pairs InternVL3.5-4B-MPO (abbreviated InternVL3.5-4B) with Qwen3-4B-Thinking. The merging comparisons hold the source checkpoints fixed to examine how the fusion rule combines the available adaptations. Appendices A and A.4 describe checkpoint compatibility, evaluation conditions, and the baseline families.

Tables 1 and 2 present benchmark point estimates. We report reasoning accuracy, run-to-run variation, and visual-task scores separately: a gain in average reasoning performance does not imply improvement on every benchmark or preservation of every visual capability. Three-run means and SDs appear in Appendix C.2.

4.2 REASONING TRANSFER ACROSS COMPATIBLE MODEL PAIRS

Table 1 reports a positive five-benchmark average change for all six pairs, ranging from $+1.1$ for Idefics3-8B to $+7.0$ for InternVL3.5-4B. The reported gains also include $+4.9$ for NVILA-15B and $+6.3$ for InternVL2.5-38B, alongside $+2.0$ for Qwen2.5-VL-7B and $+4.3$ for LLaVA-OneVision1.5-4B. The benefit therefore appears in multiple compatible constructions and scales, rather than only the InternVL3.5-4B pair. Among the two Qwen3-based 4B pairs, LLaVA-OneVision1.5 gains most on MMMU, whereas InternVL3.5 gains most on WeMath, illustrating different task-level benefit profiles within the same donor family.

MMMU, MMMU-Pro, and MathVision improve in every listed pair. The outcome is nevertheless task-dependent: MathVista decreases by 7.7 points for Idefics3 and 12.2 for NVILA, and Idefics3 also loses 0.6 on WeMath. These regressions remain visible in the main table and the complete differences in Appendix C. The result supports improved average reasoning-benchmark performance in the evaluated compatible pairs, not a uniformly improved model on every task.

Run-to-run variation. The three-run summaries give average scores of 34.5 ± 0.08 for Idefics3 and 45.6 ± 0.17 for Qwen2.5-VL (mean $\pm$ SD; Table 7 in Appendix C.2). These estimates characterize variation across ANO runs and are distinct from the point estimates in Table 1. They complement the transfer comparison but are not paired significance tests of the gains over the source MLLM.

Table 1: **Transfer across six compatible pairs.** Bold marks the higher score within each pair; ΔAvg. is relative to the source. Avg.: five-benchmark average. Scores and gains are rounded independently to one decimal place. Three-run statistics: Appendix C.2.

Method	MMMU	MMMU-Pro	MathVista	MathVision	WeMath	Avg.	ΔAvg.
Idefics3-8B / DeepSeek-R1-Distill-Llama-8B							
Source	43.9	32.6	**58.4**	20.1	**12.3**	33.5	0.0
ANO	**48.3**	**34.7**	50.7	**27.6**	11.7	**34.6**	+1.1
NVILA-15B / DeepSeek-R1-Distill-Qwen-14B							
Source	53.2	36.2	**67.6**	23.2	31.1	42.3	0.0
ANO	**61.3**	**49.4**	55.4	**37.2**	**32.3**	**47.1**	+4.9
InternVL2.5-38B / QwQ-32B							
Source	63.9	48.0	71.9	32.2	38.3	50.9	0.0
ANO	**69.2**	**56.8**	**73.1**	**39.7**	**47.0**	**57.2**	+6.3
Qwen2.5-VL-7B / OpenThought3-7B							
Source	53.9	36.3	67.8	25.4	35.2	43.7	0.0
ANO	**55.0**	**38.9**	**69.0**	**27.5**	**38.0**	**45.7**	+2.0
LLaVA-OneVision1.5-4B / Qwen3-4B-Thinking							
Source	52.7	35.3	67.6	24.2	36.7	43.3	0.0
ANO	**62.9**	**38.7**	**69.2**	**27.8**	**39.4**	**47.6**	+4.3
InternVL3.5-4B-MPO / Qwen3-4B-Thinking							
Source	65.4	60.1	71.7	48.0	39.8	57.0	0.0
ANO	**69.6**	**65.7**	**77.6**	**55.1**	**51.9**	**64.0**	+7.0

Table 2: **Matched-checkpoint merging comparisons.** Bold marks the best score in each block. Baseline settings and comparison scope: Section 4.3 and Appendix A.4. Each block uses the same source checkpoints. ANO w/o prior: magnitude-only fusion. Repeated-run statistics are separate (Appendix C.2).

Method	MMMU	MMMU-Pro	MathVista	MathVision	WeMath	Avg.
(a) InternVL3.5-4B-MPO / Qwen3-4B-Thinking						
Source MLLM	65.4	60.1	71.7	48.0	39.8	57.0
Weight averaging	67.1	60.5	76.8	54.1	50.4	61.8
Task arithmetic	67.4	60.7	76.6	54.4	50.6	61.9
TIES	67.8	61.0	76.4	54.6	50.7	62.1
DARE	68.0	61.2	76.2	54.9	50.8	62.2
SLERP	67.4	60.7	76.9	54.1	50.3	61.9
OptMerge	51.2	48.7	61.7	42.8	39.6	48.8
WUDI	51.8	48.9	62.4	43.1	40.2	49.3
DOGE	52.4	49.1	63.0	43.7	40.8	49.8
ANO w/o prior	66.6	63.1	76.9	54.9	50.3	62.4
ANO	**69.6**	**65.7**	**77.6**	**55.1**	**51.9**	**64.0**
(b) Qwen2.5-VL-7B / OpenThought3-7B						
Source MLLM	53.9	36.3	67.8	25.4	35.2	43.7
DARE	54.2	36.7	68.4	25.9	36.0	44.2
TIES	54.0	36.4	67.9	25.6	35.5	43.9
ANO	**55.0**	**38.9**	**69.0**	**27.5**	**38.0**	**45.7**

4.3 Comparison with Model-Merging Baselines

Holding the source checkpoints fixed separates the value of the donor from the choice of fusion rule. On the InternVL3.5-4B pair, weight averaging, task arithmetic, TIES (Yadav et al., 2023), and DARE (Yu et al., 2024) reach averages of 61.8–62.2, versus 57.0 for the source MLLM (Table 2). Ordinary merging already accesses useful information from the reasoning-trained donor. ANO reaches 64.0, adding 1.8 points over the strongest of these generic configurations, DARE. Its contribution is thus an improvement in how the available adaptations are combined, not the claim that generic merging cannot transfer them.

