3DM: Distill, Dynamic Drop, and Merge for Debiasing Multi-modal Large Language Models

Anonymous ACL submission

Abstract

The rapid advancement of Multi-modal Language Models (MLLMs) has significantly enhanced performance in multimodal tasks, yet these models often exhibit inherent biases that compromise their reliability and fairness. Traditional debiasing methods face a trade-off between the need for extensive labeled datasets and high computational costs. Model merging, which efficiently combines multiple models into a single one, offers a promising alternative but its usage is limited to MLLMs with the same architecture. We propose **3DM**, a novel framework integrating **Distill**, **Dynamic Drop**, and **Merge** to address these challenges. 3DM employs knowledge distillation to harmonize models with divergent architectures and introduces a dynamic dropping strategy that assigns parameter-specific drop rates based on their contributions to bias and overall performance. This approach preserves critical weights while mitigating biases, as validated on the MMSD2.0 sarcasm detection dataset. Our key contributions include architecture-agnostic merging, dynamic dropping, and the introduction of the Bias Ratio (BR) metric for systematic bias assessment. Empirical results demonstrate that 3DM outperforms existing methods in balancing debiasing and enhancing the overall performance, offering a practical and scalable solution for deploying fair and efficient MLLMs in real-world applications.

1 Introduction

Recent advances in MLLMs (Liu et al., 2023; Chen et al., 2024; GLM et al., 2024; Zhu et al., 2024a) have shown remarkable performance in various multimodal tasks, ranging from image captioning (Wang et al., 2024) and visual question answering (Li et al., 2023) to a nuanced multimodal sarcasm detection (Tang et al., 2024). Despite the progress, MLLMs are prone to biased predictions (Cui et al., 2023; Han et al., 2024). For instance, Table 1 shows

Model	Acc	Precision	Recall	F1
LLaVA-v1.5-7b	0.516	0.469	0.947	0.628
ChatGLM4-9b	0.689	0.725	0.450	0.555

Table 1: The performance of LLaVA-v1.5-7b with a positive bias, and ChatGLM4-9b with a negative bias.

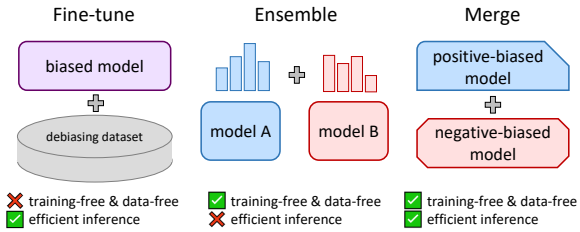

Figure 1: Conceptual comparison of model merging with fine-tuning and ensembling in the context of debiasing. Model merging is training-free and benefits from efficient inference.

that LLaVA (Liu et al., 2023) favors classifying inputs as sarcastic (positive-biased model), whereas ChatGLM (GLM et al., 2024) has the opposite tendency (negative-biased model). This may be a symptom of hallucinating answers from spurious correlations seen in the dataset (Bai et al., 2024). MLLMs' inherent biases compromise their reliability and fairness for deployment in real-world applications. Thus, enhancing MLLMs' accuracy *and* ensuring minimal bias have significant practical implications.

In this paper, we present the first attempt, to the best of our knowledge, at *merging* models (Yang et al., 2024; Ramé et al., 2023; Lin et al., 2024) to debias MLLMs and showcasing its general effectiveness. Existing debiasing or dehallucination methods have relied on labeled datasets for fine-tuning (Chen et al., 2021; Guo et al., 2022; Liu et al., 2024) or repetitive inference for ensembling predictions (Clark et al., 2019), both of which incur non-trivial computational overhead. In contrast, our approach collects a positive-biased model and a negative-biased model, then merges them in the pa-

rameter space without the need for additional training and repeated inference. Through this process, biases in opposite directions are canceled out efficiently. See Fig. 1 for the conceptual comparisons between merging and the traditional approaches.

However, merging MLLMs for debiasing faces several challenges: (1) Merging models often requires the same architecture across models to allow for parameter-wise operations, a condition rarely satisfied in the rapidly evolving ecosystem of MLLMs (Zhang et al., 2024); (2) Reducing the bias alone does not always translate to improved accuracy—debiased models may struggle with task performance. This highlights the need to refine existing merging methods (Ilharco et al., 2022; Yadav et al., 2024; Yu et al., 2024) through the lens of reducing bias and enhancing accuracy.

We propose **3DM** (**D**istill, **D**ynamic **D**rop and **M**erge), an architecture-agnostic merging framework designed to address these challenges. First, knowledge distillation (Gou et al., 2021) bridges architectural gaps between models, enabling parameter-level merging even for heterogeneous MLLMs. Second, we introduce a dynamic dropping strategy that assigns parameter-specific drop rates based on their influence on bias and accuracy. This is motivated by a recent merging method—DARE (Yu et al., 2024)—that sparsifies parameters by a uniform chance of dropout and treats all parameters equally.

