MERGEHEIR: MITIGATING MULTIMODAL HALLUCINATIONS AS THE TAX OF MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Model merging consolidates task-specialized experts into a single deployable model. However, we show that such capability consolidation incurs a merging tax of increased hallucination: across 8 model-merging methods, every merged checkpoint exhibits a higher hallucination rate than the average of its constituent experts. An intuitive approach is to adapt existing hallucination-mitigation methods to the post-merge model, yet this unconstrained adaptation disrupts inherited capabilities, creating a tension between hallucination mitigation and expertise retention. To tackle this challenge, we introduce MergeHEIR, a post-merge adaptation framework designed to reduce this merging tax while preserving expertise inherited from initial experts. Using small expert-task calibration sets, MergeHEIR constructs layer-wise null-space projectors via SVD from task-specific activations collected from the merged checkpoint, and periodically projects the accumulated post-merge displacement onto the resulting null spaces to preserve inherited expertise. Theoretically, we establish minimum-distortion and maximum-dimensionality guarantees, characterize the threshold-controlled adaptation-retention trade-off, and extend perturbation guarantees beyond finite calibration data. Across 24 paired comparisons spanning three MLLM configurations and 8 model-merging methods, MergeHEIR consistently mitigates hallucination while largely preserving inherited expertise, demonstrating a more favorable hallucination-retention trade-off.

1 INTRODUCTION

Model merging consolidates independently specialized models into a single deployable checkpoint, providing a practical way to combine complementary capabilities without jointly retraining a unified model on all constituent data (Wortsman et al., 2022; Ilharco et al., 2023). This paradigm is particularly appealing for multimodal large language models (MLLMs), for which capabilities such as chart understanding, geometry, optical character recognition, and visual question answering are often distributed across specialized checkpoints (Wei et al., 2025). However, capability consolidation alone does not guarantee deployment reliability.

We first investigate visual hallucination after model merging and reveal a systematic *merging tax*: across eight merging methods and two Qwen2.5-VL (Bai et al., 2025) scales, every evaluated merged checkpoint exhibits a higher AMBER hallucination rate than the average of its constituent experts, as shown in Figure 1. The mean rate increases from 33.48 to 54.35 for Qwen2.5-VL-3B and from 20.83 to 25.41 for Qwen2.5-VL-7B. Notably, all constituent experts share the same frozen vision encoder and differ only through language-side LoRA specialization. We attribute the resulting tax to *visual-evidence utilization collision*: expert specialization forms co-adapted language-side computations for using visual evidence, while model merging can perturb the cross-layer interactions through which this evidence influences generation. As a result, the encoded visual representation may remain intact while its downstream use becomes less effective, increasing susceptibility to visually unsupported generation. Appendix B provides a mechanistic interpretation of this effect. These results show that consolidating complementary capabilities does not guarantee reliable visual grounding, making hallucination correction a necessary post-merge stage.

A natural remedy is to adapt the merged model through hallucination-oriented multimodal preference optimization (Wang et al., 2024a). Yet this correction introduces another problem. The same

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.

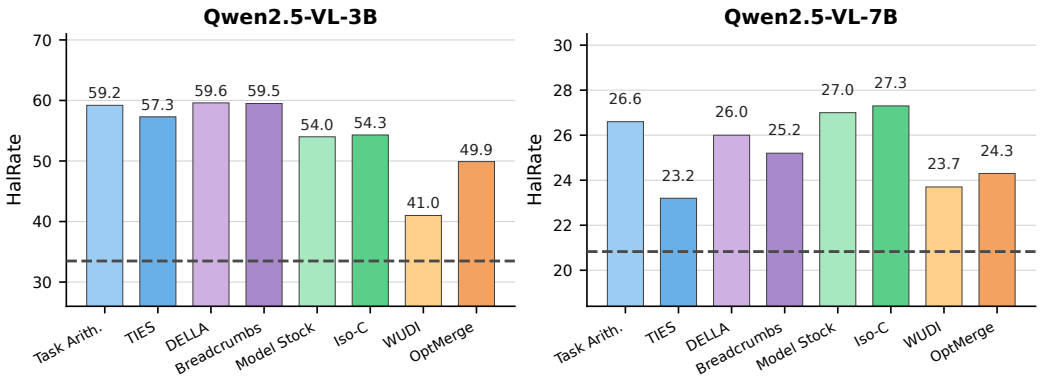


Figure 1: Merging tax manifested as elevated hallucination rates. We evaluate eight merging methods with AMBER (Wang et al., 2023) on Qwen2.5-VL-3B and Qwen2.5-VL-7B. Dashed lines indicate the corresponding expert-average baselines, with HalRates of 33.5 and 20.8, respectively. Every evaluated merged checkpoint lies above its corresponding baseline.

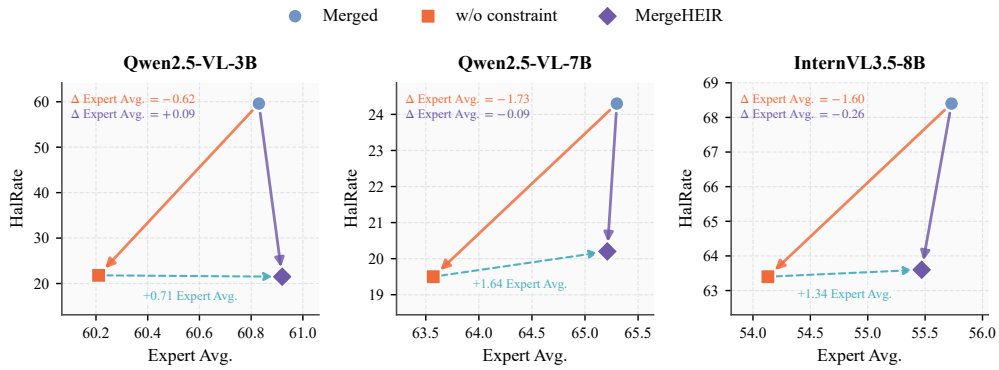


Figure 2: Matched comparison of post-merge adaptation outcomes across three model configurations. HalRate denotes the AMBER hallucination rate, while Expert Avg. is the mean performance across the seven expert-task benchmarks. Unconstrained adaptation achieves comparable hallucination mitigation but incurs a substantially larger reduction in Expert Avg., whereas MergeHEIR better preserves inherited expertise while achieving effective hallucination mitigation.

updates used to strengthen visual grounding can inadvertently disrupt the heterogeneous capabilities inherited from the constituent experts. Unconstrained adaptation that optimizes hallucination alone can therefore degrade expert-task performance. Figure 2 visualizes this tension across the three model configurations: unconstrained adaptation achieves hallucination mitigation comparable to MergeHEIR but consistently incurs a larger reduction in Expert Avg. This degradation reflects adaptation-induced interference with the expertise inherited from the constituent models. The resulting challenge is not merely to reduce hallucination, but to do so without undoing the capabilities that merging was intended to consolidate. Post-merge hallucination mitigation therefore cannot be treated as behavioral correction alone; it must improve visual grounding while limiting degradation of the inherited expert capabilities. This leads to our central question: *how can hallucination-oriented preference optimization mitigate post-merge hallucination while limiting interference with the expert capabilities consolidated through model merging?*

To address this question, we propose MergeHEIR, a post-merge adaptation framework that coordinates hallucination correction with the retention of inherited expertise. It combines hallucination-oriented preference optimization with activation-derived null-space projection to suppress adaptation-induced interference with inherited expertise. Using only a small calibration set, MergeHEIR constructs null spaces from the merged model's activation statistics and periodically projects accumulated post-merge updates onto the corresponding null spaces to preserve expert-task capabilities. We further show that this is a minimum-distortion projection onto a

maximum-dimensional admissible subspace, derive pooled and task-specific representation perturbation bounds, characterize the threshold-controlled adaptation–retention trade-off, and extend the guarantee to the corresponding activation distribution.

We evaluate MergeHEIR through 24 paired comparisons spanning three MLLM configurations and eight merging methods. Across these settings, MergeHEIR consistently mitigates hallucination while largely preserving the expert capabilities consolidated through model merging. Matched comparisons against unconstrained post-merge adaptation further show that null-space projection consistently improves expert retention while maintaining effective hallucination mitigation, yielding a more favorable hallucination–retention trade-off.

Our contributions are threefold:

- We identify a two-stage reliability problem in multimodal model merging: capability consolidation incurs systematic *merging tax* from disrupted language-side visual-evidence utilization, while unconstrained hallucination correction can interfere with inherited expertise.
- We propose MergeHEIR, a capability-preserving framework that constructs SVD-derived activation-insensitive subspaces and periodically projects accumulated updates onto them. We also establish theoretical guarantees on adaptation geometry and perturbation control.
- We validate the effectiveness and robustness of MergeHEIR through 24 paired comparisons across three MLLM configurations and eight merging methods, complemented by controlled comparisons and threshold analyses that isolate the role of the constraint.

2 Problem Formulation

Figure 1 reveals a systematic *merging tax*: across the evaluated merging methods, the resulting checkpoints exhibit consistently higher hallucination rates than their corresponding constituent-expert averages. This motivates hallucination-oriented post-merge adaptation. However, such adaptation modifies parameters that also carry the consolidated expert capabilities, creating a conflict between hallucination correction and expert retention. Post-merge adaptation thus poses a two-objective problem: improving visual grounding while retaining expert-task performance. We formulate this problem as hallucination correction under expert-retention constraints.

2.1 Merged model and retention targets

Let $\mathcal{T} = \{\mathrm{Chart}, \mathrm{Geometry}, \mathrm{OCR}, \mathrm{VQA}\}$ denote the set of expert tasks, and let θ_t denote the parameters of the expert model specialized for task $t \in \mathcal{T}$. A model-merging operation combines these expert parameters into a single checkpoint:

$$\theta_m = \mathrm{Merge}(\{\theta_t\}_{t\in\mathcal{T}})\,. \tag{1}$$

Post-merge adaptation initializes the trainable parameters θ from θ_m, while θ_m remains fixed as the reference checkpoint. For each task t, let $\mathcal{D}_t$ denote a representative expert-task distribution. The collection $\{\mathcal{D}_t\}_{t\in\mathcal{T}}$ captures the heterogeneous expert capabilities that should remain stable.

2.2 Hallucination correction under expert-retention constraints

Let $\mathcal{L}_{\mathrm{hall}}(\theta)$ denote a hallucination-oriented adaptation objective applied to the merged checkpoint. Optimizing this objective alone imposes no restriction on behavioral changes over the expert-task distributions. We therefore formulate post-merge adaptation as

$$\min_{\theta}\ \mathcal{L}_{\mathrm{hall}}(\theta) \quad \text{s.t.} \quad \mathcal{R}_t(\theta, \theta_m) \le \epsilon_t, \qquad \forall t \in \mathcal{T}, \tag{2}$$

where $\mathcal{R}_t(\theta, \theta_m)$ measures the adaptation-induced functional deviation from the original merged checkpoint on $\mathcal{D}_t$, and ϵ_t denotes the tolerated deviation. MergeHEIR instantiates these retention constraints through activation-derived null-space projection, as described in the next section.

3 Method

MergeHEIR combines hallucination-oriented multimodal preference optimization with activation-derived null-space projection. The preference objective drives hallucination mitigation, while the

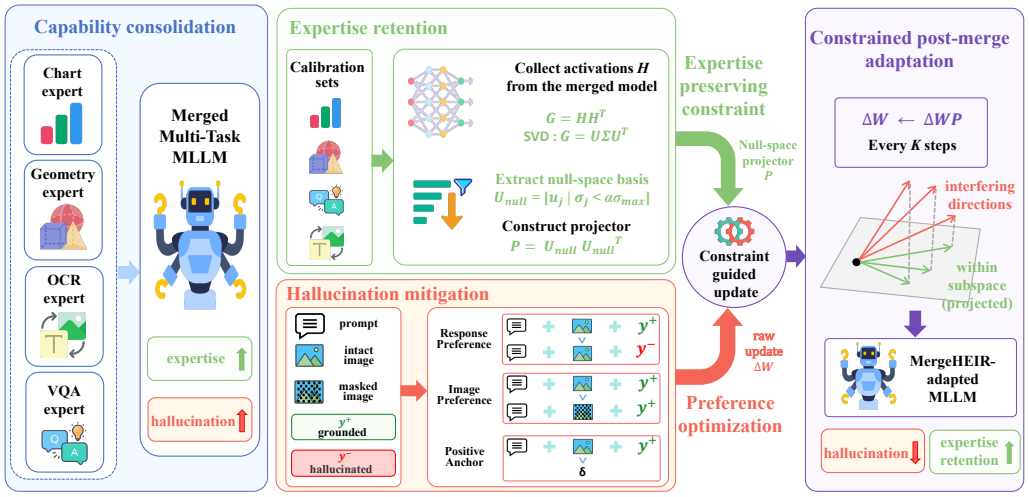


Figure 3: Overview of MergeHEIR. Four task-specialized experts are first consolidated into a merged checkpoint θ_m. Hallucination-oriented preference optimization produces corrective updates, while expert-task calibration activations define layer-wise null-space projectors. MergeHEIR periodically projects the accumulated parameter displacement onto the corresponding null spaces, reducing hallucination while retaining inherited expertise.

null-space constraint restricts accumulated parameter changes to directions that produce limited perturbation on the expert-task calibration data. Figure 3 summarizes the complete framework.