Table 2a also includes SLERP at 61.9 on the 4B pair. In panel b, on Qwen2.5-VL-7B, ANO reaches 45.7 against 44.2 for DARE and 43.9 for TIES. The advantage over these generic merges therefore also appears in the second controlled pair. These head-to-head comparisons cover two pairs; the six-pair study instead evaluates transfer relative to each source MLLM.

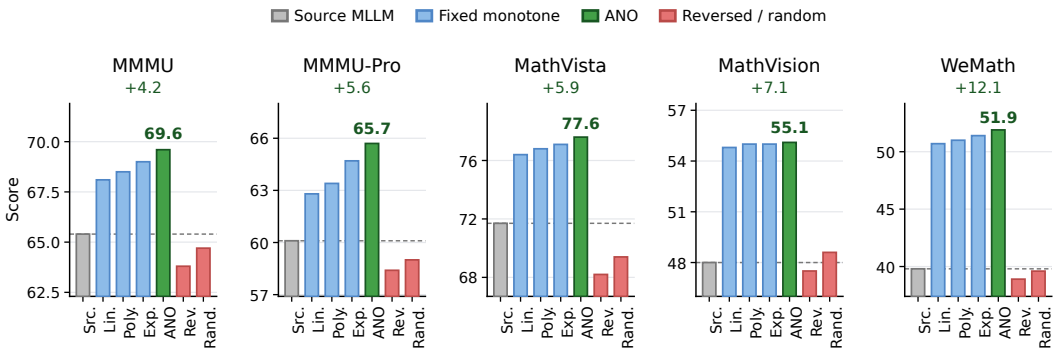


Figure 3: Contribution of the modality prior on the InternVL3.5-4B-MPO / Qwen3-4B-Thinking pair. Src. and the dashed line denote the source MLLM. Lin./Poly./Exp.: linear, polynomial, and MMMU-tuned exponential priors; Rev./Rand.: reversed and randomized priors. Exp. uses one coefficient selected on MMMU over $[0.1, 0.9]$ and shared across all five tasks. Green bars denote ANO. Values are point estimates without run uncertainty; annotations are gains over the source, not the strongest alternative schedule. Magnitude-only results appear in Table 2.

The benchmark-level comparison is consistent across these two pairs: ANO exceeds DARE on all five tasks in each. For InternVL3.5-4B, the margin is largest on MMMU-Pro ($+4.5$ points), while MathVision differs by only 0.2. For Qwen2.5-VL-7B, the largest margins are on MMMU-Pro ($+2.2$) and WeMath ($+2.0$). Thus the aggregate advantage reflects gains across the suite, with task-dependent magnitudes, rather than one isolated benchmark improvement.

OptMerge (Wei et al., 2025b), WUDI (Cheng et al., 2025), and DOGE (Wei et al., 2025c) have lower scores than the source under the evaluated configurations. These outcomes do not establish that their method families cannot work in cross-modal transfer. Our comparison covers the listed implementations and settings, not an exhaustive search of each method family (Appendix A.4). These comparisons assess transfer of the donor's full adaptation, rather than isolating reasoning from changes in language behavior, instruction following, answer formatting, or generation length.

4.4 Contribution of Attention Calibration

Beyond task-vector magnitude. The magnitude-only variant uses the same source checkpoints but removes the modality prior from the coefficient rule. Its average of 62.4 rises to 64.0 with the calibrated prior (Table 2). The 1.6-point difference directly tests the design introduced in Figure 2: adaptation scale alone does not recover the full result in this setting.

Beyond a generic depth preference. Figure 3 compares the calibrated prior with linear, polynomial, MMMU-tuned exponential, reversed, and randomized schedules on the InternVL3.5-4B pair. For the exponential prior, we sweep the decay coefficient over $[0.1, 0.9]$, select the candidate with the highest MMMU score, and hold that single coefficient fixed across all five benchmarks without task-specific retuning. ANO instead estimates its decay coefficient from the source MLLM's visual-attention profile using unlabeled calibration images, without downstream performance-based selection.

ANO has the highest point estimate in each panel, exceeding the MMMU-tuned exponential prior by approximately 0.5 points on average. The comparison therefore evaluates model-derived calibration against a shared prior selected using downstream benchmark performance. All three monotone alternatives already outperform the source MLLM across the five tasks, whereas the reversed schedule falls below it throughout. These results distinguish the benefit of a depth preference from the further improvement obtained through model-derived calibration; they do not establish statistical significance or a universally optimal schedule.

Calibration signal and sensitivity. The source-MLLM attention measurements show shallow-heavy direct visual-token use in three representative model pairs (Figure 4), consistent with earlier

observations of depth-dependent token use (Chen et al., 2024a). This supplies a measured signal for the smooth prior rather than a hard partition of visual and reasoning layers. For the InternVL3.5-4B calibration-source comparison, MSCOCO, ChartQA, and ScreenSpot yield slopes of 0.108–0.115 and MMMU scores of 69.3–69.6 (Table 8). The small variation in this three-source check complements the schedule ablation; it is not a cross-family robustness guarantee. Appendix D collects the attention plots, calibration records, and separate geometry diagnostics.

