CoMMIT: Coordinated Multimodal Instruction Tuning

Xintong Li[1*] **Junda Wu**[1*] **Tong Yu**[2] **Rui Wang**[2] **Yu Wang**[1] **Xiang Chen**[2]
Jiuxiang Gu[2] **Lina Yao**[3,4] **Julian McAuley**[1] **Jingbo Shang**[1]
[1]University of California, San Diego [2]Adobe Research
[3]The University of New South Wales [4]CSIRO's Data61
{xil240,juw069,yuw164,jmcauley,jshang}@ucsd.edu
{tyu,xiangche,jigu}@adobe.com lina.yao@data61.csiro.au

Abstract

Instruction tuning in multimodal large language models (MLLMs) generally involves cooperative learning between a backbone LLM and a feature encoder of non-text input modalities. The major challenge is how to efficiently find the synergy between the two modules so that LLMs can adapt their reasoning abilities to downstream tasks while feature encoders can adjust to provide more task-specific information about its modality. In this paper, we analyze the MLLM instruction tuning from both theoretical and empirical perspectives, where we find the unbalanced learning between the feature encoder and the LLM can cause problems of oscillation and biased learning that lead to sub-optimal convergence. Inspired by our findings, we propose a Multimodal Balance Coefficient that enables quantitative measurement of the balance of learning. Based on this, we further design a dynamic learning scheduler that better coordinates the learning between the LLM and feature encoder, alleviating the problems of oscillation and biased learning. In addition, we introduce an auxiliary regularization on the gradient to promote updating with larger step sizes, which potentially allows for a more accurate estimation of the proposed MultiModal Balance Coefficient and further improves the training sufficiency. Our proposed approach is agnostic to the architecture of LLM and feature encoder, so it can be generically integrated with various MLLMs. We conduct experiments on multiple downstream tasks with various MLLMs, demonstrating the proposed method is more effective than the baselines in MLLM instruction tuning.

1 Introduction

Multimodal instruction tuning aligns pre-trained multimodal large language models (MLLMs) with specific downstream tasks by fine-tuning MLLMs to follow arbitrary instructions (Dai et al., 2024; Zhang et al., 2023; Zhao et al., 2024; Lu et al., 2023; Han et al., 2023; Wu et al., 2024b; Wang et al., 2024b; Wu et al., 2025b,a). Leading pre-trained Multimodal Large Language Models (MLLMs) typically share similar architectures (Li et al., 2023; Liu et al., 2024; Tang et al., 2023a; Chu et al., 2023). Specifically, the non-text data (image, audio, etc) is first encoded by a feature encoder into embedding tokens. Then, these encoded embeddings are inserted into language prompts, creating multimodal sequence inputs for the LLMs. Effective multimodal understanding and reasoning in MLLMs depend on the model's ability to learn aligned multimodal features using its feature encoder (*e.g.*, (Li et al., 2023)), and on leveraging the pre-trained capabilities of its backbone LLM (*e.g.*, (Touvron et al., 2023; Chiang et al., 2023)) to interpret these multimodal inputs. This generally involves a two-prolonged learning process: (1) **LLM Adaptation.** Encoded non-text features (*e.g.*, visual and auditory) in downstream tasks may not be perfectly aligned with pre-trained text features, thus requiring the backbone LLM to adapt its pre-trained parameters to recognize these new, non-text modality tokens. (2) **Feature Encoder Adaptation.** While LLMs possess strong reasoning ability

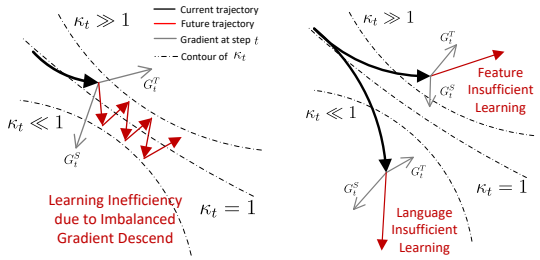

Figure 1: Illustration of (a) the oscillation problem and (b) the biased learning problem, caused by imbalanced multimodal learning. The optimization trajectories are shown in solid bold lines and the multimodal gradients at the current step t are in solid thin lines.

*These authors contributed equally to this work.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11534–11548
November 4-9, 2025

from their pre-trained, it requires the feature encoders to be fine-tuned to extract task-specific information for evidence of reasoning. Cooperative balancing of these two learning stages is crucial for effective instruction tuning of MLLMs. When the learning is biased on the LLMs with (1), the insufficiently learned feature encoder can lead to information loss (Bai et al., 2024; Tong et al., 2024; Wu et al., 2025c), hindering the LLM's ability to reason effectively due to a lack of adequate evidence from non-text modalities. Conversely, if the learning is biased on the feature encoder with (2), the LLMs can be insufficiently adapted and struggle to interpret non-text modalities. As a result, it will cause the hallucination problem (Bai et al., 2024; Rawte et al., 2024; Wu et al., 2024a,d) due to the strong language prior inherent in the backbone LLMs. Therefore, it is essential to balance the learning between the feature encoder and backbone LLM, so that the learning is not overly biased on either of the two modules.

In this paper, we first propose a multimodal balance coefficient that quantifies the learning balance between the feature encoder and the backbone LLM in MLLM instruction tuning. Based on theoretical analysis and empirical observations, we identify two types of learning dilemmas that can be quantitatively measured by our proposed multimodal balance coefficient: *i)* the oscillation problem and *ii)* the biased learning problem, as illustrated in Figure 1. Specifically, Figure 1(a) demonstrates the oscillation problem where the learning is *alternatively* favoring either the feature encoder or the LLM. This oscillation impedes the convergence of optimization and undermines learning efficiency since the learning is hardly progressing in consistent directions. On the other hand, Figure 1(b) shows the biased learning problem where the training *consistently* favors either the LLM or the feature encoder. In such cases, the gradient descent primarily only updates either the LLM or the feature encoder, resulting in insufficient learning of the other module. This diminishes the effectiveness of gradient descent since the under-trained module (LLM or feature encoder) will not be capable of contributing sufficient information to the generation outputs.

To address these challenges, we propose **Co**ordinated **M**ulti**M**odal **I**nstruction **T**uning (**CoMMIT**), which regularizes the training with a coordinated learning rate scheduler (Section 6). This scheduler dynamically adjusts the learning rates of the feature encoder and LLM according to the proposed multimodal balance coefficient, ensuring sufficient gradient descent for both the feature encoder and LLM while mitigating the oscillation problem. We also introduce a regularization loss that promotes larger update steps during training, further alleviating gradient diminishing. We theoretically analyze the convergence rate and demonstrate that we can achieve accelerated convergence when optimizing with **CoMMIT** (Section A). We summarize our main contributions as follows:

- We introduce a theoretical framework to uncover the pitfalls of the learning imbalance problem in MLLM instruction tuning, which can cause MLLM insufficient learning and the oscillation problem.
- Based on the theoretical analysis and empirical observation, we propose **CoMMIT** to balance multimodal learning progress by dynamically coordinating learning rates on the feature encoder and LLM. **CoMMIT** also enforces a gradient regularization that encourages larger step sizes and improves training efficiency.
- Applying **CoMMIT** introduces a novel term in the convergence rate analysis. Theoretical analysis proves that this term is always greater than one, leading to faster convergence. We also demonstrate that the theorem can be generalized across various optimizers.
- Empirical results on downstream tasks in vision and audio modalities with various LLM backbones show the efficiency and effectiveness of the methods. We demonstrate that **CoMMIT** can better coordinate multimodal learning progress with balanced training between the feature encoder and the LLM.

