

Harmonious Parameter Adaptation in Continual Visual Instruction Tuning for Safety-Aligned MLLMs

Ziqi Wang¹, Chang Che¹, Qi Wang², Hui Ma^{1*}, Zenglin Shi^{1*}, Cees G. M. Snoek³, Meng Wang¹
¹Hefei University of Technology ²Tsinghua University ³University of Amsterdam

Abstract

While continual visual instruction tuning (CVIT) has shown promise in adapting multimodal large language models (MLLMs), existing studies predominantly focus on models without safety alignment. This critical oversight ignores the fact that real-world MLLMs inherently require such mechanisms to mitigate potential risks. In this work, we shift our focus to CVIT for safety-aligned MLLMs and observe that during continual adaptation, the model not only suffers from task forgetting but also exhibits degradation in its safety. Achieving a harmonious balance between safety and task performance remains a crucial challenge. To address this, we propose *Harmonious Parameter Adaptation (HPA)*, a post-training framework composed of focusing-based parameter partition, harmoniously balanced parameter selection, and orthogonal parameter adjustment. Specifically, HPA partitions parameters into two types based on their focus on safety or task performance, and selects the focused ones to preserve from a balanced perspective. In addition, HPA imposes orthogonality constraints on parameter updates to further alleviate catastrophic forgetting. Extensive experiments on the CVIT benchmark and safety evaluation datasets demonstrate that HPA better maintains high safety and mitigates forgetting than existing baselines. Code is available at <https://github.com/Minato-Zackie/HPA>.

1. Introduction

Multimodal large language models (MLLMs) [17, 28], built upon large language models (LLMs) [21, 26], have recently attracted significant attention for their strong vision-language reasoning capabilities. Their impressive performance stems from a two-stage training paradigm [17, 38]: a pre-training stage that equips the model with broad cross-modal knowledge, followed by visual instruction tuning that enables the model to tackle diverse vision-centric tasks by framing them as language instructions. Recently, continual visual instruction tuning (CVIT) [3, 29] has been pro-

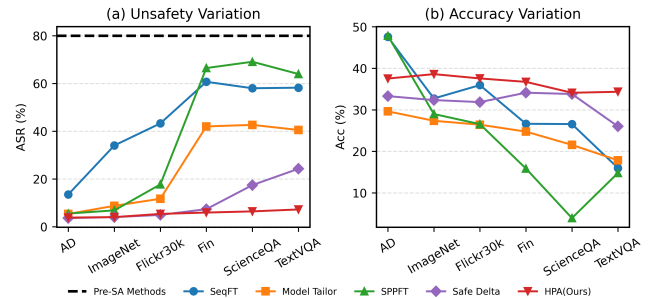


Figure 1. Unsafty Variation (on the safety benchmark) and Accuracy Variation (on the first tuning dataset AD) under Post-SA CVIT for different methods. The black dashed line indicates that the MLLM remains unsafe throughout Pre-SA CVIT due to missing safety alignment.

posed to sustain strong capability on downstream tasks by incrementally adapting the model to new task environments through sequential visual instruction tuning.

Despite the performance gains brought by CVIT, its implications for model safety have received limited attention. Current MLLMs remain vulnerable to harmful vision and language inputs [31], raising critical safety concerns. To mitigate these risks, safety alignment [32, 40] has been introduced as an additional stage after pre-training and instruction tuning. Through techniques such as supervised fine-tuning [6, 40] or preference optimization [30, 34], safety alignment ensures that MLLMs produce safe and reliable outputs even when exposed to unsafe visual or textual content. Nevertheless, existing works on CVIT [10, 39] primarily focus on MLLMs before the safety alignment stage (pre-SA CVIT), while the effect of continual adaptation after safety alignment (post-SA CVIT) remains unexplored. In real-world deployment, continual updating after safety alignment is inevitable, making the concern for safety during CVIT both practical and urgent.

In this work, we pioneer the investigation of post-SA CVIT, examining how continual adaptation impacts safety-aligned MLLMs. Pre-SA CVIT is inherently limited to addressing task forgetting, entirely neglecting the dimension of safety. Suffering from the absence of safety alignment, the MLLM perpetuates its vulnerability across the contin-

*Corresponding authors: {huima, zenglin.shi}@hfut.edu.cn

ual adaptation steps in the pre-SA CVIT setting, as illustrated by the black dashed line in Figure 1 (a). In contrast, the post-SA CVIT setting introduces a dual and more formidable challenge: as shown in Figure 1, MLLMs not only experience catastrophic forgetting but also exhibit a continual degradation of pre-established safety.