We first conduct experiments on the MMSD2.0 (Qin et al., 2023) sarcasm detection dataset and measure models' bias with our newly proposed metric, **Bias Ratio** (Sec. 3). The results demonstrate that (1) merging methods are in common effective in reducing bias, and that (2) 3DM significantly outperforms DARE and other baselines in accuracy, F1-score, and Bias Ratio. In addition, experiments on MMSD1.0 (Cai et al., 2019) further validate that 3DM generalizes well across different datasets. Compared with methods requiring hyperparameter search over the validation data, 3DM does not contain such hyperparameters, making it convenient for implementation.

In essence, our contributions are as follows:

1. **Architecture Alignment**: A distillation pipeline that aligns MLLM architectures, preserving their original bias and accuracy.

2. **Dynamic Dropping**: A merging strategy that adaptively adjusts drop rates to reduce biases and improve accuracy.

3. **Bias Ratio**: A metric for quantifying bias direction and magnitude, contributing to ongoing efforts in bias quantification.

4. **Empirical Validation**: Extensive experiments demonstrating 3DM's effectiveness in terms of both debiasing and accuracy enhancement.

2 Related Work

2.1 Model Debiasing

Existing debiasing mechanisms in the literature can be classified into two primary categories (Mehrabi et al., 2021; Pessach and Shmueli, 2022): training-based debiasing and training-free debiasing. Training-based debiasing approaches necessitate modifications to the training dataset (Li and Vasconcelos, 2019), demonstrating notable effectiveness while requiring extensively annotated training data. Conversely, training-free debiasing methodologies primarily focus on altering the output distribution (Kamiran et al., 2012), with ensembling emerging as a crucial technique in this domain (Clark et al., 2019).

A notable example of ensembling is the blindfolding strategy proposed by Zhu et al. (2024b), which involves masking specific portions of the input and computing the final output score as the difference between traditional inference, fully blindfolded inference, and partially blindfolded inference. Although ensembling methods eliminate the need for training processes, they incur substantial computational overhead due to the requirement for multiple inference operations. In light of these considerations, we propose our merging strategy as an effective compromise between these two approaches, offering the dual advantages of eliminating excessive inference requirements while maintaining a label-free training process.

2.2 Model Merging

Garipov et al. (2018); Draxler et al. (2018) demonstrated that two models trained from different initializations can be connected by a path of non-increasing loss in the loss landscape, referred to as *model connectivity*. If the two models share a significant part of the optimization trajectory (e.g., pre-trained model), they are often connected by a *linear* path (Frankle et al., 2020; Neyshabur et al., 2020; Mirzadeh et al., 2021), where interpolating along the path potentially leads to better accuracy and generalization (Izmailov et al., 2018). This

property has been exploited as simply averaging the weights of numerous models fine-tuned from different hyperparameters to improve accuracy (Wortsman et al., 2022), popularizing *model merging* as an efficient alternative to ensemble in combining models without additional instruction tuning.

The success of averaging fine-tuned models has led to a surge of merging methods, aimed at steering models' behavior in desired way. A prominent example is multi-task learning via merging, where accounting for parameter importance (Matena and Raffel, 2022) and minimizing prediction differences to the fine-tuned models (Jin et al., 2022) are shown to be effective. While these methods relies on statistics that are expensive to compute, Task Arithmetic (Ilharco et al., 2022) presents a cost-effective and scalable method of adding the weighted average of task vectors (i.e., fine-tuned part of parameters) to the pre-trained model. Subsequent studies are dedicated to pre-processing task vectors to reduce interference across models (Yadav et al., 2024; Yu et al., 2024; Deep et al., 2024). Moreover, distillation is proposed for architecture alignment by FUSECHAT (Wan et al., 2024). Our distill-merge pipeline and dynamic dropping strategy aligns with this line of research, however we are focused on editing task vectors to reduce bias and improve accuracy.

3 Bias Ratio

The metrics used to evaluate a model's bias (or fairness) remain a subject of ongoing dialogue, with no clear consensus yet (Caton and Haas, 2024). Previous studies have employed various evaluation metrics to assess bias. In this work, we introduce the Bias Ratio (BR) as a measure of a model's bias, which is based on the quantities of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN).

$$BR = \frac{FP}{FP + TN} - \frac{FN}{FN + TP} \tag{1}$$

The Bias Ratio (BR) ranges from -1 to 1, where its absolute value indicates the magnitude of bias, and its sign denotes the direction. For instance, a BR value of 0.8 reflects a relatively high degree of positive bias, whereas a BR of -0.1 suggests a relatively low degree of negative bias. While previous studies have primarily conducted qualitative analyses of bias based on TP, TN, FP, and FN, we propose a quantitative metric to systematically assess both the degree and direction of bias.

