

---

# Be Confident: Uncovering Overfitting in MLLM Multi-Task Tuning

---

Wenke Huang <sup>\*1</sup> Jian Liang <sup>\*1</sup> Guancheng Wan <sup>1</sup> Didi Zhu <sup>2</sup> He Li <sup>1</sup> Jiawei Shao <sup>3</sup> Mang Ye <sup>†1</sup> Bo Du <sup>†1</sup>  
Dacheng Tao <sup>4</sup>

## Abstract

Fine-tuning Multimodal Large Language Models (MLLMs) in multi-task learning scenarios has emerged as an effective strategy for achieving cross-domain specialization. However, multi-task fine-tuning appears performance degradation on open-response datasets. We posit that free-form answer generation primarily depends on language priors, and strengthening the integration of visual behavioral cues is critical for enhancing prediction robustness. In this work, we propose Noise Resilient Confidence Alignment to address the open-response overfitting challenge during multi-task fine-tuning. Our approach prioritizes maintaining consistent prediction patterns in MLLMs across varying visual qualities. To achieve this, we synthesize distorted visual inputs and enforce token prediction confidence alignment towards normal visual branch. By explicitly linking confidence calibration to visual robustness, this method reduces over-reliance on language priors. We conduct extensive empirical evaluations across diverse multi-task downstream via popular MLLM architectures. The comprehensive experiment demonstrates our effectiveness, showcasing its ability to alleviate open-response overfitting while maintaining satisfying multi-task performance.

## 1. Introduction

Driven by the remarkable success of Large Language Model (LLM) in the natural language processing and understand-

<sup>\*</sup>Equal contribution; Work done during an internship at TeleAI  
<sup>1</sup>National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, China <sup>2</sup>Department of Computer Science and Technology, Zhejiang University, Hangzhou, China <sup>3</sup>Institute of Artificial Intelligence (TeleAI), China Telecom, China <sup>4</sup>Nanyang Technological University, Singapore. Correspondence to: Mang Ye <ye-mang@whu.edu.cn>, Bo Du <dubo@whu.edu.cn>.

*Proceedings of the 42<sup>nd</sup> International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

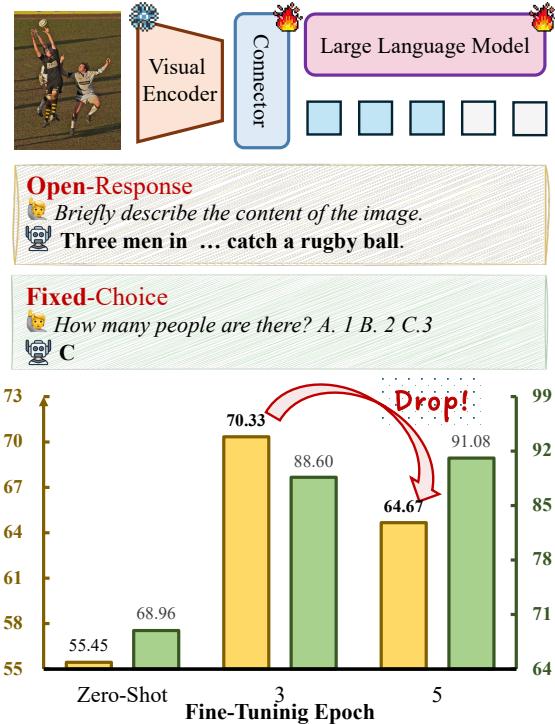


Figure 1: **Background and Problem.** Fine-tuning Multimodal Large Language Model (MLLM) on **Multi-Task** scenarios enables the model to acquire diverse specialization capabilities. This process involves training (🔥) both the connector and LLM modules for  $E$  epochs. We notice that during tuning process, **Open-Response** task reveals a notable **performance degradation**, while **Fixed-Choice** task demonstrates consistent stability. Experiments are conducted on **Flickr30k** and **ScienceQA** with VILA architecture.

ing (Devlin et al., 2018; Radford et al., 2019; Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2023). Many efforts have been made to extend LLM to Multimodal Large Language Model (MLLM), which have demonstrated remarkable capabilities in generating coherent and contextually relevant descriptions from visual inputs (Liu et al., 2023b; Dai et al., 2023; Lin et al., 2023; Liu et al., 2023a). Specifically, MLLM generally follows the paradigm to fuse the pre-trained vision encoder (Radford et al., 2021; Dosovitskiy et al., 2021) into the representation space of the Large Language Model, e.g., LLaMA (Touvron et al., 2023) and Vicuna (Chiang et al., 2023), via the connector module (Dai et al., 2023; Liu et al., 2023b; Luo et al., 2024). Consider-

ing that Multimodal Large Language Model is optimized on millions of multimodality instruction-following datasets (Lin et al., 2014; Singh et al., 2019; Mishra et al., 2019), it brings powerful generalization ability on different related tasks. Despite this, MLLM still performs poorly on different downstream tasks (Ren et al., 2023; Zhou et al., 2024).

Towards adapting to the target domain, the straightforward solution is to fine-tune on the task-specific dataset (Zhou et al., 2024; Han et al., 2024). With respect to existing explorations, they normally focus on **single-task** adaption, it would be simple to construct the personalized fine-tuning paradigm. However, facing the **multi-task** tuning requirement, we are curious what challenge would incur. To visualize our exploration, we conducted extensive experiments on two major downstream tasks: open-response and fixed-choice tasks. We randomly select datasets from these two views to combine the fine-tuning resource. As shown in Fig. 1, *with longer training epochs, the open-response task appears the overfitting tendency* (Guo et al., 2017; Lakshminarayanan et al., 2017) and brings the performance degradation on the corresponding testing dataset. We further investigate **why** Multimodal Large Language Model (MLLM) overfitting excites. As noted in previous studies (Favero et al., 2024; Wang et al., 2024a; Li et al., 2023b), MLLM tends to rely on the language priors to generate description sentences, consequently allocating less attention to visual information. Therefore, particularly in open-response tasks, MLLM tends to overfit to the current distribution, resulting in constrained specialization performance and an increased likelihood of fabricating content (Huang et al., 2023; Gekhman et al., 2024). Thus, a fundamental question naturally emerges: *How to alleviate open-response overfitting during multi-task MLLM fine-tuning?*

The naive solution is to establish asynchronous training epochs, *i.e.*, shorter for caption and longer for VQA. However, this operation would incur catastrophic forgetting on the caption task. In this work, we seek to mitigate overfitting in MLLM fine-tuning by examining the problem from an **inner visual behavior** perspective. Specifically, we find that when confronted with low-quality images, MLLM increasingly relies on textual modality prior information rather than visual cues (Leng et al., 2024; Favero et al., 2024; Wang et al., 2024a; Li et al., 2023b). This observation suggests that maximizing the mutual information between different visual input views during fine-tuning could help counterbalance the textual bias. Motivated by this insight, we propose Noise Resilient Confidence Alignment (NRCA), a method inspired by contrastive learning paradigms that encourage augmentation invariance and instance discrimination (Ye et al., 2019; 2020; He et al., 2020; Chen et al., 2020; Huang et al., 2021; Wu et al., 2022; Wang et al., 2024c). However, directly applying a standard contrastive paradigm in MLLM for visual coherence two challenges. **First**, common data

augmentation strategies (*e.g.*, color jitter, random cropping) (Tian et al., 2020; Chen et al., 2024a; Woo et al., 2024; Kim et al., 2024b) can drift from the original answer. *To address this issue*, we propose the Noisy Visual Mixup which utilizes the Gaussian noise to construct the distorted visual signal, which does not modify the semantic level knowledge. **Second**, the inherently variable prediction length of MLLM makes straightforward contrastive regularization unstable and potentially disruptive. *To overcome this problem*, we introduce the Token Confidence Alignment, which calculates the overall prediction token confidence and expects to achieve the confidence-level alignments. The rationale behind this is that beyond local empirical minimization, we encourage the noisy visual input to bring coherent prediction confidence with the original ones to alleviate textual bias and encourage visual representation ability. For a thorough examination, we conduct experiments on multi-tasking scenarios. We mix up and fine-tune on two major downstream tasks, *i.e.*, open-response and fixed-choice. The main contributions are summarized as follows:

- We focus on the multi-task Multimodal Large Language Model fine-tuning scenario and reveal that open-response tasks, such as image captioning, exhibit performance degradation due to language prior overfitting.
- We propose Noise Resilient Confidence Alignment (NRCA), a method designed to mitigate the open-response overfitting phenomenon by enhancing visual cues during Multimodal Large Language Model tuning process. Generally speaking, we construct noisy visual inputs and encourage token confidence alignment with the normal branch to reinforce visual representation.
- We perform a comprehensive analysis on multi-task scenarios, including both open-response datasets (Flickr30k (Young et al., 2014) and COCO-Cap (Lin et al., 2014)) and fixed-choice datasets (ScienceQA (Lu et al., 2022) and IconQA(Lu et al., 2021)). Experiments are conducted on the VILA (Lin et al., 2023) and LLaVA (Liu et al., 2023b) architectures. Through a series of ablation studies, the promising results empirically validate the effectiveness of NRCA in alleviating open-response overfitting and enhancing multi-task fine-tuning overall performance.

