LEARNING FROM TEACHERS WITHOUT TRAINING: FUNCTIONAL GEOMETRY FOR CAPABILITY TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training endows pretrained models with advanced capabilities but typically relies on iterative rollout generation and gradient-based optimization, making methods such as On-Policy Distillation (OPD) computationally expensive. We investigate whether a Base model can directly acquire a capability already learned by a post-trained Teacher, without reproducing the training process through which that capability was obtained. We find that parameter-space displacement poorly reflects the actual computational consequence of a Teacher update. In contrast, evaluating the same update on Base activations induces a functional geometry that better captures its nonlinear consequence and remains consistent across different activation domains. Motivated by these observations, we propose Functional Geometry Calibration (FGC), a training-free framework that calibrates channel-wise Teacher updates under the Base-induced functional geometry and restores the response magnitude of the corresponding Base channels. FGC requires neither rollout generation nor gradient-based optimization, and independently calibrated capability updates can be directly composed to integrate complementary capabilities from multiple Teachers. Experiments on multimodal reasoning and visual perception show that FGC enables a Base model to acquire capabilities from individual post-trained Teachers and to compose complementary capabilities from multiple Teachers using only checkpoints and a small set of calibration inputs.

1 INTRODUCTION

Post-training transforms pretrained models into specialized systems with capabilities acquired through supervised fine-tuning, reinforcement learning, or distillation (Wei et al., 2022; Schulman et al., 2017; Zhao et al., 2026). Once such capabilities have been acquired by a post-trained Teacher model, transferring them typically requires another iterative training process involving data or rollout generation, model inference, and repeated parameter updates. Directly reusing capabilities already encoded in post-trained checkpoints, without reproducing the underlying training process, therefore provides an appealing alternative to conventional post-training.

A natural representation for such capability transfer is the parameter difference between a post-trained Teacher and its Base model. Task arithmetic and subsequent model merging methods manipulate Base-relative updates through sparsification, conflict resolution, rescaling, or geometric structure (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024; Cheng et al., 2025; Yang et al., 2026; Zhu et al., 2026), while ParamΔ demonstrates that post-training updates can themselves support training-free knowledge transfer (Cao et al., 2025). These approaches, however, primarily characterize an update through its structure in parameter space. Parameter displacement alone does not determine computational effect: the same displacement can induce substantially different responses depending on the representations on which it operates. This motivates characterizing post-training updates through their effects on the hidden activations produced by the Base model, which we refer to as Base representations.

In this work, we characterize Teacher updates by the output changes they induce when acting on these Base representations, which we refer to as their functional responses. We find that the magnitude of these responses more faithfully characterizes the nonlinear consequence and channel-wise allocation of Teacher updates than the magnitude of the parameter changes themselves. We further find that post-training approximately preserves the response magnitude established by the Base model.

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.

Motivated by these observations, we propose Functional Geometry Calibration (FGC), which adjusts the contribution of each Teacher channel according to its induced response and then rescales the resulting parameters to restore the Base response magnitude. FGC processes each Teacher independently, allowing the resulting updates from multiple Teachers to be directly added without additional training .

Our contributions are summarized as follows: (1) To our knowledge, we are the first to characterize post-training capability transfer through the functional geometry induced by Base representations, revealing a fundamental mismatch between parameter displacement and the computational significance of Teacher updates; (2) we propose Functional Geometry Calibration (FGC), a training-free framework that enables a Base model to acquire capabilities already learned by post-trained Teachers by calibrating their updates according to their functional effects, without rollout generation or gradient-based optimization; and (3) we extend FGC to multiple Teachers by independently calibrating each Teacher's Base-referenced update and directly combining the resulting updates, enabling the integration of complementary capabilities across multimodal reasoning and visual perception.

2 Related work

2.1 On-Policy Distillation for LLMs

On-Policy Distillation (OPD) trains a Student on its own generated trajectories using dense supervision from a Teacher, reducing the distribution mismatch of offline distillation (Agarwal et al., 2024). Recent work has extended OPD beyond the conventional large-Teacher–small-Student setting. In particular, on-policy self-distillation considers equal-capacity or even parameter-sharing Teacher–Student pairs, where the Teacher provides privileged information or acquired behavior rather than additional model capacity (Zhao et al., 2026; Ye et al., 2026). These studies position distillation not only as compression, but also as a mechanism for internalizing capabilities into a target model.

Recent studies further investigate the structure of capabilities transferred through OPD. OPD^2 (Heo et al., 2026) isolates reasoning-specific behavior through the output difference between a reasoning-tuned Teacher and its Base, while Yu et al. (2026) show that OPD induces structured parameter updates that are coordinate-sparse and spectrally concentrated. Our work follows this capability-transfer perspective but focuses on capabilities already encoded in post-trained Teacher checkpoints and their direct transfer without additional iterative distillation. Rather than characterizing such capabilities through parameter changes alone, we characterize Teacher updates by their effects on Base representations and use these effects to guide training-free capability transfer.

2.2 Training-Free Parameter Arithmetic and Model Merging

Early work on model merging methods primarily relied on weight averaging to combine multiple fine-tuned models (Wortsman et al., 2022). Task Arithmetic (Ilharco et al., 2023) subsequently formalized the parameter difference between fine-tuned and pre-trained models as a task vector, enabling arithmetic operations over acquired capabilities. To mitigate parameter interference during merging, TIES-Merging (Yadav et al., 2023), DARE (Yu et al., 2024), and DELLA (Deep et al., 2024) further introduced sign-conflict resolution, sparsification, and magnitude-aware pruning. Recent work has extended parameter-space transfer beyond conventional task-vector arithmetic. ParamΔ (Cao et al., 2025) directly transfers post-training deltas between base models, while Extrapolation Merging (Lin et al., 2025) exploits extrapolation along fine-tuning trajectories to provide an optimization direction for model merging. WUDI (Cheng et al., 2025) further mitigates parameter interference through task-vector-guided subspace manipulation without requiring additional data.