4.5 Reasoning–Visual Capability Trade-offs

Reasoning gains are only one consequence of changing the multimodal decoder. We therefore compare visual scores both with the source MLLM, which measures the cost of transfer, and with alternative merges, which tests whether the calibrated rule mitigates that cost.

For the 4B MME-RW comparison, DARE decreases the source score by 4.4 points, ANO w/o decay by 3.0, and ANO by 1.7 (Table 3). The 7B results show the same ordering, with drops of 4.0, 2.3, and 1.1. Thus the calibrated construction reduces the score loss relative to both DARE and the w/o decay setting in these comparisons. This evidence concerns the relative cost of the tested merges, not lossless visual preservation. The w/o decay setting is treated separately from the magnitude-only reasoning ablation.

Table 3: **MME-RW score changes after fusion.** Bold: best merged variant; Δ: change from source. 4B: InternVL3.5-4B; 7B: Qwen2.5-VL-7B / OpenThought3-7B. Domain results: Appendix E.

	4B		7B	
Method	Score	Δ	Score	Δ
Source MLLM	56.7	0.0	57.4	0.0
DARE	52.3	-4.4	53.4	-4.0
ANO w/o decay	53.7	-3.0	55.1	-2.3
ANO	**55.0**	**-1.7**	**56.3**	**-1.1**

The finer-grained evaluations characterize what the aggregate omits. For InternVL3.5-4B, ANO is below the source in all five displayed MME-RW domains and above DARE in four; remote sensing is the exception. RefCOCO scores decrease by 0.8–1.6 points across three splits, and captioning decreases by 1.5 CIDEr. These direct-task results cover one model pair (Appendix E) and distinguish retained task performance from an unchanged visual encoder: modifying the decoder can still affect grounding and description. Alongside the MathVista regressions, they show why the practical objective is better reasoning transfer with an evaluated visual cost, rather than simultaneous improvement of every capability.

4.6 Limitations and Future Work

ANO uses shared-base, decoder-compatible checkpoints, following the standard setting of task-vector merging (e.g., task arithmetic, TIES, and DARE). Heterogeneous merging is being explored (Section 2), but reliable cross-architecture alignment remains challenging. We plan to investigate alignment-aware fusion to expand the range of reusable checkpoints. Beyond this paper's reasoning focus, the exploratory coding-donor results on Design2Code (Appendix F) suggest potential for multimodal code generation. Future work will explore other donor specializations, including domain knowledge, tool use, and multilingual capabilities.

5 Conclusion

ANO makes modality information part of the fusion rule for reusing a compatible text reasoner in an MLLM. Task-vector energy and a visual-attention-calibrated prior jointly determine layer-wise weights, while the original visual front end is retained in a single fused multimodal checkpoint. Across six compatible pairs, the reported average reasoning-benchmark score improves over the source MLLM. Matched-checkpoint comparisons and ablations support the incremental value of this construction beyond the evaluated generic merges and magnitude-only fusion. Visual evaluations characterize the accompanying cost and show reduced degradation relative to the tested fusion alternatives. With available compatible checkpoints, ANO enables reasoning transfer through unlabeled calibration and parameter arithmetic, without additional multimodal training.

AI Use Statement. We used LLM-based generative AI tools to assist with writing refinement and the creation of figures. All research ideas, methodology, experiments, analyses, and conclusions were developed and verified by the authors.

Reproducibility statement The fusion construction and coefficient derivation are specified in Section 3 and Appendix B. Model pairings and compatibility conditions appear in Appendix A; baseline and statistical protocols appear in Appendices A.4 and C.2. Reproduction depends on matching source checkpoints, parameter correspondences, calibration inputs, and evaluation settings. Point estimates and repeated-run summaries are distinguished throughout, and no claim of paired statistical significance is made from ANO run dispersion alone.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, volume 15139, pp. 19–35, Milan, Italy, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, volume 15075, pp. 370–387, Milan, Italy, 2024b.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9803–9817. PMLR, 2025. URL `https://proceedings.mlr.press/v267/chen25cm.html`.

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*, 2024c.

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 *ICML*, 2025.

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.

Yiyang Du, Xiaochen Wang, Chi Chen, Jiabo Ye, Yiru Wang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, Zhifang Sui, Maosong Sun, and Yang Liu. Adamms: Model merging for heterogeneous multimodal large language models with unsupervised coefficient optimization. In *CVPR*, pp. 9413–9422, Nashville, TN, USA, 2025.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

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

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural Tangent Kernel: Convergence and generalization in neural networks. In *NeurIPS*, pp. 8580–8589, Montreal, Canada, 2018.

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

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. In *Workshop on RBFM*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *TMLR*, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *ICLR*, 2024.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: on pre-training for visual language models. In *CVPR*, pp. 26679–26689, Seattle, WA, USA, 2024.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, volume 8693, pp. 740–755, Zurich, Switzerland, 2014.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26286–26296, Seattle, WA, USA, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024b.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024c.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

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*, Vienna, Austria, 2024.

Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li. TextCoT: Zoom in for enhanced multimodal text-rich image understanding. *arXiv preprint arXiv:2404.09797*, 2024.

Ruotian Ma, Peisong Wang, Cheng Liu, Xingyan Liu, Jiaqi Chen, Bang Zhang, Xin Zhou, Nan Du, and Jia Li. S ^{2} R: Teaching llms to self-verify and self-correct via reinforcement learning. *arXiv preprint arXiv:2502.12853*, 2025.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *NeurIPS*, New Orleans, LA, USA, 2022.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *CVPR*, pp. 14420–14431, Seattle, WA, USA, 2024.