## 2. Related Works

### 2.1. Multimodal Large Language Models

Motivated by the great progress of Large Language Model (LLM) (Radford et al., 2019; Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023; Shen et al., 2024b; Jin et al., 2024; 2025), researchers have been actively exploring ways to combine pre-trained LLM and vision encoders to construct the end-to-end Multimodal Large Language Model (MLLM) capable of processing image-text. Existing solutions follow to feed the visual features produced by a pre-trained vision encoder into the LLM module with the

visual projector, *e.g.*, Flamingo (Alayrac et al., 2022), BLIP (Li et al., 2022a; 2023a), InstructBLIP (Dai et al., 2023), QWen-VL (Bai et al., 2023), LLaVA (Liu et al., 2023b;a; Zhu et al., 2024c; Li et al., 2024a; Luo et al., 2024), VILA (Lin et al., 2023; Fang et al., 2024), Cambrian-1 (Tong et al., 2024). Although MLLM brings the powerful generalization ability on a wide range of tasks (Marino et al., 2019; Lu et al., 2021; Lin et al., 2014; Kazemzadeh et al., 2014) but appears limited performance on specific domains. Thus, fine-tuning MLLM on the target task brings a reasonable solution for boosting downstream performance (Zhou et al., 2024).

## 2.2. Fine-Tune MLLM

Pre-trained MLLM on large-scale datasets can be easily transferred to downstream vision tasks through fine-tuning (Huang et al., 2025). Existing solutions can be primarily categorized into two types. **I**) Reparameterization Fine-Tuning (Hu et al., 2022; Zhang et al., 2023; Wang et al., 2023; Hao et al., 2024; Liu et al., 2024c; Lin et al., 2024; Hu et al., 2024; Bi et al., 2025a;b; Liang et al., 2025) normally apply low-rank matrices to approximate weight changes during fine-tuning and can merge with pre-trained weights prior to inference. Despite certain advantages, it faces several limitations that hinder application. First, it normally achieves sub-optimal performance compared with direct weight updates (Biderman et al., 2024; Shuttleworth et al., 2024). Second, it introduces an additional parameter module, increasing the complexity of achieving universal architectural compatibility. **II**) Partial Fine-Tuning (Li et al., 2022b; Ansell et al., 2022; Li et al., 2023c; Yu et al., 2024; Zhang et al., 2024d;a; Zhu et al., 2024a; Lu et al., 2024b) focuses on directly optimizing selected candidate elements in MLLM, such as visual connectors and LLM blocks, to align with the requirements of downstream tasks. This paradigm enjoys the architecture-agnostic property and thus largely benefits MLLM transferability. In our work, we adhere to the partial fine-tuning and further provide detailed a learnable parameter discussion in the following Sec. 4.1.

## 2.3. Overfitting Behavior in MLLM Fine-Tuning

For downstream fine-tuning, the neural networks are default optimized based on the cross entropy (De Boer et al., 2005) which has been shown to overfit to the current distribution (Platt et al., 1999; Amodei et al., 2016; Guo et al., 2017). In the context of MLLM, over-fitting on new factual knowledge encourages MLLM prone to fabricating content, amplifying the risk of hallucinations (Huang et al., 2023; Gekhman et al., 2024). Substantial research efforts have been directed towards compensating vision shortcomings and could be mainly classified into two types: Vision Enhancement (Luo et al., 2024; Shang et al., 2024; Ghosh et al., 2024; Zhao et al., 2024; Li et al., 2024a; Shi et al., 2024;

Shen et al., 2024a; Guo et al., 2024) and Update Calibration (Panigrahi et al., 2023; Chen et al., 2024b). The former solution introduces additional **external vision signals**, *e.g.*, high resolution, object grounding. However, these solutions increase the computational burden and limit architecture compatibility due to the added parameters in pre-trained models (Huo et al., 2024). In contrast, Update Calibration methods derive two main research directions. **① Partial Update Mask** (Li et al., 2022b; Ansell et al., 2022; Li et al., 2023c; Yu et al., 2024; Zhang et al., 2024d;a; Zhu et al., 2024a; Lu et al., 2024b; Hui et al., 2024; Chen et al., 2024b) selects the candidate parameters group to alleviate the overfitting phenomenon. For example, Random Mask-Tuning randomly selects half of the learnable parameters and freezes the rest. **② Stiff Penalty Regularization** (Kirkpatrick et al., 2017; Zenke et al., 2017; Xuhong et al., 2018; Ritter et al., 2018; Buzzega et al., 2020; Li et al., 2020; Panigrahi et al., 2023) measures the parameter importance degree to construct the parameter weight updates penalty. However, existing approaches primarily focus on single-task adaptation and indiscriminately apply anti-overfitting objectives across all training samples, failing to address the fitting condition inconsistencies in multi-task setting (Dong et al., 2023) such as open-response and fixed-choice objectives. In this work, we focus on MLLM multi-task fine-tuning and reveal that open-response tasks are particularly susceptible to overfitting due to excessive dependence on textual modalities. To mitigate this limitation, we propose leveraging **intrinsic visual behavior** patterns to strengthen cross-modal representation robustness. To address this issue, we incorporate. Specifically, we synthesize perturbed visual inputs through Gaussian noise injection and enforce confidence consistency between predictions from corrupted and pristine visual branches. This dual-branch alignment mechanism effectively reduces textual bias dominance while preserving semantic fidelity, effectively counteracting overfitting tendencies during MLLM optimization.

## 3. Methodology

### 3.1. Preliminary

Given the Multimodal Large Language Model (MLLM  $\theta$ ) architecture, the MLLM model typically includes three modules: visual encoder  $f$  (Dosovitskiy et al., 2021), LLM  $g$ , and the connector module  $\varphi$  (Liu et al., 2023b; Dai et al., 2023; Liu et al., 2023a; Lin et al., 2023). For a query instance, the input consists of both a visual image  $x^v$  and a textual instruction  $x^t$ . Denote the vocabulary dictionary as  $V$  and language response  $y = \{y_1, \dots, y_t, \dots, y_T\} \in \mathbb{R}^T$ . Thus,  $y_t \in V$  and  $t \in \{1, \dots, T\}$ .  $T$  denotes the token length. To be precise, we extract the visual features  $m^v = f(x^v)$ , and then apply the trainable projection  $\varphi$  to convert  $m^v$  into language embedding tokens,  $h^v = \varphi \cdot m^v$ .

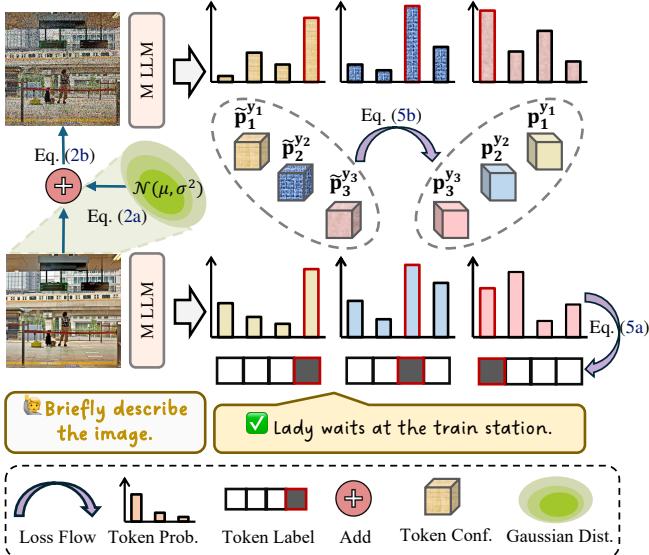


Figure 2: **Schematization of NRCA.** We propose Noisy Visual Mixup, which leverages Gaussian distribution to generate noisy visual inputs. Additionally, we introduce Token Confidence Alignment, designed to align the confidence of noisy tokens with the normal visual branch behavior, thereby enhancing visual representations during multi-task fine-tuning. The process is illustrated via logits output with the token length  $T=3$  and class number  $|V|=4$ . Best viewed in color. Zoom in for details. Refer to Sec. 3.2

And textual token as  $h^t = \text{Tokenize}(x^t)$ . Next, we combine both visual and textual tokens and pass them into the LLM module  $g$  to generate the logits output  $z = g(h^v, h^t)$ . In our work, following previous MLLM fine-tuning works and benchmarks (Zhou et al., 2024; Zhu et al., 2024a; Han et al., 2024) to select the trainable parameters. For the visual module, we freeze the vision encoder and tune visual connector module  $\varphi$ . For LLM aspect, LLM includes several transformer blocks and selects the candidate block layers set  $\mathcal{N}$  for optimization as  $g[L]$ . Thus, we obtain learnable modules as  $w = \{\varphi, g[\mathcal{N}]\}$ . Recall the fine-tuning distribution  $D$ , the default MLLM optimization procedure follows:

$$\arg \min_w \mathbb{E}_{(x^v, x^t, y) \in \mathcal{D}} \mathcal{L}(g(h^v, h^t), y). \quad (1)$$

The  $\mathcal{L}$  normally utilizes the Cross-Entropy term (De Boer et al., 2005) for regular optimization target.

### 3.2. Noise Resilient Confidence Alignment

**Motivation.** In general, features derived from the standard visual output layer are effective for reasoning across Multimodal tasks. However, due to the language prior bias (Leng et al., 2024; Favero et al., 2024; Liu et al., 2024a; Wang et al., 2024b; Zhang et al., 2024b), Multimodal Large Language Model (MLLM), particularly during the fine-tuning stage for caption generation, tend to rely on the language-based intuition. Therefore, in our work, we focus on enhancing the visual representation to mitigate overfitting and improve

downstream performance. Inspired by the success of the contrastive paradigm (Wu et al., 2018; Ye et al., 2019; 2020; Chen et al., 2020; Khosla et al., 2020; Chen et al., 2022), we aim to encourage the MLLM to *generate consistent predictive behavior for both distorted and normal visual inputs*.