Due to the coexistence of these challenges, developing an effective solution for post-SA CVIT is essential. Existing CVIT methods [10, 39] and safety alignment approaches [16, 18] struggle to achieve a harmonious balance between safety and capability during continual fine-tuning. To address this problem, we propose Harmonious Parameter Adaptation (HPA) in CVIT for MLLMs, an efficient post-training framework designed for the adaptation of safety alignment and capability learning.

HPA consists of focusing-based parameter partition, harmoniously balanced selection, and orthogonal parameter adjustment. To maintain the efficiency of MLLMs during CVIT, HPA updates model parameters only after training, without interfering with the existing training pipeline. At each fine-tuning step, HPA first partitions parameters into safety-focused and task-focused types based on their sensitivity differences before and after tuning, which enables a more effective selection of parameters that are crucial for either safety or capability. Subsequently, HPA achieves balanced parameter selection from both intra-layer and inter-layer perspectives, preventing one-sided parameter retention from disrupting the safety–capability trade-off. Given the above components lack a dedicated mechanism to alleviate catastrophic forgetting, HPA further employs orthogonal parameter adjustment to ensure that adaptations for the current task minimally interfere with representations learned from previous tasks. Through carefully balanced parameter adaptation at each stage, HPA harmonizes the trade-off between capability learning and safety guarantee. Our contributions are summarized as follows:

- We shift focus to post-SA CVIT, identifying that safety-aligned MLLMs suffer from task forgetting and safety degradation during CVIT.
- We propose HPA for post-SA CVIT, a new post-training framework that selectively updates parameters via focusing-based partition, balanced selection, and orthogonal adjustment to achieve a harmonious balance between safety and task performance.
- Extensive experiments on CVIT and safety benchmarks establish our method’s superiority, achieving a better trade-off between robust safety assurance and preserved task performance than the compared baselines.

2. Related Work

CVIT for MLLMs. CVIT enhances MLLMs by sequentially fine-tuning them on a series of vision instruction datasets, enabling adaptation to evolving task environ-

ments. A number of benchmarks [3, 29, 35] and methods [2, 10, 33, 36] have been introduced to study this field. Early work such as Eproj [10] first combined visual instruction tuning with continual learning and proposed a structure-expansion strategy to mitigate catastrophic forgetting. Fwd-Prompt [36] addressed negative forward transfer through a prompt-based intervention. CoIN [3] further established a comprehensive CVIT benchmark to evaluate MLLMs under sequential instruction tuning.

More recent approaches typically rely on additional parameter modules and modifications to the original training process. MoeLoRA [3] integrates a token-wise mixture-of-experts mechanism to selectively activate LoRA experts. SMOeLoRA [29] applies separable routing to alleviate dual forgetting. BranchLoRA [33] introduces an asymmetric framework to capture both task-invariant and task-specific factors. SEFE [4] tackles superficial forgetting via answer-style diversification and essential forgetting through constraining masked parameters to maintain key knowledge across tasks. LiLoRA [2] expands lightweight task-specific structures for each new task. However, these methods increase parameter redundancy and training overhead, and they do not account for the safety alignment requirements of MLLMs. In contrast, our work focuses on addressing both task forgetting and safety degradation during CVIT without altering the original training pipeline, aiming to maintain efficiency while preserving safety.

Safety Alignment for MLLMs. Safety alignment [13, 37] aims to ensure that model behavior remains consistent with human values, ethical norms, and intended usage constraints. With the rapid development of MLLMs, efforts toward safety alignment in multimodal settings have also grown. VLGuard [40] introduces a vision-language safety instruction-following dataset to align MLLMs toward safer responses. SPA-VL [34] automatically constructs safety preference data to enable multimodal preference alignment. ADPO [30] incorporates adversarial training into preference optimization to strengthen safety under worst-case adversarial perturbations. SEA [20] synthesizes learnable embeddings for additional modalities, making it possible to conduct multimodal safety alignment even when only textual safety data is available. However, repeatedly performing safety alignment during CVIT is constrained by privacy concerns and prohibitive computational costs. In this work, we explore a more efficient post-training strategy to maintain safety while preserving downstream task performance during continual adaptation.