3.1 Hallucination-oriented preference optimization

Let π_θ denote the trainable policy initialized from the merged checkpoint θ_m, and let π_m denote the frozen reference policy. Each preference sample is given by (x, I, y^+, y^-), where y^+ is grounded and y^- is hallucinated. Following mDPO (Wang et al., 2024a), we define the reference-relative score as $s_\theta(y \mid x, I) = \beta \log \frac{\pi_\theta(y|x,I)}{\pi_m(y|x,I)}$. To strengthen image-conditioned preference, we further construct a perturbed image $\widetilde{I}$ via random region masking following OPA-DPO (Yang et al., 2025). The preference objective consists of

$$\begin{aligned}
\mathcal{L}_{\text{DPO}} &= -\mathbb{E} \log \sigma\big(s_\theta(y^+ \mid x, I) - s_\theta(y^- \mid x, I)\big), \\
\mathcal{L}_{\text{img}} &= -\mathbb{E} \log \sigma\Big(s_\theta(y^+ \mid x, I) - s_\theta(y^+ \mid x, \widetilde{I})\Big), \\
\mathcal{L}_{\text{anc}} &= -\mathbb{E} \log \sigma\big(s_\theta(y^+ \mid x, I) - \delta\big).
\end{aligned} \tag{3}$$

Here, $\mathcal{L}_{\text{DPO}}$ favors grounded responses over hallucinated ones, $\mathcal{L}_{\text{img}}$ strengthens reliance on the original visual evidence, and $\mathcal{L}_{\text{anc}}$ prevents the preference margin from increasing solely through suppression of the rejected response. The complete objective is

$$\mathcal{L}_{\text{pref}} = \mathcal{L}_{\text{DPO}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anc}}. \tag{4}$$

3.2 Expert-task null-space projection

The hallucination-oriented objective specifies the desired behavioral correction, but does not explicitly constrain its impact on the expert capabilities consolidated in the merged model. To address this, we propose to appropriately regulate the model update during training to reduce interference with the learned expertise. Specifically, MergeHEIR constructs layer-wise null spaces from expert-task activations and constrains post-merge model updates for hallucination correction to these spaces, to preserve the acquired expert capabilities.

Preservation condition. Consider a linear layer with weight matrix W_m in the original merged checkpoint and a post-merge parameter update ΔW. For an input activation h, the induced change

in the layer output is

$$(W_m + \Delta W)h - W_m h = \Delta W h. \tag{5}$$

For each constituent expert task $t \in \mathcal{T}$, we form a small calibration set by randomly sampling task examples. We feed these examples through the merged model and collect the resulting input activations into a matrix H_t. We then concatenate the task-specific matrices along the sample dimension:

$$H = \left[H_1, H_2, \ldots, H_{|\mathcal{T}|}\right]. \tag{6}$$

The joint feature matrix H characterizes the activation subspace induced by the expert-task calibration data. An update satisfying

$$\Delta W H = 0 \tag{7}$$

leaves the original layer outputs unchanged on all collected activations. This condition motivates restricting post-merge updates to directions orthogonal to the activation subspace represented by the joint matrix H consisting of different task features.

Null-space construction. Directly extracting a left null-space basis from the activation matrix H becomes costly when a large number of token activations are collected. Since H and $HH^\top$ share the same left null space, we instead operate on the smaller activation Gram matrix and construct the projector via singular value decomposition (SVD):

$$G = HH^\top = U\Sigma U^\top, \qquad U_{\text{null}} = [u_j \mid \sigma_j < \alpha\sigma_{\max}], \qquad P = U_{\text{null}}U_{\text{null}}^\top. \tag{8}$$

Here, α defines an approximate null space by selecting singular vectors satisfying $\sigma_j/\sigma_{\max} < \alpha$. We consider $\alpha \in [5 \times 10^{-6}, 5 \times 10^{-2}]$ and use $\alpha = 5 \times 10^{-4}$ by default to balance adaptation flexibility and expert-capability preservation. MergeHEIR then projects the post-merge update as $\Delta W \leftarrow \Delta W P$, suppressing components aligned with activation-sensitive directions and confining adaptation to directions weakly represented in the expert-task activation space.

3.3 Geometric Optimality and Perturbation Control

The projection in MergeHEIR must preserve useful hallucination-correction updates while limiting changes to expert-task representations. We first establish the minimum-distortion property within a fixed subspace, then examine spectral perturbation control and threshold-dependent adaptation flexibility, and finally extend this control beyond calibration data. Let P be any fixed orthogonal projection matrix, and let $\mathcal{S}_P = \{Z : Z = ZP\}$ denote its induced admissible displacement subspace.

Proposition 3.1 (Minimum-distortion projection)**.** *For any accumulated displacement ΔW,*

$$\Delta W P = \arg\min_{Z \in \mathcal{S}_P} \|Z - \Delta W\|_F^2. \tag{9}$$

Moreover, $\Delta W = \Delta W P + \Delta W(I - P)$ is an orthogonal decomposition, so $\|\Delta W P\|_F \le \|\Delta W\|_F$.

Proposition 3.1 establishes that null-space projection enforces the subspace constraint with minimum Frobenius-norm distortion. MergeHEIR therefore retains all admissible components of the preference update, avoiding unnecessary suppression of the learned hallucination-correction update.

Theorem 3.2 (Spectral perturbation control)**.** *The range of the projector in Eq.* (8) *is maximum-dimensional among subspaces satisfying $v^\top G v < \alpha\sigma_{\max}\|v\|_2^2$ for every nonzero v. For any displacement ΔW,*

$$\|\Delta W P H_t\|_F^2 \le \|\Delta W P H\|_F^2 \le \alpha\sigma_{\max}\|\Delta W P\|_F^2, \qquad \forall t \in \mathcal{T}. \tag{10}$$

Theorem 3.2 implies that choosing an appropriately small α allows MergeHEIR to limit layer-output perturbations on expert-task calibration activations. This reduces interference with inherited task representations during hallucination correction. Meanwhile, the maximum-dimensional construction simultaneously retains as many admissible update directions as the sensitivity constraint permits.

Proposition 3.3 (Threshold-controlled adaptation geometry)**.** *For $0 \le \alpha_1 \le \alpha_2$, let $P_1 = P_{\alpha_1}$, $P_2 = P_{\alpha_2}$, and $Q = P_2 - P_1$. Then* $\text{range}(P_1) \subseteq \text{range}(P_2)$ *and* $\text{rank}(P_1) \le \text{rank}(P_2)$*. Moreover, for any accumulated displacement ΔW,*

$$\|\Delta W P_2\|_F^2 = \|\Delta W P_1\|_F^2 + \|\Delta W Q\|_F^2, \qquad \|\Delta W P_2 H\|_F^2 = \|\Delta W P_1 H\|_F^2 + \|\Delta W Q H\|_F^2. \tag{11}$$

Table 1: Expert-task results on Qwen2.5-VL models.

Methods	Stage	Chart	Geometry		OCR		VQA		Avg.
		ChartQA	MATH-Vision	Geometry3K	TextVQA	DocVQA	GQA	OKVQA	
Task Arithmetic	Merged	83.96	20.72	48.25	78.67	92.66	60.98	40.48	60.82
	MergeHEIR	84.20	21.38	48.75	78.72	92.66	60.90	40.93	61.08
TIES Merging	Merged	83.92	21.71	48.75	78.53	92.54	61.15	40.18	60.97
	MergeHEIR	83.96	20.72	47.92	78.67	92.67	60.96	40.09	60.71
DELLA Merging	Merged	83.68	21.05	48.75	78.55	92.57	60.86	40.35	60.83
	MergeHEIR	84.00	21.05	48.42	78.75	92.69	60.70	40.84	60.92
Breadcrumbs Merging	Merged	83.96	20.39	48.42	78.87	92.59	61.18	40.34	60.82
	MergeHEIR	84.08	21.05	48.09	78.80	92.71	61.00	40.63	60.91
Model Stock	Merged	83.56	20.72	48.09	78.15	92.44	60.35	41.13	60.63
	MergeHEIR	83.48	21.05	47.75	78.20	92.39	59.87	41.70	60.64
Iso-C	Merged	83.44	21.05	48.42	78.14	92.48	60.35	41.04	60.70
	MergeHEIR	83.76	21.38	47.25	78.38	92.46	60.37	41.36	60.71
WUDI Merging	Merged	84.08	21.05	48.59	79.24	92.75	62.28	39.41	61.06
	MergeHEIR	84.36	20.07	48.25	79.56	93.23	62.23	39.58	61.04
OptMerge	Merged	84.08	21.71	48.09	78.85	92.55	61.70	40.19	61.02
	MergeHEIR	84.48	21.71	47.59	79.05	92.62	61.28	40.69	61.06

(a) Qwen2.5-VL-3B

Methods	Stage	Chart	Geometry		OCR		VQA		Avg.
		ChartQA	MATH-Vision	Geometry3K	TextVQA	DocVQA	GQA	OKVQA	
Task Arithmetic	Merged	86.32	26.32	54.41	83.26	94.62	60.94	44.42	64.32
	MergeHEIR	85.76	26.32	53.58	82.71	94.76	60.90	44.21	64.03
TIES Merging	Merged	84.12	25.00	53.41	82.87	94.73	60.60	43.79	63.50
	MergeHEIR	82.52	23.36	52.91	82.21	94.66	60.59	43.04	62.76
DELLA Merging	Merged	87.24	25.99	54.91	83.46	94.50	61.52	45.44	64.72
	MergeHEIR	87.16	24.34	55.41	83.84	94.82	61.54	44.81	64.56
Breadcrumbs Merging	Merged	87.24	26.32	54.58	83.45	94.44	61.40	45.40	64.69
	MergeHEIR	87.16	24.67	55.07	83.96	94.80	61.62	44.64	64.56
Model Stock	Merged	83.72	23.03	52.41	82.84	94.71	60.54	43.45	62.96
	MergeHEIR	81.56	23.03	52.25	81.95	94.70	60.41	43.30	62.46
Iso-C	Merged	83.40	23.03	50.75	82.87	94.68	60.52	43.52	62.68
	MergeHEIR	82.00	22.70	52.25	82.03	94.72	60.52	43.04	62.47
WUDI Merging	Merged	87.52	27.30	54.08	83.12	94.41	63.40	48.74	65.51
	MergeHEIR	86.52	26.97	55.41	82.33	94.51	63.48	48.39	65.37
OptMerge	Merged	87.44	27.30	54.24	83.18	94.41	62.55	47.99	65.30
	MergeHEIR	87.00	27.63	53.74	82.86	94.74	62.73	47.75	65.21

(b) Qwen2.5-VL-7B

Proposition 3.3 shows that tuning α allows MergeHEIR to balance preference-update retention against calibration perturbation. An appropriate threshold thus avoids overly restricting hallucination-correction updates while limiting perturbations to expert-task representations. We evaluate this adaptation–retention balance in Figure 5.

Corollary 3.4 (Population extension)**.** *Let N be the number of columns in H, let $\widehat{G} = \frac{1}{N} H H^\top$ with $\widehat{\sigma}_{\max} = \lambda_{\max}(\widehat{G})$, and let $\Gamma = \mathbb{E}[hh^\top]$. If $\|\widehat{G} - \Gamma\|_2 \le \varepsilon$, then*

$$\mathbb{E}\big[\|\Delta W P h\|_2^2\big] \le (\alpha \widehat{\sigma}_{\max} + \varepsilon)\, \|\Delta W P\|_F^2. \tag{12}$$

Corollary 3.4 indicates that a projector constructed from representative expert-task calibration data controls expected perturbations on unseen activations from the same distribution. MergeHEIR thus limits adaptation-induced changes to expert-task representations beyond the calibration set.

Together, these results provide a theoretical account of how null-space projection in MergeHEIR controls perturbations to expert-task representations during hallucination-oriented adaptation. The projection minimizes update distortion within the selected subspace, which is maximum-dimensional under the prescribed sensitivity constraint. The threshold analysis characterizes the balance between adaptation flexibility and representation stability, while the population extension carries perturbation control beyond calibration data under the stated spectral approximation condition. Complete proofs and fixed-prefix output analysis are provided in Appendix C.