4 Method

Focusing on a two-way classification task (e.g., sarcasm detection), suppose we are given two MLLMs, a **positive-biased model** and a **negative-biased model**: A positive-biased model tends to classify an input as positive sample, represented by high recall and low precision (Table 1). Likewise, a negative-biased model is inclined to classify an input as negative sample, represented by low recall and high precision (Table 1).

Then we apply our proposed **3DM** framework following three steps, as illustrated in Fig. 2: (1) knowledge distillation for architecture alignment; (2) dynamic dropping strategy that filters out delta parameters based on the contribution to accuracy and bias; (3) merging the positive-biased delta parameters and negative-biased delta parameters to cancel out predictive bias.

4.1 Architecture Alignment via Distillation

An intuitive way to mitigate bias is to merge a positive-biased model and a negative-biased model to cancel out the bias. However, the diverse ecosystem of MLLMs makes it challenging to guarantee those two models to share the same architecture, blocking them from being merged through parameter-wise operations. Knowledge distillation provides a viable solution by reshaping the two models into the same architecture, while preserving the predictive accuracy and bias of each model. Hence we start by distilling the two types of models and proceed to model merging (Sec. 4.2, 4.3) on the basis of compatible architecture.

Knowledge distillation (Gou et al., 2021) typically follows a teacher-student structure, where the teacher model's output (generated by the prompt proposed in Sec. 5.1.2) supervises the student model such that the student model inherits the behavior of the teacher model. Note that the student model is not required to be smaller than the teacher model in our case, as our goal of knowledge distillation does not lie in compression.

Specially, we fine-tune the pre-trained model using pseudo labels generated by a teacher model (i.e., either positive-biased model or negative-biased model). We minimize cross-entropy loss evaluated on the pseudo labels:

$$\mathcal{L}_{ce} = -\sum_{t=1}^{m} \log P(\hat{y}_t \mid x, \hat{y}_{<t}) \tag{2}$$

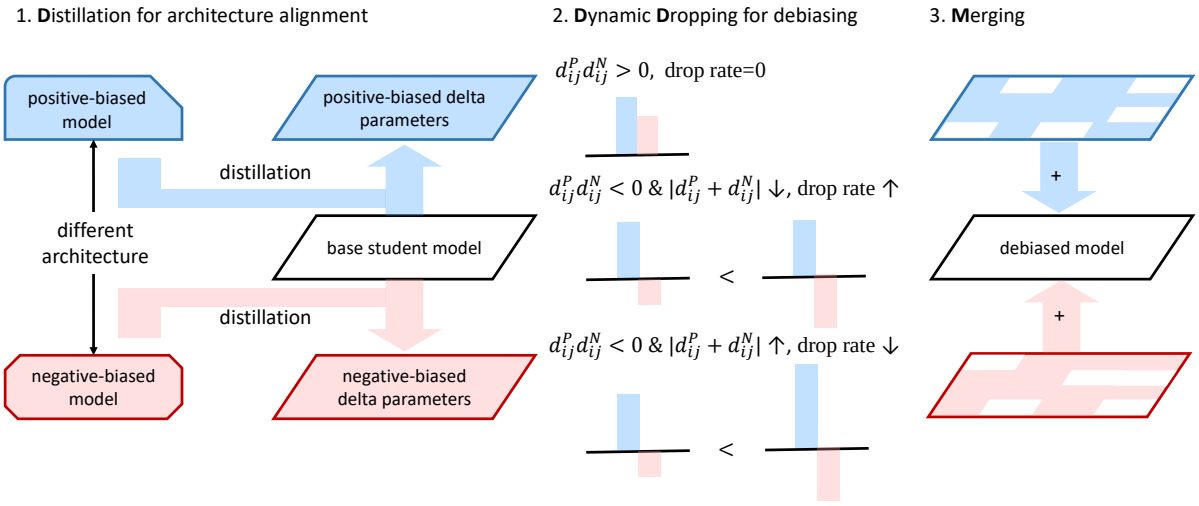

Figure 2: Overview of **3DM** framework. First, positive-biased model and negative-biased model are distilled to a base student model to share an identical architecture. Second, dynamic dropping assigns a drop rate to each delta parameter based on the discrepancy between the positive-biased model and the negative-biased model. Then, sparse task vectors after dropping are added to the base model to build a debiased model.

where $\{\hat{y}_i\}_{i=1}^m$ is the pseudo label of length m generated by teacher model. In the context of sarcasm detection, x is a pair of input text and image and $\{\hat{y}_i\}_{i=1}^m$ is an answer sequence indicating whether the input pair contains sarcasm.