More recent approaches move beyond element-wise parameter operations toward structured transformations of weight matrices. OrthoMerge (Yang et al., 2026) performs merging through orthogonal components and their underlying manifold structure, while ISO-Merger (Zhu et al., 2026) exploits the spectral structure of RLVR-trained models by preserving the base singular-value spectrum and combining changes in the singular frames. These studies demonstrate the potential of exploiting structural properties of post-training parameter changes. In contrast, our work characterizes these changes by the output responses they induce on Base model activations and uses these responses to guide capability transfer.

3 FUNCTIONAL GEOMETRY OF POST-TRAINING UPDATES

We begin with a linear projection, the basic parameterized operation considered in our analysis. Let $W_B \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ denote its weights in a Base model and W_T the corresponding weights in a specialized Teacher obtained by post-training the Base model. Their difference $\Delta W = W_T - W_B$ is the parameter update introduced by post-training. Running calibration examples through the Base model provides N activation vectors at the input of the projection, which we stack as $A \in \mathbb{R}^{N \times d_{\text{in}}}$. Since the projection computes $AW^\top$, the Teacher update induces the output change

$$\Delta Y = AW_T^\top - AW_B^\top = A\Delta W^\top.$$

We refer to this induced output change as the *functional response* of the update on Base activations. Each row $W_j \in \mathbb{R}^{d_{\text{in}}}$ produces one output dimension, or channel j, so ΔW_j denotes the update to that channel and $A\Delta W_j^\top$ its functional response.

3.1 FUNCTIONAL SIGNIFICANCE AND UPDATE ALLOCATION

For channel j, the parameter magnitude $P_j = \|\Delta W_j\|_2$ measures how far its weights move in parameter space. We instead define the functional magnitude

$$F_j = \frac{1}{\sqrt{N}} \left\| A\Delta W_j^\top \right\|_F,$$

which measures how strongly the same update changes the projection output on Base activations. The two quantities differ because

$$F_j^2 = \Delta W_j G_A \Delta W_j^\top, \qquad G_A = \frac{1}{N} A^\top A.$$

The activation covariance G_A therefore weights parameter directions according to how they affect representations encountered by the Base model. We refer to the geometry induced by this metric as the *Base-induced functional geometry*. Unlike Euclidean magnitude, which weights all parameter directions equally, this geometry distinguishes updates according to their computational effects on Base activations.

Functional magnitude predicts nonlinear consequence. To determine whether this distinction reflects actual model computation, we compare P_j and F_j against an intervention-based measure of channel consequence. Our models use SwiGLU feed-forward networks (FFNs), in which Gate and Up projections jointly form an intermediate representation that is subsequently mapped back to the model dimension by a Down projection. We remove individual Teacher channel updates from a combined two-Teacher FFN and measure the resulting change in the complete nonlinear FFN output.

Specifically, let $\Delta\Theta_1$ and $\Delta\Theta_2$ denote the FFN parameter updates of two specialized Teachers derived from the same Base FFN parameters Θ_B, and let $\Theta_M = \Theta_B + \Delta\Theta_1 + \Delta\Theta_2$. For a channel j in the projection being analyzed, $\Theta_{M,-j}$ denotes the same combined FFN with only the corresponding row of the second Teacher update removed. On N_{test} Base activations A_{test} collected at the FFN input from held-out examples, we define its nonlinear consequence as

$$S_j = \frac{1}{\sqrt{N_{\text{test}} d}} \left\| f_{\text{FFN}}(A_{\text{test}}; \Theta_M) - f_{\text{FFN}}(A_{\text{test}}; \Theta_{M,-j}) \right\|_F,$$

where d is the FFN output dimension. Thus, F_j measures the immediate response of a linear channel update, whereas S_j measures its consequence after the complete nonlinear FFN computation. The activations used to estimate F_j and S_j are collected from disjoint input examples.

The left panel of Table 1 shows a clear separation between the two measures: across both Gate and Up projections, functional magnitude is substantially more predictive of intervention consequence than parameter magnitude. This relationship also persists when functional responses and intervention consequences are measured across different calibration domains (Appendix B.1). The Base-induced metric therefore captures computational structure that is largely obscured by parameter magnitude alone.

Table 1: Functional characterization of Teacher updates. Left: correlation with the intervention-based nonlinear consequence. Right: alignment with the channel-wise allocation of nonlinear consequence; allocation results are medians across layers.

	P		F	
Proj.	Pear.	Spear.	Pear.	Spear.
Gate	0.295	0.595	**0.635**	**0.918**
Up	0.313	0.542	**0.482**	**0.925**

Proj.	Measure	Spear. ↑	TV ↓	JS ↓
Gate	P	0.429	0.353	0.096
	F	0.639	0.283	0.064
	E	**0.690**	**0.282**	**0.063**
Up	P	0.459	0.266	0.059
	E	0.687	0.225	0.046
	F	**0.753**	**0.203**	**0.038**

Functional magnitude captures update allocation. Capability transfer depends not only on identifying consequential channels, but also on how contribution is distributed across them. In addition to the absolute functional magnitude F_j, we consider the Base-relative functional magnitude

$$E_j = \frac{\|A\Delta W_j^\top\|_F}{\|AW_{B,j}^\top\|_F},$$

which measures the response induced by the update relative to the response magnitude already produced by the corresponding Base channel.