Table 2: Hallucination-task results on Qwen2.5-VL models.

Methods	Stage	AMBER			Object Hal		POPE Adversarial		GPT-4 Eval
		CHAIR↓	HalRate↓	Cog↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓	Acc.↑	F1↑	SHR↓
Task Arithmetic	Merged	8.5	59.2	7.0	31.7	8.5	82.2	83.5	28.3
	MergeHEIR	4.7	20.9	1.8	23.3	6.1	85.8	86.0	25.4
TIES Merging	Merged	8.5	57.3	6.4	34.0	9.0	82.2	83.5	28.3
	MergeHEIR	5.1	23.0	2.0	21.7	5.9	85.8	86.1	26.5
DELLA Merging	Merged	8.8	59.6	7.1	36.0	9.0	82.1	83.4	29.9
	MergeHEIR	4.8	21.5	1.9	21.7	6.4	85.8	86.0	27.3
Breadcrumbs Merging	Merged	8.7	59.5	7.0	34.3	9.2	81.9	83.2	28.5
	MergeHEIR	5.0	22.4	1.9	22.0	6.7	85.7	86.0	25.2
Model Stock	Merged	8.4	54.0	6.1	34.3	8.6	82.9	83.9	27.8
	MergeHEIR	4.9	20.4	1.8	22.7	6.0	86.6	86.5	25.6
Iso-C	Merged	8.4	54.3	6.1	36.3	8.6	82.8	83.9	28.4
	MergeHEIR	5.4	22.6	2.1	21.3	5.9	85.7	86.0	26.0
WUDI Merging	Merged	8.0	41.0	4.5	40.3	9.9	85.4	85.2	31.0
	MergeHEIR	4.9	21.2	2.0	23.7	5.6	89.1	87.9	27.2
OptMerge	Merged	8.3	49.9	5.7	37.3	10.6	83.5	84.2	28.5
	MergeHEIR	5.2	25.2	2.5	22.0	6.9	87.3	87.0	25.1

(a) Qwen2.5-VL-3B

Methods	Stage	AMBER			Object Hal		POPE Adversarial		GPT-4 Eval
		CHAIR↓	HalRate↓	Cog↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓	Acc.↑	F1↑	SHR↓
Task Arithmetic	Merged	4.9	26.6	1.6	39.3	8.4	83.4	80.7	28.8
	MergeHEIR	4.0	21.5	1.1	23.3	6.2	87.0	86.0	25.9
TIES Merging	Merged	4.8	23.2	1.1	39.3	8.3	85.5	84.4	29.7
	MergeHEIR	4.1	20.3	0.9	24.0	5.9	89.2	88.0	27.5
DELLA Merging	Merged	4.9	26.0	1.5	42.3	9.2	84.0	81.8	28.5
	MergeHEIR	4.3	21.4	1.0	24.7	6.7	87.0	86.5	26.1
Breadcrumbs Merging	Merged	4.7	25.2	1.6	41.0	8.8	83.9	81.7	29.4
	MergeHEIR	4.2	21.6	1.2	24.3	6.4	87.5	86.9	25.8
Model Stock	Merged	5.0	27.0	1.6	42.7	8.9	83.2	80.4	28.3
	MergeHEIR	4.2	22.3	1.2	25.0	7.1	86.8	85.7	26.0
Iso-C	Merged	5.2	27.3	1.7	39.7	8.8	83.4	80.5	28.7
	MergeHEIR	3.7	20.8	1.0	25.3	6.9	86.7	85.5	26.2
WUDI Merging	Merged	4.9	23.7	1.3	37.3	8.7	86.0	85.1	29.8
	MergeHEIR	3.9	20.4	1.1	22.7	6.6	89.4	88.9	26.9
OptMerge	Merged	4.9	24.3	1.4	41.0	9.0	86.1	84.9	30.1
	MergeHEIR	4.2	20.2	1.2	25.0	6.2	89.7	89.3	26.8

(b) Qwen2.5-VL-7B

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models, merging methods, and data. We conduct experiments on three MLLM configurations: Qwen2.5-VL-3B, Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3.5-8B (Wang et al., 2025). For each configuration, we merge four task-specialized experts covering chart understanding, geometry reasoning, OCR, and visual question answering using eight representative methods: Task Arithmetic, TIES, DELLA, Breadcrumbs, Model Stock, Iso-C, WUDI, and OptMerge. Post-merge adaptation uses 10,000 multimodal preference pairs sampled from RLAIF-V (Yu et al., 2025), while layer-wise null-space projectors are constructed from 500 calibration examples per expert task.

Evaluation. We evaluate both the retention of inherited expertise and the mitigation of hallucination. Expert-task performance is measured on ChartQA (Masry et al., 2022) for chart understanding; MATH-Vision (Wang et al., 2024b) and Geometry3K (Lu et al., 2021) for geometry reasoning; TextVQA (Singh et al., 2019) and DocVQA (Mathew et al., 2021) for OCR; and GQA (Hudson & Manning, 2019) and OKVQA (Marino et al., 2019) for visual question answering. We report the unweighted mean over these seven benchmarks as *Expert Avg.* Hallucination behavior is evaluated using AMBER (Wang et al., 2023) for multidimensional generative hallucination, Object HalBench (Rohrbach et al., 2018) for object hallucination in open-ended image descriptions, POPE (Li et al., 2023) for object-existence judgments, and GPT-4-assisted evaluation (Zhao et al., 2023) for fine-grained hallucinations in generated image descriptions. The latter reports the Sentence-level Hallucination Ratio (SHR), where lower values indicate fewer hallucinated sentences. To reduce sampling-induced variance and ensure consistent checkpoint-level comparisons across all benchmarks, we generate every model response with greedy decoding throughout evaluation.

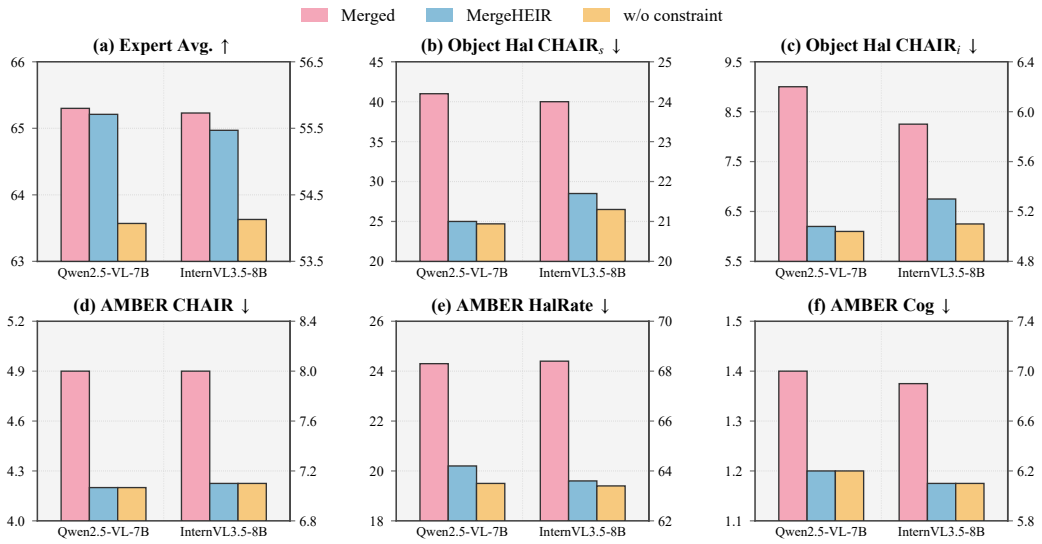


Figure 4: Constraint ablation on Qwen2.5-VL-7B and InternVL3.5-8B. MergeHEIR better preserves inherited expertise while achieving hallucination mitigation comparable to unconstrained adaptation.

Implementation details. We perform LoRA-based post-merge adaptation for two epochs with a global batch size of 32 and a DPO scaling coefficient of $\beta = 0.1$. LoRA is applied only to non-visual modules, while all visual-side parameters remain frozen. For Qwen2.5-VL-3B, we use $(r, \alpha_{\text{LoRA}}) = (32, 64)$ and a learning rate of 2×10^{-4}; for Qwen2.5-VL-7B and InternVL3.5-8B, we use $(r, \alpha_{\text{LoRA}}) = (16, 32)$ and a learning rate of 1×10^{-4}. We set $\lambda_{\text{img}} = 0.2$, $\lambda_{\text{anc}} = 1.0$, and $\delta = 0$, with a random image-masking ratio of 0.3. The accumulated parameter displacement is projected every five optimizer steps using a relative spectral threshold of $\alpha = 5 \times 10^{-4}$. Additional experimental details are provided in Appendix D.

4.2 Main Results

Tables 1 and 2 report the complete Qwen2.5-VL results across all eight merging methods. Each method is evaluated before and after post-merge adaptation, enabling direct paired comparison. Appendix E reports complete InternVL3.5-8B evaluations, POPE results across all three splits, and compatibility experiments with RLAIF-V (Yu et al., 2025) and OPA-DPO (Yang et al., 2025). Together, these results examine whether the retention benefit of null-space projection generalizes across model families, evaluation settings, and alternative preference objectives.

Consistent hallucination mitigation. MergeHEIR reduces AMBER HalRate in all 24 matched comparisons across three model configurations and eight merging methods. The improvement is most pronounced on Qwen2.5-VL-3B, for which HalRate decreases from 41.0–59.6 before adaptation to 20.4–25.2 afterward. Consistent reductions are also observed on Qwen2.5-VL-7B and InternVL3.5-8B. Object HalBench further exhibits substantial reductions in both CHAIR$_s$ and CHAIR$_i$ across most configurations. POPE accuracy and F1 improve across all evaluated splits and all 24 comparisons. Finally, the GPT-4-assisted evaluation provides complementary judge-based evidence: SHR decreases in all 24 matched comparisons, with mean reductions of 2.80, 2.76, and 3.26 points on Qwen2.5-VL-3B, Qwen2.5-VL-7B, and InternVL3.5-8B, respectively. Together, these results demonstrate that MergeHEIR consistently mitigates hallucination across complementary generative, discriminative, and judge-based evaluations.

Preservation of inherited expertise. The broad hallucination improvements are achieved with limited changes to the expert capabilities inherited by the merged checkpoints. On Qwen2.5-VL-3B, MergeHEIR achieves substantial hallucination mitigation while leaving Expert Avg essentially unchanged. Qwen2.5-VL-7B and InternVL3.5-8B exhibit the same favorable pattern, with broad hallucination improvements accompanied by mean Expert Avg decreases of only 0.28 and 0.31 points, re-

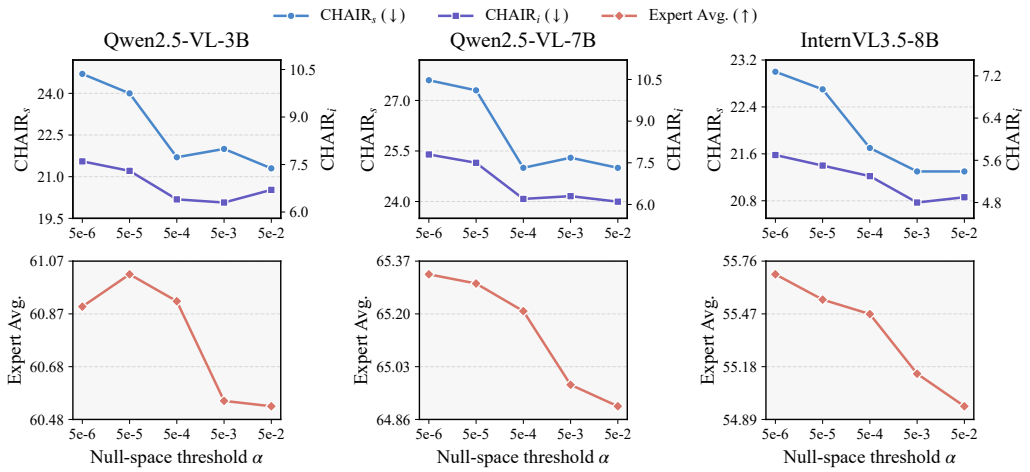


Figure 5: Sensitivity to the null-space threshold α across three model configurations. The upper panels report Object HalBench CHAIR$_s$ and CHAIR$_i$, while the lower panels report Expert Avg.

spectively. These results demonstrate that MergeHEIR delivers broad hallucination mitigation across model families without substantially compromising the capabilities consolidated during merging. The matched ablation in Figure 4 tests whether null-space projection improves expert retention without materially weakening hallucination mitigation on Qwen2.5-VL-7B and InternVL3.5-8B.