4.2 Dynamic Dropping

Merging a positive-biased model and a negative-biased model is in general effective in alleviating the bias. In this section, we further propose dynamic dropping, aiming to improve accuracy and F1-score while simultaneously reducing bias.

In model merging, delta parameters are defined as the subtraction of parameters of base model from the fine-tuned model, and they can be understood as task vectors (Ilharco et al., 2022). Findings by Yu et al. (2024) suggest that one could randomly zero-out delta parameters of an LLM with a drop rate of p and re-scale the remaining ones by $1/(1-p)$ without impacting the model's performance. This sparsification strategy—coined as DARE—has been shown to be helpful in reducing parameter interference among the models to be merged. However, DARE assigns the same drop rate for all delta parameters. Conversely, the drop rate of a delta parameter should ideally be determined by its contribution to improving accuracy and reducing bias. That is, "important" delta parameters should be preserved by a higher probability.

Delta Parameters We merge the distilled positive-biased model and negative-biased model by editing their respective delta parameters and combining those to the base student model. Delta parameters are defined as:

$$d_{ij}^P = W_{ij}^P - W_{ij}^{base} \tag{3}$$

$$d_{ij}^N = W_{ij}^N - W_{ij}^{base} \tag{4}$$

where $W^{base} \in \mathbb{R}^{m \times n}$ is a parameter matrix of the base model and W^P and W^N are the ones distilled from positive-biased model and negative-biased model, respectively. i and j denotes position (i, j) of the parameter in W.

Classification of Delta Parameters. In terms of which delta parameters are more responsible for boosting model's accuracy and suppressing bias, we suggest the following criteria for classifying delta parameters into three categories:

1. **Bias-free Delta** (Fig. 3(a)), where d_{ij}^P and d_{ij}^N have the same sign, i.e. $d_{ij}^P d_{ij}^N > 0$.

2. **Unidirectional Delta** (Fig. 3(b)), where d_{ij}^P and d_{ij}^N have the opposite sign, and the magnitude of one dominates the magnitude of the other, i.e. $d_{ij}^P d_{ij}^N < 0$ and $|d_{ij}^P + d_{ij}^N| > c$ where c is a threshold.

3. **Bidirectional Delta** (Fig. 3(c)), where d_{ij}^P and d_{ij}^N have the opposite signs, and the magnitudes of both are comparable, i.e. $d_{ij}^P d_{ij}^N < 0$ and $|d_{ij}^P + d_{ij}^N| < c$.

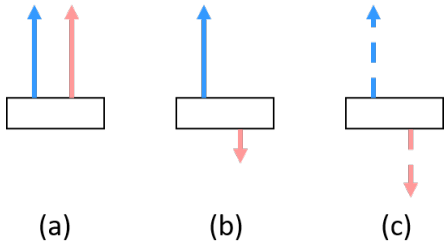

Figure 3: Configurations of delta parameters under different conditions. The delta parameter from the positive-biased model (blue) and the negative-biased model (pink) can exhibit (a) the same sign, (b) opposite signs with a large magnitude difference, or (c) opposite signs with comparable magnitudes (dashed).

The above criteria follows from our hypothesis about the roles of delta parameters: (1) Delta parameters with the same sign indicates a consistent direction in parameter updates by the positive-biased model and negative-biased model, potentially implying salient deltas that are associated with accuracy; (2) Given that positive-biased model and negative-biased model are best distinguished by their bias, those delta parameters with the opposite sign have greater contribution to bias, in which bidirectional delta may lead to severer interference while merging than unidirectional delta.

Towards Adaptive Drop Rate via Dynamic Dropping. Our classification of delta parameters motivates us to assign increasing drop rates for bias-free delta, unidirectional delta, and bidirectional delta. In light of this, we present **dynamic dropping**, a strategy of applying adaptive drop rate p_{ij} at a parameter-level:

$$p_{ij} = \begin{cases} 0 & \text{if } d_{ij}^P d_{ij}^N \geq 0 \\ 1 - \frac{|d_{ij}^P + d_{ij}^N|}{|d_{ij}^P| + |d_{ij}^N|} & \text{if } d_{ij}^P d_{ij}^N < 0 \end{cases} \quad (5)$$

Here, p_{ij} is the drop rate betwen 0 and 1. Intuitively, Eq. 5 excludes bias-free delta from dropout operation, and for $d_{ij}^P d_{ij}^N < 0$, Eq. 5 imposes higher drop rate on bidirectional delta than on unidirectional delta. Noted, we implement a synchronized dropping mechanism where delta parameters at the same position are either dropped or retained simultaneously.