OpenAI. OpenAI o1. https://openai.com/o1/, 2024.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. LMM-R1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

Qwen. Qwq. https://qwenlm.github.io/blog/qwq-32b-preview/, 2024.

Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2Code: Benchmarking multimodal code generation for automated front-end engineering. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3956–3974, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.199. URL https://aclanthology.org/2025.naacl-long.199/.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *ICLR*, Vienna, Austria, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025a.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi K1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025b.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087. URL https://doi.org/10.1109/CVPR.2015.7299087.

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*, Vancouver, BC, Canada, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

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.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *NeurIPS*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, New Orleans, LA, USA, 2022.

Lai Wei, Yuting Li, Kaipeng Zheng, Chen Wang, Yue Wang, Linghe Kong, Lichao Sun, and Weiran Huang. Advancing multimodal reasoning via reinforcement learning with cold start. *arXiv preprint arXiv:2505.22334*, 2025a.

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

Yongxian Wei, Anke Tang, Li Shen, Chun Yuan, and Xiaochun Cao. Modeling multi-task model merging as adaptive projective gradient descent. In *ICML*, 2025c.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, volume 162, pp. 23965–23998, 2022.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

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

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

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pp. 69–85, 2016. doi: 10.1007/978-3-319-46475-6_5. URL https://doi.org/10.1007/978-3-319-46475-6_5.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, pp. 9556–9567, Seattle, WA, USA, 2024a.

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. *arXiv preprint arXiv:2409.02813*, 2024b.

Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. Composing parameter-efficient modules with arithmetic operation. In *NeurIPS*, New Orleans, LA, USA, 2023.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025a.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024.

Yifan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *ICLR*, Singapore, 2025b.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In *NAACL*, pp. 2299–2314, Mexico City, Mexico, 2024.

Yuyan Zhou, Liang Song, Bingning Wang, and Weipeng Chen. Metagpt: Merging large language models using model exclusive task arithmetic. In *EMNLP*, pp. 1711–1724, Miami, FL, USA, 2024.

A Experimental Setup and Implementation

This appendix specifies the checkpoint construction, fusion procedure, and comparison scope underlying Section 4. Mathematical details follow in Appendix B; task-level results and diagnostics are organized in Appendices C–E.

A.1 Model Pairings and Compatible Parameters

Table 4 lists the six model constructions. Idefics3 (Laurençon et al., 2024) is paired with DeepSeek-R1-Distill-Llama (Guo et al., 2025); NVILA (Liu et al., 2024c) is paired with DeepSeek-R1-Distill-Qwen (Guo et al., 2025); and InternVL2.5 (Chen et al., 2024c) is paired with QwQ (Qwen, 2024). The two Qwen3-derived MLLMs are InternVL3.5 (Wang et al., 2025) and LLaVA-OneVision1.5 (Li et al., 2024). The remaining pair combines Qwen2.5-VL (Bai et al., 2025) with OpenThought3 (Guha et al., 2025).

Table 4: **MLLM, reasoning donor, and shared base.** The MLLM supplies the retained visual modules; the two decoder adaptations are measured relative to the shared base.

MLLM	Reasoning donor	Shared base
Idefics3-8B	DeepSeek-R1-Distill-Llama-8B	LLaMA3.1-8B
NVILA-15B	DeepSeek-R1-Distill-Qwen-14B	Qwen2.5-14B
InternVL2.5-38B	QwQ-32B	Qwen2.5-32B
InternVL3.5-4B-MPO	Qwen3-4B-Thinking	Qwen3-4B
LLaVA-OneVision1.5-4B	Qwen3-4B-Thinking	Qwen3-4B
Qwen2.5-VL-7B	OpenThought3-7B	Qwen2.5-7B-Instruct

The MLLM supplies the visual encoder, projector, tokenizer, and image preprocessing. Fusion operates on decoder tensors with corresponding names, shapes, and parameter semantics. Embeddings and output heads are eligible only when their vocabulary and tokenizer mappings agree. The 4B controlled comparison uses InternVL3.5-4B-MPO / Qwen3-4B-Thinking; the LLaVA pair uses the same named donor and base. Compatibility is checked at the tensor level rather than inferred from nominal model size alone: additional visual tokens, tied weights, or adapter parameterizations can require separate treatment.

Asset use and redistribution. Use and redistribution of the public checkpoints and evaluation benchmarks (Yue et al., 2024a;b; Lu et al., 2024; Wang et al., 2024a; Qiao et al., 2024; Zhang et al., 2025b) remain subject to their respective terms. Any release of fused weights, task vectors, or per-item outputs must respect the applicable source-model and dataset licenses and benchmark access restrictions.

A.2 Calibration and Checkpoint Construction

ANO measures visual-key attention in the source MLLM during generation, averages over heads, generated query positions, and calibration examples, and fits the depth-dependent prior using unlabeled MSCOCO images. It then forms decoder task vectors relative to the shared base, computes squared norms and the normalized coefficients, and saves one fused decoder with the retained visual modules. No gradient updates or labeled downstream examples are used to choose the fusion coefficients.

Task-vector arithmetic uses floating-point parameters, and the fused checkpoint is saved in the corresponding MLLM evaluation precision. Attention calibration is performed before fusion, so the fitted prior reflects the source MLLM rather than behavior induced by the merge. The exponential amplitude is discarded; its fitted depth slope enters the normalized prior. Appendix B explains the zero-energy and flat-fit cases of this construction.

A.3 Evaluation Protocol and Metrics

Matched comparisons use the same source checkpoints and require a common benchmark split, prompt, image processing, decoding budget, answer parser, and evaluator. Point estimates and repeated-run summaries are kept separate: Appendix C.2 reports dispersion across three ANO runs without treating those summaries as paired significance estimates against the source.