4.3 Ablation Studies

Effect of null-space projection. We conduct matched ablations on two MLLM configurations. Within each configuration, both variants use the same training setup, differing only in null-space projection. Across these settings, removing null-space projection decreases Expert Avg. by 1.64 and 1.34 points relative to MergeHEIR on Qwen2.5-VL-7B and InternVL3.5-8B, respectively, while yielding comparable hallucination performance. This pronounced asymmetry shows that unconstrained adaptation incurs substantially greater damage to inherited expertise for only marginal behavioral gains. By restricting the accumulated post-merge displacement to the expert-task null spaces, MergeHEIR achieves a markedly better hallucination–retention trade-off, preserving expert performance without forfeiting the primary hallucination improvements.

Sensitivity to the null-space threshold. Figure 5 illustrates the effect of the relative spectral threshold α on hallucination mitigation and expert retention. Across the three model configurations, larger values of α generally yield lower CHAIR$_s$ and CHAIR$_i$, whereas smaller values preserve higher Expert Avg. This trend is consistent with Proposition 3.3, showing that α controls the trade-off between hallucination mitigation and inherited-expertise preservation. We therefore use $\alpha = 5 \times 10^{-4}$ as the shared default because it achieves substantial hallucination mitigation while limiting degradation of inherited expert capabilities across model configurations.

5 Conclusion

This work investigates hallucination in multimodal model merging. We begin by identifying two coupled challenges: a systematic merging tax and expertise interference during post-merge correction, based on which we propose MergeHEIR, combining multimodal preference optimization with activation-derived null-space projection. We further analyze the mechanism underlying the merging tax and establish guarantees on adaptation geometry and representation perturbation control. Extensive experiments across three MLLM configurations, eight merging methods, and diverse benchmarks validate its effectiveness and generality. We highlight the importance of coordinating hallucination mitigation with inherited-expertise preservation for reliable model merging.

REFERENCES

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

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301, 2019.

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th international conference on computational linguistics*, pp. 1511–1520, 2022.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7824–7846, 2024. URL https://proceedings.mlr.press/v235/chen24bi.html.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10121–10143, 2025. URL https://proceedings.mlr.press/v267/cheng25h.html.

MohammadReza Davari and Eugene Belilovsky. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *Computer Vision – ECCV 2024*, pp. 270–287, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/09596.pdf.

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. URL https://arxiv.org/abs/2406.11617.

Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. MIRAGE: Assessing hallucination in multimodal reasoning chains of MLLM. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/b238324b309da12c7446d92c14db9f7e-Abstract-Conference.html.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9413–9422, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Du_AdaMMS_Model_Merging_for_Heterogeneous_Multimodal_Large_Language_Models_with_CVPR_2025_paper.html.

Hao Fang, Jinyu Li, Jiawei Kong, Tianqu Zhuang, Kuofeng Gao, Bin Chen, Shu-Tao Xia, and Yaowei Wang. Seeing through the chain: Mitigate hallucination in multimodal reasoning models via CoT compression and contrastive preference optimization. *arXiv preprint arXiv:2602.03380*, 2026. URL https://arxiv.org/abs/2602.03380.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations*, volume 2025, pp. 16366–16396, 2025.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16563–16577, 2025. doi: 10.18653/v1/2025.findings-acl.850. URL https://aclanthology.org/2025.findings-acl.850/.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. In *International Conference on Learning Representations*, volume 2025, pp. 3490–3511, 2025.

Zhehao Huang, Yuhang Liu, Baijiong Lin, Yixin Lou, Zhengbao He, Hanling Tian, Tao Li, and Xiaolin Huang. Rain-merging: A gradient-free method to enhance instruction following in large reasoning models with preserved thinking format. *arXiv preprint arXiv:2602.22538*, 2026.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_Question_Answering_CVPR_2019_paper.html.

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

Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. Model stock: All we need is just a few fine-tuned models. In *Computer Vision – ECCV 2024*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/06044.pdf.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5648–5656, 2018.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pp. 498–517. Springer, 2022.

Jiawei Kong, Hao Fang, Shunxiang Liao, Jinyu Li, Bin Chen, Hao Wu, Shu-Tao Xia, and Min Zhang. Reasoning matters: Mitigate hallucination in multimodal large reasoning models via reasoning-conditioned preference optimization. *arXiv preprint arXiv:2605.27906*, 2026.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 6774–6786, 2021. doi: 10.18653/v1/2021.acl-long.528. URL https://aclanthology.org/2021.acl-long.528/.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost Van De Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43177–43199, 2025. URL https://proceedings.mlr.press/v267/marczak25a.html.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, 2019. doi: 10.1109/CVPR.2019.00331. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013.

Yifan Niu, Han Xiao, Dongyi Liu, Nuo Chen, and Jia Li. Mitigating the safety alignment tax with null-space constrained policy optimization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GFyVxtyMvq.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024a. doi: 10.18653/v1/2024.emnlp-main.460. URL https://aclanthology.org/2024.emnlp-main.460/.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. URL https://arxiv.org/abs/2311.07397.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with the MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://openreview.net/forum?id=QWTCcxMpPA.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 184–193. IEEE, 2021.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024c.

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.

Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, Lu Hou, Sinan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. OptMerge: Unifying multimodal LLM capabilities and modalities via model merging. *arXiv preprint arXiv:2505.19892*, 2025. URL https://arxiv.org/abs/2505.19892.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya 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 *Proceedings of the 39th International Conference on Machine Learning*, pp. 23965–23998, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Jiulong Wu, Zhengliang Shi, Shuaiqiang Wang, Jizhou Huang, Dawei Yin, Lingyong Yan, Min Cao, and Min Zhang. Mitigating hallucinations in large vision-language models via entity-centric multimodal preference optimization. *arXiv preprint arXiv:2506.04039*, 2025. URL https://arxiv.org/abs/2506.04039.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/62868cc2fc1eb5cdf321d05b4b88510c-Abstract-Conference.html.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10610–10620, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024. doi: 10.1007/s11432-023-4118-6.

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. *arXiv preprint arXiv:2311.03099*, 2023. URL https://arxiv.org/abs/2311.03099.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19985–19995, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yu_RLAIF-V_Open-Source_AI_Feedback_Leads_to_Super_GPT-4V_Trustworthiness_CVPR_2025_paper.html.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

A Related Work

Model merging. Model merging consolidates independently fine-tuned models into a shared checkpoint without incurring the inference cost of an ensemble. Static merging methods include interpolation- and arithmetic-based approaches, such as model soups (Wortsman et al., 2022) and Task Arithmetic (Ilharco et al., 2023); sparsification-based approaches, including TIES (Yadav et al., 2023), DARE (Yu et al., 2023), DELLA (Deep et al., 2024), and Model Breadcrumbs (Davari & Belilovsky, 2024); and geometry-based or subspace-based approaches, such as Model Stock (Jang et al., 2024) and Iso-C (Marczak et al., 2025). Optimization-based methods instead learn merging coefficients or optimize task-vector combinations, including AdaMerging (Yang et al., 2024), WUDI (Cheng et al., 2025), and OptMerge (Wei et al., 2025). Recent methods further extend model merging to heterogeneous MLLM capabilities and modalities (Du et al., 2025; Wei et al., 2025). The merging methods evaluated in our study span these methodological families. MergeHEIR takes the resulting merged checkpoints as its starting point and studies how to mitigate the hallucination degradation introduced by capability consolidation while retaining inherited expertise.

Hallucination mitigation in MLLMs. Multimodal hallucination refers to generated content that is inconsistent with the visual input. Existing mitigation methods can be broadly divided into training-free and training-based approaches. Training-free methods intervene during inference through contrastive or adaptive decoding, prompting, and post-hoc correction (Leng et al., 2024; Chen et al., 2024; Yin et al., 2024). Training-based approaches instead improve visual grounding using hallucination-aware supervision or preference data. RLAIF-V (Yu et al., 2025) constructs multimodal preferences from AI feedback, while mDPO (Wang et al., 2024a) strengthens image dependence through image-conditioned preference optimization. HDPO (Fu et al., 2025) and EMPO (Wu et al., 2025) further target heterogeneous hallucination causes and fine-grained modality alignment. Recent work also studies hallucination within multimodal reasoning chains (Dong et al., 2025; Fang et al., 2026; Kong et al., 2026).

Null-space constrained adaptation. Null-space constraints have been used to mitigate interference in continual learning and model editing (Wang et al., 2021; Fang et al., 2025). Recent work extends this idea to LLM alignment and merging: NSPO projects safety gradients using general-task representations (Niu et al., 2026), while RAIN-Merging projects task vectors using reasoning-related forward features (Huang et al., 2026). MergeHEIR instead targets hallucination-oriented adaptation after heterogeneous multimodal expert merging, constructing a shared protected subspace from pooled expert-task activations and periodically projecting the accumulated post-merge displacement. We further provide geometric and perturbation guarantees for this projection.

B Mechanistic Interpretation of the Merging Tax

We examine the mechanism underlying the *merging tax*, where model merging increases multimodal hallucination. All constituent experts share the same frozen visual encoder and differ only in their language-side adaptations. For a fixed image, the encoded visual representation is therefore unchanged before and after merging, locating the source of the degradation downstream in how the language-side network uses visual evidence during generation.

Multimodal hallucination occurs when a response makes claims about the visual input that are unsupported by, or inconsistent with, the available visual evidence. During expert specialization, language-side parameters are optimized jointly for each task, forming co-adapted computations that govern how visual cues are propagated, integrated with textual context, and translated into token predictions. Model merging recombines these independently learned adaptations into a new parameter configuration. This recomposition can alter the cross-layer interactions that determine how strongly visual evidence influences generation, even when the visual representation itself remains unchanged. We refer to this disruption in the downstream use of visual evidence as *visual-evidence utilization interference*.

The constituent experts inherit the same pretrained semantic and linguistic structure, while task specialization adapts how this shared knowledge is conditioned on visual input. When merging weakens these image-conditioned adaptations, generation becomes less tightly coupled to the current image and more susceptible to semantic regularities that are not visually supported. Unsupported objects, attributes, quantities, relations, or text can therefore enter the response more readily. The systematic

increase in hallucination observed in Figure 1 is consistent with this interpretation, suggesting a shared failure mode in which visual evidence exerts weaker control over generation despite differences among merging methods.

C Additional Theoretical Analysis

This section provides complete proofs for the results in Section 3.3, clarifying how null-space projection in MergeHEIR balances adaptation flexibility and expert-task representation stability. We begin with the projection update and spectral subspace selection, then examine the role of the threshold. The spectral perturbation bound is further extended in two directions: to the corresponding activation distribution under spectral approximation, and to fixed-prefix outputs under local smoothness.

Notation and normalization. Consider one adapted linear module with merged weight $W_m \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and accumulated displacement ΔW. Let $H = [H_t]_{t \in \mathcal{T}} \in \mathbb{R}^{d_{\text{in}} \times N}$ concatenate the expert-task calibration activations, and write

$$G = HH^\top = U \operatorname{diag}(\sigma_1, \ldots, \sigma_{d_{\text{in}}}) U^\top, \qquad 0 \le \sigma_1 \le \cdots \le \sigma_{d_{\text{in}}} = \sigma_{\max}. \tag{13}$$

Define

$$\sigma_{\text{cut}} = \alpha \sigma_{\max}, \qquad \mathcal{I}_\alpha = \{i : \sigma_i < \sigma_{\text{cut}}\}, \qquad U_{\text{null}} = [u_i \mid i \in \mathcal{I}_\alpha], \qquad P = U_{\text{null}} U_{\text{null}}^\top. \tag{14}$$

Thus $P = P^\top = P^2$ and $PG = GP$. For $\widehat{G} = \frac{1}{N} G$, its eigenvalues satisfy $\widehat{\sigma}_i = \frac{\sigma_i}{N}$ and $\widehat{\sigma}_{\max} = \frac{\sigma_{\max}}{N}$. Hence

$$\sigma_i < \sigma_{\text{cut}} \quad \Longleftrightarrow \quad \widehat{\sigma}_i < \alpha \widehat{\sigma}_{\max}, \tag{15}$$

so normalization leaves P unchanged, allowing the same projector to be analyzed using the empirical second moment in the population extension.

C.1 Proof of Proposition 3.1

We first justify the projection of the accumulated displacement onto a fixed admissible subspace $\mathcal{S}_P = \{Z : Z = ZP\}$. The orthogonal decomposition induced by P establishes that ΔWP minimizes update distortion in Frobenius norm.