For each nonnegative measure $M \in \{P, F, E, S\}$, we define its normalized channel allocation as

$$q_j^M = \frac{M_j}{\sum_k M_k}.$$

We compare the allocations induced by P, F, and E with the intervention-based allocation q^S: Spearman correlation measures agreement in channel ordering, while total-variation (TV) and Jensen–Shannon (JS) distances measure agreement between the normalized allocations.

The right panel of Table 1 shows that functional measures align substantially more closely with the nonlinear consequence allocation than parameter magnitude. The relative measure E performs best for Gate projections, while the absolute functional magnitude F performs best for Up projections. Together with the intervention results above, this shows that Base activations reveal not only which Teacher updates materially affect model computation, but also how their computational consequence is distributed across channels.

3.2 Base-Anchored Functional Geometry

The same Base-induced metric also defines the response scale of the parameters themselves. Recall that the Base activations induce $G_A = \frac{1}{N}A^\top A$. For any weight vector $w \in \mathbb{R}^{d_{\text{in}}}$, we define its *functional radius* as

$$\|w\|_A = \sqrt{w^\top G_A w} = \frac{1}{\sqrt{N}}\|Aw^\top\|_F,$$

which is precisely the magnitude of the output produced by w on the Base activations. Weight vectors with the same functional radius therefore lie on the functional level set

$$\mathcal{E}_A(c) = \{w : \|w\|_A = c\} = \left\{w : w^\top G_A w = c^2\right\}, \qquad c > 0.$$

Because G_A weights parameter directions according to the Base activations, weight vectors with equal Euclidean norms need not have equal functional radii. When G_A is positive definite, $\mathcal{E}_A(c)$ forms an ellipsoid rather than a Euclidean sphere, with its shape determined by the covariance of the Base activations. The functional level set thus describes equal response magnitude under the representation-dependent geometry of the Base model.

Post-training preserves the Base functional radius. We now locate the post-trained Teacher in this geometry. For each output channel j, we compare the functional radius of its Teacher weight

vector with that of the corresponding Base weight vector:

$$\rho_j = \frac{\|W_{T,j}\|_A}{\|W_{B,j}\|_A}.$$

Here, $\rho_j = 1$ means that the Teacher and Base channels produce the same response magnitude on Base activations despite differing in parameter space. Across Gate, Up, and Down projections, these ratios are tightly concentrated around one: the median is 1.0000 for all three projection types, with median absolute log-deviations below 10^{-3}.

Thus, post-training changes the parameters and their induced computation while leaving the overall response magnitude of each channel nearly unchanged. Equivalently, the Teacher weight $W_{T,j}$ remains close to the Base-anchored functional level set

$$\mathcal{E}^{(j)}_{A,B} = \{w : \|w\|_A = \|W_{B,j}\|_A\}.$$

This reveals a complementary structure to the functional allocation above: Teacher updates vary substantially in where their computational consequence is concentrated, yet their endpoints remain anchored to the response scale of the Base model. Together, these observations motivate capability transfer that calibrates the contribution of Teacher updates in the Base-induced functional geometry while preserving the Base functional radius.

4 Functional Geometry Calibration

The analysis above reveals two complementary properties of post-training updates. First, their responses on Base activations better capture nonlinear consequence and its channel-wise allocation than parameter magnitude. Second, despite these functional changes, post-trained Teacher channels approximately preserve the functional radius of their corresponding Base channels. These findings suggest that capability transfer should account for where the Teacher update has greater computational effect while preserving the response scale established by the Base model.

Based on these findings, we propose **Functional Geometry Calibration (FGC)**, a training-free framework consisting of two operations: *Functional Calibration* and *Functional Rebase*. Functional Calibration rescales individual channel updates according to their Base-relative functional magnitude while preserving their original update directions. Because this rescaling can alter the functional radius of the resulting parameters, Functional Rebase subsequently restores the functional radius of the corresponding Base channel. FGC requires only model checkpoints and a small set of calibration inputs, without rollout generation or gradient-based optimization. Figure 1 illustrates the overall framework.

4.1 Functional Calibration

Consider a target linear projection l with Teacher update $\Delta W^{(l)} = W_T^{(l)} - W_B^{(l)}$, and let A_l contain N_l Base activation vectors collected at its input. For these activations, the functional norm $\|w\|_{A_l} = \frac{1}{\sqrt{N_l}}\|A_l w^\top\|_F$ measures the response magnitude of a weight vector w. Following the functional allocation identified in Section 3.1, we characterize each channel update by its Base-relative functional magnitude

$$E_{l,j} = \frac{\|\Delta W_j^{(l)}\|_{A_l}}{\|W_{B,j}^{(l)}\|_{A_l}}.$$

Thus, $E_{l,j}$ measures how strongly the Teacher update changes the projection output relative to the response magnitude already produced by the corresponding Base channel.