Reasoning benchmark scores and MME-RW scores are kept separate. RefCOCO uses accuracy at an IoU threshold of 0.5; COCO Caption uses CIDEr. These measurements test different aspects of multimodal behavior and are not combined into a single preservation score. MME-RW domain scores provide a complementary breakdown of OCR, remote sensing, diagram/table understanding, monitoring, and autonomous driving. The aggregate MME-RW score is not treated as the unweighted mean of those five columns.

A.4 Baseline Settings and Comparison Coverage

Table 2 compares eight merging alternatives on the InternVL3.5-4B pair, plus ANO's magnitude-only ablation, and DARE/TIES on Qwen2.5-VL-7B. This coverage is narrower than the six-pair source-versus-ANO study. Table 5 summarizes the configurations that the reported results can distinguish.

Table 5: **Baseline families and comparison coverage.** Each family combines the same source adaptations using a different fusion construction.

Method	Model pair	Fusion construction
Weight averaging / task arithmetic	InternVL3.5-4B	Direct parameter averaging or combination of updates relative to a shared base.
TIES / DARE	InternVL3.5-4B; Qwen2.5-VL-7B	Task-vector trimming and sign agreement / dropping and rescaling updates.
SLERP	InternVL3.5-4B	Spherical interpolation of parameter vectors.
OptMerge / WUDI / DOGE	InternVL3.5-4B	Surrogate-objective or geometric approaches to reducing interference.

A fixed weight average is one LERP coefficient. Task arithmetic is equivalent to LERP only under the sum-to-one construction in Section 3.3; SLERP alone does not determine a layer schedule. The comparisons evaluate the listed configurations rather than an exhaustive search over each method family. Accordingly, the OptMerge/WUDI/DOGE outcomes characterize these cross-modal constructions, not an inability of those methods to transfer capabilities in other settings.

MMMU-tuned exponential prior. For the exponential-prior comparison in Figure 3, we search decay coefficients over $[0.1, 0.9]$ and select the candidate achieving the highest MMMU score. The selected coefficient is then fixed across MMMU, MMMU-Pro, MathVista, MathVision, and WeMath; the five tasks do not receive separately optimized coefficients. This baseline uses MMMU performance for parameter selection, whereas ANO estimates the decay coefficient from source-MLLM attention measurements on unlabeled calibration images.

ANO uses unlabeled calibration inputs to set its depth prior. This differs from a strictly data-free rule, even though neither uses downstream labels. Holding source checkpoints fixed isolates the choice of fusion construction; it does not make all methods identical in their data access or coefficient-selection procedure.

B Mathematical Details of ANO

This appendix separates the exact solution of the coefficient objective from the empirical choice of salience. It also describes degenerate inputs and the relationship to local loss approximations.

B.1 Salience-Normalized Fusion

For one layer, the optimization in Eq. 5 is

$$J = s_V(1-\lambda_V)^2 + s_R(1-\lambda_R)^2, \qquad \lambda_V + \lambda_R = 1, \quad \lambda_V, \lambda_R \geq 0. \tag{8}$$

Proposition 1 (Salience-normalized two-branch fusion)**.** *For $s_V, s_R \geq 0$ and $s_V + s_R > 0$, the unique minimizer is $\lambda_V = s_V/(s_V + s_R)$ and $\lambda_R = s_R/(s_V + s_R)$.*

Proof. Substituting $\lambda_R = 1 - \lambda_V$ gives $J = s_V(1-\lambda_V)^2 + s_R\lambda_V^2$. Its derivative is $-2s_V + 2(s_V + s_R)\lambda_V$, and its second derivative is $2(s_V + s_R) > 0$. The stationary point lies in $[0, 1]$, including an endpoint when one salience is zero, and is therefore the unique feasible minimizer. □

Multiplying both saliences by the same positive scalar leaves the coefficients unchanged. Neither exact task-vector orthogonality nor a transformer linearization is required for this proposition. The saliences combine task-vector energy and a calibrated modality prior; their empirical utility is tested in Section 4.4.

B.2 Prior Construction and Degenerate Inputs

For $y_l = \log(a_l + \epsilon)$, the least-squares slope gives

$$\hat{\alpha} = -\frac{\sum_l (l - \bar{l})(y_l - \bar{y})}{\sum_l (l - \bar{l})^2}. \tag{9}$$

The amplitude is not used in the modality prior because it can depend on image-token count and prompt format. For numerical evaluation, the normalized exponential can be written

$$\frac{\exp[-\hat{\alpha}(l-1)] - \exp[-\hat{\alpha}(L-1)]}{1 - \exp[-\hat{\alpha}(L-1)]}, \tag{10}$$

whose limit at $\hat{\alpha} \to 0$ is $(L-l)/(L-1)$ for $L > 1$. Thus the exponential profile approaches a linear depth schedule as its fitted decay vanishes. The main construction uses a positive fitted decay; a nondecaying profile does not provide the same empirical motivation for this prior.

When both energies vanish, $\theta_V = \theta_R = \theta_0$ for that block, so retaining the base is unambiguous despite the coefficient formula being $0/0$. A zero energy in just one branch gives zero salience for that branch when the prior is positive. The main algorithm describes blocks with positive total salience.