2 Related Works

MLLMs have become a new paradigm to empower multimodal learning with advanced language reasoning capabilities, such as with vision (Li et al., 2023; Liu et al., 2024; Wang et al., 2024c; Maaz et al., 2023; Zhang et al., 2023; Huang et al., 2023a; Yan et al., 2024), and audio (Huang et al., 2023b; Tang et al., 2023a; Gardner et al., 2023). Despite good generalizability and zero-shot performance of existing large language models (LLMs), the discrepancy between different modalities can be one of the greatest challenges for LMMs to achieve

comparable reasoning performance as LLMs. To bridge the multimodality gap and align with downstream tasks, several works focus on two-fold considerations: feature (modality) alignment and reasoning alignment. The most common approach for feature alignment is to encode the source modality feature to semantic tokens within the LLMs' embedding feature space. By adding the modality-specific tokens (Wang et al., 2024a; Liu et al., 2021a; Zhang et al., 2024) as soft prompt inputs (Liu et al., 2021b; Xie et al., 2023; Wu et al., 2023, 2024c), the backbone LLMs can process these tokens with language tokens as a unified sequence. However, the newly added semantic tokens cannot be understood by LLMs directly for language reasoning, due to the limited text-only pretraining of LLMs. Such misalignment problems will lead to textual hallucination problems, namely *linguistic bias* (Ko et al., 2023; Tang et al., 2023b), in which the language models reason only based on their language prior. Thus, multimodal alignment should be achieved by additional adaptation of the LLM itself with multimodal instruction tuning.

3 Preliminaries

Given a pair of non-text input I^S (images, audio, etc) and instruction prompt I^X of nature language, the instruction-tuned MLLM should comprehend the semantics of I^S and generate outputs that comply with the instruction specified in I^X. State-of-the-art MLLM instruction tuning generally adopts similar diagrams of training (Gardner et al., 2023; Li et al., 2023; Liu et al., 2024), which involves cooperative training between a feature encoder S and a pretrained LLM X. Specifically, S first encodes the multimedia input I^S into the embedding space of X. Then, the encoded I^S is inserted into the instruction prompt I^X as input that conditions the output generation of X,

$$P_{S,X}(\hat{y}_k|I^S, I^X, y_{j<k}) = X(\hat{y}_k|S(I^S), I^X, y_{j<k}), \tag{1}$$

where $\hat{y}_k$ is the kth predicted token and $y_{j<k}$ denotes the first $k-1$ of the expected ground truth tokens in auto-regressive generation, $k = 1, \cdots, K$. The training loss is the cross-entropy defined by the predicted distribution on $\hat{y}_k$ and the kth ground truth token y_k(Liu et al., 2024; Ouyang et al., 2022),

$$L(Y = \{y_k\}_{k=1}^K \mid I_S, I^T) = \\ -\frac{1}{K}\sum_{k=1}^{K} y_k \log P_{S,X}(\hat{y}_k|I^S, I^X, y_{j<k}), \tag{2}$$

The learning objective is to find the optimal X and S by minimizing the loss function. As mentioned in Section 1, the learning can either be inefficient by oscillating between the optimization of the two modules or insufficient by biasing on one of S and X. Our goal is to find a balance between the learning of X and S, so to accelerate the convergence while ensuring that both S and X are sufficiently trained.

4 Measurement of Learning Balance in MLLM Instruction Tuning

To assess the balance between the updates on X and S, we first measure the significance of each update separately with X and S. Formally, for the t-th step of training, we define $d(P_{X_t,S_t}||P_{X_{t+1},S_t})$ and $d(P_{X_t,S_t}||P_{X_t,S_{t+1}})$ that quantify the significance of updates on X and S, respectively, by measuring the shift in output distributions,

$$d(P_{X_t,S_t}||P_{X_{t+1},S_t}) = \\ \frac{1}{K}\sum_k \mathbb{KL}\left(P_{S_t,X_t}(\hat{y}_k|I^S, I^X)||P_{S_{t+1},X_t}(\hat{y}_k|I^S, I^X)\right), \tag{3}$$

$$d(P_{X_t,S_t}||P_{X_t,S_{t+1}}) = \\ \frac{1}{K}\sum_k \mathbb{KL}\left(P_{S_t,X_t}(\hat{y}_k|I^S, I^X)||P_{S_t,X_{t+1}}(\hat{y}_k|I^S, I^X)\right) \tag{4}$$

where we use the subscript t to index the trained steps and $\mathbb{KL}(\cdot||\cdot)$ is the KL divergence. Based on (3) and (4), we define the Multimodal Balance Coefficient that measures the balance between training on X and S.

Definition 4.1 (Multimodal balance coefficient)**.** *For time step t of joint training on X and S, the Multimodal Balance Coefficient κ_t is measured with the respective learning steps on the feature encoder S_t and the LLM X_t.*

$$\kappa_t = \frac{d(P_{X_t,S_t}||P_{X_{t+1},S_t})}{d(P_{X_t,S_t}||P_{X_t,S_{t+1}})}. \tag{5}$$

κ_t with a value always near 1 indicates that the learning is balanced between the feature encoder and the LLM. On the contrary, $\kappa_t >> 1$ and $\kappa_t \to 0$ suggests that the learning biased by leaning toward X and S, respectively. κ_t with high

variance corresponds to learning that oscillates between optimizing on X or S. To further illustrate this, we derive Theorem 1 which estimated the gradient on X and S during training.

Theorem 4.2. *Let G_t^X and G_t^S be the gradient on X and S at time step t, we can derive the multimodal gradient estimated bounds,*

$$\|G_t^X\| \leq (\kappa_t + 1)H_t^S, \quad \|G_t^S\| \leq (\frac{1}{\kappa_t} + 1)H_t^X, \tag{6}$$

where H_t^S and H_t^T represent the individual learning steps of S and X, respectively. These are given by,

$$H_t^S = \left(\|I_t^X\| + \|S_t(I_t^S)\|\right)^{-1} \|logits(P_{X_t,S_{t+1}})\|,$$
$$H_t^X = \left\|I_t^S\right\|^{-1} \|logits(P_{X_{t+1},S_t})\|. \tag{7}$$

The detailed proof is provided in Appendix B.1.

Within the metric space of probability distribution $(P_{S,X}, d)$, the values of H_t^S and H_t^T are bounded by a finite norm. H_t^S and H_t^T are also lower-bounded, assuming multimodal gradients are not diminishing which we alleviate by proposing a regularization on the gradient in Section 6. Therefore, κ_t can account the most for the gradient upper bound in (6) so it is suitable to measure the learning imbalance problem.

5 Empirical Observations of Learning Dilemmas in MLLM Instruction Tuning

In this section, we illustrate the learning dilemmas described in Section 1 by observing the dynamics of κ_t is different experiment settings. The experiment is conducted by fine-tuning the BLIP-2 (Li et al., 2023) model on TextVQA(Singh et al., 2019), a widely used dataset for visual question question answering (Dai et al., 2024; Yin et al., 2024). To probe the problems of oscillation and learning bias, we consider the following three learning strategies: (1) **Synced LR** is trained by setting the learning rate of both X and S to $1e-4$; (2) **Language LR** $\uparrow$ increases learning rate of X to $1e-3$; (2) **Encoder LR** $\uparrow$ increases the learning rate of S to $1e-3$.

5.1 The Dilemma of Oscillation

To quantitatively understand the oscillation problem in MLLM instruction tuning, we show the learning curves (Figure 2) of the measurement variables H_t^S, H_t^X, and κ_t proposed in (6).