We normalize this quantity by its within-projection median, $m_l = \mathrm{median}_{j'} E_{l,j'}$, and define Functional Calibration as

$$\widetilde{\Delta W}_j^{(l)} = \left(\frac{E_{l,j}}{m_l}\right)^{\alpha} \Delta W_j^{(l)},$$

where α controls the calibration strength. Median normalization makes the scaling depend on the relative functional magnitude of channels within each projection rather than on its overall response scale. Since the scaling factor $(E_{l,j}/m_l)^\alpha$ is positive, Functional Calibration preserves the direction

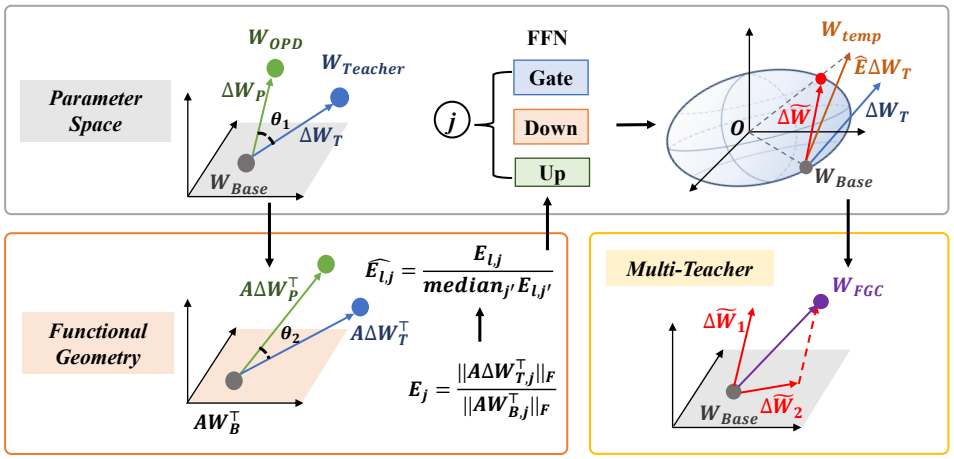

Figure 1: Overview of Functional Geometry Calibration (FGC). Functional Calibration rescales channel-wise Teacher updates according to their responses on Base activations, while Functional Rebase restores the functional radius of the corresponding Base channels. Teacher updates processed separately by FGC can be directly combined for multi-Teacher capability transfer.

of each Teacher update while changing its magnitude: channels with larger Base-relative responses are amplified, whereas those with smaller responses are attenuated. The calibrated update defines the temporary endpoint

$$W_{\text{temp},j}^{(l)} = W_{B,j}^{(l)} + \widetilde{\Delta W}_j^{(l)}.$$

Because Functional Calibration changes the magnitude of individual channel updates, $W_{\text{temp},j}^{(l)}$ does not in general retain the Base functional radius. Functional Rebase restores this property.

4.2 Functional Rebase

Section 3.2 shows that post-trained Teacher channels remain close to the functional level set defined by the response magnitude of their corresponding Base channels. For projection l and channel j, this Base-anchored level set is

$$\mathcal{E}_{A_l,B}^{(j)} = \left\{ w : \|w\|_{A_l} = \|W_{B,j}^{(l)}\|_{A_l} \right\}.$$

We use this empirical structure as the reference for correcting the endpoint produced by Functional Calibration. Specifically, we rescale $W_{\text{temp},j}^{(l)}$ along its radial direction until it reaches the functional level set of the corresponding Base channel. Writing $W_{\text{FGC},j}^{(l)} = r_{l,j} W_{\text{temp},j}^{(l)}$, the radius constraint and its closed-form solution are

$$\left\| W_{\text{FGC},j}^{(l)} \right\|_{A_l} = \left\| W_{B,j}^{(l)} \right\|_{A_l}, \qquad r_{l,j} = \frac{\|W_{B,j}^{(l)}\|_{A_l}}{\|W_{\text{temp},j}^{(l)}\|_{A_l}}, \qquad W_{\text{FGC},j}^{(l)} = r_{l,j} W_{\text{temp},j}^{(l)}.$$

This radial rescaling preserves the direction of $W_{\text{temp},j}^{(l)}$ while exactly restoring the Base response magnitude.

The two operations therefore serve complementary roles. Functional Calibration reallocates the contribution of Teacher updates according to their effects on Base activations, while Functional Rebase restores the response magnitude that post-training empirically preserves.

Vision-Side Instantiation. For single-Teacher transfer, we additionally instantiate the same Base-activation principle on selected linear projections in the visual encoder. We apply a vision-specific variant of Functional Calibration followed by Functional Rebase using the corresponding Base visual activations. Details are provided in Appendix C.1.

4.3 MULTI-TEACHER COMPOSITION

FGC processes multiple Teachers separately relative to the same Base model. For each Teacher t, we apply Functional Calibration and Functional Rebase to its Base-relative parameter update using the Base activations $A^{(t)}$ collected from its corresponding calibration inputs, obtaining

$$W_{\mathrm{FGC}}^{(t)} = \mathrm{FGC}\left(W_B, W_T^{(t)}; A^{(t)}\right), \qquad \Delta W_{\mathrm{FGC}}^{(t)} = W_{\mathrm{FGC}}^{(t)} - W_B.$$

Each Teacher is calibrated using only its own parameter update and Base activations, without jointly estimating calibration coefficients across Teachers. The resulting updates are therefore independently calibrated relative to the same Base model.