B.3 Relationship to Local Loss Approximations

For a conceptual source loss $\mathcal{L}_t$ and $h_t = \theta_F - \theta_t$, a second-order expansion gives

$$\mathcal{L}_t(\theta_F) - \mathcal{L}_t(\theta_t) = \nabla\mathcal{L}_t(\theta_t)^\top h_t + \tfrac{1}{2} h_t^\top H_t h_t + o(\|h_t\|^2). \tag{11}$$

Dropping the gradient would require a stationary-checkpoint approximation, and replacing H_t by $\kappa_t I$ would require an isotropic-curvature approximation. Neither follows from a nearly linear output trajectory. Under the interpolation constraint, let $d = \tau_R - \tau_V$. Then

$$h_V = \lambda_R d, \qquad h_R = -\lambda_V d. \tag{12}$$

Even if both additional approximations held, the sum of the two local loss increases would be

$$\tfrac{1}{2}\|d\|^2[\kappa_V\lambda_R^2 + \kappa_R\lambda_V^2]. \tag{13}$$

For $d \neq 0$ and positive total curvature, these weights depend on κ_V, κ_R, not automatically on $\|\tau_V\|^2, \|\tau_R\|^2$. If $d = 0$, both adapted blocks coincide and this expression is independent of the interpolation coefficient. Equating curvature with task-vector energy would therefore be an additional modeling assumption, not a consequence of the expansion.

With anisotropic curvature, the directional terms are $d^\top H_t d$. A small Euclidean inner product $\tau_V^\top \tau_R$ does not control $\tau_V^\top H_t \tau_R$. The coefficient objective thus supplies a precise optimization interpretation without asserting global accuracy, capability independence, or optimality for the original model losses.

C Extended Transfer Results and Run Variability

C.1 Task-Level Changes

Table 6 complements the absolute scores in Table 1. MMMU, MMMU-Pro, and MathVision improve in all six pairs. The MathVista decreases for Idefics3 and NVILA, and the WeMath decrease for Idefics3, show why a positive average is not uniform improvement across tasks.

Table 6: **Changes relative to the source MLLM.** Entries are score differences; benchmark scores and aggregate changes are rounded independently.

MLLM	MMMU	MMMU Pro	Math Vista	Math Vision	WeMath	ΔAvg.
Idefics3-8B	+4.4	+2.1	-7.7	+7.5	-0.6	+1.1
NVILA-15B	+8.1	+13.2	-12.2	+14.0	+1.2	+4.9
InternVL2.5-38B	+5.3	+8.8	+1.2	+7.5	+8.7	+6.3
Qwen2.5-VL-7B	+1.1	+2.6	+1.2	+2.1	+2.8	+2.0
LLaVA-OneVision1.5-4B	+10.2	+3.4	+1.6	+3.6	+2.7	+4.3
InternVL3.5-4B	+4.2	+5.6	+5.9	+7.1	+12.1	+7.0

C.2 Repeated Runs and Their Interpretation

Table 7 gives three-run means and SDs. These summaries describe variation across ANO runs; they are distinct from the point estimates in Tables 1 and 2. The table reports ANO repeats, not paired repeated evaluations of every source model and merging baseline.

Table 7: **Three-run ANO statistics, mean ± SD.** Benchmark-level summaries complement the point estimates in the main tables. $n = 3$ runs per entry. SD describes run dispersion, not a confidence interval for a paired gain. —: aggregate not reported.

Benchmark	Idefics3-8B	Qwen2.5-VL-7B	InternVL3.5-4B
MMMU	48.2 ± 0.17	54.9 ± 0.21	69.53 ± 0.12
MMMU-Pro	34.7 ± 0.10	38.8 ± 0.26	65.60 ± 0.10
MathVista	50.6 ± 0.06	69.0 ± 0.25	77.53 ± 0.06
MathVision	27.5 ± 0.10	27.5 ± 0.06	55.07 ± 0.15
WeMath	11.6 ± 0.06	38.0 ± 0.10	51.83 ± 0.12
Avg.	34.5 ± 0.08	45.6 ± 0.17	—

For B benchmark datasets, a macro-average source-versus-ANO contrast is

$$\widehat{\Delta} = \frac{1}{B} \sum_{b=1}^{B} [\widehat{m}_b(\text{ANO}) - \widehat{m}_b(\text{reference})], \tag{14}$$

A paired item-level bootstrap interval for this contrast would use source and ANO predictions on the same examples, with common metric weights. Other interval constructions require their own sampling assumptions. SD across ANO runs alone is not such a paired interval, and benchmarks are not independent replicates of the construction process. We therefore use these summaries to describe run dispersion rather than to assert statistical significance or non-inferiority.

C.3 Aggregation and Comparison Scope

Point estimates and repeated estimates. Avg. summarizes the five reasoning benchmarks, while ΔAvg. compares that aggregate with the source. Independently rounded averages and changes need not subtract exactly at the displayed precision. For example, the displayed NVILA source and ANO benchmark cells average to 42.26 and 47.12, giving a change of 4.86; these round to 42.3, 47.1, and +4.9, respectively. The displayed InternVL3.5-4B ANO point scores average to 63.98, which rounds to 64.0. Point estimates and three-run means are not interchanged when computing the corresponding gains.

Coverage and settings. Six pairs assess transfer, two pairs have merging-baseline comparisons, three pairs have attention/geometry diagnostics, and direct grounding/captioning covers one pair. These complementary evaluations answer different questions: transfer breadth, relative fusion quality, checkpoint geometry, and direct visual behavior. Baseline coverage appears in Appendix A.4; the four-task stress test is discussed in Appendix D.3; task regressions appear in Appendix E.3.

D Calibration and Task-Vector Diagnostics

The main ablations evaluate the fusion rule. This appendix describes the attention measurements used to motivate calibration and the geometry of the evaluated adaptations. These empirical observations are separate from the coefficient-space proof in Appendix B.