**Noisy Visual Mixup.** Compared to popular contrastive augmentation such as color jitter and random cropping (Xiao et al., 2021; Zhao et al., 2021; Wang et al., 2022; 2024c), would largely bring the hallucination prediction and fail to align with the original feedback. Thus, we model the inherent characteristics of the current image by introducing a random Gaussian distribution based on its own feature space. This approach leverages the image unique feature distribution to generate stochastic variations, capturing both its intrinsic structure and potential perturbations.

$$\mu = f_\mu(x^v), \quad \sigma = f_\sigma(x^v), \quad (2a)$$

$$\tilde{x}^v = \underbrace{\delta \mathcal{N}(\mu, \sigma^2)}_{\text{Noisy Part}} + \underbrace{(1 - \delta)x^v}_{\text{Normal Part}}. \quad (2b)$$

$\mu$  and  $\sigma^2$  respectively denote the mean and variance of the distribution.  $\mathcal{N}$  specifies the Gaussian distribution. The  $\delta$  denotes the noisy ratio and is default set as 0.5. Then, we feed both normal  $x^v$  and noisy  $\tilde{x}^v$  visual information with the corresponding text prompt  $x^t$  into the MLLM and respectively obtain the contextual word prediction logits as  $z = g(h^v, h^t)$  and  $\tilde{z} = g(\tilde{h}^v, h^t)$ .

**Token Confidence Alignment.** Compared with traditional classification tasks, the prediction  $z \in \mathbb{R}^{T \times c}$  for MLLM appears high output dimensions attributed to vocabulary scale, e.g.,  $c = 32000$  in LLaMA. As a result, directly conducting distribution alignment faces challenges related to the imbalance between head-tail knowledge variations (Gu et al., 2024; 2025; Kim et al., 2024a; Ko et al., 2024), and fails to adequately reflect the reliability of the ground-truth signal. Therefore, in this work, we propose aligning token confidence rather than token distribution to mitigate overfitting during the fine-tuning stage of MLLM. Specifically, we first measure the prediction probability for the token  $t$  via the following formulation:

$$\mathbf{p}_t = \sigma(z_t), \quad \tilde{\mathbf{p}}_t = \sigma(\tilde{z}_t) \in \mathbb{R}^c, \\ \mathbf{p} = [\mathbf{p}_t]_{t=1}^T, \quad \tilde{\mathbf{p}} = [\tilde{\mathbf{p}}_t]_{t=1}^T \in \mathbb{R}^{T \times C}. \quad (3)$$

$\sigma$  means the softmax function. Then, we further obtain the overall token prediction confidence based on the ground-truth labels  $\mathbf{y} = \{\mathbf{y}_t\}_{t=1}^T$  as:

$$\mathcal{I} = \frac{1}{T} \sum_{t=1}^T \mathbf{p}_t^{\mathbf{y}_t}, \quad \tilde{\mathcal{I}} = \frac{1}{T} \sum_{t=1}^T \tilde{\mathbf{p}}_t^{\mathbf{y}_t}. \quad (4)$$

Naturally, we encourage the noisy prediction confidence  $\tilde{\mathcal{I}}$  to align with the normal ones  $\mathcal{I}$ . Furthermore, applying the same penalty strength across all samples fails to account for sample-specific fitting conditions. To address this, we introduce the empirical loss as a guiding factor, allowing for tailored penalty allocation for each query sample. We introduce the following regularization term.

$$\mathcal{L}_{CE} = \frac{1}{T} \sum_{t=1}^T -\mathbf{1}_{y_t} \log(\mathbf{p}_t), \quad (5a)$$

$$\mathcal{L}_{NRCA} = |1 - \frac{\tilde{\mathcal{I}}}{\mathcal{I}}| \times \mathcal{L}_{CE}. \quad (5b)$$

where  $\mathbf{1}$  denotes the one-hot encoding of  $y_t$ . We detach the normal branch signals in Eq. (5b) to avoid the MLLM optimization degradation. Empirically, we identify open-response samples based on their label length, applying this operation separately to these samples. Specifically, we observe that fixed-choice sample labels typically consist of a single word followed by the [EOS] token, with the label length as 2. Finally, we carry out the following optimization objective in fine-tuning phase:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{NRCA}. \quad (6)$$

$\lambda$  represents the penalization hyper-parameter that controls the strength of confidence alignment. We set  $\lambda = 2$  and provide the corresponding ablation analysis in Sec. 4.2. Overall, the MLLM model is regularized to preserve the downstream distribution while promoting invariance in distorted confidence. As a result, our method effectively mitigates downstream open-response overfitting by enhancing visual behavior alignment. We provide a algorithm description in Algorithm 1 and the methodological framework in Fig. 2.

### 3.3. Discussion and Limitation

**Gaussian Noise Meets MLLM.** Existing MLLM research often employs Gaussian noise (Goodman, 1963; Tüske et al., 2015; Variani et al., 2015; Hayashi & Uchida, 2019; Ardizzone et al., 2020) to generate low-quality visual inputs, thereby amplifying language priors (Xiao et al., 2024). A common approach to mitigating MLLM hallucinations is to adjust next-token logits in a contrastive manner using the normal prediction distribution (Leng et al., 2024; Woo et al., 2024; Zhu et al., 2024b; Chen et al., 2024a; Huo et al., 2024; Zhang et al., 2024b; Xiao et al., 2024; Liu et al., 2024b). The rationale behind this approach is that disturbed inputs significantly exacerbate hallucinations, while contrastive decoding mitigates this effect by subtracting hallucinated concepts from the original distribution, thereby reducing confusion. Consequently, existing methodologies often treat noisy visual behavior as an optimization trap and disregard the corresponding feedback. In contrast, our work introduces a novel MLLM multi-task tuning approach to address open-response overfitting by encouraging prediction confidence alignment between normal and noisy visual signals. This strategy compels the network to fully leverage its visual processing capabilities, achieving consistent prediction confidence even when faced with degraded visual inputs.

**Conceptual Difference.** Fine-tuning the MLLM for downstream tasks is a straightforward approach to developing

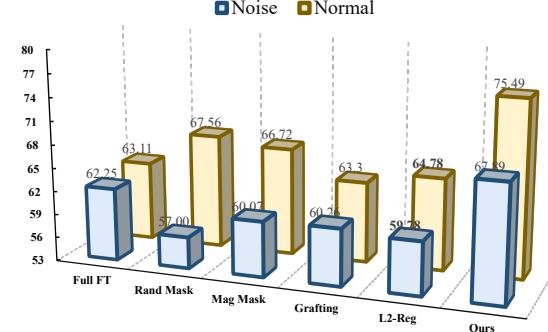


Figure 3: **Noise Evaluation Comparison.** We evaluate noisy and normal samples on Flickr30k using LLaVA with  $E = 3$ . Normal Full Fine-Tuning shows a limited performance gap between noisy and normal visual signals, revealing a tendency of language prior reliance. In contrast, our approach demonstrates a notable performance improvement by emphasizing the importance of visual information. See Sec. 3.3 for details.

Table 1: **Attention Allocation and Performance Comparison.** We conduct the evaluation on the Flickr30k with noisy visual input. We tuning on the Flickr30k+ScienceQA datasets based on LLaVA architecture. Please see Sec. 4.2 for detailed discussion.

Metrics	Full FT	Ran Mask	Mag Mask	Grafting	L2-Reg	Ours
System (All)	0.6837	0.6966	0.6927	0.6673	0.6791	0.6677
Prompt (All)	0.1229	0.1297	0.1280	0.1241	0.1181	0.1225
Visual (All)	0.1933	0.1737	0.1793	0.2084	0.2029	<b>0.2098</b>
System (Mid)	0.5865	0.5874	0.5859	0.5726	0.5826	0.5574
Prompt (Mid)	0.1733	0.1843	0.1832	0.1664	0.1623	0.1576
Visual (Mid)	0.2402	0.2282	0.2309	0.2609	0.2552	<b>0.2850</b>

domain-specialized experts. However, the limited size of downstream datasets poses a significant challenge, often leading to overfitting on target-specific behaviors and catastrophic forgetting of pre-trained knowledge. Existing studies predominantly focus on mitigating catastrophic forgetting in MLLM, where fine-tuning risks compromise the generality achieved during pre-training, thus causing upstream performance degradation. Consequently, prior research has primarily aimed to balance generalization and specialization, ensuring robust performance on both seen and unseen tasks. In contrast, our work addresses a more practical scenario by fine-tuning MLLM for multi-task applications. We argue that leveraging MLLM to achieve multi-task specialization is a more efficient approach than the conventional one-to-one fine-tuning paradigm. However, when dealing with multi-task distributions, we identify a critical issue: *MLLM frequently experiences performance degradation in open-response tasks*, e.g., image caption understanding, as shown in Fig. 1. We hypothesize that this degradation arises because open-response tasks tend to rely excessively on language priors rather than leveraging visual information, as confirmed in Fig. 3. Full Fine-Tuning brings limited performance gap towards the input visual quality. Furthermore, as for Multimodal Large Language Model (MLLM), input tokens include System Tokens, Prompt Tokens and Visual

Table 2: Incremental Resource Cost with Overall Performance

**Comparison.** Accuracies are derived from Flickr30k & ScienceQA based on VILA for  $E = 3$ .  $\mathcal{O}$  denotes the complexity degree. - indicates unchanged values.  $\uparrow$  means improved accuracy compared with Full FT. Please refer to Sec. 3.3.

Metrics	Full FT	L2-Reg	Ran Mask	Our NRCA
Resource Cost	-	$\mathcal{O}( \theta )$	$\mathcal{O}( \theta )$	$\mathcal{O}(B \times ( \tilde{x}^v  +  \tilde{z} ))$
Performance	79.47	79.52	83.02	<b>83.24</b> $\uparrow 3.77$

 Table 3: Ablative Experiments for  $\mathcal{L}_{NRCA}$  Formulation in Eq. (5b). Please see Sec. 4.2 for detailed discussion.