This common Base reference provides a direct way to combine the calibrated updates. For fixed Base activations A, their locally induced responses preserve the additive structure of the parameter updates:

$$A\left(\sum_{t=1}^{T} \Delta W_{\mathrm{FGC}}^{(t)}\right)^{\top} = \sum_{t=1}^{T} A \Delta W_{\mathrm{FGC}}^{(t)\top}.$$

We therefore construct the multi-Teacher model by adding the calibrated updates to the shared Base:

$$W_{\mathrm{comp}} = W_B + \sum_{t=1}^{T} \Delta W_{\mathrm{FGC}}^{(t)}.$$

This composition requires neither joint calibration across Teachers nor additional interpolation coefficients. Appendix B.3 further analyzes the role of cross-Teacher functional overlap and shows that unit-weight addition is recovered in the orthogonal limit.

In our multimodal experiments, multi-Teacher composition is applied to the language-side FGC updates, while the visual backbone remains that of the Base model. Algorithm 1 summarizes the complete FGC procedure.

Algorithm 1: Functional Geometry Calibration (FGC)

Input: Base model W_B, post-trained Teacher model W_T, Base activations $\{A_l\}$ for target projections, and calibration strength α.
Output: Calibrated model W_{FGC}.
for *each target projection* l **do**
  $\Delta W^{(l)} \leftarrow W_T^{(l)} - W_B^{(l)}$
  Compute the Base-relative functional magnitude $E_{l,j}$ for each channel j using A_l.
  $m_l \leftarrow \mathrm{median}_{j'} E_{l,j'}$
  foreach *channel* j **do**
    $\widetilde{\Delta W}_j^{(l)} \leftarrow (E_{l,j}/m_l)^{\alpha} \Delta W_j^{(l)}$
    $W_{\mathrm{temp},j}^{(l)} \leftarrow W_{B,j}^{(l)} + \widetilde{\Delta W}_j^{(l)}$
    Radially rescale $W_{\mathrm{temp},j}^{(l)}$ to the Base-anchored functional level set, obtaining $W_{\mathrm{FGC},j}^{(l)}$.
return W_{FGC}

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Benchmarks. We evaluate FGC on benchmarks covering multimodal mathematical reasoning and fine-grained visual perception. For mathematical reasoning, we use MathVista (Lu et al., 2024), MathVerse (Zhang et al., 2024), and MathVision (Wang et al., 2024). For visual perception, we evaluate on V*Bench (Wu & Xie, 2024) and HR-Bench (Wang et al., 2025b), reporting HR-Bench results at both 4K and 8K resolutions.

Table 2: Performance comparison on mathematical reasoning and visual perception benchmarks. Vista and MV denote MathVista and MathVision; TD, VD, TL, VI, and VO denote the five MathVerse subsets. Best and second-best results across all methods are shown in bold and underlined.

Method	Math Reasoning								Visual Perception		
	Vista	MV	MathVerse						V*	HR-Bench	
			avg.	TD	VD	TL	VI	VO		4K	8K
Base	67.9	26.0	43.6	53.4	40.1	39.9	43.5	41.0	66.5	58.5	46.1
VLRethinker	**74.8**	31.6	<u>50.9</u>	**58.9**	<u>49.1</u>	51.3	<u>48.6</u>	**46.7**	66.0	60.6	49.0
ZwZ	71.5	26.6	46.8	56.1	44.3	47.6	44.5	41.2	**75.9**	**65.8**	**58.1**
LinearMerge	74.0	<u>33.6</u>	49.5	56.9	48.0	51.3	47.1	44.4	72.8	61.9	54.5
OrthoMerge	73.4	34.2	48.7	57.2	44.5	49.4	47.1	45.2	70.7	61.8	54.5
ISO-Merger	73.5	<u>33.6</u>	50.4	58.5	48.0	<u>51.8</u>	47.6	<u>46.3</u>	73.8	<u>64.1</u>	<u>56.3</u>
FGC (Ours)	<u>74.4</u>	**37.5**	**51.3**	<u>58.6</u>	**50.3**	**52.2**	**49.8**	45.6	<u>74.9</u>	63.3	55.9

Models & Baselines. We use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as the Base model and two specialized models derived from the same backbone as Teachers. VL-Rethinker-7B (Wang et al., 2025a) serves as the reasoning-specialized Teacher, while ZwZ-7B (Wei et al., 2026) serves as the fine-grained perception Teacher. The two Teachers provide complementary capabilities while sharing the same Base architecture and initialization. We compare FGC with representative training-free model merging approaches, including Linear Merge (Ilharco et al., 2023), OrthoMerge (Yang et al., 2026), and Iso-Merger (Zhu et al., 2026). All methods operate on the same Base and specialized Teacher checkpoints without additional training.

Implementation Details. For the language backbone, we apply FGC to the Gate, Up, and Down projections of all 28 FFN layers and set the calibration exponent to $\alpha = 0.25$. We collect Base activations from eight calibration examples per Teacher, sampled from MMFineReason for VL-Rethinker and ZwZ-RL-VQA for ZwZ. For multi-Teacher composition, we compose the independently calibrated language-side updates while retaining the visual backbone of the Base model. The vision-side instantiation used in the single-Teacher experiments is described in Appendix C.1, and sensitivity to α is reported in Appendix C.2.

Evaluation. We conduct all evaluations using VLMEvalKit (Duan et al., 2024). Following the evaluation configuration of VL-Rethinker-7B, we set the minimum and maximum numbers of input pixels to $256 \times 28 \times 28$ and $1280 \times 28 \times 28$, respectively. For generation, we enable sampling with a temperature of 1.0 and top-p of 0.95, with a maximum generation length of 8,192 tokens. We disable model-specific custom prompts and use the standard benchmark prompts provided by VLMEvalKit. For benchmarks requiring LLM-based answer extraction, we use GPT-4o as the evaluator. The same evaluation configuration and protocols are used for the Base, Teachers, merging baselines, and FGC.