Observation 5.1. *As shown in Figure 2(c), the multimodal learning process can suffer from significant oscillation problems with highly variant κ_t in the **Synced LR** setting.*

Specifically, the learning curve of κ_t in the **Synced LR** setting exhibits high variance near the value of 1 , which is a showcase of the oscillation problems that signify training instability. This will cause inefficient training since the learning is alternating between X and S, instead of progressing in consistent directions. We demonstrate such inefficiency in Figures 5 and 6, showing that its convergence is slower compared to our proposed approach, which balances the learning process with a more stable κ_t. Further, it is interesting to find in Figure 2 that the feature encoder S is more unstable than the language model X, by comparing H_t^S in Figure 2(a) and H_t^X in Figure 2(b). By increasing the learning rate either X (**Encoder LR**) or S (**Language LR**), we observe that the three metrics in Figure 2 are stabilized. In the next section, we show that such stabilization is at the expense of biased learning, causing insufficient training on X or S.

5.2 The Dilemma of Biased Learning

The **Encoder LR** $\uparrow$ and **Language LR** $\uparrow$ in Figure 2 demonstrated the biased learning problem, where the training is biased on either S or X. Let θ_t^S and θ_t^X be the parameters of S and X at time step t. We further show three metrics with the same backbone MLLM and training data as in Section 5.1: (1) the normalized learning gradient $\|G_t^S\|/\|\theta_t^S\|$ of the feature encoder S in Figure 3(a), (2) the normalized learning gradient $\|G_t^X\|/\|\theta_t^X\|$ of the language model X in Figure 3(b), and (3) the cross-entropy loss in Figure 3(c). These metrics help understand the impact of biased learning on either X or S in MLLM instruction tuning.

Observation 5.2. *In Figure 3(c), **Encoder LR** $\uparrow$ and **language LR** $\uparrow$, biased on either X or S, can slow the convergence of the MLLM with gradient diminishing and inferior training performance.*

Such diminishing gradient would result in insufficient training on X or S with gradient descent. For example, we can observe in Figure 3(a) and Figure 3(b) that the **Encoder LR** can simultaneously cause the gradient diminishing in both X and S, with the cross-entropy converging to a higher value in Figure 3(c). In such cases, it is necessary to strategically balance the learning between

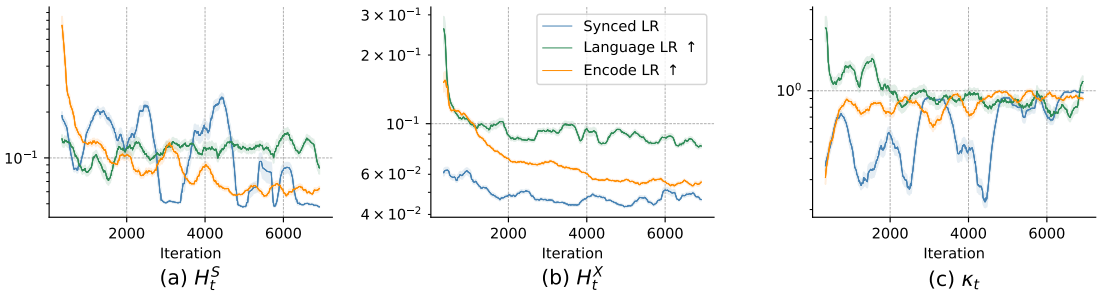

Figure 2: Learning curves of the variables H_t^S, H_t^X, and κ_t to measure learning balance in multimodal instruction tuning.

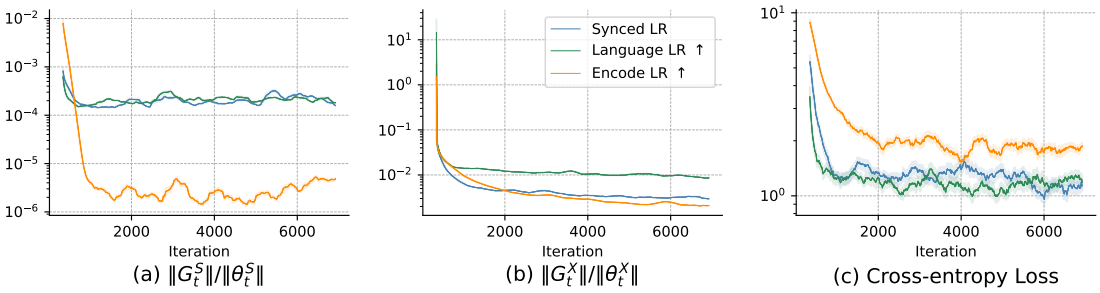

Figure 3: The learning curves of normalized learning gradient $\|G_t^S\|/\|\theta_t^S\|$ and $\|G_t^X\|/\|\theta_t^X\|$ for the feature encoder and language model respectively, as well as the cross-entropy training losses.

different modules, so the training is not leaning toward either X or S. In Figure 4, we show that our proposed approach can reduce such bias while not inducing oscillation, *i.e.*, by learning with less variant κ_t that is valued close to 1.

6 CoMMIT: Coordinated Multimodal Instruction Tuning

Based on the observations in Section 5, the learning rate on X should be boosted when the training is leaning toward S ($\kappa_t \to 0$), and vice versa. So motivated, we propose a dynamic learning rate scheduling method to coordinate multimodal learning between X and S, which alleviates the oscillation problems while ensuring the training is not biased on either X or S.

Inspired by damping strategies in optimization (Lucas et al., 2018; Tanaka and Kunin, 2021; Wei et al., 2021), we use the proposed learning balance metric κ_t in (5) as the damping parameter that facilitates balanced multimodal learning. Specifically, we track the N_κ moving average of κ_t through the learning process,

$$\tilde{\kappa}_t = \frac{1}{N_\kappa} \sum_{i=1}^{N_\kappa} \kappa_{t-i+1}, \tag{8}$$

then dynamically adjust learning rates of X and S in accordance. Let β_t^X and β_t^S be the learning rates on X and S at time step t. We adjust the learning rates by,

$$\beta_t^T = \frac{2\alpha}{\tilde{\kappa}_t + 1}, \ \beta_t^S = \frac{2\alpha}{1/\tilde{\kappa}_t + 1}, \tag{9}$$

where α is the base learning rate.

During training, the diminishing H_t^S and H_t^T as observed in Figure 3 can cause higher estimation errors in $\tilde{\kappa}_t$. To address these, we propose an auxiliary regularization that encourages large step sizes for both X and S, which mitigates gradient diminishing. Specifically, we want to encourage larger distribution drifts $d(P_{X_t,S_t}||P_{X_{t+1},S_t})$ for X and $d(P_{X_t,S_t}||P_{X_t,S_{t+1}})$ for S, apart from gradient descending on the cross-entropy loss in (2). The gradient update for our proposed CoMMIT at time step t is,

$$\begin{aligned} \theta_{t+1}^X \leftarrow \theta_t^X &- \beta_t^X \cdot \nabla_{\theta^X} L(S_t(\tilde{X}_{\theta_t^X})) \\ &+ \beta_t^X \cdot \nabla_{\theta^X} d(P_{X_t,S_t}||P_{X_{t+1},S_t}), \end{aligned} \tag{10}$$

$$\begin{aligned} \theta_{t+1}^S \leftarrow \theta_t^S &- \beta_t^S \cdot \nabla_{\theta^S} L(T_t(S_t(X)); \theta_t^S) \\ &+ \beta_t^S \cdot \nabla_{\theta^S} d(P_{X_t,S_t}||P_{X_t,S_{t+1}}). \end{aligned} \tag{11}$$

Note that the distribution drifts $d(P_{X_t,S_t}||P_{X_{t+1},S_t})$ and $d(P_{X_t,S_t}||P_{X_t,S_{t+1}})$ does not involve ground truth labels.