After dynamic dropping, each remaining delta parameter is rescaled by $1/(1 - p_{ij})$ to preserve the expectation of input embeddings, as elaborated in Yu et al. (2024).

MMSD1.0	All	Positive	Negative
Train	19816	8642	11174
Validation	2410	959	1451
Test	2409	959	1450

Table 2: Composition of MMSD1.0 dataset.

MMSD2.0	All	Positive	Negative
Train	19816	9572	10240
Validation	2410	1042	1368
Test	2409	1037	1372

Table 3: Composition of MMSD2.0 dataset.

4.3 Parameter Merging

Let the delta parameters after dynamic dropping and re-scaling be $\hat{d}_{ij}^P$ and $\hat{d}_{ij}^N$. Then the average of $\hat{d}_{ij}^P$ and $\hat{d}_{ij}^N$ is added to the base model parameter to derive the merged parameter W_{ij}^*:

$$W_{ij}^* = 0.5\hat{d}_{ij}^P + 0.5\hat{d}_{ij}^N + W_{ij}^{base} \quad (6)$$

W^* is the final model weights of our 3DM method, where bias is reduced and the overall perforamnce is boosted.

5 Experiments

In this section, we first introduce the experimental setup, including the datasets, prompts, base models, and baselines. Then, we design experiments to validate our method. Distillation (5.2), merging (5.3), ensembling (5.5), and generalizability (5.6) are analyzed in this section.

5.1 Experimental Setup

5.1.1 Dataset

We validate our approach on MMSD2.0 (Qin et al., 2023), a multi-modal sarcasm detection dataset whose test set contains 2409 sentences along with images, and we test the generalizability on MMSD1.0 (Cai et al., 2019). See Table 3 for dataset statistics.

5.1.2 Implementation Details

Prompt Template. We use a fixed template to format the prompt. The template is carefully designed to ensure consistency across all samples and to minimize any potential bias introduced by expression. The following prompt template is used:

"<image>This is an image with: " as the caption. Is the image-text pair sarcastic?First answer the question with yes or no, then explain your reasons."

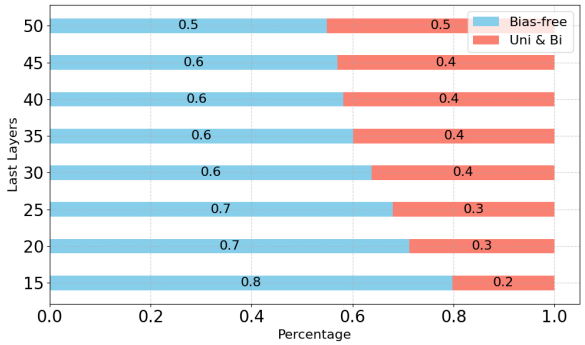

Figure 4: Visualization of the percentage of different types of parameters we encounter when merging. Blue and red represent the sign of d_{ij}^P and d_{ij}^N being same and different. The y axis represents the last y layers in the model.

Knowledge Distillation. To examine the validity of knowledge distillation in transferring both accuracy and bias from the teacher models, we choose LLaVA-1.5-7b (Liu et al., 2023) and InternVL-2.5-8b (Chen et al., 2024) as student models (base models), and select LLaVA-1.5-7b and ChatGLM4-9b (GLM et al., 2024) as teacher models. Our choices of small-sized MLLMs are intended to show that 3DM does not necessitate any pre-existing sarcasm detection capabilities in the student models.

Dynamic Dropping. To assess the effectiveness of dynamic dropping, we fix InternVL as the base model and obtain positive and negative delta parameters distilled from LLaVA and ChatGLM, respectively. Choosing InternVL is informed by empirical observations (See Table 4), indicating that the pre-trained InternVL exhibits weak bias ($BR = 0.185$) relatively, and has no pre-existing knowledge of sarcasm detection ($Acc \approx 0.5$), making it an appropriate candidate for our experiment.

Hyperparameter Searching The fixed drop rate of DARE and our ablation study (unused in 3DM) is set to 0.7, which is the result of tuning on the validation set of MMSD2.0 (Table 8).

5.1.3 Baselines

We compare 3DM with merging baselines including **Average Merging** (Wortsman et al., 2022), **TIES** (Yadav et al., 2024) and **DARE** (Yu et al., 2024), in addition to ensembling. **TIES** merges models by drop-elect-merge operations, where the "drop" step mitigates interference by removing redundant delta parameters based on their magnitudes.

5.2 Distillation Experiments

Table 4 and Table 5 present the performance and bias of both teacher models and student models after distillation. As observed, the student models effectively inherit the bias direction of their respective teacher models, while also achieving improved performance, except for the F1-score of LLaVA base model distilled from ChatGLM4. These results demonstrate that we can successfully prepare models for merging-based debiasing through distillation, without the need for elaborate training labels.