Proof. For any $Z \in \mathcal{S}_P$,

$$Z - \Delta W = (Z - \Delta WP) - \Delta W(I - P). \tag{16}$$

Since $Z - \Delta WP = (Z - \Delta WP)P$, the two terms are Frobenius-orthogonal:

$$\langle Z - \Delta WP, \Delta W(I-P) \rangle_F = \operatorname{tr}\big(P(Z - \Delta WP)^\top \Delta W(I-P)\big) \tag{17}$$

$$= \operatorname{tr}\big((I-P)P(Z - \Delta WP)^\top \Delta W\big) = 0. \tag{18}$$

Therefore,

$$\|Z - \Delta W\|_F^2 = \|Z - \Delta WP\|_F^2 + \|\Delta W(I-P)\|_F^2. \tag{19}$$

The second term is independent of Z, whereas the first is uniquely minimized at $Z = \Delta WP$. Hence, ΔWP is the unique orthogonal projection of ΔW onto $\mathcal{S}_P$ under the Frobenius inner product. Setting $Z = 0$ in the preceding decomposition yields

$$\|\Delta W\|_F^2 = \|\Delta WP\|_F^2 + \|\Delta W(I-P)\|_F^2, \tag{20}$$

and therefore $\|\Delta WP\|_F \le \|\Delta W\|_F$. □

C.2 Proof of Theorem 3.2

The preceding result justifies the projection within a fixed subspace. To relate this geometric property to expert-task stability, we analyze the spectral construction, establishing maximum dimensionality under the sensitivity constraint and then deriving pooled and task-specific perturbation bounds.

Proof. Step 1: Maximum-dimensional admissible subspace. Let $k = |\mathcal{I}_\alpha| = \mathrm{rank}(P)$. Every nonzero vector $v \in \mathrm{range}(P)$ can be written as

$$v = \sum_{i \in \mathcal{I}_\alpha} c_i u_i. \tag{21}$$

Using the orthonormality of the eigenvectors,

$$v^\top G v = \sum_{i \in \mathcal{I}_\alpha} \sigma_i c_i^2 < \sigma_{\mathrm{cut}} \sum_{i \in \mathcal{I}_\alpha} c_i^2 = \sigma_{\mathrm{cut}} \|v\|_2^2. \tag{22}$$

Thus, $\mathrm{range}(P)$ satisfies the prescribed activation-sensitivity condition.

Now consider any subspace $\mathcal{V} \subseteq \mathbb{R}^{d_{\mathrm{in}}}$ with $\dim(\mathcal{V}) > k$, and define

$$\mathcal{U}_{\mathrm{high}} = \mathrm{span}\{u_i : \sigma_i \geq \sigma_{\mathrm{cut}}\}. \tag{23}$$

Since $\dim(\mathcal{U}_{\mathrm{high}}) = d_{\mathrm{in}} - k$, the dimension inequality

$$\dim(\mathcal{V}) + \dim(\mathcal{U}_{\mathrm{high}}) > d_{\mathrm{in}} \tag{24}$$

implies that $\mathcal{V} \cap \mathcal{U}_{\mathrm{high}}$ contains a nonzero vector. For any such vector v,

$$v^\top G v \geq \sigma_{\mathrm{cut}} \|v\|_2^2. \tag{25}$$

Hence, no subspace of dimension greater than k can satisfy the strict sensitivity condition for every nonzero vector. Since $\mathrm{range}(P)$ has dimension k and satisfies this condition, it is maximum-dimensional.

Step 2: Pooled and task-specific perturbation bounds. We now translate the spectral sensitivity constraint into a bound on the representation perturbation induced by the projected displacement. For the pooled activation matrix,

$$\|\Delta W P H\|_F^2 = \mathrm{tr}\big(\Delta W P H H^\top P \Delta W^\top\big) \tag{26}$$

$$= \mathrm{tr}\big(\Delta W P G P \Delta W^\top\big). \tag{27}$$

By the spectral definition of P,

$$PGP = \sum_{i \in \mathcal{I}_\alpha} \sigma_i u_i u_i^\top \preceq \sigma_{\mathrm{cut}} \sum_{i \in \mathcal{I}_\alpha} u_i u_i^\top = \sigma_{\mathrm{cut}} P. \tag{28}$$

Therefore,

$$\|\Delta W P H\|_F^2 \leq \sigma_{\mathrm{cut}}\, \mathrm{tr}\big(\Delta W P \Delta W^\top\big) \tag{29}$$

$$= \sigma_{\mathrm{cut}} \|\Delta W P\|_F^2. \tag{30}$$

To obtain task-specific bounds from the pooled construction, let $G_t = H_t H_t^\top$. Since H concatenates all task-specific activation matrices,

$$G = \sum_{t \in \mathcal{T}} G_t, \qquad 0 \preceq G_t \preceq G. \tag{31}$$

Applying the congruence transformation induced by P gives

$$P G_t P \preceq PGP \preceq \sigma_{\mathrm{cut}} P. \tag{32}$$

It follows that

$$\|\Delta W P H_t\|_F^2 = \mathrm{tr}\big(\Delta W P G_t P \Delta W^\top\big) \tag{33}$$

$$\leq \sigma_{\mathrm{cut}}\, \mathrm{tr}\big(\Delta W P \Delta W^\top\big) \tag{34}$$

$$= \sigma_{\mathrm{cut}} \|\Delta W P\|_F^2. \tag{35}$$

Substituting $\sigma_{\mathrm{cut}} = \alpha \sigma_{\max}$ yields the stated pooled and task-specific bounds. □

C.3 Proof of Proposition 3.3

The spectral threshold determines both the admissible displacement space and its perturbation bound. We characterize this balance by comparing the retained update and calibration perturbation for a fixed displacement across nested spectral subspaces.

Proof. Let

$$\mathcal{I}_\alpha = \{i : \sigma_i < \alpha\sigma_{\max}\}. \tag{36}$$

Since $\alpha_1 \le \alpha_2$,

$$\mathcal{I}_{\alpha_1} \subseteq \mathcal{I}_{\alpha_2}. \tag{37}$$

Therefore,

$$Q = \sum_{i \in \mathcal{I}_{\alpha_2} \setminus \mathcal{I}_{\alpha_1}} u_i u_i^\top \tag{38}$$

is an orthogonal projector satisfying

$$P_2 = P_1 + Q, \qquad P_1 Q = Q P_1 = 0. \tag{39}$$

The range inclusion, projector identities, Loewner ordering, and rank inequality follow immediately.

Using the orthogonality of P_1 and Q,

$$\|\Delta W P_2\|_F^2 = \|\Delta W (P_1 + Q)\|_F^2 = \|\Delta W P_1\|_F^2 + \|\Delta W Q\|_F^2. \tag{40}$$

Likewise, since

$$I - P_1 = (I - P_2) + Q \tag{41}$$

and $(I - P_2)Q = 0$, we obtain

$$\|\Delta W (I - P_1)\|_F^2 = \|\Delta W (I - P_2)\|_F^2 + \|\Delta W Q\|_F^2. \tag{42}$$

Finally, P_1 and Q project onto disjoint eigenspaces of G, so

$$P_1 G Q = Q G P_1 = 0. \tag{43}$$

Hence,

$$\begin{aligned} \|\Delta W P_2 H\|_F^2 &= \operatorname{tr}\big(\Delta W (P_1 + Q) G (P_1 + Q) \Delta W^\top\big) && (44)\\ &= \operatorname{tr}\big(\Delta W P_1 G P_1 \Delta W^\top\big) + \operatorname{tr}\big(\Delta W Q G Q \Delta W^\top\big) && (45)\\ &= \|\Delta W P_1 H\|_F^2 + \|\Delta W Q H\|_F^2. && (46) \end{aligned}$$

□

By relating the retained update, projection-induced distortion, and calibration perturbation, these identities characterize how α balances adaptation flexibility and representation stability.

Because the retained index set changes only when $\alpha\sigma_{\max}$ crosses an eigenvalue of G, the projector P_α is piecewise constant between the normalized spectral breakpoints $\{\sigma_i/\sigma_{\max}\}_{i=1}^{d_{\text{in}}}$. This spectral structure also explains why threshold sensitivity can exhibit plateau-like regions.

C.4 Proof of Corollary 3.4

The calibration bound controls perturbations through the empirical second moment $\widehat{G}$. To extend this control to the corresponding activation distribution, we compare $\widehat{G}$ with Γ and account for their spectral discrepancy.

Proof. The condition

$$\|\widehat{G} - \Gamma\|_2 \le \varepsilon \tag{47}$$

implies the Loewner-order bounds

$$-\varepsilon I \preceq \Gamma - \widehat{G} \preceq \varepsilon I. \tag{48}$$

Applying the congruence transformation induced by P and using $P^2 = P$ gives

$$P\Gamma P \preceq P\widehat{G}P + \varepsilon P. \tag{49}$$

Since P retains only eigenvectors of $\widehat{G}$ whose eigenvalues are below $\alpha\widehat{\sigma}_{\max}$,

$$P\widehat{G}P \preceq \alpha\widehat{\sigma}_{\max} P. \tag{50}$$

Consequently,

$$P\Gamma P \preceq (\alpha\widehat{\sigma}_{\max} + \varepsilon)\, P. \tag{51}$$

For an activation vector h drawn from the corresponding distribution, the definition $\Gamma = \mathbb{E}[hh^\top]$ gives

$$\mathbb{E}\big[\|\Delta W P h\|_2^2\big] = \operatorname{tr}\big(\Delta W P \Gamma P \Delta W^\top\big) \tag{52}$$

$$\leq (\alpha\widehat{\sigma}_{\max} + \varepsilon)\operatorname{tr}\big(\Delta W P \Delta W^\top\big) \tag{53}$$

$$= (\alpha\widehat{\sigma}_{\max} + \varepsilon)\,\|\Delta W P\|_F^2. \tag{54}$$

□

Corollary 3.4 extends the empirical perturbation bound to the corresponding activation distribution, with the spectral discrepancy entering the bound additively through ε. Thus, the thresholded null-space construction provides representation-level perturbation control beyond the collected calibration activations under the stated spectral condition.

C.5 From Representation Perturbations to Fixed-Prefix Outputs

To relate representation stability to model behavior, we combine the module-level bound in Theorem 3.2 with local smoothness assumptions. This yields perturbation bounds for logits and next-token distributions under fixed calibration prefixes. Let θ_m denote the merged checkpoint, and let $\mathbf{\Delta}_P$ denote the model-wide projected displacement whose ℓ-th adapted block is $\Delta W_\ell P_\ell$. Define

$$\|\mathbf{\Delta}_P\|_{\mathrm{blk}}^2 = \sum_{\ell=1}^{L} \|\Delta W_\ell P_\ell\|_F^2, \tag{55}$$

where L is the number of adapted modules. Let H_ℓ collect the inputs to module ℓ under the merged checkpoint θ_m at the evaluated positions, and let Z_θ denote the corresponding stacked logits.

Local output smoothness. To quantify how module-level perturbations propagate to the logits, we use a first-order Taylor expansion with a quadratic remainder (Nesterov, 2013) and bounded local sensitivities (Fazlyab et al., 2019). For a fixed collection of calibration prefixes, assume that the logit map admits the following local expansion around θ_m:

$$Z_{\theta_m + \mathbf{\Delta}_P} - Z_{\theta_m} = \sum_{\ell=1}^{L} \mathcal{J}_\ell[\Delta W_\ell P_\ell H_\ell] + \mathcal{R}(\mathbf{\Delta}_P), \tag{56}$$

where $\mathcal{J}_\ell$ is a bounded linear map describing the local response of the stacked logits to the output perturbation at module ℓ, with

$$\|\mathcal{J}_\ell[X]\|_F \leq \kappa_\ell \|X\|_F. \tag{57}$$

Assume further that the nonlinear remainder satisfies

$$\|\mathcal{R}(\mathbf{\Delta}_P)\|_F \leq C\|\mathbf{\Delta}_P\|_{\mathrm{blk}}^2. \tag{58}$$

For compactness, define

$$\mathcal{B}_\alpha(\mathbf{\Delta}_P) = \sum_{\ell=1}^{L} \kappa_\ell \sqrt{\sigma_{\mathrm{cut}}^{(\ell)}}\,\|\Delta W_\ell P_\ell\|_F + C\|\mathbf{\Delta}_P\|_{\mathrm{blk}}^2, \tag{59}$$

where $\sigma_{\mathrm{cut}}^{(\ell)} = \alpha\sigma_{\max}^{(\ell)}$.