Theoretical Convergence Analysis. We further provide a theoretical analysis of our proposed CoM-

MIT framework in Appendix A. Specifically, we establish a new convergence bound under sufficient assumptions for the coordinated learning rate and regularization strategy, demonstrating that the proposed method can achieve a provably faster convergence rate compared to standard imbalanced instruction tuning approaches. We further provide the detailed proof of the proposed theorems in Appendix B. Theoretical results show that dynamic adjustment of learning rates, guided by the multimodal balance coefficient, consistently accelerates optimization and ensures sufficient learning across modalities. We also discuss how results can be generalized to various gradient-based optimizers.

7 Experiment

Experiment Setup We conduct experiments on two non-text modalities, vision and audio, with multiple instruction-tuning downstream tasks: (1) for **Vision**, we evaluate the backbone MLLMs including BLIP-2 (Li et al., 2023), InternVL2 (Chen et al., 2024), and LLaVA-1.5 (Liu et al., 2023), on three visual question-answering tasks: TextVQA (Singh et al., 2019), IconQA (Lu et al., 2021), and A-OKVQA (Schwenk et al., 2022), which focus on text recognition and reasoning, knowledge-intensive QA, and abstract diagram understanding, respectively; (2) for **Audio**, we leverage the SALMONN (Tang et al., 2023a) model and evaluate one audio question-answering task and two audio captioning tasks: ClothoAQA (Lipping et al., 2022), MACS (Morato and Mesaros, 2021), and SDD (Manco et al., 2023), which focus respectively on crowdsourced audio question-answering, acoustic scene captioning, and text-to-music generation. We include detailed implementation details in Appendix B.4 for individual backbone MLLM.

We follow the instruction tuning diagram (Dai et al., 2024; Tang et al., 2023a; Huang et al., 2023a), where the parameters of backbone LLMs are finetuned with LoRAs (Hu et al., 2021) and the feature encoders are finetuned directly. We set the learning rate to $1e-4$ for all the feature encoders and backbone LLMs in our baseline methods **Constant LR** (Dai et al., 2024; Tang et al., 2023a), **Feature CD**, **Language CD** (Wright, 2015). For Feature CD, we first update the feature encoder until its weights stabilize, then update the backbone LLMs. For Language CD, the process is reversed, with the LLMs being trained first. We also use $1e-4$ as the base learning rates for our CoMMIT variants. There are two CoMMIT variants: **CoMMIT** and **CoMMIT-CLR**. **CoMMIT** is out proposed method in this paper, while **CoMMIT-CLR** is an ablation on **CoMMIT**, without the last regularization terms in (10) and (11).

LLM Usage In this paper, LLMs are only used for refining the writing of natural language.

Mitigating Oscillation and Biased Learning. In Figure 4, we inspect the oscillation problem in Section 5.1 by showing the value of κ_t with our proposed CoMMIT and CoMMIT-CLR, comparing to the three learning rate scheduling methods described in Section 5. We report with the BLIP-2 (Li et al., 2023) backbone model on the task of TextVQA (Singh et al., 2019). It can be observed that both CoMMIT and CoMMIT-CLR can stabilize multimodal learning with smaller standard deviations of κ_t over time, thus alleviating the oscillation problem in Observation 5.1 (in Section 5.1).

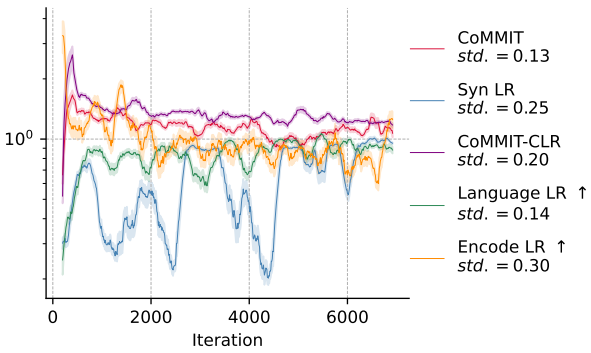

Figure 4: Learning curves of the multimodal learning balance coefficient κ_t for multiple methods. In addition to the learning curve, we also report the standard deviation of κ_t of each method.

Though Language LR ↑ also yields high stability on κ, such learning rate adjustment method suffers from the problems of biased learning as described in Observation 5.2 (in Section 5.2), which potentially causes insufficient training on the feature encoder and worse performance of instruction tuning. Comparing CoMMIT and CoMMIT-CLR, we can observe that CoMMIT achieves less biased learning with the value of κ_t closer to 1 while demonstrating relatively milder learning oscillation with less variant κ_t during training. Further, we find that the proposed loss regularization in Section 6 also improves the training balance. Accompanied by the loss regularization and learning balance coefficient κ_t, CoMMIT more effectively adapts the optimization process for better training between the feature encoder and LLM.

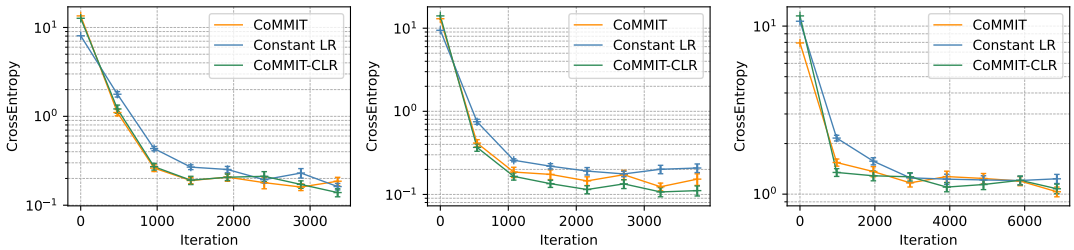

Figure 5: Instruction-tuning learning curves of BLIP-2 on three vision-based downstream tasks.

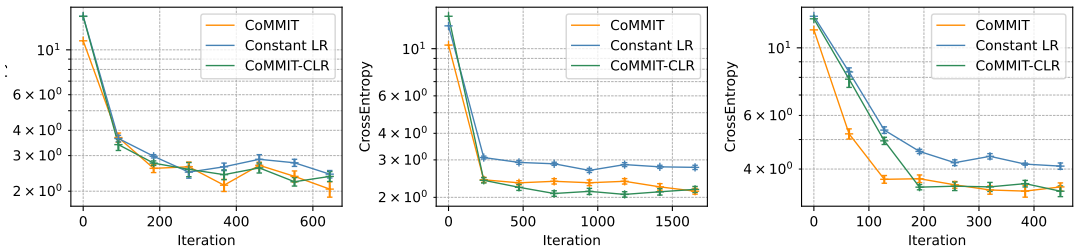

Figure 6: Instruction-tuning learning curves of SALMONN on three audio-based downstream tasks.

Note that SynLR with the oscillated value of κ_t is our baseline Constant LR. In Figure 5 and 6, it is shown that Constant LR generally results in a slower training rate characterized by a flatter descent in loss values during the early stages. This is consistent with Observation 5.1 in Section 5.1, suggesting that the oscillation problem can slow the convergence of MLLMs in instruction tuning.

Improved Convergence in MLLM Instruction Tuning. We evaluate the convergence performance of the proposed methods CoMMIT-CLR and CoMMIT in comparison with Constant LR in Figure 5 and 6. For visual question-answering tasks in Figure 5, we observe that CoMMIT-CLR and CoMMIT can accelerate the instruction tuning of BLIP-2 in the early stage. This is especially evident in the case of IconQA, which is out-of-domain for BLIP-2's pretraining. Specifically, IconQA requires regional-level and spatial visual understanding, which differs from the tasks BLIP-2 was pre-trained on (Chen et al., 2023). In addition, CoMMIT-CLR and CoMMIT achieve lower training losses than Constant LR, validating CoMMIT's improved training efficiency. This efficiency gain is attributed to CoMMIT's ability to mitigate the oscillation (Section 5.1), accelerating convergence.