5.2 Main Results

Table 2 compares FGC with training-free model merging baselines for integrating the complementary reasoning and visual capabilities of VL-Rethinker and ZwZ. FGC shows its strongest gains on mathematical reasoning, achieving the best performance on MathVision and average MathVerse, as well as on the VD, TL, and VI subsets. Notably, FGC improves MathVision to 37.5, compared with 34.2 for the strongest merging baseline, while reaching 51.3 on average MathVerse and remaining competitive with VL-Rethinker on its strongest reasoning metrics.

FGC also retains substantial fine-grained visual capability from the perception-specialized Teacher. It achieves 74.9 on V*, close to ZwZ (75.9) and higher than all merging baselines, while reaching 63.3 and 55.9 on HR-Bench 4K and 8K, respectively. Although ISO-Merger performs slightly better on HR-Bench, FGC provides substantially stronger mathematical reasoning while preserving strong visual perception. These results show that separately calibrated Teacher updates can integrate complementary capabilities within a shared Base without additional training.

Table 3: Ablation of language- and vision-side adaptations in single-Teacher capability transfer. Best results within each Teacher setting are shown in bold.

Method	Math Reasoning								Visual Perception		
	Vista	MV	MathVerse						V*	HR-Bench	
			avg.	TD	VD	TL	VI	VO		4K	8K
FGC (Ours)											
-*VLRe*^{Teacher}	**74.0**	**34.5**	**49.8**	57.0	47.1	**52.2**	**49.6**	43.0	68.6	59.8	47.9
w/o Vision	73.9	33.9	49.4	**58.0**	**48.0**	50.1	46.6	**44.2**	67.5	59.8	47.8
w/o LLM	70.3	27.6	44.2	53.9	39.3	45.3	43.0	39.6	66.0	58.8	45.4
-*ZwZ*^{Teacher}	70.8	29.6	42.3	53.0	42.0	44.7	41.9	29.9	**75.9**	**64.1**	**56.1**
w/o Vision	71.4	26.6	45.2	54.8	41.8	45.4	44.2	40.0	75.4	63.8	55.4
w/o LLM	70.6	27.0	43.5	52.3	38.8	45.1	42.4	38.7	65.5	58.5	44.8

Table 4: Left: ablation of rebase geometry after functional calibration. Right: time (min:s).

Method	Vista	MV	V*	HR-4K	HR-8K
Calibration Only	73.4	33.2	**74.9**	63.5	55.8
+ Euclidean Rebase	73.5	35.2	73.8	63.5	**55.9**
+ Functional Rebase	**74.4**	**37.5**	**74.9**	**63.8**	**55.9**

Method	Time
OrthoMerge	25:55
Iso-Merger	34:20
FGC (Ours)	**5:18**

5.3 Ablation Study

We examine the language- and vision-side components of FGC in single-Teacher transfer. As shown in Table 3, removing language-side FGC substantially weakens each Teacher's specialized capability: MathVision drops from 34.5 to 27.6 for VL-Rethinker, while V* drops from 75.9 to 65.5 for ZwZ. In contrast, removing the vision-side component produces much smaller changes. These results identify language-side FGC as the primary mechanism of capability transfer.

We next isolate the role of Functional Rebase after Functional Calibration. Starting from the same calibrated endpoint, we compare no rebase, Euclidean norm restoration, and Functional Rebase. As shown in the left panel of Table 4, MathVision improves from 33.2 without rebase to 35.2 with Euclidean Rebase and 37.5 with Functional Rebase, while visual performance is preserved. The advantage over Euclidean Rebase supports the Base-induced functional radius, rather than generic norm restoration, as the appropriate reference for rebase.

5.4 Computational Efficiency

FGC requires only activation–weight products and channel-wise operations. For N calibration activations and target modules $\mathcal{S}$, its dominant computational complexity is $O\Big(N \sum_{l \in \mathcal{S}} d_{\text{in}}^{(l)} d_{\text{out}}^{(l)}\Big)$, scaling linearly with the calibration size and the number of target parameters without requiring matrix factorization or inversion. As shown in Table 4, FGC also incurs substantially lower end-to-end merging time under the same hardware and implementation environment.

6 Conclusion

We study training-free capability transfer from post-trained Teacher checkpoints through the functional geometry induced by Base activations. Building on the findings that this geometry better captures the computational significance of Teacher updates and that post-training approximately preserves the Base functional radius, we propose Functional Geometry Calibration (FGC) for direct single- and multi-Teacher capability transfer. Experiments on multimodal reasoning and visual perception demonstrate effective capability transfer without rollout generation or gradient-based optimization.

AI USE STATEMENT

Generative AI tools were used to assist with writing and language polishing. All AI-assisted content were reviewed and verified by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The code for reproducing our method and experiments is available at the anonymous repository `https://anonymous.4open.science/r/FGC-7027/`.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

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, 2025. URL `https://arxiv.org/abs/2502.13923`.

Sheng Cao, Mingrui Wu, Karthik Prasad, Yuandong Tian, and Zechun Liu. Param δ for direct weight mixing: Post-train large language model at zero cost. *arXiv preprint arXiv:2504.21023*, 2025.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10121–10143. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/cheng25h.html`.

Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. Della-merging: Reducing interference in model merging through magnitude-based sampling. *arXiv preprint arXiv:2406.11617*, 2024.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-policy delta distillation. *arXiv preprint arXiv:2607.15161*, 2026.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=6t0Kwf8-jrj`.