For the subsequent merging experiments, we apply our proposed distill-merge pipeline for debiasing when LLaVA serves as the student model. For InternVL as the student model, we further compare our proposed dynamic dropping method with baseline merging approaches, as InternVL itself exhibits a weak bias and can therefore be used as the base model.

5.3 Merging Experiments

This section analyzes the results on the testing set of MMSD2.0 (Qin et al., 2023).

For the LLaVA base model, we compare the performance of merged models against their original counterparts. As shown in Table 5, the average merging strategy fails to surpass the negative-biased model. However, applying DARE (fixed dropping) leads to significant improvements, with both accuracy and F1-score approximating those of the zero-shot inference of teacher models, alongside a substantial reduction in the absolute value of BR. This highlights the potential of our distill-merge pipeline for debiasing tasks when combined with a well-designed merging method.

Similarly, Table 4 implies that all merging strategies significantly reduce the absolute value of BR, compared to student base models distilled from biased models into InternVL, further demonstrating the effectiveness of our distill-merge pipeline. Moreover, 3DM, which introduces a tailored dropping mechanism in the "merge" phase, achieves state-of-the-art performance in accuracy, F1-score and BR across all merging approaches. This underscores the effectiveness and superiority of dynamic dropping.

We provide insights into why 3DM outperforms other merging approaches. While TIES mitigates interference between delta parameters through sign selection, it struggles in cases like Fig. 3(c), where

Model	Method	Strategy	Acc	F1	TP	FP	TN	FN	Bias direction	Bias Ratio
LLaVA-v1.5-7b	/	zero-shot inference	0.516	0.628	982	1110	262	55	+	0.765
ChatGLM4-9b	/	zero-shot inference	0.689	0.555	466	177	1195	571	-	-0.422
InternVL-2.5-8b	/	zero-shot inference	0.499	0.509	625	796	576	412	weak	0.183
InternVL-2.5-8b	Distillation	positive learning	0.543	0.629	934	998	374	103	+	0.628
		negative learning	0.644	0.428	321	141	1231	716	-	-0.588
	Merging	average merging	0.688	0.614	599	314	1058	438	weak	-0.194
		TIES	0.648	0.484	397	208	1164	640	-	-0.466
		DARE	0.684	0.609	592	316	1056	445	weak	-0.199
		3DM	0.697	0.643	658	351	1021	379	weak	-0.110
	Ensembling	ensembling	0.663	0.516	431	205	1159	605	-	-0.434

Table 4: Results of applying multiple debiasing methods, including average merging, fixed dropping (DARE), our proposed 3DM and ensembling methods. "+" and "-" indicate that the model tends to give positive and negative answers.

Model	Method	Strategy	Acc	F1	TP	FP	TN	FN	Bias direction	Bias Ratio
LLaVA-v1.5-7b	Distillation	positive learning	0.516	0.628	982	1110	262	55	+	0.765
		negative learning	0.710	0.666	572	233	1139	465	-	-0.279
	Merging	average merging	0.671	0.474	357	113	1259	680	-	-0.573
		DARE	0.714	0.649	617	290	1082	400	weak	-0.189
	Ensembling	ensembling	0.716	0.693	773	421	951	264	weak	0.05

Table 5: Results of applying debiasing methods on LLaVA-based models. Because LLaVA itself has a positive bias, we apply the original model to the "positive learning" row.

Model	Ablation type	Acc	F1	TP	FP	TN	FN	Bias direction	Bias Ratio
InternVL-2.5-8b	Bias-free	0.663	0.661	790	564	808	247	weak	0.173
	Uni. & Bi.	0.649	0.477	386	195	1177	651	-	-0.486
	Async.	0.691	0.637	654	362	1010	383	weak	-0.105
	3DM	0.697	0.643	658	351	1021	379	weak	-0.110

Table 6: Ablation study. In 3DM, we classify each position into two categories based on their signs. In this part, we remove one of them, and test the method's performance. We also examine the performance of the synchronized dropping mechanism.

it may remain on the wrong side. DARE, on the other hand, applies a uniform drop rate to all delta parameters, disregarding their distinct roles. However, as illustrated in Fig. 4, the proportion of bias-free delta parameters (blue) is comparable to that of unidirectional and bidirectional delta parameters (red), highlighting the necessity of dynamically assigning drop rates based on their roles (Sec. 4.2) and merging conditions (Fig. 3).

5.4 Ablation Study

To better understand the role of dynamic dropping in 3DM, we conduct an ablation study by modifying key components of the mechanism.