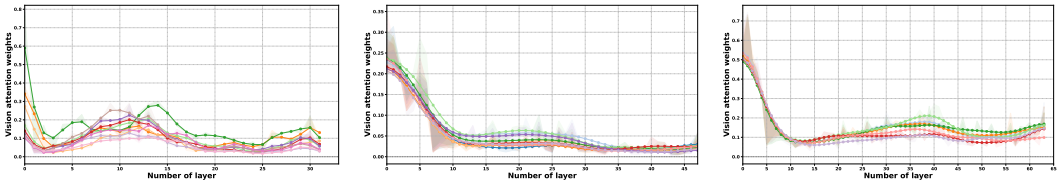
D.1 Visual Attention and Calibration Source

Figure 4 shows visual-attention profiles for Idefics3-8B, NVILA-15B, and InternVL2.5-38B. The measurement in Eq. 2 concerns direct attention to visual keys. Shallow-heavy profiles motivate a continuous prior, not a division into exclusively visual and reasoning layers: deeper layers can use visual information already integrated into nonvisual representations.

The three calibration sources yield similar slopes and MMMU scores on InternVL3.5-4B. The fitted slopes range from 0.108 to 0.115, while MMMU varies by 0.3 points. This comparison suggests that the prior need not be fitted on a single image domain to remain useful in this setting. It does not establish invariance across model families or arbitrary calibration conditions. Image tokenization, prompts, and generated-query selection can affect the measured attention profile.

Table 8: **Calibration sources.** InternVL3.5-4B; fitted slope and MMMU score.

Calibration source	$\hat{\alpha}$	MMMU
MSCOCO	0.108	69.6
ChartQA	0.115	69.3
ScreenSpot	0.111	69.4



(a) Idefics3-8B. (b) NVILA-15B. (c) InternVL2.5-38B.

Figure 4: **Visual-attention depth profiles.** Direct attention to visual keys is concentrated toward shallow layers in these source MLLMs, with model-dependent behavior at greater depth. Attention to visual keys is not a measure of all visual information carried by the decoder. Shading is not interpreted as a confidence interval.

D.2 Task-Vector Geometry

For task vectors defined relative to the shared base, the layer-wise cosine is

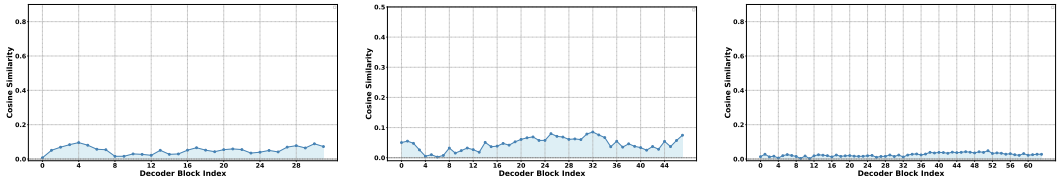
$$c_l = \frac{\langle \tau_V^{(l)}, \tau_R^{(l)} \rangle}{\|\tau_V^{(l)}\|_2 \|\tau_R^{(l)}\|_2}. \tag{15}$$

Figure 5 and Table 9 characterize the three measured model pairs. Small Euclidean cosine describes parameter directions, not capability independence: $\tau_V^\top H \tau_R$ can be nonzero for anisotropic curvature even when $\tau_V^\top \tau_R$ is small. Cosine is undefined for zero-energy task vectors.

Table 9: **Geometry and interpolation summaries.** Cosine concerns decoder task vectors; projection error concerns the base-to-branch path. Values are approximate checkpoint-level summaries. Donors and bases are given in Table 4. Small Euclidean cosine is not a guarantee of independent capabilities.

			Layers	
MLLM	Mean $\lvert\cos\rvert$	Max $\lvert\cos\rvert$	$\lvert\cos\rvert < 0.1$	Mean $\lvert s-\gamma\rvert$
Idefics3-8B	$\approx$0.050	$\approx$0.095	$\approx$100%	$\approx$0.045
NVILA-15B	$\approx$0.048	$\approx$0.090	$\approx$100%	$\approx$0.025
InternVL2.5-38B	$\approx$0.028	$\approx$0.050	$\approx$100%	$\approx$0.018

These summaries characterize Euclidean parameter geometry in the evaluated checkpoints. Small cosine alone cannot distinguish task-specific separation from a contribution of high dimensionality, and it does not define a universal acceptance threshold for merging. The coefficient-space solution in Appendix B does not require such a threshold.



(a) Idefics3-8B. (b) NVILA-15B. (c) InternVL2.5-38B.

Figure 5: **Cross-branch task-vector cosine by decoder block.** Model pairs are given in Table 4; ordinate scales differ across panels. The shaded area extends from the curve to zero, not a confidence interval. Small Euclidean cosine describes parameter directions rather than capability independence.

D.3 Correlation Stress Test

Table 10: **Correlation stress test on Idefics3-8B / DeepSeek-R1-Distill-Llama-8B.** Avg. covers the four displayed tasks, excluding MathVista. Perturbations change more than cosine. FT: fine-tuning. Avg. is the mean of the four displayed task scores.

			MMMU			
Setting	Mean $\lvert\cos\rvert$	MMMU	Pro	MathVision	WeMath	Avg.
Default	≈ 0.05	48.3	34.7	27.6	11.7	30.6
Shared reasoning FT	≈ 0.10	47.5	34.2	27.4	11.7	30.2
Pre-blended	≈ 0.18	44.2	32.9	22.4	11.4	27.7

The Idefics3-8B comparison associates larger task-vector cosine with lower four-task performance. Shared reasoning fine-tuning and pre-blending also alter donor quality, update norms, and the interpolation path; this is not an intervention on cosine alone. The results do not identify a safe cosine threshold or a causal effect of orthogonality.

The four-task average decreases from 30.6 to 30.2 and then to 27.7 across the three settings. This mean excludes MathVista and is therefore not interchangeable with the five-benchmark average in Table 1. The task scores describe the joint effect of each perturbation rather than a performance law determined by cosine alone.