Fixed-prefix output control. Under the local smoothness condition above, the logit perturbation satisfies

$$\|Z_{\theta_m+\mathbf{\Delta}_P} - Z_{\theta_m}\|_F \le \mathcal{B}_\alpha(\mathbf{\Delta}_P). \tag{60}$$

Moreover, if p_j and q_j denote the next-token distributions induced by the merged and post-projection models, respectively, at fixed prefix j, then

$$\sum_j D_{\mathrm{KL}}(p_j \| q_j) \le \frac{1}{4}\mathcal{B}_\alpha(\mathbf{\Delta}_P)^2. \tag{61}$$

Derivation. Applying the triangle inequality, the local operator bounds, and Theorem 3.2 module-wise gives

$$\|Z_{\theta_m+\mathbf{\Delta}_P} - Z_{\theta_m}\|_F \le \sum_{\ell=1}^{L} \kappa_\ell \|\Delta W_\ell P_\ell H_\ell\|_F + \|\mathcal{R}(\mathbf{\Delta}_P)\|_F \tag{62}$$

$$\le \sum_{\ell=1}^{L} \kappa_\ell \sqrt{\sigma_{\mathrm{cut}}^{(\ell)}} \|\Delta W_\ell P_\ell\|_F + C\|\mathbf{\Delta}_P\|_{\mathrm{blk}}^2 \tag{63}$$

$$= \mathcal{B}_\alpha(\mathbf{\Delta}_P), \tag{64}$$

which establishes Equation (60).

To translate logit perturbations into changes in next-token distributions, we use the log-partition function

$$f(z) = \log \sum_i \exp(z_i). \tag{65}$$

For categorical exponential families, the KL divergence can be expressed as the Bregman divergence induced by the log-partition function (Banerjee et al., 2005; Wainwright & Jordan, 2008). For $\rho = \mathrm{softmax}(z)$, its Hessian is

$$\nabla^2 f(z) = \mathrm{diag}(\rho) - \rho\rho^\top. \tag{66}$$

For any vector x,

$$x^\top \nabla^2 f(z) x = \mathrm{Var}_{i\sim\rho}(x_i) \le \frac{(\max_i x_i - \min_i x_i)^2}{4} \le \frac{1}{2}\|x\|_2^2. \tag{67}$$

Hence $\|\nabla^2 f(z)\|_2 \le 1/2$, so f is $1/2$-smooth. For $p = \mathrm{softmax}(z)$ and $q = \mathrm{softmax}(z+u)$,

$$D_{\mathrm{KL}}(p\|q) = f(z+u) - f(z) - \nabla f(z)^\top u \tag{68}$$

$$\le \frac{1}{4}\|u\|_2^2. \tag{69}$$

Applying this inequality at each fixed prefix and using the fact that Z_θ stacks the corresponding logit vectors gives

$$\sum_j D_{\mathrm{KL}}(p_j \| q_j) \le \frac{1}{4}\|Z_{\theta_m+\mathbf{\Delta}_P} - Z_{\theta_m}\|_F^2 \tag{70}$$

$$\le \frac{1}{4}\mathcal{B}_\alpha(\mathbf{\Delta}_P)^2, \tag{71}$$

which establishes Equation (61).

Leading-order implication. The first term in $\mathcal{B}_\alpha(\mathbf{\Delta}_P)$ is of order $\mathcal{O}(\|\mathbf{\Delta}_P\|_{\mathrm{blk}})$. Therefore, as $\|\mathbf{\Delta}_P\|_{\mathrm{blk}} \to 0$, expanding Equation (61) yields

$$\sum_j D_{\mathrm{KL}}(p_j \| q_j) \le \frac{1}{4}\left(\sum_{\ell=1}^{L} \kappa_\ell \sqrt{\sigma_{\mathrm{cut}}^{(\ell)}} \|\Delta W_\ell P_\ell\|_F\right)^2 + \mathcal{O}\big(\|\mathbf{\Delta}_P\|_{\mathrm{blk}}^3\big). \tag{72}$$

Thus, the fixed-prefix distribution shift admits a quadratic leading-order upper bound in the retained displacement, jointly governed by the downstream sensitivities $\{\kappa_\ell\}_{\ell=1}^L$ and the layer-wise spectral cutoffs. This extends the representation-level perturbation analysis to the model's local output behavior under shared calibration contexts.

Theoretical implications. Taken together, these results connect the geometry of MergeHEIR's projected updates to representation stability and local output behavior. The projection minimizes update distortion within a maximum-dimensional subspace under the prescribed sensitivity constraint, while α balances adaptation flexibility and calibration perturbation. The population extension accounts for spectral approximation error, and the fixed-prefix analysis propagates module-level perturbation bounds to logits and next-token distributions under local smoothness. These results explain how projection limits interference with expert-task representations during hallucination-oriented adaptation, complementing the empirical evaluation of expert retention.

D Experimental Details

D.1 Training Task-Specialized Experts

Training data. For each backbone, we independently train four task-specialized experts for chart understanding, geometry reasoning, OCR, and visual question answering. Table 3 summarizes the datasets and training-set sizes used for each expert.

Table 3: Summary of the training datasets used to construct the four task-specialized experts.

Task Category	Size	Training Datasets
VQA	589K	GQA (Hudson & Manning, 2019), CogVLM-Multiround (Wang et al., 2024c), CogVLM-Singleround (Wang et al., 2024c)
Geometry	177K	GeoQA+ (Cao & Xiao, 2022), G-LLaVA (Gao et al., 2025)
Chart	219K	ChartQA (Masry et al., 2022), DVQA (Kafle et al., 2018)
OCR	239K	OCRVQA (Mishra et al., 2019), SynthDoG (Kim et al., 2022), DocVQA (Mathew et al., 2021), TextVQA (Singh et al., 2019), LLaVAR (Zhang et al., 2023), ST-VQA (Biten et al., 2019)

Training configuration. All experts are trained with LoRA applied only to non-visual modules, while all visual-side parameters remain frozen, using $(r, \alpha_{\mathrm{LoRA}}) = (32, 64)$ for Qwen2.5-VL-3B and $(r, \alpha_{\mathrm{LoRA}}) = (16, 32)$ for Qwen2.5-VL-7B and InternVL3.5-8B. The learning rate is set to 2×10^{-5} for Qwen2.5-VL-3B and 1×10^{-5} for the two larger configurations, with a cosine learning-rate schedule. Each expert is trained for one epoch to limit deviation from the shared initialization and preserve compatibility for subsequent model merging.

D.2 Evaluation Benchmarks

Expert-task benchmarks. Chart understanding. ChartQA (Masry et al., 2022) evaluates question answering over charts, requiring models to extract visualized information and perform logical or arithmetic reasoning over chart content. **Geometry reasoning.** MATH-Vision (Wang et al., 2024b) contains competition-sourced visual mathematics problems that require joint reasoning over visual contexts and textual problem statements, while Geometry3K (Lu et al., 2021) focuses on solving geometry problems by integrating textual conditions with diagram structure and geometric relations. **OCR.** TextVQA (Singh et al., 2019) evaluates the ability to read and reason about text embedded in natural images, whereas DocVQA (Mathew et al., 2021) requires question answering over document images using their textual content and document structure. **Visual question answering.** GQA (Hudson & Manning, 2019) emphasizes compositional reasoning over real-world scenes, while OKVQA (Marino et al., 2019) contains questions that require knowledge beyond the information directly visible in the image. We report the unweighted mean over these seven benchmarks as *Expert Avg.*

Hallucination benchmarks. AMBER. AMBER (Wang et al., 2023) evaluates generative and discriminative hallucinations; we use its generative image-description setting and report CHAIR, Hal-Rate, and Cog. **Object HalBench.** Object HalBench (Rohrbach et al., 2018) measures object hallucination in image descriptions and reports response-level CHAIR_s and mention-level CHAIR_i. **POPE.** POPE (Li et al., 2023) evaluates object hallucination through binary object-existence questions under Random, Popular, and Adversarial splits; we report accuracy and F1. **GPT-4-assisted evaluation.** Following Zhao et al. (2023), we use Visual Genome (Krishna et al., 2017) annotations and GPT-4o-mini to assess fine-grained hallucinations in generated descriptions, reporting Sentence-level Hallucination Ratio (SHR).

Table 4: Expert-task results on InternVL3.5-8B.

Merging method	Stage	Chart	Geometry		OCR		VQA		Avg.
		ChartQA	MATH-Vision	Geometry3K	TextVQA	DocVQA	GQA	OKVQA	
Task Arithmetic	Merged	76.20	23.68	60.23	63.83	56.46	59.96	42.17	54.65
	MergeHEIR	76.08	23.03	60.57	63.62	56.26	60.84	39.61	54.29
TIES Merging	Merged	76.52	24.67	60.07	64.21	56.85	60.93	43.87	55.30
	MergeHEIR	76.76	23.03	59.73	64.50	57.29	61.78	42.71	55.11
DELLA Merging	Merged	77.08	24.67	60.07	64.09	56.63	60.57	43.39	55.21
	MergeHEIR	77.12	23.68	59.73	63.97	56.54	61.50	42.04	54.94
Breadcrumbs Merging	Merged	76.80	24.34	60.07	64.14	56.69	60.41	43.12	55.08
	MergeHEIR	76.92	24.34	60.73	64.07	56.77	61.40	40.44	54.95
Model Stock	Merged	75.72	23.68	59.90	63.76	55.90	59.49	40.67	54.16
	MergeHEIR	75.52	22.04	59.23	63.62	56.19	59.76	39.52	53.70
Iso-C	Merged	75.76	24.01	60.07	63.61	56.13	59.53	40.83	54.28
	MergeHEIR	75.68	22.04	59.23	63.49	56.06	59.79	39.07	53.63
WUDI Merging	Merged	77.36	25.33	60.57	64.45	56.89	61.11	44.42	55.73
	MergeHEIR	77.28	23.68	60.07	64.40	57.49	61.63	43.77	55.47
OptMerge	Merged	77.68	25.33	60.57	64.29	57.09	61.64	46.33	56.13
	MergeHEIR	77.60	25.00	61.23	64.30	57.04	62.78	44.11	56.01

Table 5: Hallucination-task results on InternVL3.5-8B.

Merging method	Stage	AMBER			Object HalBench		POPE Adversarial		GPT-4 Eval
		CHAIR↓	HalRate↓	Cog↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓	Acc.↑	F1↑	SHR↓
Task Arithmetic	Merged	8.3	66.5	7.3	25.3	6.7	73.9	78.6	36.2
	MergeHEIR	7.6	63.1	6.3	24.3	6.7	77.5	80.6	32.8
TIES Merging	Merged	8.1	68.7	7.1	24.0	6.2	76.5	80.0	36.3
	MergeHEIR	7.2	62.3	6.1	20.3	5.3	80.6	82.5	33.6
DELLA Merging	Merged	8.1	68.7	7.3	23.7	6.2	75.3	79.4	36.9
	MergeHEIR	7.3	63.4	6.4	20.7	5.2	79.1	81.5	33.1
Breadcrumbs Merging	Merged	8.2	69.9	7.1	22.7	5.9	75.1	79.2	36.2
	MergeHEIR	6.9	63.0	6.3	20.3	5.1	79.2	81.5	32.4
Model Stock	Merged	8.0	65.6	6.8	27.7	6.8	73.3	78.2	36.7
	MergeHEIR	7.7	63.3	6.2	23.3	6.4	77.2	80.5	34.0
Iso-C	Merged	8.4	66.6	6.9	27.7	6.8	73.5	78.3	36.7
	MergeHEIR	7.8	62.8	6.0	23.7	6.5	77.1	80.3	33.8
WUDI Merging	Merged	8.0	68.4	6.9	24.0	5.9	78.2	80.9	35.9
	MergeHEIR	7.1	63.6	6.1	21.7	5.3	82.4	83.3	33.2
OptMerge	Merged	7.6	65.0	6.8	18.0	4.2	80.8	82.6	36.5
	MergeHEIR	6.8	61.2	6.0	17.3	4.8	86.6	88.2	32.4

E Additional Results

Complete InternVL3.5-8B results. Tables 4 and 5 provide the complete matched results for InternVL3.5-8B, complementing the Qwen2.5-VL results in the main text. Across the eight model-merging methods, MergeHEIR consistently improves hallucination performance while largely maintaining expert-task performance. This pattern is consistent with the intended role of MergeHEIR in coordinating hallucination correction with the retention of inherited expertise.

Complete POPE results. Tables 6–8 report accuracy and F1 on all three POPE splits. Across all 24 cases, MergeHEIR improves both metrics on every split.

Compatibility with alternative preference methods. To examine whether the retention benefit of MergeHEIR depends on its specific preference objective, we integrate its null-space projection mechanism with RLAIF-V (Yu et al., 2025) and OPA-DPO (Yang et al., 2025). Across all six matched comparisons, direct adaptation with the underlying preference methods causes substantially greater degradation of inherited expert capabilities, whereas incorporating MergeHEIR's null-space projection preserves expert performance while maintaining essentially the same hallucination-mitigation benefit. These results indicate that the retention benefit of MergeHEIR is not tied to its specific preference objective and support its use as a plug-and-play mechanism for capability-preserving post-merge adaptation.

Table 6: Complete POPE results for Qwen2.5-VL-3B.