$ 1 - \frac{\tilde{z}}{z} $	$\mathcal{L}_{CE}$	Flickr30k	SQA	AVG	COCO-Cap	SQA	AVG
<i>Fine-Tune with VILA architecture</i>							
Zero-shot	55.45	68.96	62.20	72.57	68.96	70.76	
✗ (Full FT)	70.33	88.60	79.47	109.58	90.33	99.96	
✓	70.87	90.43	80.65	114.83	89.59	102.21	
✓ ✓	75.10	90.23	<b>82.67</b>	120.05	89.94	<b>105.00</b>	
<i>Fine-Tune with LLaVA architecture</i>							
Zero-shot	25.31	69.56	47.44	40.28	69.56	54.92	
✗ (Full FT)	63.11	88.3	75.71	101.76	88.65	95.21	
✓	66.82	87.56	77.19	104.39	87.80	96.10	
✓ ✓	72.39	86.61	<b>79.50</b>	114.54	86.32	<b>100.43</b>	

Tokens and appear different contribution for the prediction output. Thus, we utilize the attention map between the first output token and the input token under the noisy visual input to visualize the contribution of different input token. Besides, recent works (Che et al., 2025; Zhang et al., 2025; 2024c) have shown that the MLLM tends to extract object information from the image at the middle layers. Thus, we plot the all layers and middle layers attention allocation for the first output token in the following Tab. 1. The results reveal that our method allocates more attention weights on the visual cue. Therefore, both performance comparisons and attention analyses show that our method reduces reliance on language priors and improves overall model performance by encouraging consistent predictions, even under distorted visual inputs—highlighting the value of posterior guidance over prior dependence.

**Limitations.** Our method, NRCA constructs additional noisy visual inputs to obtain the corresponding prediction confidence and expects confidence alignment with the normal branch. As a result, our method incurs extra storage requirements for the noisy input  $\tilde{x}^v$  and the logits output signals  $\tilde{z}$ . However, existing methods typically require storing the full pre-trained parameters to apply parameter stiffness restrictions or sparse mask updating strategies, which heavily depend on the MLLM model scale ( $\mathcal{O}(|\theta|)$ ). As shown in Table 2, while NRCA introduces an incremental storage complexity as  $\mathcal{O}(B \times (|\tilde{x}^v| + |\tilde{z}|))$ , this cost is relatively small compared to existing methods. Additionally, NRCA achieves the highest performance of 83.24, demonstrating that it effectively balances resource efficiency and downstream performance. Detailed performance comparison

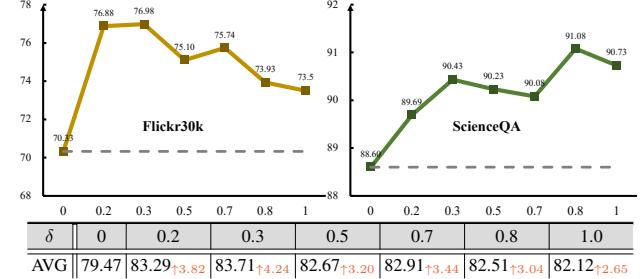


Figure 4: Ablation on Noisy Ratio  $\delta$  in Eq. (2b) effect for **image captioning** (Left), **visual question-answering** (Middle), and overall performance (Right) with VILA.  $\delta = 0$  degrades into the Full Fine-Tuning.  $\uparrow$  means improved accuracy compared with Full FT. Refer to Sec. 4.2 for the explanation.

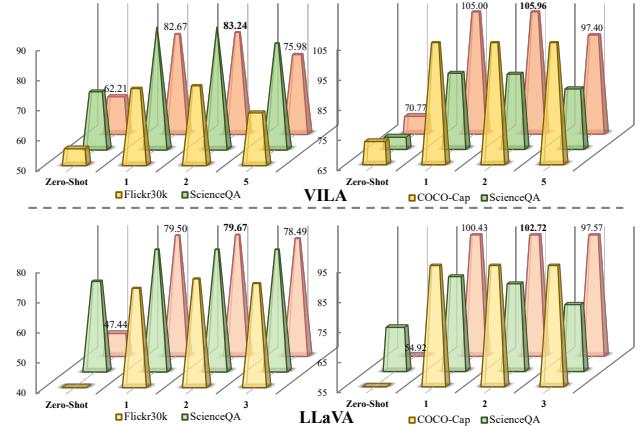


Figure 5: Ablation on the hyper-parameter  $\lambda$  in Eq. (6) for the proposed method with different MLLM architectures. We default set  $\lambda = 2$  in the following experiments. See Sec. 4.2. for details.

refers to Sec. 4.3. Furthermore, when the downstream distribution consists solely of fixed-choice tasks, MLLM does not encounter significant performance degradation. However, open-response tasks are crucial for unlocking MLLM ability to generate free-form expressions, which go beyond the limitations of traditional classification. In the context of multi-task settings, we argue that open-response tasks play an essential role in showcasing the full potential of dialogue experience of MLLM, even though they introduce challenges such as performance drops due to overfitting on language information priors.

## 4. Experiments

### 4.1. Experimental Setup

**Architecture and Datasets.** Adhering to the Multimodal Large Language Model (MLLM) paradigm, we evaluate the effectiveness of our proposed methods using two widely adopted models, LLaVA(Liu et al., 2023b) and VILA(Lin et al., 2023), as the foundational MLLM for our experiments. We categorize the four downstream datasets into two task types: **image captioning** (open-response) and **visual**

Table 4: Comparison with the state-of-the-art Multimodal Large Language Model (MLLM) Fine-Tuning Solutions on the multi-task setting: **image captioning** (Flickr30k and COCO-Cap) and **visual question-answering** (ScienceQA and IconQA) based on VILA and LLaVA architectures with tuning epoch  $E = 3$ . We mark the Best in bold and Second in underline across different tuning methods.  $+$  means improved accuracy compared with Zero-shot. Please refer to Sec. 4.3 for relative explanations.

Methods	Flickr30k	ScienceQA	AVG	COCO-Cap	ScienceQA	AVG	Flickr30k	IconQA	AVG	COCO-Cap	IconQA	AVG
<i>Fine-Tune with VILA architecture</i>												
Zero-shot	55.45	68.96	62.20	72.57	68.96	70.76	55.45	55.57	55.51	72.57	55.57	64.07
Full FT	70.33	88.60	79.47	109.58	90.33	99.96	68.20	86.62	77.41	108.16	86.91	97.53
Ran Mask	74.77	91.27	83.02	113.86	90.68	102.27	73.16	86.57	79.87	113.59	86.64	100.12
Mag Mask	75.37	91.08	83.23	114.12	90.28	102.20	73.13	87.32	80.23	113.29	87.11	100.20
Grafting	70.23	89.14	79.69	109.02	90.93	99.98	69.01	86.67	77.84	107.73	87.44	97.59
L2-Reg	68.71	90.33	79.52	108.94	90.48	99.71	69.3	87.22	78.26	108.69	87.00	97.85
NRCA	76.00	90.48	<b>83.24</b>	122.27	89.64	<b>105.96</b>	75.98	86.99	<b>81.49</b>	120.82	86.19	<b>103.51</b>
	+ 20.55	+ 21.52	+ 21.04	+ 49.70	+ 20.68	+ 35.20	+ 20.53	+ 31.42	+ 25.98	+ 48.25	+ 30.62	+ 39.44
<i>Fine-Tune with LLaVA architecture</i>												
Zero-shot	25.31	69.56	47.44	40.28	69.56	54.92	25.31	48.42	36.87	40.28	48.42	44.35
Full FT	63.11	88.30	75.71	101.76	88.65	95.21	64.14	82.68	73.41	101.68	83.06	92.37
Ran Mask	67.56	85.62	76.59	105.24	85.82	95.53	66.64	79.65	73.15	104.86	79.69	92.28
Mag Mask	66.72	85.72	76.22	105.67	85.57	95.62	66.70	79.54	73.12	105.94	79.56	92.75
Grafting	63.30	89.44	76.37	100.47	89.64	95.06	66.35	83.91	75.13	101.20	83.76	92.48
L2-Reg	64.78	90.03	77.41	101.05	89.94	95.50	65.24	83.95	74.60	101.92	83.72	92.82
NRCA	75.49	83.84	<b>79.67</b>	121.50	83.94	<b>102.72</b>	74.31	77.60	<b>75.96</b>	118.74	77.83	<b>98.29</b>
	+ 50.18	+ 14.28	+ 32.23	+ 81.22	+ 14.38	+ 47.80	+ 49.00	+ 29.18	+ 39.09	+ 78.46	+ 29.41	+ 53.94

**question-answering** (fixed-choice), as follows:

- **Flickr30k** (Young et al., 2014): Design for understanding the visual content of images associated with linguistic expressions, it contains 31,000 images from Flickr, each paired with 5 human-annotated reference sentences.
- **COCO-Cap** (Lin et al., 2014): Comprise over 330,000 images with more than 1.5 million captions, this dataset provides 5 human-generated captions per image for both training and validation.
- **ScienceQA** (Lu et al., 2022): Collect from elementary and high school science curricula. This dataset includes approximately 21,000 multimodal multiple-choice questions spanning diverse science topics.
- **IconQA** (Lu et al., 2021): Focus on abstract diagram understanding and cognitive reasoning in real diagram.

Specifically, we follow the training settings from prior works (Zhou et al., 2024; Zhu et al., 2024a) and utilize official optimization scripts<sup>1</sup>. For fixed-choice tasks (ScienceQA and IconQA), we use the textual prompt: “*Answer with the option’s letter from the given choices directly.*” For open-response tasks (Flickr30k and COCO-Cap), we collect five manually written instructions and randomly sample one as the prompt for each caption, as follows:

- *“Describe the image as simply as possible with a sentence or phrase”*
- *“Give a brief summary of what you see”*
- *“Provide a short description of the image”*
- *“Write a short description for the image”*
- *“Briefly describe the content of the image”*

<sup>1</sup><https://huggingface.co/datasets/BAAI/DataOptim>

For each dataset, we randomly sample 10k instances from the training set. We randomly select datasets from the above mentioned captioning and VQA tasks to construct the multi-task training datasets.