Yiguan Lin, Bin Xu, Yinghao Li, and Yang Gao. Extrapolation merging: Keep improving with extrapolation and merging. *arXiv preprint arXiv:2503.04834*, 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 *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. 38, Main Conference:30865–30891, 2025a. doi: 10.52202/085713-1037. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2c84844a559e4f962752570bff456ae4-Paper-Conference.pdf.

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. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025b.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gEZrGCozdqR.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

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

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

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

Sihan Yang, Kexuan Shi, and Weiyang Liu. Orthogonal model merging. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=pCaeMsqpCa.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Guo Yu, Wenlin Liu, Yulan Hu, Hao-Xuan Ma, Jun-Peng Jiang, and Han-Jia Ye. Dense supervision, sparse updates: On the sparsity and geometry of on-policy distillation. *arXiv preprint arXiv:2606.13657*, 2026.

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 Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yu24p.html.

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 *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

Hanqing Zhu, Wenyan Cong, Zhizhou Sha, Sagnik Mukherjee, Xinyuan Song, David González-Martínez, Xiaoxia Wu, Yuandong Tian, Shiwei Liu, David Z Pan, et al. Iso: An rlvr-native optimization stack. *arXiv preprint arXiv:2607.19331*, 2026.

A Appendix

B Observations and Analysis

B.1 Cross-Domain Robustness of Functional Geometry

We examine whether the Base-induced functional geometry depends strongly on the calibration domain. Let A_1 and A_2 denote Base activations collected from two distinct multimodal domains. We compute functional magnitudes using each activation set and evaluate their ability to predict nonlinear intervention consequences measured on either domain, including the cross-domain settings $F(A_1) \to S(A_2)$ and $F(A_2) \to S(A_1)$.

Table 5: Cross-domain robustness of functional magnitude. Values are pooled Spearman correlations across channels.

Projection	$F_{A_1} \to S_{A_1}$	$F_{A_2} \to S_{A_2}$	$F_{A_1} \to S_{A_2}$	$F_{A_2} \to S_{A_1}$	$\rho(F_{A_1}, F_{A_2})$
Gate	0.918	0.900	0.885	0.871	0.939
Up	0.925	0.874	0.878	0.848	0.919

Functional magnitude remains strongly predictive when the activation domain used to compute F differs from the domain used to evaluate S. Moreover, A_1 and A_2 induce highly consistent channel orderings. These results show that the Base-induced functional geometry is robust to the choice of calibration domain.

B.2 Functional Correspondence across Independent Post-Training Paths

As supplementary evidence for the Base-induced functional geometry, we compare two independently obtained post-training paths from the same Base model: a specialized Teacher and an OPD-trained model. This analysis is not used by FGC; it tests whether updates that appear weakly related in parameter space exhibit stronger correspondence through their effects on the same Base activations.

Let W_B, W_T, and W_P denote the Base, Teacher, and OPD parameters, respectively, with

$$\Delta W_T = W_T - W_B, \qquad \Delta W_P = W_P - W_B.$$

Across the 196 linear projections in the language backbone, these updates have a median parameter-space cosine similarity of only 0.155. On the same Base activations A, their functional responses are

$$R_T = A\Delta W_T^\top, \qquad R_P = A\Delta W_P^\top.$$

As shown in Table 6(a), the correspondence becomes substantially stronger when the updates are evaluated through their functional responses. The median cosine similarity increases from 0.155 in parameter space to 0.420 in functional space, with higher functional alignment in all 196 projections. The effect is strongest for FFN projections, where the median increases from 0.147 to 0.453.

This increase depends on the full covariance structure of the Base activations. For parameter perturbations U and V,

$$\left\langle AU^\top, AV^\top \right\rangle_F = \mathrm{tr}\left(UA^\top AV^\top\right),$$

Table 6: Correspondence between independent post-training updates under the Base-induced functional geometry. Left: parameter and functional alignment. Right: controls on the activation covariance.

Family	Param.	Functional
All	0.155	**0.420**
Attention	0.167	**0.398**
FFN	0.147	**0.453**

(a) Parameter vs. functional alignment

Family	Param.	Diag.	Mismatch	Full
Attention	0.163	0.210	0.169	**0.359**
FFN	0.146	0.210	0.142	**0.441**

(b) Activation-covariance controls

so functional alignment depends on cross-feature covariance rather than activation magnitude alone. Consistently, retaining only diagonal activation statistics or disrupting feature correspondences substantially weakens the alignment (Table 6(b)).

These results provide independent support for the Base-induced functional geometry: post-training updates that are weakly aligned in parameter coordinates exhibit substantially stronger correspondence when evaluated through their effects on Base activations. FGC does not require the OPD checkpoint; this comparison serves only as a diagnostic of shared post-training structure.

B.3 Functional-Overlap Interpretation of Multi-Teacher Composition

The main method combines independently calibrated Teacher updates by direct addition. Here, we provide a more general interpretation of this choice in terms of functional overlap. This analysis is used only to interpret the composition rule and introduces no additional optimization into FGC.