Merging method	Stage	POPE Adversarial		POPE Popular		POPE Random	
		Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
Task Arithmetic	Merged	82.17	83.47	85.83	86.41	91.96	92.03
	MergeHEIR	85.80	86.05	88.63	88.51	92.23	92.08
TIES Merging	Merged	82.17	83.46	85.77	86.34	91.89	91.96
	MergeHEIR	85.80	86.06	88.70	88.58	92.20	92.05
DELLA Merging	Merged	82.07	83.37	85.97	86.50	91.89	91.96
	MergeHEIR	85.77	86.04	88.73	88.62	92.23	92.09
Breadcrumbs Merging	Merged	81.90	83.21	85.70	86.25	91.82	91.88
	MergeHEIR	85.70	85.99	88.67	88.57	92.30	92.16
Model Stock	Merged	82.87	83.91	87.17	87.44	91.96	91.97
	MergeHEIR	86.60	86.45	89.93	89.52	92.20	92.02
Iso-C	Merged	82.77	83.87	87.23	87.53	91.99	92.02
	MergeHEIR	85.73	86.05	88.70	88.62	92.30	92.18
WUDI Merging	Merged	85.37	85.24	88.00	87.57	90.82	90.47
	MergeHEIR	89.13	87.92	90.63	89.67	91.41	90.99
OptMerge	Merged	83.53	84.16	86.80	86.89	91.75	91.62
	MergeHEIR	87.33	86.95	89.73	89.17	92.13	91.78

Table 7: Complete POPE results for Qwen2.5-VL-7B.

Merging method	Stage	POPE Adversarial		POPE Popular		POPE Random	
		Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
Task Arithmetic	Merged	83.37	80.70	83.87	81.17	84.09	81.84
	MergeHEIR	86.97	86.00	88.20	87.09	88.93	88.14
TIES Merging	Merged	85.53	84.41	87.27	86.02	87.84	86.91
	MergeHEIR	89.17	88.05	91.07	90.19	92.34	92.06
DELLA Merging	Merged	84.00	81.85	84.67	82.47	85.26	83.46
	MergeHEIR	86.97	86.50	88.33	87.68	89.86	89.31
Breadcrumbs Merging	Merged	83.93	81.69	84.57	82.28	85.12	83.24
	MergeHEIR	87.47	86.87	88.70	88.02	89.66	89.45
Model Stock	Merged	83.20	80.36	83.63	80.77	83.78	81.37
	MergeHEIR	86.80	85.65	87.77	86.56	88.52	87.62
Iso-C	Merged	83.37	80.53	83.73	80.88	83.81	81.42
	MergeHEIR	86.70	85.54	87.70	86.48	88.52	87.60
WUDI Merging	Merged	86.00	85.10	87.90	86.85	88.63	87.87
	MergeHEIR	89.43	88.94	91.77	91.12	93.16	92.94
OptMerge	Merged	86.10	84.92	87.30	86.04	88.01	87.06
	MergeHEIR	89.70	89.29	91.40	90.73	92.61	92.33

Table 8: Complete POPE results for InternVL3.5-8B.

Merging method	Stage	POPE Adversarial		POPE Popular		POPE Random	
		Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
Task Arithmetic	Merged	73.93	78.61	78.43	81.62	92.20	92.68
	MergeHEIR	77.50	80.64	82.20	84.04	93.40	93.61
TIES Merging	Merged	76.53	80.02	81.50	83.56	92.85	93.13
	MergeHEIR	80.63	82.52	85.73	86.35	94.60	94.48
DELLA Merging	Merged	75.33	79.36	79.73	82.40	92.41	92.79
	MergeHEIR	79.13	81.51	83.77	85.02	93.81	93.90
Breadcrumbs Merging	Merged	75.07	79.21	79.40	82.18	92.34	92.74
	MergeHEIR	79.17	81.49	83.83	84.96	94.02	94.01
Model Stock	Merged	73.30	78.23	77.43	80.96	92.03	92.54
	MergeHEIR	77.23	80.46	81.70	83.67	93.47	93.67
Iso-C	Merged	73.47	78.33	77.50	81.00	92.03	92.54
	MergeHEIR	77.07	80.35	81.73	83.70	93.40	93.61
WUDI Merging	Merged	78.23	80.93	83.13	84.56	92.78	92.96
	MergeHEIR	82.37	83.31	87.60	87.44	94.47	94.37
OptMerge	Merged	80.80	82.61	85.07	85.93	93.02	93.09
	MergeHEIR	86.60	88.18	90.67	91.46	94.60	94.87

Table 9: Compatibility of MergeHEIR with alternative hallucination-oriented preference methods. "+ HEIR" denotes augmenting the corresponding method with the null-space projection mechanism of MergeHEIR.

Model	Setting	Expert	AMBER			Object HalBench	
		Avg.↑	CHAIR↓	HalRate↓	Cog↓	$CHAIR_s$ ↓	$CHAIR_i$ ↓
Qwen2.5-VL-3B	Merged	60.83	8.8	59.6	7.1	36.0	9.0
	RLAIF-V	60.28	5.3	32.2	3.5	27.7	7.9
	RLAIF-V + HEIR	60.86	5.5	30.8	3.6	28.0	7.7
	OPA-DPO	60.15	4.5	21.2	1.8	21.3	6.1
	OPA-DPO + HEIR	60.90	4.6	21.5	2.0	21.7	6.2
Qwen2.5-VL-7B	Merged	65.30	4.9	24.3	1.4	41.0	9.0
	RLAIF-V	63.61	4.6	22.4	1.3	30.3	7.5
	RLAIF-V + HEIR	65.23	4.5	22.5	1.3	30.7	7.8
	OPA-DPO	63.32	4.1	19.6	1.0	24.0	6.3
	OPA-DPO + HEIR	65.18	4.2	20.0	1.0	24.7	6.1
InternVL3.5-8B	Merged	55.73	8.0	68.4	6.9	24.0	5.9
	RLAIF-V	54.30	7.6	65.8	6.8	23.0	5.6
	RLAIF-V + HEIR	55.58	7.6	66.2	6.6	22.7	5.7
	OPA-DPO	54.21	7.0	62.4	6.0	20.7	5.2
	OPA-DPO + HEIR	55.52	7.2	63.2	6.1	21.0	5.1

F Case Study

We provide qualitative comparisons among the merged model, MergeHEIR, and unconstrained post-merge adaptation on both expert tasks and hallucination benchmarks. As shown in Figs. 6–9, unconstrained adaptation can substantially impair the capabilities inherited from the constituent experts, producing incorrect chart interpretations, flawed geometric reasoning, inaccurate text recognition, or erroneous visual answers. In contrast, MergeHEIR retains the task-specific knowledge of the merged model and produces responses that remain consistent with the visual evidence.

Figs. 10 and 11 further compare the three model variants on hallucination evaluation examples. The merged model tends to introduce nonexistent objects or visually unsupported scene details, whereas both post-merge adaptation methods effectively suppress such hallucinations. Notably, MergeHEIR achieves this behavioral correction while avoiding the pronounced expertise degradation exhibited by unconstrained adaptation. These examples qualitatively corroborate the quantitative results, demonstrating that MergeHEIR achieves a more favorable trade-off between hallucination mitigation and expertise retention.

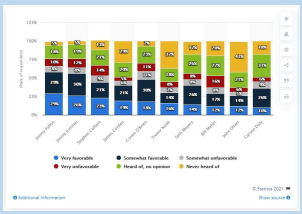
G Impact Statement

This paper takes an early step toward mitigating hallucinations in merged MLLMs while preserving inherited expert capabilities. The goal is to improve the reliability of post-merge adaptation and support more trustworthy multimodal systems. The broader societal impacts are largely consistent with those associated with improving the robustness and reliability of machine learning systems, and we do not identify additional consequences that require specific discussion.

H Statement on LLM assistance

An LLM assistant was used for writing refinement and formatting adjustments.

Expert-task Chart



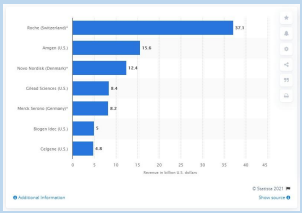
Which late night host had the highest favorability ratings?

Merged ✓: For Jimmy Fallon, the "Very favorable" is 29% and the "Somewhat favorable" is 29%. Adding these gives 58%. For Jimmy Kimmel, the "Very favorable" is 26% and the "Somewhat favorable" is 30%. Adding these gives 56%. For Stephen Colbert, the "Very favorable" is 23% and the "Somewhat favorable" is 21%. Adding these gives 44%. The total favorable ratings are 58% for Jimmy Fallon, 56% for Jimmy Kimmel, and 44% for Stephen Colbert. Therefore, Jimmy Fallon has the highest overall favorability rating.

MergeHEIR ✓: For Jimmy Fallon, the "Very favorable" is 29% and the "Somewhat favorable" is 29%. Adding these together gives a total of 58%. For Jimmy Kimmel, the "Very favorable" is 26% and the "Somewhat favorable" is 30%. Adding these together gives a total of 56%. For Stephen Colbert, the "Very favorable" is 23% and the "Somewhat favorable" is 21%. Adding these together gives a total of 44%. Comparing the totals, Jimmy Fallon has the highest favorability rating with 58%.

w/o constraint ✗: For Jimmy Fallon, the "Very favorable" is 29% and "Somewhat favorable" is 26%. Adding these together gives 55%. For Jimmy Kimmel, the "Very favorable" is 26% and "Somewhat favorable" is 30%. Adding these together gives 56%. For Stephen Colbert, the "Very favorable" is 23% and "Somewhat favorable" is 21%. Adding these together gives 44%. The total of favorable ratings for Jimmy Kimmel (56%) is higher than those for Jimmy Fallon (55%) and Stephen Colbert (44%).

Expert-task Chart



What is the average of all produces less than 15 billion?

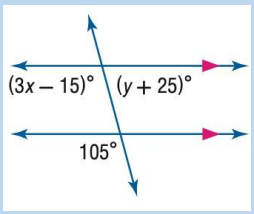
Merged ✓: The products whose values are strictly below 15 billion U.S. dollars are: 1. Novo Nordisk (Denmark) with 12.4 billion U.S. dollars. 2. Gilead Sciences (U.S.) with 8.4 billion U.S. dollars. 3. Merck Serono (Germany) with 8.2 billion U.S. dollars. 4. Biogen Idec (U.S.) with 5 billion U.S. dollars. 5. Celgene (U.S.) with 4.8 billion U.S. dollars. There are five values that qualify. Their sum is 38.8 billion U.S. dollars. Dividing this sum by the count gives an average of 7.76 billion U.S. dollars.

MergeHEIR ✓: The products with revenue strictly below 15 billion U.S. dollars are: - Novo Nordisk (Denmark): 12.4 billion U.S. dollars - Gilead Sciences (U.S.): 8.4 billion U.S. dollars - Merck Serono (Germany): 8.2 billion U.S. dollars - Biogen Idec (U.S.): 5 billion U.S. dollars - Celgene (U.S.): 4.8 billion U.S. dollars There are 5 values that qualify. The sum of these values is 38.8 billion U.S. dollars. Dividing this sum by the count gives an average of 7.76 billion U.S. dollars.

w/o constraint ✗: The products with revenue strictly below 15 billion U.S. dollars are Biogen Idec (U.S.) with $5 billion and Celgene (U.S.) with $4.8 billion. There are two values that qualify, summing to $9.8 billion. Dividing this sum by the count gives an average of $4.9 billion. Conclusion: The average revenue of products with revenue strictly below 15 billion U.S. dollars is $4.9 billion. Final answer: 4.9

Figure 6: Visualization results comparing the merged model, MergeHEIR, and unconstrained post-merge adaptation on chart understanding task. Incorrect or visually unsupported response spans are highlighted in red.

Expert-task
Geometry



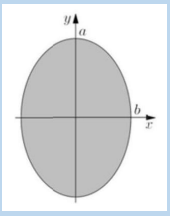
Find the value of the variable y in the figure.

Merged ✓: The angle vertically opposite the marked 105° angle is (3x - 15)°. Since vertically opposite angles are equal, we have: 3x - 15 = 105 Solving for x : 3x = 120 x = 40 Now, using the relationship between consecutive interior angles formed by a transversal with parallel lines, we know that: (y + 25) + 105 = 180 Simplifying: y + 130 = 180 y = 50 Conclusion: The calculated value of y aligns with the geometric properties and relationships in the figure. Final answer: 50

MergeHEIR ✓: The angle vertically opposite the marked 105° angle is (3x - 15)°. Since vertically opposite angles are equal, we have: 3x - 15 = 105 Solving for x : 3x = 120 x = 40 Next, we use the fact that the sum of interior angles on the same side of the transversal is 180°: (y + 25) + 105 = 180 Solving for y : y + 130 = 180 y = 50 Conclusion: The calculation shows that y = 50 , which fits the geometric relationships in the diagram. Final answer: 50

w/o constraint ✗: The angle vertically opposite the marked 105° angle is (3x - 15)°. Since vertically opposite angles are equal, we have: 3x - 15 = 105 Solving for x : 3x = 120 x = 40 Next, using the corresponding angle relationship with the parallel lines and transversal, (y + 25)° must be equal to 105° because they are corresponding angles: y + 25 = 105 Solving for y : y = 80

Expert-task
Geometry



Compare the ellipsoids and their volumes obtained by rotating an ellipse (\(a>b\)) about the \(x\)- and \(y\)-axes.