Similar to the vision-based tasks, we can find in Figure 6 that CoMMIT-CLR and CoMMIT can also converge to lower loss values in audio tasks. Specifically, we observe that the CoMMIT-CLR and CoMMIT can achieve better accelerations on the audio captioning tasks of MACS and SDD, compared to training on the audio question-answering task of ClothoAQA. Since audio captioning tasks need more adaptation in MLLMs to generate relatively longer context and align the generation distribution with specific tasks, the coordinated learning rate scheduling method in Section 6 can more dynamically adjust the learning rate for less learned components at each model update step. In addition, we show that the proposed loss regularization method adopted in CoMMIT can actively promote the difference in MLLM's generation distribution between optimization steps, which can better benefit tasks, such as audio captioning, that require the model to generate longer contexts. Such improved convergence is associated with our proposed balancing strategy that improves on the oscillation and biased learning problem as discussed.

Improved Downstream Performance across Modalities. In Table 1, We evaluate the performance of the proposed methods CoMMIT-CLR and CoMMIT, comparing with three baselines Constant LR, Feature CD, and Language CD. Among the three baselines, we observe that coordinate gradient descend methods have the most improvement compared to the constant learning rate methods that show significant learning tendencies towards a certain modality (*e.g.*, Language CD in SDD, and Feature CD in A-OKVQA and ClothoAQA). However, since such learning balance varies in downstream tasks, coordinate descend methods cannot consis-

Model	Task	Constant LR	Feature CD	Language CD	CoMMIT CLR	CoMMIT
BLIP-2	A-OKVQA	54.06	57.99	49.87	60.44	**64.37**
	IconQA	37.16	35.48	34.47	**39.09**	38.65
	TextVQA	26.48	18.00	19.44	27.66	**28.12**
SALMONN	ClothoAQA	42.49	45.80	38.52	**52.86**	50.55
	MACS	24.60	22.41	23.64	23.81	**25.06**
	SDD	15.10	5.70	**15.74**	15.07	15.33
InternVL2	A-OKVQA	76.59	73.19	79.47	78.00	**80.52**
	IconQA	80.94	**83.20**	81.60	80.85	82.87
	TextVQA	65.22	65.60	65.08	65.18	**67.00**
LLaVA-1.5	A-OKVQA	79.20	77.64	76.94	77.82	**79.55**
	IconQA	64.09	64.16	58.17	65.78	**69.60**
	TextVQA	41.98	43.34	**49.32**	47.80	49.30

Table 1: Instruction tuning results for four MLLMs: BLIP-2, SALMONN, InternVL2-8B, and LLaVA-1.5-7B. These are pre-trained LLMs in vision and audio respectively. For questions-answering tasks like A-OKVQA, IconQA, TextVQA, and ClothoAQA, we report the accuracy score of the generated answers. For audio captioning tasks (MACS and SDD), we report the Rouge-L metric that compares the generated caption with candidate captions. We highlight the best method in bold font for each downstream task of instruction tuning.

Method	A-OKVQA	TextVQA	IconVQA
LR=1e-5	50.30	27.58	35.45
LR=1e-4	54.06	26.48	37.16
LR=1e-3	45.24	20.60	34.93
CoMMIT	**64.37**	**28.12**	**38.65**

Table 2: Comparison on A-OKVQA, TextVQA, and IconVQA with BLIP-2 backbone model. Baselines are Constant LR that direct fine tune the backbone model with various learning rate.

tently improve MLLM instruction tuning, while arbitrarily inclining towards only a certain modality can result in inferior model performance (*e.g.*, Feature CD in SDD and Language CD in A-OKVQA).

Different from the fixed learning tendency which needs to be predetermined by coordinate descend methods, the proposed coordinated learning rate scheduling method can dynamically adapt learning rates for multimodal components and balance the multimodal joint training. With better coordinated multimodal learning, CoMMIT-CLR and CoMMIT consistently improve Constant LR across modalities and downstream tasks. In addition, the proposed regularization in CoMMIT can promote larger step sizes in gradient descent, which enlarges differences in the generated output distributions between different time steps. This prevents learning from being stuck at local optima, which can be especially beneficial for modality-specific captioning tasks whose optimization space can be relatively larger than question-answering tasks.

Comparing various learning rates. In Table 2, we compare CoMMIT with results of constant LR with different learning rates. We report the results on tasks of A-OKVQA, TextVQA, and Icon-VQA with BLIP-2 backbone model. We can observe that our proposed CoMMIT outperforms the Constant LR baselines by a significant margin. In addition, we can also find that no fixed value of learning rate consistently yields the best performance for Constant LR, while our proposed CoMMIT can dynamically adjust its learning rate. These results demonstrate the necessity and effectiveness of dynamic learning rate adjustment for balanced learning between the feature encoder and LLM in multimodal instruction tuning.

8 Conclusion

In this work, we address the challenge of imbalanced learning between the feature encoder and the backbone LMM during MLLM instruction tuning. Through theoretical analysis and empirical observations, we uncovered how this imbalance can lead to the dilemmas of oscillation and biased learning. To mitigate these problems, we proposed **CoMMIT**, a novel approach that dynamically coordinates the learning rates of the feature encoder and LLM backbone. Our **CoMMIT** also included regularization on the gradient gradients that promotes training sufficiency. Our theoretical and empirical analyses demonstrate that **CoMMIT** improves the balance between the training of the feature encoder and the LLM. Experiments across multiple vision and audio downstream instruction tuning tasks illustrate that the training with **CoMMIT** for MLLMs is more effective compared to baselines.

9 Limitation

While our framework demonstrates compatibility with open-source LLMs to ensure reproducibility and controlled experimentation, its integration with API-based models remains unexplored. This focus aligns with common research practices prioritizing transparent benchmarking, though we acknowledge that real-world deployment scenarios may involve complex model ecosystems. Future work could investigate adapter mechanisms to bridge this gap, but such extensions lie beyond our current scope of foundational multimodal learning dynamics.

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*.

Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. 2003. *Convex analysis and optimization*, volume 1. Athena Scientific.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2023. Lion: Empowering multimodal large language model with dual-level visual knowledge. *arXiv preprint arXiv:2311.11860*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. 2020. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. 2023. Llark: A multimodal foundation model for music. *arXiv preprint arXiv:2310.07160*.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, and 1 others. 2023. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2023a. Vtimellm: Empower llm to grasp video moments. *arXiv preprint arXiv:2311.18445*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, and 1 others. 2023b. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. 2023. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747*.

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. *arXiv preprint arXiv:2301.12597*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Tianyi Liu, Zuxuan Wu, Wenhan Xiong, Jingjing Chen, and Yu-Gang Jiang. 2021a. Unified multimodal pre-training and prompt-based tuning for vision-language understanding and generation. *arXiv preprint arXiv:2112.05587*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021b. P-tuning v2: Prompt tuning can be comparable to

fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. 2023. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*.

James Lucas, Shengyang Sun, Richard Zemel, and Roger Grosse. 2018. Aggregated momentum: Stability through passive damping. In *International Conference on Learning Representations*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bodganov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, and 1 others. 2023. The song describer dataset: A corpus of audio captions for music-and-language evaluation. *arXiv preprint arXiv:2311.10057*.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning*, pages 23803–23828. PMLR.

IM Morato and A Mesaros. 2021. Macs-multi-annotator captioned soundscapes.

Yurii Nesterov. 2013. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Vipula Rawte, Anku Rani, Harshad Sharma, Neeraj Anand, Krishnav Rajbangshi, Amit Sheth, and Amitava Das. 2024. Visual hallucination: Definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2403.17306*.

J Reddi Sashank, Kale Satyen, and Kumar Sanjiv. 2018. On the convergence of adam and beyond. In *International conference on learning representations*, volume 5.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

J Michael Steele. 2004. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press.