D.4 Output Interpolation

For $\theta(\gamma) = (1-\gamma)\theta_0 + \gamma\theta_t$, let $\Delta = f(\theta_t) - f(\theta_0)$ at aligned token positions. The projection and residual are

$$s(\gamma) = \frac{[f(\theta(\gamma)) - f(\theta_0)]^\top \Delta}{\|\Delta\|_2^2}, \qquad \rho(\gamma) = \frac{\|f(\theta(\gamma)) - f(\theta_0) - \gamma\Delta\|_2}{\|\Delta\|_2}. \tag{16}$$

Table 9 reports mean $|s - \gamma|$ on the base-to-branch path. These values are consistent with a local-approximation perspective (Jacot et al., 2018; Zhou et al., 2024), but projection agreement does not bound the orthogonal residual: an output change can project close to $\gamma\Delta$ while still having a substantial component perpendicular to Δ. The constructed-model path would instead be

$$\theta(u) = \theta_V + u(\theta_F - \theta_V), \qquad u \in [0, 1], \tag{17}$$

with the source visual modules retained. This path has not been evaluated here. Neither source-loss stationarity nor isotropic curvature follows from the available projection summaries; none is needed for the coefficient-space solution.

E Visual-Capability Trade-offs

Visual evaluation uses two references: the unfused MLLM measures the cost of transfer, while alternative merges indicate whether ANO reduces that cost. The following results expand the task types in Section 4.5; they do not establish lossless preservation or a multi-family grounding guarantee.

E.1 MME-RW Domain Results

Table 11: **MME-RW domains on InternVL3.5-4B.** Domain scores and the aggregate are shown separately; Overall is not the unweighted mean of the five domains. RS: remote sensing; DT: diagram/table; MO: monitoring; AD: autonomous driving.

Method	OCR	RS	DT	MO	AD	Overall
Source	81.5	45.0	70.6	37.6	34.9	56.7
DARE	72.4	44.4	66.9	32.0	30.4	52.3
ANO	79.8	44.1	67.5	36.8	31.4	55.0
ANO − source	−1.7	−0.9	−3.1	−0.8	−3.5	−1.7

On InternVL3.5-4B, ANO is below the source in all five displayed domains and above DARE in four; remote sensing is the exception. The largest source-relative decreases are in autonomous driving and diagram/table understanding. OCR and monitoring show smaller drops, of 1.7 and 0.8 points. The domain breakdown therefore describes which visual tasks change, while the aggregate alone cannot localize the cost of transfer.

E.2 Grounding and Captioning

The RefCOCO and COCO Caption results cover one model pair. RefCOCO decreases by 0.8, 1.6, and 0.9 points across its splits; captioning decreases by 1.5 CIDEr. These tasks directly test localization and image description, complementing the MME-RW comparisons. Their decreases show that retaining the visual encoder and projector does not guarantee unchanged visual behavior after decoder fusion. Grounding accuracy and captioning quality use different units and are not combined into an average preservation score; the comparisons are descriptive rather than a non-inferiority test.

Table 12: **Direct visual tasks.** InternVL3.5-4B; Δ is relative to source.

Evaluation split	Source	ANO	Δ
Grounding: RefCOCO (Acc@0.5)			
val	90.4	89.6	−0.8
test-A	91.8	90.2	−1.6
test-B	85.5	84.6	−0.9
Description: COCO Caption (CIDEr)			
Captioning	130.2	128.7	−1.5

E.3 Task-Level Regressions

The MathVista changes are already visible in Tables 1 and 6: Idefics3 changes from 58.4 to 50.7 (−7.7), NVILA from 67.6 to 55.4 (−12.2), and InternVL2.5 from 71.9 to 73.1 (+1.2). Aggregate task scores do not identify OCR, counting, localization, or formatting as the cause. No item-level transition analysis or subtask-level failure decomposition is established by these measurements.

The opposite MathVista outcomes across pairs also caution against treating a benchmark as a pure measure of either perception or reasoning. A successful answer may require both accurate extraction of visual evidence and subsequent inference. The pair-level scores establish heterogeneous task outcomes, while finer attribution would require item-level evaluation of those stages.

Practical interpretation. The available comparisons support reasoning transfer with an evaluated visual cost, and reduced degradation relative to the tested alternatives. They do not establish preserved visual capability for arbitrary inputs. Domain-specific validation remains necessary; adaptive routing, query-specific fusion strength, and selective-module recovery are not evaluated ANO components. More fluent explanations alone are not evidence of more accurate visual grounding.

F Exploratory Transfer with a Coding Donor

The exploratory comparison pairs OpenCodeReasoning-Nemotron-1.1-7B with Qwen2.5-VL-7B-Instruct on Design2Code (Si et al., 2025), a screenshot-to-webpage-code benchmark. It asks whether checkpoint reuse can support a different downstream task by changing the donor specialization. The coding donor is derived from Qwen2.5-7B-Instruct, providing a compatible decoder family for this exploration. This single-pair result is separate from the main reasoning experiments and is not used to establish capability-agnostic transfer.

Table 13: **Exploratory screenshot-to-code comparison on Design2Code.** Scores are computed using the official Design2Code evaluation metrics and reported as Low-Level and High-Level summaries.

Model	Low-Level	High-Level
Qwen2.5-VL-7B-Instruct	83.4	87.6
ANO with coding donor	89.4	90.9

Both summary scores increase for the coding-donor merge. The comparison motivates further study of screenshot-to-code generation, which combines visual interpretation with structured program synthesis. It does not isolate coding from the donor's broader reasoning adaptation, or establish transfer to tool use, multilingual tasks, video, or audio. These remain directions for extending the checkpoint-reuse approach.