Merged ✓: The ellipse has semi-axis lengths a (along the y-axis) and b (along the x-axis). When rotated about the x-axis, the ellipsoid E_x has semi-axes b, a, a. When rotated about the y-axis, the ellipsoid E_y has semi-axes a, b, b. The volume of an ellipsoid with semi-axes r_1, r_2, r_3 is given by 4/3π r_1 r_2 r_3. Thus, Vol(E_x) = 4/3π b a^2 and Vol(E_y) = 4/3π a b^2. The ratio of volumes is Vol(E_x)/Vol(E_y) = b a^2/a b^2 = a/b. Since a > b, Vol(E_x) > Vol(E_y).

MergeHEIR ✓: The ellipse has semi-axis lengths a and b. When rotated about the x-axis, the ellipsoid E_x has semi-axes a, a, b. When rotated about the y-axis, the ellipsoid E_y has semi-axes b, b, a. The volume of E_x is 4/3π a^2 b, while the volume of E_y is 4/3π b^2 a. Simplifying their ratio gives Vol(E_x)/Vol(E_y) = a/b. Given a > b, Vol(E_x) > Vol(E_y).

w/o constraint ✗: The ellipse has semi-axis lengths a (along the y-axis) and b (along the x-axis). When rotated about the x-axis, the ellipsoid E_x has semi-axes b, a, a. When rotated about the y-axis, the ellipsoid E_y has semi-axes a, b, b. The volume of an ellipsoid with semi-axes r_1, r_2, r_3 is given by 4/3π r_1 r_2 r_3. For E_x, Vol(E_x) = 4/3π b a^2. For E_y, Vol(E_y) = 4/3π a b^2. Given a > b, it follows that a^2 > b^2, thus Vol(E_x) < Vol(E_y).

Figure 7: Visualization results comparing the merged model, MergeHEIR, and unconstrained post-merge adaptation on geometry reasoning task. Incorrect or visually unsupported response spans are highlighted in red.

Expert-task
OCR



What brand is this appliance?

Merged ✓: The appliance has the word "babycook" prominently displayed on its front panel. This text is in lowercase letters and is positioned below a circular control panel with symbols indicating settings. Additionally, the name "BEABA" is illuminated at the bottom of the appliance's base, which likely represents the brand. The combination of these labels clearly identifies the brand of the appliance as BEABA.

MergeHEIR ✓: The appliance has the word "babycook" prominently displayed on its front panel, indicating its purpose. Below this, there's a circular button with symbols that suggest settings or controls for the device. The brand name "BEABA" is illuminated at the bottom of the appliance, confirming its manufacturer. The combination of these elements clearly identifies the brand of the appliance as BEABA. Conclusion: The appliance's branding and design indicate that it is manufactured by BEABA. Final answer: BEABA

w/o constraint ✗: The appliance has a label that reads "babycook" on its front panel, indicating the brand of the product. This label is prominently displayed, making it clear that the appliance is branded as "babycook." The presence of this label on the front of the appliance directly answers the question about the brand of the appliance. Conclusion: The label "babycook" on the front panel identifies the brand of the appliance. Final answer: Babycook

Expert-task
OCR



What is the name of the company?

Merged ✓: The company's name is prominently displayed at the top right corner of the page as "ITC Limited." Below the logo, there is a description of Sunfeast's product offerings. The text states that Sunfeast straddles all segments in the biscuit category and offers high-quality products in exciting and innovative formats. This description emphasizes Sunfeast's role within the company but does not directly state the company name. However, the company name "ITC Limited" is clearly visible and is the most direct answer to the question.

MergeHEIR ✓: The text at the top right corner of the image reads "ITC Limited REPORT AND ACCOUNTS 2013." This indicates that ITC Limited is the company being referred to in the context of the report and accounts for the year 2013. The presence of the Sunfeast logo suggests that it is a subsidiary or product line under ITC Limited. Conclusion: The text at the top right corner of the image provides the company's name. Final answer: ITC Limited

w/o constraint ✗: The image features a black and white photograph of various biscuit products, with a logo that reads "Sunfeast" prominently displayed in the top left corner. This logo is the most direct answer to the question of what the name of the company is, as it is the only identifiable branding present in the image. Conclusion: The presence of the "Sunfeast" logo indicates the company's branding. Final answer: Sunfeast

Figure 8: Visualization results comparing the merged model, MergeHEIR, and unconstrained post-merge adaptation on OCR task. Incorrect or visually unsupported response spans are highlighted in red.

Expert-task VQA



How many sides do these signs normally have?

Merged ✓: The stop sign in the image is an octagonal sign, which typically has eight sides. This can be determined by observing the straight edges and corners of the sign, which form a regular polygon with eight sides. The octagonal shape is characteristic of stop signs in many countries, including the United States, where they are used to control traffic at intersections or other road junctions. Conclusion: An octagon has eight sides. Final answer: 8

MergeHEIR ✓: The image shows a stop sign, which is typically an octagonal sign with a straight edge at each corner. The stop sign has eight sides because it is designed to be easily identifiable from any angle due to its octagonal shape. Each corner of the octagon is sharp and defined by the edges, making it easy to count as having eight sides. Conclusion: A stop sign typically has eight sides. Final answer: 8

w/o constraint ✗: The image shows a stop sign, which is typically an octagonal shape with four corners and straight edges defining its shape. The sign appears to be slightly distorted due to the reflection on the windshield, but it maintains the characteristic octagonal form. The four corners and the straight edges suggest that the sign has four sides. Conclusion: The stop sign in the image has four sides. Final answer: 4

Expert-task VQA



What type of animal is walking across the grassy pasture?

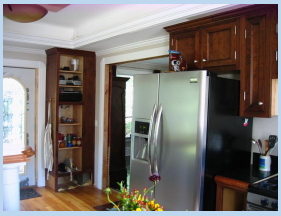
Merged ✓: The image shows a man on horseback and a herd of sheep grazing in a pasture. The man is riding the horse, which is in motion, while the sheep are scattered across the field, some standing and others grazing. The horse's movement suggests it is walking across the grassy pasture, as indicated by its forward stance and the position of its legs. The sheep, however, are stationary, focused on feeding. Conclusion: Horse Final answer: Horse

MergeHEIR ✓: In the image, there is a man riding a horse amidst a flock of sheep grazing on the grassy pasture. The man, dressed in red, sits atop the dark-colored horse, which is in motion. The sheep are scattered across the field, some standing while others graze. The horse's movement suggests it is walking across the grassy pasture as described in the question. Conclusion: Horse Final answer: Horse

w/o constraint ✗: In the image, there is a man riding a horse through a grassy pasture. The horse is moving at a steady pace, and the rider appears to be guiding it. Additionally, there is a flock of sheep grazing on the grass nearby. The sheep are scattered throughout the field, some standing close together while others are more spread out. The movement that best matches the question "What type of animal is walking across the grassy pasture?" is the sheep, as they are the primary animals seen grazing in the field. Their presence indicates that they are indeed walking across the grassy pasture.

Figure 9: Visualization results comparing the merged model, MergeHEIR, and unconstrained post-merge adaptation on VQA task. Incorrect or visually unsupported response spans are highlighted in red.

Hallucination-task
POPE



Is there a suitcase in the image?

Merged ✗: Yes. The image shows a kitchen with wooden cabinets, a stainless steel refrigerator, and an open shelving area. The shelving contains various items like jars and containers. On the bottom shelf, there is a dark rectangular object that resembles a suitcase due to its shape and size. It appears to be placed among other items, blending somewhat with the surroundings. Conclusion: The dark rectangular object on the bottom shelf resembles a suitcase based on its shape and placement.

MergeHEIR ✓: No. The image depicts a kitchen with wooden cabinets and a stainless steel refrigerator. To the left, there is an open shelving area with various items, including jars and kitchen supplies. Among these items, there appears to be a dark rectangular object on the bottom shelf. It resembles a suitcase due to its shape and size, but it's more likely a storage container given its placement in a kitchen setting. Conclusion: The dark rectangular object resembles a storage container rather than a suitcase.

w/o constraint ✓: No. The image shows a kitchen with wooden cabinets, a stainless steel refrigerator, and an open shelving area. The shelving contains various items like bottles and containers. In the lower part of this shelving unit, there is a dark rectangular object that resembles a suitcase due to its shape and placement among other items. However, it appears more like a storage box or cabinet rather than a typical suitcase. Conclusion: The dark rectangular object does not clearly resemble a suitcase.

Hallucination-task
POPE



Is there a dog in the image?

Merged ✗: Yes, there is a dog in the image. The dog is located on the right side of the image, near the bottom. It appears to be a small breed with light-colored fur. The dog is wearing a collar and has a tag attached to it. Conclusion: There is a dog present in the image, but it is not clearly identifiable due to the angle and distance.

MergeHEIR ✓: No. In the foreground, there is a slice of chocolate cake with a layer of caramel on top, accompanied by a fork. The cake is placed on a decorative tray with a floral pattern. In the background, there is another slice of cake on a plate, and a bottle of Guinness Extra Stout beer is visible. The presence of the beer suggests that it might be used as a pairing for the cake, but there is no direct visual evidence of a dog in the image. Conclusion: There is no clear visual evidence of a dog in the image.

w/o constraint ✓: No. The main objects in the foreground are a slice of chocolate cake on a metal tray with a fork resting on it. The surrounding scene includes another slice of cake on a plate and a bottle of Guinness Extra Stout beer. There is no clear visual evidence in the image that suggests the presence of a dog. Conclusion: No, there is no dog present in the image.

Figure 10: Visualization results comparing the merged model, MergeHEIR, and unconstrained post-merge adaptation on POPE. Hallucinated or visually unsupported content is highlighted in red.

Hallucination-task
Object Hal



Provide a thorough description of the given picture.

Merged ✗: In the image, there is a group of people standing in a living room, seemingly engaged in an activity that involves holding and using remote controls. There are at least five individuals visible in the scene, with some of them holding remotes in their hands. The living room has a couch positioned against one wall, and a chair can be seen on the right side of the room. A television set is mounted on the wall, located towards the left side of the room. The atmosphere appears to be casual and social, as the group seems to be enjoying their time together.

MergeHEIR ✓: In the image, there is a group of people gathered in a living room, with some standing and others sitting on couches. They appear to be engaged in playing a video game together. The individuals are holding game controllers and seem to be enjoying their time together. The living room has a cozy atmosphere, with multiple couches and chairs arranged for comfortable seating. There are also some books visible in the background, suggesting that this might be a space where people come to relax and spend quality time with friends or family.

w/o constraint ✓: In the image, there is a group of people gathered in a living room, with some standing and others sitting on couches. The individuals appear to be engaged in an activity involving video games, as they are holding game controllers and seem to be focused on something off-camera. The room has a cozy atmosphere with multiple couches and chairs scattered throughout. There are also several books visible in the background, adding to the casual and relaxed setting of the gathering.

Hallucination-task
Object Hal



What is this photo about? Please answer in great detail.

Merged ✗: The image depicts a train station platform with a silver passenger train parked on the right side of the platform. The train has several windows and doors, and it appears to be a modern design. There are multiple people visible on the platform, some walking and others standing near the train. One person is carrying a suitcase, indicating that they might be either arriving or departing from the train. Additionally, there are two traffic lights positioned above the platform, likely used for controlling the movement of trains and vehicles at the station.

MergeHEIR ✓: The image captures a scene at a train station where a silver train is parked on the tracks. A person, possibly a passenger, is walking along the platform near the train. The platform has a distinctive white and gray color scheme, with a noticeable ramp for accessibility. There are also some other people visible in the background, suggesting that the station is not deserted. The overall atmosphere of the image conveys a sense of everyday life at a train station.

w/o constraint ✓: The image captures a scene at a train station, where a silver and white passenger train is parked on the tracks. A person can be seen walking along the platform near the train, possibly preparing to board or having just disembarked. There are several other people visible in the background, some of whom might be waiting for their trains or simply passing by. The presence of multiple individuals suggests that this is a busy time at the station, with passengers arriving and departing from the train.

Figure 11: Visualization results comparing the merged model, MergeHEIR, and unconstrained post-merge adaptation on ObjHalBench. Hallucinated objects and visually unsupported scene details are highlighted in red.