Hidenori Tanaka and Daniel Kunin. 2021. Noether's learning dynamics: Role of symmetry breaking in neural networks. *Advances in Neural Information Processing Systems*, 34:25646–25660.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023a. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, and 1 others. 2023b. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dongsheng Wang, Miaoge Li, Xinyang Liu, MingSheng Xu, Bo Chen, and Hanwang Zhang. 2024a. Tuning multi-mode token-level prompt alignment across modalities. *Advances in Neural Information Processing Systems*, 36.

Jianing Wang, Junda Wu, Yupeng Hou, Yao Liu, Ming Gao, and Julian McAuley. 2024b. Instructgraph: Boosting large language models via graph-centric instruction tuning and preference alignment. *arXiv preprint arXiv:2402.08785*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and 1 others. 2024c. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Fuchao Wei, Chenglong Bao, and Yang Liu. 2021. Stochastic anderson mixing for nonconvex stochastic optimization. *Advances in Neural Information Processing Systems*, 34:22995–23008.

Stephen J Wright. 2015. Coordinate descent algorithms. *Mathematical programming*, 151(1):3–34.

Junda Wu, Jessica Echterhoff, Kyungtae Han, Amr Abdelraouf, Rohit Gupta, and Julian McAuley. 2025a. Pdb-eval: An evaluation of large multimodal models for description and explanation of personalized driving behavior. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pages 242–248. IEEE.

Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A Rossi, Franck Dernoncourt, and 1 others. 2024a. Personalized multimodal large language models: A survey. *arXiv preprint arXiv:2412.02142*.

Junda Wu, Zachary Novack, Amit Namburi, Jiaheng Dai, Hao-Wen Dong, Zhouhang Xie, Carol Chen, and Julian McAuley. 2024b. Futga: Towards fine-grained music understanding through temporally-enhanced generative augmentation. *arXiv preprint arXiv:2407.20445*.

Junda Wu, Rui Wang, Handong Zhao, Ruiyi Zhang, Chaochao Lu, Shuai Li, and Ricardo Henao. 2023. Few-shot composition learning for image retrieval with prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4729–4737.

Junda Wu, Yu Xia, Tong Yu, Xiang Chen, Sai Sree Harsha, Akash V Maharaj, Ruiyi Zhang, Victor Bursztyn, Sungchul Kim, Ryan A Rossi, and 1 others. 2025b. Doc-react: Multi-page heterogeneous document question-answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 67–78.

Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan A Rossi, Lina Yao, and 1 others. 2025c. Mitigating visual knowledge forgetting in mllm instruction-tuning via modality-decoupled gradient descent. *arXiv preprint arXiv:2502.11740*.

Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. 2024c. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in Neural Information Processing Systems*, 36.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, and 1 others. 2024d. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*.

Kaige Xie, Tong Yu, Haoliang Wang, Junda Wu, Handong Zhao, Ruiyi Zhang, Kanak Mahadik, Ani Nenkova, and Mark Riedl. 2023. Few-shot dialogue summarization via skeleton-assisted prompt transfer in prompt tuning. *arXiv preprint arXiv:2305.12077*.

An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, and 1 others. 2024. List items one by one: A new data source and learning paradigm for multimodal llms. *arXiv preprint arXiv:2404.16375*.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, and 1 others. 2024. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. 2024. Vpgtrans: Transfer visual prompt generator across llms. *Advances in Neural Information Processing Systems*, 36.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Hang Zhao, Yifei Xin, Zhesong Yu, Bilei Zhu, Lu Lu, and Zejun Ma. 2024. Slit: Boosting audio-text pre-training via multi-stage learning and instruction tuning. *arXiv preprint arXiv:2402.07485*.

A Theoretical Analysis

In this section, we present the computation and proof of a new convergence bound with our proposed method CoMMIT. Our theoretical analysis demonstrates that it achieves a faster convergence rate compared to the imbalanced MLLM instruction tuning.

A.1 Setup and Notations

Consider a non-convex random objective function $F : \mathbb{R}^d \to \mathbb{R}$ expressed as the average of K component functions, $F(x) = \frac{1}{K}\sum_{k=1}^{K} f_k(x)$, where each $f_k(x)$ is an i.i.d sample. Our goal is to minimize $\mathbb{E}\left[F(x)\right]$ over $x \in \mathbb{R}^d$. We also define $\mathbb{E}_{k-1}$ as the conditional expectation with respect to $f_1, f_2, \cdots, f_k$. Following the notation in Adam (Kingma and Ba, 2014), let m_k, v_k, $x_k \in \mathbb{R}^d$ be vectors at iteration k. The i-th component of each vector is denoted by $m_{k,i}$, $v_{k,i}$, $x_{k,i}$. Building upon the insight of Défossez et al. (Défossez et al., 2020) regarding the presence of two bias correction terms m_k and v_k, we define $\alpha_{k,i} = \alpha_i\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}$. Notably, we opt to drop the correction term for m_k due to its faster convergence compared to v_k.

Aligned with our proposed methodology, we incorporate two additional terms into the original Adam algorithm. To prevent learning from oscillating too heavily toward either S or X, we adapt λ according to the moving average $\tilde{\kappa}$. To mitigate the risk of vanishing or exploding gradients, we introduce $h(x)$ as an extra term in the gradient updates defined in Section 6 to enhance training stability and support the overall robustness of the learning process. We fix $\beta_1 = 0$, $0 < \beta_2 \leq 1$, $\alpha_{k,i} > 0$, $\epsilon = 10^{-8}$, and initialize $m_0 = 0$ and $v_0 = 0$. The updated rules follow,

$$v_{k,i} = \beta_2 v_{k-1,i} + \left(\lambda\nabla_i f_k(x_{k-1}) + \nabla_i h_k(x_{k-1})\right)^2 \tag{12}$$

$$x_{k,i} = x_{k-1,i} - \alpha_k \frac{\lambda\nabla_i f_k(x_{k-1}) + \nabla_i h_k(x_{k-1})}{\sqrt{v_{k,i} + \epsilon}} \tag{13}$$

Throughout the proof, we assume the norm of the gradients $\|\nabla f(x) + \nabla h(x)\|$ is bounded by $R - \sqrt{\epsilon}$. The small constant ϵ is used for numerical stability.

A.2 Necessary Assumptions

We state the necessary assumptions (Bertsekas et al., 2003) commonly used when analyzing the convergence of stochastic algorithms for non-convex problems:

Assumption A.1. *The minimum value of $f(x)$ is lower-bounded,*

$$\forall x \in \mathbb{R}^d,\ f^* = \min f(x).$$

Assumption A.2. *The gradient of the non-convex objective function f is L-Liptchitz continuous (Nesterov, 2013). Then $\forall x, y \in \mathbb{R}^d$, the following inequality holds,*

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|x - y\|_2^2.$$

A.3 Convergence Analysis

Following the second Theorem outlined by Défossez et al. (Défossez et al., 2020), we calculate the convergence bound of our algorithm with a dynamic learning rate and loss function.

Theorem A.3. *Given the assumptions from Appendix A.2 and applying Lemma B.2, for all components of the step sizes and gradients, update α_i with the corresponding value from H_t^S and H_t^T. Let $\{x_k\}$ be a sequence generated by the optimizer, with $0 < \beta_2 \leq 1$, and $\alpha_i > 0$. For any time step K, we have,*

$$\mathbb{E}\left[\|\nabla F(x_k)_2^2\|\right] \leq 2R\frac{F(x_0) - f^*}{\lambda\alpha_i K} + C \tag{14}$$

where

$$C = \frac{1}{K}\left(\frac{2\alpha_i R}{\sqrt{1-\beta_2}} + \frac{\alpha_i^2 L}{2(1-\beta_2)}\right)\ln\left(\frac{(1-\beta_2^k)R^2}{(1-\beta_2)\epsilon\beta_2}\right)$$

The detailed proof is provided in Appendix B.3.