As shown in Table 6, *Bias-free*, which replaces the adaptive drop rates of unidirectional and bidirectional deltas in Eq. 5 with a fixed rate, results in lower accuracy, along with a higher absolute value of BR. This suggests that a fixed drop rate fails to effectively leverage the variations in d_{ij}^P and d_{ij}^N. Similarly, *Uni. & Bi.*, which follows DARE by applying a fixed drop rate to bias-free deltas instead of fully preserving them, also performs suboptimally compared to 3DM.

Additionally, we evaluate a less aggressive strategy than 3DM (synchronized dropping), called *Async.*, which drops delta parameters individually based on Eq. 5. This reduces the likelihood of simultaneously eliminating delta parameters[1] in the scenario shown in Fig. 3(c). While this approach achieves a slightly lower BR, it suffers a small drop in accuracy and F1-score. This could be due to it tends to retain a delta parameter in a single wrong direction, thus degenerating into TIES. This reinforces the effectiveness of the synchronized dropping mechanism, which not only preserves

[1]The final parameter at that position defaults to the base model's.

Model	Method	Strategy	Acc	F1	TP	FP	TN	FN	Bias direction	Bias Ratio
LLaVA-v1.5-7b	/	zero-shot inference	0.445	0.587	952	1331	119	7	+	0.911
ChatGLM4-9b	/	zero-shot inference	0.713	0.587	492	225	1225	467	-	-0.332
InternVL-2.5-8b	/	zero-shot inference	0.483	0.473	559	846	604	400	weak	0.166
InternVL-2.5-8b	Distillation	positive learning	0.501	0.592	871	1113	337	88	+	0.676
		negative learning	0.667	0.466	350	193	1257	609	-	-0.502
	Merging	average merging	0.691	0.619	605	390	1060	354	weak	-0.100
		TIES	0.676	0.519	422	244	1206	537	-	-0.392
		DARE	0.686	0.613	600	397	1053	359	weak	-0.101
		3DM	0.691	0.636	651	436	1014	308	weak	-0.020
	Ensembling	ensembling	0.680	0.530	433	241	1200	526	-	-0.381

Table 7: Performance of methods on MMSD1.0 dataset.

flexibility in handling unidirectional deltas but also forces the dropping of delta parameters in the bidirectional delta condition, where they may introduce greater bias or interference.

5.5 Comparison with Ensemble

We conduct a systematic comparison between our 3DM method and ensemble approaches. For sarcasm detection, ensemble methods generate individual probability distributions and aggregate them for final predictions. While achieving acceptable performance, these methods incur substantial computational overhead, with inference costs scaling as $O(n)$, compared to $O(1)$ for merging methods. This establishes a fundamental advantage for merging approaches.

In our experiments, we implement basic averaging ensemble, where model distributions are arithmetically averaged. As shown in Table 4, Table 5, and Table 7, this approach demonstrates limited effectiveness on the testing dataset. Although more sophisticated ensemble techniques might surpass 3DM's performance, they cannot overcome the inherent computational limitations of all ensemble methods, which remain a fundamental constraint compared to merging approaches.

5.6 Generalizability Analysis

In order to test the generalizability of our method, we validate our method on the testing set of MMSD1.0 (Cai et al., 2019). We retain the checkpoints in Sec. 5.2, and apply average merging, TIES, DARE and 3DM in exactly the same way as Sec. 5.3, but on the MMSD1.0 dataset. Table 7 presents the results of multiple methods, where 3DM exhibits the highest accuracy, the highest F1-score, and the lowest absolute value of BR. Moreover, all merging-based methods reduce the absolute value of BR. The results in Table 7 imply comparable tendency with Table 4, demonstrating the advanced generalizability of 3DM.

6 Conclusion

In this study, we present a comprehensive analysis of biases in MLLMs, empirically demonstrating that the majority of existing MLLMs exhibit significant biases in sarcasm detection tasks, with varying directional tendencies. Our work represents the first systematic effort to develop an architecture-agnostic merging framework specifically designed to address and mitigate biases in models with divergent bias orientations, particularly in debiasing tasks.

The core contributions of our research include: (1) a generalized distill-merge pipeline applicable to both black-box and white-box MLLMs, and (2) a novel dynamic dropping mechanism that assigns individualized drop rates to delta parameters based on each parameter's functional role in the model. Notably, our distill-merge pipeline serves as a general, plug-and-play component that can be seamlessly integrated into various merging methodologies.

This research establishes a new paradigm for bias mitigation in MLLMs through advanced merging techniques, while simultaneously introducing a parameter-specific analytical framework for understanding and utilizing delta parameters. We anticipate that our findings will stimulate further research in this emerging area of MLLM optimization and bias reducing.