**Counterparts.** We focus on exploring anti-overfitting MLLM fine-tuning methods and mainly compare with the **1 Partial Update Mask** and **2 Stiff Penalty Regularization** solutions as the following formulations:

- Full Fine-Tuning (Full FT) [arXiv’05] (De Boer et al., 2005): Default optimize full candidate parameters towards the downstream distribution.
- L2-Regularization (L2-Reg) [PNAS’17] (Kirkpatrick et al., 2017): Add  $\mathcal{L}_2$  regularization term with the regularization hyper-parameter, i.e., 1e-3, to the original loss function. It focuses on keeping the fine-tuning model closer to the pre-trained model, thereby mitigating forgetting.
- Grafting [ICML’23] (Panigrahi et al., 2023): Localize newly acquired skills inside fine-tuned language models, which could be regarded as  $\mathcal{L}_1$  regularization with the penalty weigh, i.e., 1e-6.
- Random Mask-Tuning (Ran Mask) [arXiv’24] (Hui et al., 2024): Randomly update half ratio of parameters within each transformer layer at each iteration while freezing the other elements. We default set the sampling ratio as 50%.
- Magnitude Tuning (Mag Mask) [NeurIPS’15] (Han et al., 2015)): Select and maintain elements with relative large weight magnitude with the pre-defined sampling ratio. The default updating proportion is set as 50%.

**Implementation Details.** We follow the official reposito-

Table 5: Comparison with the state-of-the-art Multimodal Large Language Model (MLLM) Fine-Tuning Solutions on the multi-task setting: **image captioning** (Flickr30k and COCO-Cap) and **visual question-answering** (ScienceQA and IconQA) based on VILA and LLaVA architectures with tuning epoch  $E = 5$ . We mark the Best in bold and Second in underline across different tuning methods.  $+$  means improved accuracy compared with Zero-shot. Please refer to Sec. 4.3 for relative explanations.

Methods	Flickr30k	ScienceQA	AVG	COCO-Cap	ScienceQA	AVG	Flickr30k	IconQA	AVG	COCO-Cap	IconQA	AVG
<i>Fine-Tune with VILA architecture</i>												
Zero-shot	55.45	68.96	62.20	72.57	68.96	70.76	55.45	55.57	55.51	72.57	55.57	64.07
Full FT	64.67	91.08	77.88	102.23	91.22	96.73	65.42	87.68	76.55	100.40	87.29	93.85
Ran Mask	66.50	90.98	78.74	106.52	90.98	98.75	67.43	86.95	77.19	105.4	87.21	96.31
Mag Mask	69.17	91.77	80.47	105.16	91.22	98.19	67.34	87.24	77.29	104.98	86.87	95.92
Grafting	64.76	91.03	77.90	100.16	90.13	95.15	62.4	87.82	75.11	99.43	87.43	93.43
L2-Reg	63.13	91.22	77.18	101.98	91.67	96.83	61.72	87.86	74.79	100.34	87.46	93.90
NRCA	70.87 + 15.42	90.58 + 21.62	<b>80.73</b> + 18.53	117.54 + 44.97	90.78 + 21.82	<b>104.16</b> + 33.40	74.95 + 19.50	87.22 + 31.65	<b>81.08</b> + 25.57	117.99 + 45.42	86.92 + 31.35	<b>102.46</b> + 38.39
<i>Fine-Tune with LLaVA architecture</i>												
Zero-shot	25.31	69.56	47.44	40.28	69.56	54.92	25.31	48.42	36.87	40.28	48.42	44.35
Full FT	61.46	91.08	76.27	92.82	90.88	91.85	60.51	84.69	72.60	92.19	84.48	88.34
Ran Mask	62.31	88.84	75.58	95.91	88.30	92.11	62.14	82.66	72.40	91.98	81.62	86.80
Mag Mask	63.10	88.55	75.83	96.88	88.50	92.69	61.37	82.43	71.90	95.23	82.79	89.01
Grafting	61.40	91.82	76.61	92.91	91.92	92.42	101.2	83.76	92.48	90.84	85.16	88.00
L2-Reg	59.77	91.52	75.65	91.76	91.67	91.72	59.18	85.99	72.59	91.52	85.32	88.42
NRCA	73.04 + 47.73	87.11 + 17.55	<b>80.08</b> + 32.64	120.81 + 80.53	82.80 + 13.24	<b>101.81</b> + 46.89	74.51 + 49.20	81.65 + 33.23	<b>78.08</b> + 41.21	120.31 + 80.03	78.34 + 29.92	<b>99.33</b> + 54.98

ries<sup>2,3</sup> to conduct the fine-tuning procedure. With respect to the training scenarios, we first randomly sample 10,000 instances from each dataset and combine them as multi-task tuning datasets. As for the optimization details, the learning rate  $lr$  in LLaVA (Liu et al., 2023b) is  $2e - 4$  for LLM and  $2e - 5$  for the visual projector. For VILA(Lin et al., 2023), we uniformly set the learning rate to  $1e - 4$ . The training epochs are set to  $E = 3$  and  $E = 5$ . The training batch size  $B$  is set to 16 by default. The fine-tuning block for LLM is the *last*  $N = 2$  layers. Regarding the experimental conditions, all experiments are conducted on 8 NVIDIA 4090 GPUs, each with 24GB of memory. Due to the limited computational resources, we select the LLaVA-1.5-7B for LLaVA and VILA1.5-3B for VILA.

**Evaluation Metrics.** To evaluate the performance of MLLM across different downstream task types, we utilize CIDEr (Vedantam et al., 2015) and Top-1 Accuracy metrics for image captioning and visual question-answering tasks, respectively. Additionally, we compute the mean results to represent the overall downstream performance.

## 4.2. Diagnostic Analysis

We perform ablation studies on two scenarios: Flickr30k & ScienceQA and COCO-Cap & ScienceQA, utilizing both the VILA and LLaVA architectures with a tuning epoch of  $E=3$  to facilitate an in-depth analysis.

**Proposed Training Objective.** We quantitatively analyze the proposed Noise Resilient Confidence Alignment

(NRCA) in Tab. 3. The ablation results demonstrate that directly encouraging prediction behavior consistency effectively mitigates the open-response overfitting phenomenon. Moreover, incorporating empirical loss guidance, which accounts for sample difficulty, further enhances the overall performance. Thus, combining Noisy Visual Mixup and Token Confidence Alignment acquires satisfying downstream multi-task performance that coincides with our motivation of exploiting the noisy resilience to alleviate the open-response overfitting during Multimodal Large Language Model tuning.

**Noisy MixUp Ratio.** The parameter  $\delta$ , introduced in Eq. (2b), controls the Gaussian distortion ratio. As illustrated in Fig. 4, increasing the distortion ratio poses challenges for confidence alignment in open-response tasks, resulting in limited performance improvements. Generally, as  $\delta$  increases, a trade-off between open-response and fixed-choice performance becomes more pronounced, with improvements diminishing under higher noise ratios. For consistency and convenience, we set  $\delta = 0.5$  across different scenarios in the subsequent experiments.

**Control Hyper-Parameter  $\lambda$**  in Eq. (6). The Sec. 4 quantifies the effect of hyper-parameter  $\lambda$ , which measures the strength of token confidence difference penalization on different scenarios. Specifically, the open-response and fixed-task trade-off performance progressively mounts as  $\lambda$  enlarges, and the improvement presents marginal under strict parameter stiffness. For convenience, we choose the  $\lambda = 2$  for different scenarios in the following experiments.

<sup>2</sup><https://github.com/haotian-liu/LLaVA>

<sup>3</sup><https://github.com/NVlabs/VILA>

### 4.3. Comparison to State-of-the-Arts

We compare our NRCA against related approaches on the multi-task scenarios: image-captioning & visual question-answering tasks. As shown in Tab. 4 and Tab. 5, several key observations can be made. **First**, *Partial Update Mask* does not fundamentally mitigate overfitting but merely slows the fitting process of MLLM. This is evident from its limited performance gains, even with extended tuning epochs  $E$  and larger model architectures, *e.g.*, scaling from 3B to 7B. Specifically, Random Mask-Tuning applies stochastic masks to parameters, injecting noise to reduce reliance on specific sub-networks and implicitly regularize by limiting parameter co-adaptation. While effective on smaller models (VILA-3B), its performance drops on larger ones (LLaVA-7B), suggesting its effect comes from constrained optimization rather than addressing language prior dominance. **Second**, *Stiff Penalty Regularization* is highly sensitive to regularization strength, making it difficult to maintain stable performance improvements across varying settings with consistent hyper-parameters. For instance, in VILA with COCO-Cap and IconQA under  $E = 5$ , Grafting demonstrates inferior performance compared to Full Fine-Tuning, highlighting its limitations in achieving robust and consistent results. **Third**, our method NRCA promotes prediction confidence alignment and consistently delivers robust results across various downstream tasks and model architectures. Moreover, as demonstrated in Tab. 2, our approach requires fewer computational resources compared to existing methods, highlighting its efficiency.

## 5. Conclusion

In our work, we focus on fine-tuning Multimodal Large Language Model (MLLM) in multi-task scenarios to achieve multi-domain specialization. However, we observe a critical issue: open-response distributions appear performance degradation during the training process. We attribute this degradation to over language prior reliance. To address this, we propose Noise Resilient Confidence Alignment (NRCA), which aims to enhance the effect of visual cues in the prediction process. Generally speaking, we leverage Gaussian distortions to construct noisy visual inputs and encourage overall token confidence alignment with the normal visual branch behavior. Compared to existing methods, our approach offers two key advantages: First, *Resource Cost Decline*: NRCA conducts output calibration without requiring storage of pre-trained Multimodal Large Language Model weights to counteract overfitting phenomenon, making ours computationally efficient compared to existing methods. Second, *Hyper-Parameter Stability*: Related methods often rely on pre-defined mask ratios or decorate parameter stiffness control, which are highly correlated with the architecture and scale of MLLM. In contrast, our method

introduces prediction confidence calibration that shows the model-agnostic property. NRCA has been validated across various scenarios, demonstrating its effectiveness and highlighting its potential for broader applications.