3. Method

3.1. Problem Formulation

Given a MLLM $f(x; \theta)$, CVIT seeks to incrementally fine-tune f to acquire a sequence of vision tasks $\mathcal{T} =$

$\{\tau_1, \tau_2, \dots, \tau_n\}$. At each time step t , the model is exposed to a new task τ_t together with its associated dataset $\mathcal{D}_t$, in order to integrate the new knowledge into its existing capabilities. Typically, the dataset $\mathcal{D}_t$ follows the format $\{X^{ins}, X^{vis}, X^{ans}\}$, where X^{ins} , X^{vis} , and X^{ans} denote the textual instruction input, the visual input, and the linguistic answer, respectively. For pre-SA CVIT, the final optimization objective can be formulated as:

$$\min_{\{\theta_t\}_{t=1}^n} \sum_{t=1}^n \left(\sum_{i=1}^t \mathcal{L}_{C_i}(f(x; \theta_t)) \right), \quad (1)$$

where $\mathcal{L}_{C_i}$ denotes the performance loss for the i -th task τ_i . This objective focuses solely on the performance changes of the current task τ_t and the previous $t - 1$ tasks.

Now consider a safety-aligned MLLM $f(x; \theta_0)$. After the first task-specific tuning, the model becomes $f(x; \theta_1)$. Due to privacy concerns and high resource consumption, it is infeasible to repeatedly re-align the model with the original safety dataset during CVIT. Under this constraint, the key challenge for post-SA CVIT is to prioritize preserving safety while minimizing the impact on both current-task learning and previously learned tasks.

Formally, during post-SA CVIT at each step t , the optimization objective can be expressed as:

$$\min_{\{\theta_t\}_{t=1}^n} \sum_{t=1}^n \left(\mathcal{L}_S(f(x; \theta_t)) + \sum_{i=1}^t \mathcal{L}_{C_i}(f(x; \theta_t)) \right), \quad (2)$$

where $\mathcal{L}_S$ denotes the safety loss.

3.2. Overview of HPA Framework

Assume that at time step t , we have the fine-tuned model $f(x; \theta_t)$ and the pre-tuned model $f(x; \theta_{t-1})$. Our goal is to design a post-training parameter adaptation scheme that achieves an optimal balance between safety and task performance. Specifically, the final adapted parameter at the l -th layer is defined as

$$\hat{W}_t^l = \mathcal{F}(W_{t-1}^l, W_t^l), \quad (3)$$

where $W_{t-1}^l, W_t^l \in \mathbb{R}^{r \times c}$ denote the weights of the l -th layer in $f(x; \theta_{t-1})$ and $f(x; \theta_t)$, respectively.

To achieve this goal, our framework is built upon three key components: (1) **Focusing-based parameter partition**, which partitions the parameters in W_{t-1}^l and W_t^l based on their distinct focuses on safety preservation and task adaptation (Sec 3.3); (2) **Harmoniously balanced parameter selection**, which harmonizes old and new parameters via layer-wise safety-focused selection and dynamically adjusted retention across layers (Sec 3.4); and (3) **Orthogonal parameter adjustment**, which constrains parameter updates through orthogonal optimization to further mitigate the problem of catastrophic forgetting (Sec 3.5). The overall framework of HPA is shown in Figure 2.

3.3. Focusing-Based Parameter Partition

Due to the inherent over-parameterization of deep neural networks, not all parameters contribute equally to fitting the target distribution [12]. In other words, certain parameters are more essential in knowledge representation or safety preservation, while others remain relatively redundant. In this work, parameters are considered focused when they strongly contribute to either safety or task performance. Accordingly, we partition two types of focused parameters: (1) **Safety-focused parameters**, defined as the top- $k\%$ parameters in W_{t-1}^l that contribute most to safety preservation; and (2) **Task-focused parameters**, defined as the top- $k\%$ parameters in W_t^l that are most important for the current task.

Each parameter's focus is primarily determined by its relative importance in a particular scenario. Previous methods [9, 12] that estimate parameter importance through magnitude or gradient variations are often too coarse-grained to accurately capture this focusing behavior. Inspired by Hessian-based pruning methods [7, 8], we estimate the importance of each parameter $w_{i,j}$ by the loss increase incurred upon its removal, given by $(w_{i,j})^2 / [H^{-1}]_{jj}$, where H denotes the Hessian of the loss with respect to the model parameters, H^{-1} is its inverse, and $[H^{-1}]_{jj}$ represents the j -th diagonal element of H^{-1} .