Consider T Base-relative Teacher updates $\Delta W_1, \ldots, \Delta W_T$. For fixed Base activations A, let

$$R_i = A\Delta W_i^\top$$

denote the functional response of Teacher i. To account for redundancy among these responses, consider a composed response R that preserves each Teacher's contribution along its own functional direction while minimizing total response energy:

$$\min_R \frac{1}{2}\|R\|_F^2 \qquad \text{s.t.} \qquad \langle R_i, R\rangle_F = \|R_i\|_F^2, \quad i = 1, \ldots, T.$$

The minimum-norm solution lies in the span of the Teacher responses, so let

$$R = \sum_{i=1}^{T} \alpha_i R_i.$$

Defining

$$\mathbf{K}_{ij} = \langle R_i, R_j\rangle_F, \qquad d_i = \|R_i\|_F^2,$$

the constraints yield

$$\alpha = \mathbf{K}^\dagger d,$$

where $\mathbf{K}^\dagger$ is the Moore–Penrose pseudoinverse. Thus, the coefficients depend directly on cross-Teacher functional overlap.

Direct addition as the orthogonal limit. If the Teacher responses are functionally orthogonal, $\langle R_i, R_j\rangle_F = 0$ for $i \neq j$, then

$$\mathbf{K} = \mathrm{diag}(d_1, \ldots, d_T), \qquad \alpha_i = 1.$$

The minimum-norm response therefore reduces to

$$R = \sum_{i=1}^{T} R_i = A\left(\sum_{i=1}^{T} \Delta W_i\right)^\top.$$

Hence, unit-weight parameter addition is recovered exactly when the Teacher responses are functionally orthogonal.

For two Teachers with equal response magnitudes, let

$$c = \frac{\langle R_1, R_2 \rangle_F}{\|R_1\|_F \|R_2\|_F}$$

denote their functional cosine similarity. Symmetry gives

$$\alpha_1 = \alpha_2 = \frac{1}{1+c}.$$

Thus, orthogonal responses ($c = 0$) recover unit-weight addition, while positive overlap progressively reduces the coefficients; identical responses ($c = 1$) yield $\alpha_1 = \alpha_2 = 1/2$.

This analysis places the composition rule used by FGC in a broader functional-overlap framework: direct addition is the low-overlap limit, while stronger overlap would call for redundancy-aware weighting. FGC retains the simpler unit-weight rule and introduces no cross-Teacher coefficient estimation.

C Methods and Experiments

C.1 Vision-Side Instantiation

For single-Teacher transfer, we additionally apply the Base-activation principle to selected attention output projections in the visual encoder. For projection l, let A_l denote the corresponding Base visual activations. We compute the Base-relative functional magnitude

$$E_{l,j} = \frac{\|\Delta W_{l,j}\|_{A_l}}{\|W_{B,l,j}\|_{A_l}}.$$

Unlike the language-side Functional Calibration, which amplifies differences in Base-relative functional magnitude, the visual instantiation performs response equalization to prevent a small number of highly responsive channels from dominating the transferred update. We normalize $E_{l,j}$ by its within-projection median and define

$$s_{l,j} = \left(\frac{E_{l,j}}{\mathrm{median}_k \, E_{l,k}} \right)^{\beta - 1}.$$

We set $\beta = 0.5$ and clip $s_{l,j}$ to $[0.5, 2.0]$. The visual update is then calibrated as

$$\widetilde{\Delta W}_{l,j} = s_{l,j} \Delta W_{l,j},$$

followed by Functional Rebase. We apply this vision-specific instantiation to the attention output projections of the first eight visual blocks. It is used only for single-Teacher transfer; multi-Teacher composition retains the Base visual backbone.

C.2 Sensitivity to Calibration Strength

We evaluate the sensitivity of FGC to the calibration exponent α, which controls how strongly differences in Base-relative functional magnitude affect channel-wise update scaling. Figure 2 reports results for $\alpha \in \{0.125, 0.25, 0.5, 0.75, 1.0\}$ with all other settings fixed.

Mathematical reasoning favors moderate calibration, with MathVista and MathVision peaking at $\alpha = 0.25$, whereas V* and HR-Bench vary less across the evaluated range. We therefore use $\alpha = 0.25$ throughout the main experiments.

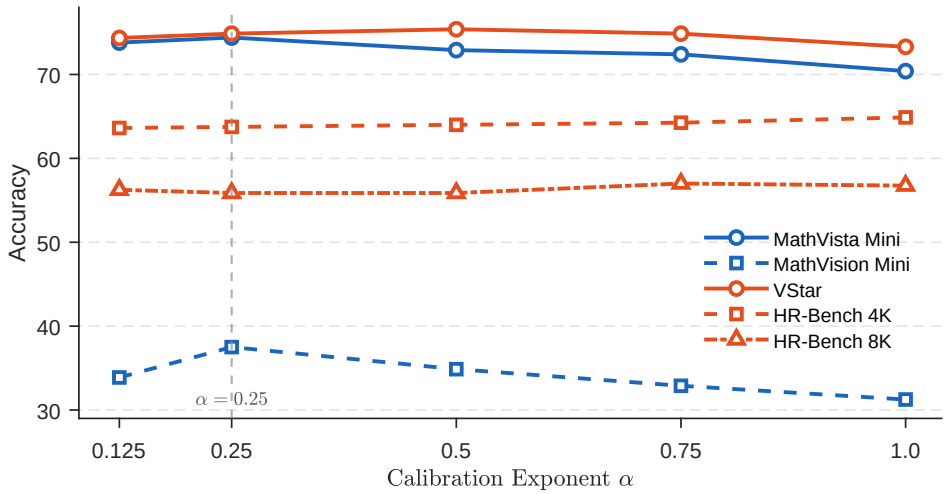

Figure 2: Sensitivity to the calibration exponent α. Mathematical reasoning favors moderate calibration, while visual perception is less sensitive across the evaluated range.