## Acknowledgement

This work is supported by National Natural Science Foundation of China under Grant (62225113, 62361166629, 62176188, 623B2080), the National Key Research and Development Program of China (2023YFC2705700, 2024YFC3308400), and the Wuhan University Undergraduate Innovation Research Fund Project. The supercomputing system at the Supercomputing Center of Wuhan University supported the numerical calculations in this paper. Dr Tao's research is supported by NTU RSR and Start Up Grants.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## References

- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J. L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Bińkowski, M. a., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, pp. 23716–23736, 2022.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- Ansell, A., Ponti, E. M., Korhonen, A., and Vujić, I. Composable sparse fine-tuning for cross-lingual transfer. In *ACL*, 2022.
- Ardizzone, L., Mackowiak, R., Rother, C., and Köthe, U. Training normalizing flows with the information bottleneck for competitive generative classification. In *NeurIPS*, volume 33, pp. 7828–7840, 2020.
- Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- Bi, J., Wang, Y., Chen, H., Xiao, X., Hecker, A., Tresp, V., and Ma, Y. Llava steering: Visual instruction tun-

- ing with 500x fewer parameters through modality linear representation-steering. In *ACL*, 2025a.
- Bi, J., Wang, Y., Yan, D., Xiao, X., Hecker, A., Tresp, V., and Ma, Y. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*, 2025b.
- Biderman, D., Ortiz, J. G., Portes, J., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., et al. Lora learns less and forgets less. *TMLR*, 2024.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *NeurIPS*, pp. 1877–1901, 2020.
- Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*, pp. 15920–15930, 2020.
- Che, L., Liu, T. Q., Jia, J., Qin, W., Tang, R., and Pavlovic, V. Eazy: Eliminating hallucinations in lmlms by zeroing out hallucinatory image tokens. *arXiv preprint arXiv:2503.07772*, 2025.
- Chen, B., Lyu, X., Gao, L., Song, J., and Shen, H. T. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *arXiv preprint arXiv:2405.15356*, 2024a.
- Chen, M., Fu, D. Y., Narayan, A., Zhang, M., Song, Z., Fatahalian, K., and Ré, C. Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *ICML*, pp. 3090–3122, 2022.
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607, 2020.
- Chen, Y., Wang, S., Lin, Z., Qin, Z., Zhang, Y., Ding, T., and Sun, R. Mofo: Momentum-filtered optimizer for mitigating forgetting in llm fine-tuning. *arXiv preprint arXiv:2407.20999*, 2024b.
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.
- De Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. A tutorial on the cross-entropy method. *Ann. Oper. Res.*, pp. 19–67, 2005.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Dong, G., Yuan, H., Lu, K., Li, C., Xue, M., Liu, D., Wang, W., Yuan, Z., Zhou, C., and Zhou, J. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*, 2023.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.
- Fang, Y., Zhu, L., Lu, Y., Wang, Y., Molchanov, P., Cho, J. H., Pavone, M., Han, S., and Yin, H. vila<sup>2</sup>: Vila augmented vila. *arXiv preprint arXiv:2407.17453*, 2024.
- Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., and Soatto, S. Multi-modal hallucination control by visual information grounding. In *CVPR*, pp. 14303–14312, 2024.
- Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., and Herzig, J. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*, 2024.
- Ghosh, S., Evuru, C. K. R., Kumar, S., Tyagi, U., Nieto, O., Jin, Z., and Manocha, D. Visual description grounding reduces hallucinations and boosts reasoning in lmlms. *arXiv preprint arXiv:2405.15683*, 2024.
- Goodman, N. R. Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *The Annals of mathematical statistics*, 34(1):152–177, 1963.
- Gu, Y., Dong, L., Wei, F., and Huang, M. Minillm: Knowledge distillation of large language models. In *ICLR*, 2024.
- Gu, Y., Zhou, H., Meng, F., Zhou, J., and Huang, M. Miniplm: Knowledge distillation for pre-training language models. In *ICLR*, 2025.

- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, pp. 1321–1330, 2017.
- Guo, Z., Xu, R., Yao, Y., Cui, J., Ni, Z., Ge, C., Chua, T.-S., Liu, Z., and Huang, G. LLava-UHD: an lmm perceiving any aspect ratio and high-resolution images. In *ECCV*, 2024.
- Han, S., Pool, J., Tran, J., and Dally, W. Learning both weights and connections for efficient neural network. In *NeurIPS*, 2015.
- Han, Z., Gao, C., Liu, J., Zhang, S. Q., et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.
- Hao, Y., Cao, Y., and Mou, L. Flora: Low-rank adapters are secretly gradient compressors. In *ICML*, 2024.
- Hayashi, H. and Uchida, S. A discriminative gaussian mixture model with sparsity. *arXiv preprint arXiv:1911.06028*, 2019.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.
- Hu, J., Zhang, J., Qi, L., Shi, Y., and Gao, Y. Learn to preserve and diversify: Parameter-efficient group with orthogonal regularization for domain generalization. In *ECCV*, 2024.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM TIS*, 2023.
- Huang, W., Yi, M., and Zhao, X. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021.
- Huang, W., Liang, J., Guo, X., Fang, Y., Wan, G., Rong, X., Wen, C., Shi, Z., Li, Q., Zhu, D., et al. Keeping yourself is important in downstream tuning multimodal large language model. *arXiv preprint arXiv:2503.04543*, 2025.
- Hui, T., Zhang, Z., Wang, S., Xu, W., Sun, Y., and Wu, H. Hft: Half fine-tuning for large language models. *arXiv preprint arXiv:2404.18466*, 2024.
- Huo, F., Xu, W., Zhang, Z., Wang, H., Chen, Z., and Zhao, P. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024.
- Jin, C., Che, T., Peng, H., Li, Y., Metaxas, D., and Pavone, M. Learning from teaching regularization: Generalizable correlations should be easy to imitate. In *NeurIPS*, 2024.
- Jin, C., Peng, H., Zhang, Q., Tang, Y., Metaxas, D. N., and Che, T. Two heads are better than one: Test-time scaling of multi-agent collaborative reasoning. *arXiv preprint arXiv:2504.09772*, 2025.
- Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pp. 787–798, 2014.
- Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In *NeurIPS*, 2020.
- Kim, G., Jang, D., and Yang, E. Promptkd: Distilling student-friendly knowledge for generative language models via prompt tuning. *arXiv preprint arXiv:2402.12842*, 2024a.
- Kim, S., Cho, B., Bae, S., Ahn, S., and Yun, S.-Y. Vacode: Visual augmented contrastive decoding. *arXiv preprint arXiv:2408.05337*, 2024b.
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, pp. 3521–3526, 2017.
- Ko, J., Kim, S., Chen, T., and Yun, S.-Y. Distillm: Towards streamlined distillation for large language models. In *ICML*, 2024.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.
- Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pp. 13872–13882, 2024.
- Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., and Li, C. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.
- Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900. PMLR, 2022a.

- Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pp. 19730–19742. PMLR, 2023a.
- Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In *MLSys*, 2020.
- Li, Y., Luo, F., Tan, C., Wang, M., Huang, S., Li, S., and Bai, J. Parameter-efficient sparsity for large language models fine-tuning. In *IJCAI*, 2022b.
- Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023b.
- Li, Y., Yu, Y., Zhang, Q., Liang, C., He, P., Chen, W., and Zhao, T. Losparse: Structured compression of large language models based on low-rank and sparse approximation. In *ICML*, pp. 20336–20350. PMLR, 2023c.
- Li, Z., Chen, C., Xu, T., Qin, Z., Xiao, J., Sun, R., and Luo, Z.-Q. Entropic distribution matching in supervised fine-tuning of llms: Less overfitting and better diversity. *arXiv preprint arXiv:2408.16673*, 2024b.
- Liang, J., Huang, W., Wan, G., Yang, Q., and Ye, M. Lorasculpt: Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models. In *CVPR*, 2025.
- Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., and Han, S. Vila: On pre-training for visual language models. In *CVPR*, 2023.
- Lin, T., Liu, J., Zhang, W., Li, Z., Dai, Y., Li, H., Yu, Z., He, W., Li, J., Jiang, H., et al. Teamlora: Boosting low-rank adaptation with expert collaboration and competition. *arXiv preprint arXiv:2408.09856*, 2024.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.
- Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *CVPR*, 2023a.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023b.
- Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., and Peng, W. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.
- Liu, S., Zheng, K., and Chen, W. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*, 2024b.
- Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. Dora: Weight-decomposed low-rank adaptation. In *ICML*, 2024c.
- Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Yu, Z., Liang, X., and Zhu, S.-C. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS*, 2021.
- Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.
- Lu, Y., Qian, B., Yuan, C., Jiang, H., and Wang, X. Controlled low-rank adaptation with subspace regularization for continued training on large language models. *arXiv preprint arXiv:2410.16801*, 2024a.
- Lu, Z., Fan, C., Wei, W., Qu, X., Chen, D., and Cheng, Y. Twin-merging: Dynamic integration of modular expertise in model merging. In *NeurIPS*, 2024b.
- Luo, G., Zhou, Y., Zhang, Y., Zheng, X., Sun, X., and Ji, R. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024.
- Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.
- Mishra, A., Shekhar, S., Singh, A. K., and Chakraborty, A. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pp. 947–952. IEEE, 2019.
- OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Panigrahi, A., Saunshi, N., Zhao, H., and Arora, S. Task-specific skill localization in fine-tuned language models. In *ICML*, pp. 27011–27033. PMLR, 2023.
- Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, pp. 61–74, 1999.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