This metric captures the sensitivity of model performance to the removal of each parameter. During fine-tuning, we can analogously evaluate a parameter's importance based on its variation before and after tuning. Specifically, for each entry (i, j) of W_{t-1}^l and W_t^l , we define two importance scores in layer l : the safety-focus score ε and task-focus score ζ :

$$\begin{aligned} \varepsilon_{i,j}^l &= \frac{(W_{t-1}^l(i, j) - W_t^l(i, j))^2}{[H_{s,l}^{-1}]_{ii}}, \\ \zeta_{i,j}^l &= \frac{(W_t^l(i, j) - W_{t-1}^l(i, j))^2}{[H_{t,l}^{-1}]_{ii}}, \end{aligned} \quad (4)$$

where $H = 2X^\top X$, and $X \in \mathbb{R}^{d \times r}$ is the activation matrix computed from the corresponding safety or task calibration dataset $\mathcal{D}_s^*$ and $\mathcal{D}_t^*$.

After obtaining $\varepsilon_{i,j}^l$ and $\zeta_{i,j}^l$, we aggregate the importance scores column-wise by averaging across r rows:

$$\bar{\varepsilon}_j^l = \frac{1}{r} \sum_{i=1}^r \varepsilon_{i,j}^l, \quad \bar{\zeta}_j^l = \frac{1}{r} \sum_{i=1}^r \zeta_{i,j}^l, \quad (5)$$

where the resulting vectors $\bar{\varepsilon}^l, \bar{\zeta}^l \in \mathbb{R}^c$ indicate the column-wise focusing strength in layer l . By selecting the top- $k\%$ indices from $\bar{\varepsilon}^l$, we can identify the safety-focused and task-focused parameters in the parameter space.

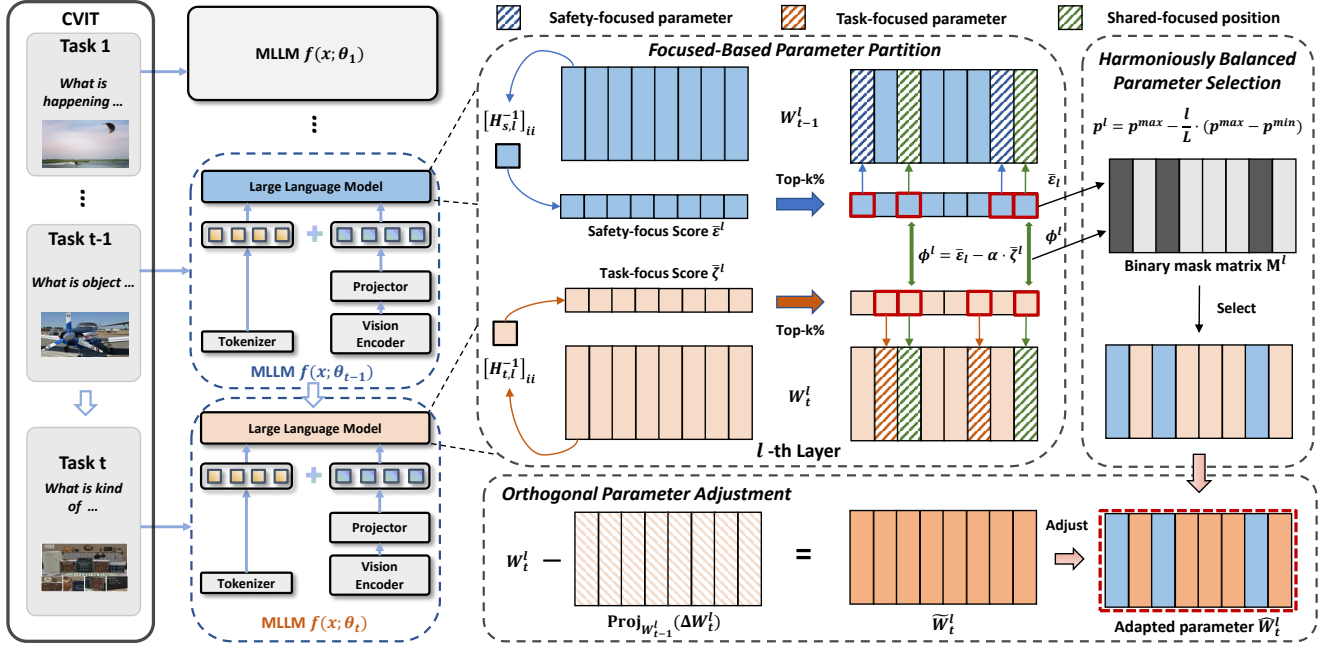


Figure 2. Overview of HPA Framework. HPA consists of three components: (1) **Focusing-based parameter partition** which partitions parameters into safety- and task-focused types; (2) **Harmoniously balanced parameter selection** which balances the selection of focused parameters; (3) **Orthogonal parameter adjustment** which applies orthogonal constraints during parameter updates to mitigate forgetting.