7 Limitations

In this study, we introduce a distill-merge pipeline designed for architectural alignment, alongside a dynamic merging mechanism that assigns a unique drop rate for each delta parameter. Nonetheless,

the current implementation of assigning drop rates overlooks the intricate interplay of synergistic and antagonistic interactions among multiple delta parameters, which could potentially influence the optimization process and outcomes. For instance, several delta parameters altogether contributes to biases, while any one of them individually can not. This limitation suggests a fertile ground for future research to explore and integrate these complex parameter interactions, thereby refining the mechanism's efficacy and robustness.

Ethics Statement

Use of AI Assistants We have employed ChatGPT as a writing assistant, primarily for polishing the text after the initial composition.

8 Appendix

Hyperparameter Tuning for DARE We search the hyperparameter on the validation set of MMSD2.0, and report the result in Table 8. Based on the result, we select 0.7 as the drop rate for DARE in our experiment.

References

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Simon Caton and Christian Haas. 2024. Fairness in machine learning: A survey. ACM Computing Surveys, 56(7):1–38.

Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. 2021. Autodebias: Learning to debias for recommendation. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 21–30.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24185–24198.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. arXiv preprint arXiv:1909.03683.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. arXiv preprint arXiv:2311.03287.

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

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. 2018. Essentially no barriers in neural network energy landscape. In International conference on machine learning, pages 1309–1318. PMLR.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. 2020. Linear mode connectivity and the lottery ticket hypothesis. In International Conference on Machine Learning, pages 3259–3269. PMLR.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. 2018. Loss surfaces, mode connectivity, and fast ensembling of dnns. Advances in neural information processing systems, 31.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. Preprint, arXiv:2406.12793.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. International Journal of Computer Vision, 129(6):1789–1819.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1012–1023.

Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang.

Method	Hyperparameter	Acc	F1	TP	FP	TN	FN	Bias direction	Bias Ratio
DARE	drop rate = 0.1	0.676	0.592	566	304	1064	476	weak	-0.235
	drop rate = 0.3	0.682	0.588	547	272	1096	495	weak	-0.276
	drop rate = 0.5	0.686	0.610	591	306	1062	451	weak	-0.209
	drop rate = 0.7	0.694	0.613	584	279	1089	458	weak	-0.236

Table 8: Hyperparameter sensitivity of DARE on the validation set of MMSD2.0

2024. The instinctive bias: Spurious images lead to hallucination in mllms. arXiv preprint arXiv:2402.03757.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. arXiv preprint arXiv:2212.04089.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. 2022. Dataless knowledge fusion by merging weights of language models. arXiv preprint arXiv:2212.09849.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In 2012 IEEE 12th international conference on data mining, pages 924–929. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR.

Yi Li and Nuno Vasconcelos. 2019. Repair: Removing representation bias by dataset resampling. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9572–9581.

Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Yun-Nung Chen. 2024. DogeRM: Equipping reward models with domain knowledge through model merging. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15506–15524, Miami, Florida, USA. Association for Computational Linguistics.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. Mitigating hallucination in large multi-modal models via robust instruction tuning. In The Twelfth International Conference on Learning Representations.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In Advances in Neural Information Processing Systems, volume 36, pages 34892–34916. Curran Associates, Inc.

Michael Matena and Colin Raffel. 2022. Merging models with fisher-weighted averaging. Preprint, arXiv:2111.09832.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6):1–35.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. 2021. Linear mode connectivity in multitask and continual learning. In International Conference on Learning Representations.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. 2020. What is being transferred in transfer learning? Advances in neural information processing systems, 33:512–523.

Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. ACM Computing Surveys (CSUR), 55(3):1–44.

Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. MMSD2.0: Towards a reliable multi-modal sarcasm detection system. In Findings of the Association for Computational Linguistics: ACL 2023, pages 10834–10845, Toronto, Canada. Association for Computational Linguistics.

Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. 2023. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In International Conference on Machine Learning, pages 28656–28679. PMLR.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1732–1742, Mexico City, Mexico. Association for Computational Linguistics.

Fanqi Wan, Longguang Zhong, Ziyi Yang, Ruijun Chen, and Xiaojun Quan. 2024. Fusechat: Knowledge fusion of chat models. arXiv preprint arXiv:2408.07990.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. Visionllm:

Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems, 36.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. Preprint, arXiv:2203.05482.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2024. Ties-merging: Resolving interference when merging models. Advances in Neural Information Processing Systems, 36.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. arXiv preprint arXiv:2408.07666.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In Forty-first International Conference on Machine Learning.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent advances in MultiModal large language models. In Findings of the Association for Computational Linguistics: ACL 2024, pages 12401–12430, Bangkok, Thailand. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024a. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In The Twelfth International Conference on Learning Representations.

Zhihong Zhu, Xianwei Zhuang, Yunyan Zhang, Derong Xu, Guimin Hu, Xian Wu, and Yefeng Zheng. 2024b. Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing. In Proc. of IJCAI.