- Ren, Y., Guo, S., Bae, W., and Sutherland, D. J. How to prepare your task head for finetuning. In *ICLR*, 2023.
- Ritter, H., Botev, A., and Barber, D. Online structured laplace approximations for overcoming catastrophic forgetting. In *NeurIPS*, 2018.
- Shang, Y., Zeng, X., Zhu, Y., Yang, X., Fang, Z., Zhang, J., Chen, J., Liu, Z., and Tian, Y. From pixels to tokens: Revisiting object hallucinations in large vision-language models. *arXiv preprint arXiv:2410.06795*, 2024.
- Shen, L., Chen, G., Shao, R., Guan, W., and Nie, L. Mome: Mixture of multimodal experts for generalist multimodal large language models. *arXiv preprint arXiv:2407.12709*, 2024a.
- Shen, Y., Shao, J., Zhang, X., Lin, Z., Pan, H., Li, D., Zhang, J., and Letaief, K. B. Large language models empowered autonomous edge ai for connected intelligence. *IEEE Communications Magazine*, 62(10):140–146, 2024b.
- Shi, M., Liu, F., Wang, S., Liao, S., Radhakrishnan, S., Huang, D.-A., Yin, H., Sapra, K., Yacoob, Y., Shi, H., et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024.
- Shuttleworth, R., Andreas, J., Torralba, A., and Sharma, P. Lora vs full fine-tuning: An illusion of equivalence. *arXiv preprint arXiv:2410.21228*, 2024.
- Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *CVPR*, pp. 8317–8326, 2019.
- Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning? In *NeurIPS*, volume 33, pp. 6827–6839, 2020.
- Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., Iyer, A., Pan, X., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Tüske, Z., Tahir, M. A., Schlüter, R., and Ney, H. Integrating gaussian mixtures into deep neural networks: Softmax layer with hidden variables. In *ICASSP*, pp. 4285–4289. IEEE, 2015.
- Variani, E., McDermott, E., and Heigold, G. A gaussian mixture model layer jointly optimized with discriminative features within a deep neural network architecture. In *ICASSP*, pp. 4270–4274. IEEE, 2015.
- Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *CVPR*, pp. 4566–4575, 2015.
- Wang, F., Xiang, X., Cheng, J., and Yuille, A. L. Normface: L2 hypersphere embedding for face verification. In *ACM MM*, pp. 1041–1049, 2017.
- Wang, J., Ming, Y., Shi, Z., Vineet, V., Wang, X., and Joshi, N. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. In *NeurIPS*, 2024a.
- Wang, P., Panda, R., Hennigan, L. T., Greengard, P., Karlinsky, L., Feris, R., Cox, D. D., Wang, Z., and Kim, Y. Learning to grow pretrained models for efficient transformer training. In *ICLR*, 2023.
- Wang, X., Pan, J., Ding, L., and Biemann, C. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *ACL*, 2024b.
- Wang, Y., Zhang, Q., Wang, Y., Yang, J., and Lin, Z. Chaos is a ladder: A new understanding of contrastive learning. In *ICLR*, 2022.
- Wang, Y., Zhang, J., and Wang, Y. Do generated data always help contrastive learning? In *ICLR*, 2024c.
- Woo, S., Jang, J., Kim, D., Choi, Y., and Kim, C. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. *arXiv preprint arXiv:2405.17821*, 2024.
- Wu, C., Wu, F., and Huang, Y. Rethinking infonce: How many negative samples do you need? In *IJCAI*, 2022.
- Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pp. 3733–3742, 2018.
- Xiao, T., Wang, X., Efros, A. A., and Darrell, T. What should not be contrastive in contrastive learning. In *ICLR*, 2021.
- Xiao, Y., Lin, Y., and Chiu, M.-C. Behavioral bias of vision-language models: A behavioral finance view. *arXiv preprint arXiv:2409.15256*, 2024.
- Xuhong, L., Grandvalet, Y., and Davoine, F. Explicit inductive bias for transfer learning with convolutional networks. In *ICML*, pp. 2825–2834. PMLR, 2018.
- Ye, M., Zhang, X., Yuen, P. C., and Chang, S.-F. Unsupervised embedding learning via invariant and spreading instance feature. In *CVPR*, pp. 6210–6219, 2019.
- Ye, M., Shen, J., Zhang, X., Yuen, P. C., and Chang, S.-F. Augmentation invariant and instance spreading feature for softmax embedding. *IEEE PAMI*, 2020.

- Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.
- Yu, L., Yu, B., Yu, H., Huang, F., and Li, Y. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *ICML*, 2024.
- Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *ICML*, pp. 3987–3995. PMLR, 2017.
- Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., and Zhao, T. Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.
- Zhang, W., Janson, P., Aljundi, R., and Elhoseiny, M. Overcoming generic knowledge loss with selective parameter update. In *CVPR*, pp. 24046–24056, 2024a.
- Zhang, X., Quan, Y., Shen, C., Yuan, X., Yan, S., Xie, L., Wang, W., Gu, C., Tang, H., and Ye, J. From redundancy to relevance: Information flow in lmlms across reasoning tasks. In *NAACL*, 2025.
- Zhang, Y.-F., Yu, W., Wen, Q., Wang, X., Zhang, Z., Wang, L., Jin, R., and Tan, T. Debiasing large visual language models. *arXiv preprint arXiv:2403.05262*, 2024b.
- Zhang, Z., Yadav, S., Han, F., and Shutova, E. Cross-modal information flow in multimodal large language models. In *CVPR*, 2024c.
- Zhang, Z., Zhang, Q., Gao, Z., Zhang, R., Shutova, E., Zhou, S., and Zhang, S. Gradient-based parameter selection for efficient fine-tuning. In *CVPR*, pp. 28566–28577, 2024d.
- Zhao, L., Deng, Y., Zhang, W., and Gu, Q. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*, 2024.
- Zhao, N., Wu, Z., Lau, R. W., and Lin, S. What makes instance discrimination good for transfer learning? In *ICLR*, 2021.
- Zhou, X., He, J., Ke, Y., Zhu, G., Gutiérrez-Basulto, V., and Pan, J. Z. An empirical study on parameter-efficient fine-tuning for multimodal large language models. In *ACL*, 2024.
- Zhu, D., Sun, Z., Li, Z., Shen, T., Yan, K., Ding, S., Kuang, K., and Wu, C. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. In *ICML*, 2024a.
- Zhu, T., Liu, Q., Wang, F., Tu, Z., and Chen, M. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*, 2024b.
- Zhu, Y., Zhu, M., Liu, N., Ou, Z., Mou, X., and Tang, J. Llava-phi: Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*, 2024c.

## APPENDIX

### A. Notation Table

We provide the notation table in Tab. 6.

Table 6: Notations table.

Description	Description
$\theta$	MLLM
$\varphi$	Connector
$\mathbf{y}$	Language response
$V$	Vocabulary dictionary
$x^v$	Visual image
$\tilde{x}^v$	Noisy visual image
$m^v$	Visual feature
$h^t$	Text token embedding
$\mathbf{p}_t$	Token Probability
$\mathcal{I}$	Overall token prediction
$E$	Fine-Tuning epoch
$\mathcal{L}$	Loss function
$\eta$	Learning rate
$f$	Vision encoder
$g$	Large Language Model
$T$	Label Length
$N$	LLM tuning blocks
$\mathcal{N}$	Gaussian distribution
$x^t$	Textual instruction
$h^v$	Visual token embedding
$z$	Logits output
$\mathbf{p}_t^{y_t}$	Token Confidence
$\sigma$	softmax function
$D$	Downstream dataset
$\lambda$	Control hyper-parameter
$B$	Training batch size

### B. Algorithm

We provide the algorithm description in Algorithm 1.

#### Algorithm 1 NRCA

**Input:** Fine-Tuning epoch  $E$ , Downstream dataset  $D$   
 Overall MLLM Network  $\theta$ , Trainable parameter module  $w$   
**Output:** The optimized selected MLLM module  $w$

```

for e = 1, 2, ..., E do
    for  $(x^v, x^t, \mathbf{y}) \in D$  do
        /* Construct Noisy Vision View */
         $\mu, \sigma \leftarrow (x^v)$  via Eq. (2a)
         $\tilde{x}^v \leftarrow (x^v, \mu, \sigma)$  by Eq. (2b)
         $h^v = \varphi(f(x^v))$  and  $\tilde{h}^v = \varphi(f(\tilde{x}^v))$ 
         $h^t = \text{Tokenize}(x^t)$ 
         $z = g(h^v, h^t)$  and  $\tilde{z} = g(\tilde{h}^v, h^t)$ 
        /* Token Confidence Alignment */
         $\mathbf{p}_t = \sigma(z_t), \tilde{\mathbf{p}}_t = \sigma(\tilde{z}_t);$  // Token Prob.
         $\mathbf{p} = [\mathbf{p}_t^{y_t}]_{t=1}^T, \tilde{\mathbf{p}} = [\tilde{\mathbf{p}}_t^{y_t}]_{t=1}^T;$  // Token Conf.
         $\mathcal{L}_{NRCA} \leftarrow (\mathbf{p}, \tilde{\mathbf{p}})$  through Eq. (5b)
         $\mathcal{L}_{CE} \leftarrow (\mathbf{p}, \mathbf{y})$  in Eq. (5a)
         $\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{NRCA}$ 
         $w = w - \eta \nabla \mathcal{L};$  // Update Param.
    end
end

```

### C. Rationale Analysis

We conduct a theoretical discussion to investigate the rationale for regularizing the prediction confidence in the Multimodal Large Language Model (MLLM) to mitigate