However, a fundamental challenge arises from the overlap phenomenon between the two focused parameter sets. Specifically, for a given position (i, j) in $\hat{W}_t^l$, if the corresponding column in W_{t-1}^l belongs to the safety-focused set while the same position in W_t^l belongs to the task-specific set, these positions are defined as the **shared-focused positions**, which jointly influence both safety and task performance. Therefore, in addition to preserving safety-focused parameters, the selection process must also decide whether to retain W_{t-1}^l or W_t^l parameters at the shared-focused positions.

3.4. Harmoniously Balanced Parameter Selection

During the CVIT process, after each task-specific fine-tuning step, we aim for the updated model $f(x; \theta_t)$ to retain those parameters from the previous model $f(x; \theta_{t-1})$ that are most critical for maintaining safety, while minimally affecting parameters essential to the current task. Thus, the primary goal of parameter adaptation is to determine how to select parameters from W_{t-1}^l and W_t^l to achieve an optimal balance between model safety and performance on the current task. Let $M^l \in \mathbb{R}^{r \times c}$ be a binary mask matrix, the parameter selection process can be formulated as:

$$\hat{W}_t^l = M^l \odot W_{t-1}^l + (1 - M^l) \odot W_t^l, \quad (6)$$

where $p\%$ of the columns in M^l are set to 1, indicating the parameters retained from W_{t-1}^l . In practice, p is typically

set below the focused-parameter ratio k , ensuring selected parameters form a tighter subset of the initial focused set.

For the l -th layer, the most straightforward strategy is to directly retain $p\%$ of the safety-focused parameters. However, such a naive selection may discard task-focused parameters located in shared-focused positions, thereby interfering with the current task performance. To avoid this interference, we first retain from W_{t-1}^l the safety-focused parameters that do not lie in the shared-focused positions, accounting for $p_s\%$ of all parameters. Considering that the number of safety-focused parameters may not reach the desired proportion p , we further select a portion of parameters located at the shared-focused positions from W_{t-1}^l to be preserved.

To determine the relative contribution of these parameters to safety and task performance, we define a balancing score $\phi^l \in \mathbb{R}^{\lfloor c \cdot (k\% - p_s\%) \rfloor}$ for shared-focused positions as

$$\phi^l = \bar{\epsilon}^l - \alpha \cdot \bar{\zeta}^l, \quad (7)$$

where α adaptively balances the contributions of safety and task importance during shared-focused parameter selection, is defined as:

$$\alpha = \alpha_1 - \frac{1}{2}(\alpha_1 - \alpha_0) \left[\tanh(\mathbb{E}[\log \frac{\bar{\epsilon}^l}{\bar{\zeta}^l}]) + 1 \right], \quad (8)$$

where α_0 and α_1 denote the lower and upper bounds of α , respectively; $\mathbb{E}[\cdot]$ represents the expectation operator; and $\tanh(\cdot)$ denotes the hyperbolic tangent function. A larger

ϕ^l indicates a stronger safety-focus tendency. Based on ϕ^l , we select the top- $(p - p_s)\%$ indices from all c columns as the remaining safety parameters from W_{t-1}^l .

During parameter selection, the retention ratio p is also determined to control how many parameters of W_{t-1}^l are preserved. Increasing p improves safety retention but may interfere more with current task adaptation. Given the inherent heterogeneity across layers, our method adaptively adjusts p according to the layer index. During fine-tuning of large models, higher layers closer to the output typically encode more task-specific knowledge [14]; hence we assign a smaller retention ratio to deeper layers to favor current-task adaptation over strict preservation of prior safety constraints. Formally, p^l decreases linearly with layer depth:

$$p^l = p^{max} - \frac{l}{L} \cdot (p^{max} - p^{min}), \quad (9)$$

where p^{max} and p^{min} denote the parameter retention ratios for the first and last layers, respectively, and L is the total number of layers.