CoMMIT adjusts both λ and $h(x)$ to balance multimodal learning progress. The parameter λ, which measures the balance between feature and language learning, remains greater than 1 during training, driven by the balance metric $\tilde{\kappa}$. By avoiding learning oscillations, λ can grow even larger, contributing to faster learning. When $\tilde{\kappa} > 1$, the model suffers from insufficient feature learning. CoMMIT reduces the learning rate of the LLM to balance learning, ensuring $\lambda = \frac{\tilde{\kappa}_t+1}{\tilde{\kappa}_{t-1}+1} > 1$. Conversely, when $\tilde{\kappa} < 1$, the model suffers from insufficient language learning, ensuring $\lambda = \frac{1/\tilde{\kappa}_t+1}{1/\tilde{\kappa}_{t-1}+1} > 1$. Notably, $\nabla h(x)$ is directly added to $\nabla f(x)$ to induce gradient changes, which further contributes to the increase of λ, resulting in a faster convergence rate.

In this section, we show the proof using Adam as the base optimizer. Due to the reason that CoMMIT does not modify the optimization algorithm itself, the theorem can also be extended to any gradient-based optimization method such as the stochastic gradient descent.

B Proof

B.1 Learning balance in multimodal joint training

Proof. According to Lipschitz continuity in cross-entropy loss function (Mao et al., 2023), there exists a sequence of T_t and S_t during MLLM instruction tuning, where multimodal components are jointly trained. Given the two metric spaces, $(\mathbb{R}, l_2)$ of the cross-entropy losses and (H, d) of the generation distributions, there exists $0 < \gamma < 1$ such that, at each optimization step t,

$$\left\| L\left(P_{X_t,S_t}\right) - L\left(P_{X_{t+1},S_{t+1}}\right) \right\|_2 \leq \gamma d\left(P_{X_{t+1},S_{t+1}} || P_{X_t,S_t}\right), \tag{15}$$

where the metric d measures the change in the prediction distribution $P_{X_t,S_t} \in H$ as the multimodal components X and S are updated. Based on the triangle inequality in metric space, a joint step of multimodal learning is bounded by the combination of two components' separate step forward,

$$d\left(P_{X_{t+1},S_{t+1}} || P_{X_t,S_t}\right) \leq d\left(P_{X_{t+1},S_t} || P_{X_t,S_t}\right) + d\left(P_{X_t,S_{t+1}} || P_{X_t,S_t}\right), \tag{16}$$

where the first term represents the change due to updating the X-component while keeping S fixed, and the second term represents the change due to updating the S-component.

Then we can derive the multimodal gradient estimated bounds based on MLLM's generative performance in its metric space d shown in the Theorem 4.2. □

B.2 Controlling Deviation from Descent Direction

Following the first Lemma outlined by Défossez et al. (Défossez et al., 2020), where the expected update direction can positively correlate with the gradient (Sashank et al., 2018), we aim to control the deviation from the descent direction to enhance convergence.

Lemma B.1. *For all $k \in \mathbb{N}^*$ and $R \geq \|\nabla f(x) + \nabla h(x)\| + \sqrt{\epsilon}$, the gradient update follows a descent direction,*

$$\begin{aligned} &\mathbb{E}_{k-1}\left[\nabla_i F(x_{k-1}) \frac{\lambda \nabla_i f_k(x_{k-1}) + \nabla_i h_k(x_{k-1})}{\sqrt{\epsilon + v_{k,i}}}\right] - \frac{\lambda(\nabla_i F(x_{k-1}))^2}{2\sqrt{\epsilon + \tilde{v}_{k,i}}} \\ &\geq \frac{\nabla_i F(x_{k-1}) \nabla_i h_k(x_{k-1})}{2\sqrt{\epsilon + \tilde{v}_{k,i}}} - 2R\mathbb{E}_{k-1}\left[\frac{(\lambda \nabla_i f_k(x_{k-1}) + \nabla_i h_k(x_{k-1}))^2}{\epsilon + v_{k,i}}\right]. \end{aligned} \tag{17}$$

Proof. Denote $F = \nabla_i F(x_{k-1})$, $f = \lambda \nabla_i f_k(x_{k-1})$, $h = \nabla_i h_k(x_{k-1})$, and $\tilde{v}_{k,i} = \beta_2 v_{k-1,i} + \mathbb{E}_{k-1}\left[(\lambda \nabla_i f_k(x_{k-1}) + \nabla_i h_k(x_{k-1}))^2\right]$, we get:

$$\mathbb{E}_{k-1}\left[\frac{F(f+h)}{\sqrt{\epsilon + v_{k,i}}}\right] = \mathbb{E}_{k-1}\left[\frac{F(f+h)}{\sqrt{\epsilon + \tilde{v}_{k,i}}}\right] + \mathbb{E}_{k-1}\left[F(f+h)\left(\frac{1}{\sqrt{\epsilon + v_{k,i}}} - \frac{1}{\sqrt{\epsilon + \tilde{v}_{k,i}}}\right)\right] \tag{18}$$

We know that g and $\tilde{v}_{k,i}$ are independent given $f_1, f_2, \cdots, f_{n-1}$. h and $\tilde{v}_{k,i}$ are also independent based on our settings which do not affect the momentum, we have,

$$\mathbb{E}_{k-1}\left[\frac{F(f+h)}{\sqrt{\epsilon+\tilde{v}_{k,i}}}\right] = \frac{\lambda F^2}{\sqrt{\epsilon+\tilde{v}_{k,i}}} + \frac{Fh}{\sqrt{\epsilon+\tilde{v}_{k,i}}} \tag{19}$$

The only thing we need to do is control the deviation of the second term in (18). Applying Cauchy-Schwarz (Steele, 2004),

$$\begin{aligned} RHS &= F(f+h)\frac{\mathbb{E}_{k-1}\left[(f+h)^2\right] - (f+h)^2}{\sqrt{\epsilon+v_{k,i}}\sqrt{\epsilon+\tilde{v}_{k,i}}(\sqrt{\epsilon+v_{k,i}}+\sqrt{\epsilon+\tilde{v}_{k,i}})} \\ &\leq F(f+h)\frac{\mathbb{E}_{k-1}\left[(f+h)^2\right]}{\sqrt{\epsilon+v_{k,i}}(\epsilon+\tilde{v}_{k,i})} + F(f+h)\frac{(f+h)^2}{\sqrt{\epsilon+v_{k,i}}(\epsilon+\tilde{v}_{k,i})}. \end{aligned} \tag{20}$$

By applying the inequality $ab \leq \frac{1}{2\lambda}b^2 + \frac{\lambda}{2}a^2$ with $\lambda = \frac{\sqrt{\epsilon+\tilde{v}_{k,i}}}{2}$, a = $\frac{F}{\sqrt{\epsilon+\tilde{v}_{k,i}}}$, and $b = \frac{(f+h)\mathbb{E}_{k-1}\left[(f+h)^2\right]}{\sqrt{\epsilon+\tilde{v}_{k,i}}\sqrt{\epsilon+v_{k,i}}}$, the conditional expectation of the first term in (20) can be bounded as,

$$\begin{aligned} \mathbb{E}_{k-1}\left[F(f+h)\frac{\mathbb{E}_{k-1}\left[(f+h)^2\right]}{\sqrt{\epsilon+v_{k,i}}(\epsilon+\tilde{v}_{k,i})}\right] &\leq \mathbb{E}_{k-1}\left[\frac{F^2}{4\sqrt{\epsilon+\tilde{v}_{k,i}}} + \frac{(f+h)^2\mathbb{E}_{k-1}\left[(f+h)^2\right]^2}{\sqrt{\epsilon+\tilde{v}_{k,i}}^3(\epsilon+v_{k,i})}\right] \\ &\leq \frac{F^2}{4\sqrt{\epsilon+\tilde{v}_{k,i}}} + \mathbb{E}_{k-1}\left[\frac{(f+h)^2\mathbb{E}_{k-1}\left[(f+h)^2\right]}{\sqrt{\epsilon+\tilde{v}_{k,i}}(\epsilon+v_{k,i})}\right] \\ &\leq \frac{F^2}{4\sqrt{\epsilon+\tilde{v}_{k,i}}} + R\mathbb{E}_{k-1}\left[\frac{(f+h)^2}{\epsilon+v_{k,i}}\right], \end{aligned} \tag{21}$$

with respect to the fact that $\epsilon+\tilde{v}_{k,i} \geq \mathbb{E}_{k-1}\left[(f+h)^2\right]$ and $\mathbb{E}_{k-1}\left[(f+h)^2\right] \leq R$.