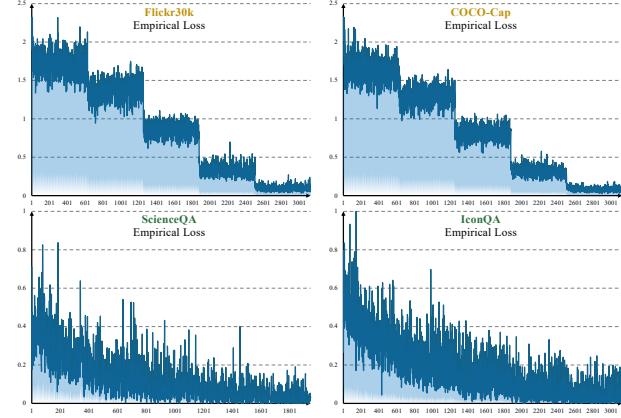


Figure 6: **Empirical Loss** on both **Open-Response** and **Fixed-Choice** tasks. Experiments are based on the LLaVA with tuning epoch  $E=5$ . We observe a *stepwise loss* tendency in the former. The open-response distribution compels MLLM to over-memorize each sample during every training epoch, resulting in a jumping loss behavior. Please refer to Appendix C.

overfitting. As a preliminary step, we revisit the classical MLLM fine-tuning optimization loss. Specifically, we suppose  $(x^v, x^t, \mathbf{y})$  as the query instance from the training dataset, where  $\mathbf{y} \in \mathbb{R}^{T \times |V|}$  is the ground-truth label. We obtain the prediction token logits as  $z = \theta(x^v, x^t)$ . Then, we formulate the cross-entropy (De Boer et al., 2005) as:

$$\begin{aligned}
 \mathcal{L}_{CE} &= \frac{1}{T} \sum_{t=1}^T -\mathbf{1}_{\mathbf{y}_t} \log(\sigma(z)), \\
 &= \frac{1}{T} \sum_{t=1}^T -\mathbf{1}_{\mathbf{y}_t} \log(\mathbf{p}_t), \\
 &= \frac{1}{T} \sum_{t=1}^T \log(\mathbf{p}_t^{y_t}),
 \end{aligned} \tag{7}$$

where  $\mathbf{1}_{\mathbf{y}_t}$  denotes the one-hot encoding of  $\mathbf{y}_t$ , and  $\sigma$  represents the softmax operation (Wang et al., 2017). Thus, to some extent, minimizing  $\mathcal{L}_{CE}$  can be interpreted as maximizing the prediction confidence for each token,  $\mathbf{p}_t^{y_t}$ .

Consequently, MLLM, with its large-scale network capacity, often encounters less challenging scenarios when handling downstream tasks with limited scope, making it prone to memorizing the training samples. In open-response tasks, a long ground-truth label sequence is typically involved. As a result, during empirical loss minimization, each token prediction confidence naturally increases to align with the target distribution. In contrast, for fixed-choice tasks, where only a single ground-truth token label usually exists (*i.e.*,  $T = 1$ ), responses are constrained to predefined options such as A, B, C, or D. This constrained candidate label space makes it more challenging to enforce memorization of each individual sample.

Therefore, we respectively plot the training loss curve  $\mathcal{L}_{CE}$  for both **open-response** and **fixed-choice** tasks. As shown

Table 7: Comparison with the state-of-the-art Multimodal Large Language Model (MLLM) Fine-Tuning Solutions on the multi-task setting: **image captioning** (Flickr30k and COCO-Cap) and **visual question-answering** (ScienceQA and IconQA) based on the VILA architectures with tuning epoch  $E = 3$ . We mark the Best in bold and Second in underline across different tuning methods.  $+$  means improved accuracy compared with Zero-shot. Please refer to Appendix D for relative explanations.

Methods	Flickr30k	ScienceQA	AVG	COCO-Cap	ScienceQA	AVG	Flickr30k	IconQA	AVG	COCO-Cap	IconQA	AVG
<i>Fine-Tune with <math>B = 32</math></i>												
Zero-shot	55.45	68.96	62.20	72.57	68.96	70.76	55.45	55.57	55.51	72.57	55.57	64.07
Full FT	71.73	90.28	81.01	111.22	90.13	100.68	71.67	86.68	79.18	110.86	87.46	99.16
Ran Mask	76.90	90.98	83.94	117.07	90.88	103.98	75.21	86.65	80.93	116.84	87.11	101.98
Mag Mask	77.89	90.78	84.34	117.35	91.27	104.31	76.13	86.97	81.55	116.93	86.91	101.92
Grafting	73.29	90.63	81.96	112.28	90.18	101.23	70.58	87.11	78.85	110.72	87.30	99.01
L2-Reg	73.92	90.68	82.30	111.5	90.48	100.99	71.73	87.06	79.40	110.76	87.35	99.06
NRCA	79.63	90.08	84.86	122.50	90.73	106.62	80.96	86.62	83.79	122.60	86.11	104.35
	<b>+ 24.18</b>	<b>+ 21.12</b>	<b>+ 22.66</b>	<b>+ 49.93</b>	<b>+ 21.77</b>	<b>+ 35.86</b>	<b>+ 25.51</b>	<b>+ 31.05</b>	<b>+ 28.28</b>	<b>+ 50.03</b>	<b>+ 30.54</b>	<b>+ 40.28</b>
<i>Fine-Tune with <math>B = 24</math></i>												
Zero-shot	55.45	68.96	62.20	72.57	68.96	70.76	55.45	55.57	55.51	72.57	55.57	64.07
Full FT	71.57	90.68	81.13	110.31	90.48	100.40	71.63	87.08	79.35	109.72	86.49	98.11
Ran Mask	75.62	90.23	82.93	115.83	90.73	103.28	75.55	87.03	81.29	116.13	87.27	101.70
Mag Mask	75.16	90.63	82.90	115.35	90.68	103.02	74.27	87.27	80.77	115.10	87.16	101.13
Grafting	71.64	90.03	80.84	110.09	90.53	100.31	68.88	87.1	77.99	110.21	86.73	98.47
L2-Reg	71.31	90.08	80.70	109.81	90.38	100.10	69.75	87.38	78.57	109.65	86.78	98.22
NRCA	81.20	90.18	85.69	123.22	90.43	106.83	80.93	86.15	83.54	123.84	86.89	105.37
	<b>+ 25.75</b>	<b>+ 21.22</b>	<b>+ 23.49</b>	<b>+ 50.65</b>	<b>+ 21.47</b>	<b>+ 36.07</b>	<b>+ 25.48</b>	<b>+ 30.58</b>	<b>+ 28.03</b>	<b>+ 51.27</b>	<b>+ 31.32</b>	<b>+ 41.30</b>

in Appendix C, the training loss for open-response tasks does not exhibit a progressive loss tendency but instead demonstrates a stepwise loss behavior. We hypothesize that this phenomenon is because MLLM tends to memorize each sample rather than explore the underlying distribution patterns, also observed in (Li et al., 2024b; Lu et al., 2024a). In contrast, it is challenging to map the same choice to different multi-modal inputs for fixed-choice settings. Consequently, MLLM is encouraged to analyze the multi-modal inputs and make informed decisions.

With recall of the empirical open-response task minimization, it would blindly increase the prediction confidence,  $p_t^{y_t}$ . Therefore, in our work, we aim to reduce the reliance of MLLM on language priors and instead emphasize the role of visual features in prediction behavior. Generally, we encourage noisy visual inputs to achieve consistent prediction confidence comparable to the normal branch, thereby mitigating learning shortcuts.

To formalize this, we denote  $I$  and  $\tilde{I}$  as the average prediction confidence over ground-truth tokens under clean and perturbed visual inputs (see Eq. (4)), respectively. Let  $z_t$  and  $\tilde{z}_t$  be the predicted logits at position  $t$ , and let  $\sigma(\cdot)$  denote the softmax function. The NRCA regularization seeks to minimize the relative deviation:

$$\min \left| 1 - \frac{\tilde{I}}{I} \right| = \min \left| \frac{1}{T} \sum_{t=1}^T \frac{\sigma_{y_t}(z_t) - \sigma_{y_t}(\tilde{z}_t)}{I} \right|. \quad (8)$$

Since  $I$  is detached from the computational graph, this simplifies to:

$$\min \left| \sum_{t=1}^T \sigma_{y_t}(z_t) - \sigma_{y_t}(\tilde{z}_t) \right|. \quad (9)$$

Assuming that  $\sigma_{y_t}(z_t) \geq \sigma_{y_t}(\tilde{z}_t)$  holds (a natural assumption since clean inputs tend to yield higher confidence), the optimization objective becomes:

$$\min \sum_{t=1}^T (\sigma_{y_t}(z_t) - \sigma_{y_t}(\tilde{z}_t)), \quad (10)$$

which encourages the perturbed-input confidence  $\sigma_{y_t}(\tilde{z}_t)$  to approach that of the clean input  $\sigma_{y_t}(z_t)$ . Since  $\tilde{z}_t = g(\phi(f(\tilde{x}^v)), h_t)$ , where  $f$  is the visual encoder and  $\phi$  and  $g$  denote the projection and decoding modules, this objective drives the network to produce robust visual representations that are invariant to input perturbations. As a result, our model is encouraged to ground its predictions more faithfully in visual evidence rather than language priors, thereby mitigating overfitting by preventing overconfidence on individual training samples. We provide the full implementation and formulation in Sec. 3.

## D. Additional Experiments

We provide the experiment comparison with different training batch sizes of  $B \in \{32, 24\}$  in Tab. 7. We conduct experiments on the VILA architectures on the four downstream datasets with the tuning epochs as  $E = 3$ . Compared with existing Multimodal Large Language Model methods, our method consistently demonstrates effectiveness.