Similarly, applying the inequality $ab \leq \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2$ with $\lambda = \frac{\sqrt{\epsilon+\tilde{v}_{k,i}}}{2\mathbb{E}_{k-1}[(f+h)^2]}$, $a = \frac{F(f+h)}{\sqrt{\epsilon+\tilde{v}_{k,i}}}$, and $b = \frac{(f+h)^2}{\epsilon+v_{k,i}}$, the conditional expectation of the second term in (20) can be bounded as,

$$\begin{aligned} \mathbb{E}_{k-1}\left[F\frac{(f+h)^2(f+h)}{\sqrt{\epsilon+v_{k,i}}(\epsilon+\tilde{v}_{k,i})}\right] &\leq \mathbb{E}_{k-1}\left[\frac{F^2}{4\sqrt{\epsilon+\tilde{v}_{k,i}}}\frac{(f+h)^2}{\mathbb{E}_{k-1}\left[(f+h)^2\right]} + \frac{\mathbb{E}_{k-1}\left[(f+h)^2\right]}{\sqrt{\epsilon+\tilde{v}_{k,i}}}\frac{(f+h)^4}{(\epsilon+v_{k,i})^2}\right] \\ &\leq \frac{F^2}{4\sqrt{\epsilon+\tilde{v}_{k,i}}} + \mathbb{E}_{k-1}\left[\frac{\mathbb{E}_{k-1}\left[(f+h)^2\right]}{\sqrt{\epsilon+\tilde{v}_{k,i}}}\frac{(f+h)^2}{(\epsilon+v_{k,i})}\right] \\ &\leq \frac{F^2}{4\sqrt{\epsilon+\tilde{v}_{k,i}}} + R\mathbb{E}_{k-1}\left[\frac{(f+h)^2}{\epsilon+v_{k,i}}\right], \end{aligned} \tag{22}$$

given again $\mathbb{E}_{k-1}\left[(f+h)^2\right] \leq R$.

Putting inequalities (21) and (22) back into (20) gives,

$$\mathbb{E}_{k-1}\left[F(f+h)\left(\frac{1}{\sqrt{\epsilon+v_{k,i}}} - \frac{1}{\sqrt{\epsilon+\tilde{v}_{k,i}}}\right)\right] \leq \frac{F^2}{2\sqrt{\epsilon+\tilde{v}_{k,i}}} + 2R\mathbb{E}_{k-1}\left[\frac{(f+h)^2}{\epsilon+v_{k,i}}\right] \tag{23}$$

And, therefore, adding (23) and (19) into (18) finishes the proof. □

B.3 Proof of Convergence

In this section, we prove the theorem A.3.

Proof. Given $\alpha_k = \alpha\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}$, we apply the Assumption A.2 and get,

$$F(x_k) \leq F(x_{k-1}) - \alpha_k \nabla F(x_{k-1})^T u_k + \frac{\alpha_k^2 L}{2}\|u_k\|_2^2. \tag{24}$$

Since we define the bound $R \geq \|\nabla f(x) + \nabla h(x)\| + \sqrt{\epsilon}$, it follows that $\sqrt{\epsilon + \tilde{v}_{k,i}} \leq R\sqrt{\sum_{j=0}^{n-1} \beta_2^j}$. By applying this inequality, we obtain,

$$\alpha_k \left(\frac{(\lambda \nabla_i F(x_{k-1}))^2}{2\sqrt{\epsilon + \tilde{v}_{k,i}}} + \frac{\nabla_i F(x_{k-1}) \nabla_i h_k(x_{k-1})}{2\sqrt{\epsilon + \tilde{v}_{k,i}}} \right) \geq \alpha \left(\frac{(\lambda \nabla_i F(x_{k-1}))^2}{2R} + \frac{\nabla_i F(x_{k-1}) \nabla_i h_k(x_{k-1})}{2R} \right). \tag{25}$$

By taking the conditional expectation, we apply (25) to Lemma B.1 to derive results from (24),

$$\mathbb{E}_{k-1}\left[F(x_K)\right] \leq \mathbb{E}_{k-1}\left[F(x_{k-1})\right] - \frac{\alpha\lambda}{2R}\|\nabla F(x_k)_2^2\| - \frac{\alpha}{2R}(\nabla F(x_k)^T \nabla h(x_k)) + \left(2\alpha_k R + \frac{\alpha_k^2 L}{2}\right) \mathbb{E}\left[\|u_k\|_2^2\right] \tag{26}$$

Summing the previous inequality over all k and taking the full expectation with respect to the fact that $\alpha \geq \alpha_k \sqrt{1-\beta_2}$. By applying Lemma 5.2 from Défossez et al. (Défossez et al., 2020), we get the final bound,

$$\mathbb{E}\left[\|\nabla F(x_k)_2^2\|\right] \leq 2R\frac{F(x_0) - f^*}{\alpha(1+\lambda)K} + \left(\frac{2\alpha R}{\sqrt{1-\beta_2}} + \frac{\alpha^2 L}{2(1-\beta_2)}\right)\left(\frac{1}{K}\ln\left(\frac{(1-\beta_2^n)R^2}{(1-\beta_2)\epsilon}\right) - ln(\beta_2)\right) \tag{27}$$

□

B.4 Implementation Details

We include specific implementation details for each backbone MLLM we used:

- **BLIP-2** consists of a vision Q-Former and a backbone OPT-2.7B LLM (Zhang et al., 2022). We fine-tune the Q-Former module as the feature encoder and freeze the visual encoder. We apply LoRA to the BLIP-2 model using a rank of 16, an alpha of 32, and a dropout rate of 0.05.
- **InternVL2** consists of the InternViT model as the visual encoder and a backbone InternLM-7B-Chat (Chen et al., 2024). We fine-tune the cross-attention layer as the feature encoder, which encodes the encoded visual representations into multimodal tokens, and freeze the visual encoder. We apply LoRA to the BLIP-2 model using a rank of 16, an alpha of 32, and a dropout rate of 0.05.
- **LLaVA-1.5** consists of the CLIP-L-336px as the visual encoder and a backbone Vicuna-7B (Chiang et al., 2023). We fine-tune the projection layer as the feature encoder, which is a linear layer connecting the encoded image and LLM's word embedding space, and freeze the visual encoder. We apply LoRA to the BLIP-2 model using a rank of 16, an alpha of 32, and a dropout rate of 0.05.
- **SALMONN** extracts both speech and audio features from waveforms and composes these low-level features by a learnable audio Q-Former structure as the feature encoder. The audio tokens generated by the audio Q-Former are prefixed to the language instruction tokens, which are then input to the backbone Vicuna-7B LLM (Chiang et al., 2023).

Computation Resources Our model is trained on 4 A100 GPUs with 40GB memory. The average training time is about 8 hours.