3.5. Orthogonal Parameter Adjustment

The harmoniously balanced parameter selection effectively balances model safety and task performance at step t by selectively retaining focused parameters from W_{t-1}^l . However, for the parameters retained from W_t^l , the update process should avoid interfering with previously learned representations. Therefore, we enforce an approximate orthogonality between the new update directions and the subspace spanned by past parameters, effectively reducing catastrophic forgetting during continual adaptation. At step t , the parameter update in layer l is defined as

$$\Delta W_t^l = W_t^l - W_{t-1}^l. \quad (10)$$

To reduce interference with past knowledge, we perform an orthogonal parameter update, ensuring that the new update direction remains as independent as possible from the existing parameter space. Specifically, we first compute the projection of the update ΔW_t^l onto the previous parameters W_{t-1}^l :

$$\text{Proj}_{W_{t-1}^l}(\Delta W_t^l) = \frac{\langle \Delta W_t^l, W_{t-1}^l \rangle}{\|W_{t-1}^l\|_F^2} W_{t-1}^l, \quad (11)$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product and $\|\cdot\|_F$ the Frobenius norm. $\text{Proj}_{W_{t-1}^l}(\Delta W_t^l)$ represents the projection of the current update ΔW_t^l onto the subspace spanned by the previous parameters W_{t-1}^l . This term quantifies the component of the new update that aligns with previously learned knowledge. By subtracting this projection, we remove the part of the update that interferes with old task representations, ensuring that the resulting update is orthogonal

Algorithm 1 Implementation of HPA

Input: Tuning datasets $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n\}$; safety-aligned MLLM $f(x; \theta_0)$; calibration set $\mathcal{D}^* = \{\mathcal{D}_s^*, \mathcal{D}_1^*, \mathcal{D}_2^*, \dots, \mathcal{D}_n^*\}$

Output: Adapted MLLM $f(x; \theta_n)$

```

1: for  $t = 1$  to  $n$  do
2:   Train  $f(x; \theta_t)$  on  $\mathcal{D}_t$  initialized from  $f(x; \theta_{t-1})$ 
3:   for each layer  $l$  of  $f$  do
4:     Extract weights  $W_{t-1}^l$  and  $W_t^l$ 
5:     Compute importance scores  $\bar{\varepsilon}^l, \bar{\zeta}^l$  using Eq. 4 and Eq. 5 with  $\mathcal{D}_s^*$  and  $\mathcal{D}_t^*$ 
6:     Derive  $\phi^l$  and  $p^l$  via Eq. 7 and Eq. 9
7:     Select parameters using  $\bar{\varepsilon}^l, \phi^l, p^l$  to form the binary mask matrix  $M^l$ 
8:     Compute orthogonal update  $\tilde{W}_t^l$  using Eq. 12
9:     Obtain adapted weights  $\hat{W}_t^l$  via Eq. 13
10:    Update  $\theta_t$  by replacing  $W_t^l$  with  $\hat{W}_t^l$ 
11:   end for
12: end for
13: return  $f(x; \theta_n)$ 

```

to W_{t-1}^l :

$$\tilde{W}_t^l = W_{t-1}^l + \Delta W_t^l - \text{Proj}_{W_{t-1}^l}(\Delta W_t^l). \quad (12)$$

The final parameter adaptation can be formulated as:

$$\hat{W}_t^l = M^l \odot W_{t-1}^l + (1 - M^l) \odot \tilde{W}_t^l. \quad (13)$$

The overall implementation procedure for HPA under CVIT is summarized in Algorithm 1.

4. Experiments

4.1. Benchmark Setup

In this section, we outline the benchmark setup used in our study. To evaluate how safety-aligned MLLMs evolve during the CVIT process, we construct a continual training and testing pipeline with six visual instruction tuning datasets and measure safety performance using two established benchmarks.

CVIT Datasets. We employ a diverse set of datasets covering VQA, image classification, and visual reasoning tasks: AD [24], ImageNet [5], Flickr30k [22], Fin [35], ScienceQA [19], and TextVQA [25]. This variety ensures broad task coverage in continual adaptation.

Safety Datasets. Safety is assessed with VLGard [40] and Ch3EF [23], two widely used benchmarks for multi-modal alignment. For VLGard, we focus on two subsets: VLGard-safe-unsafe (VLG-1) and VLGard-unsafe (VLG-2) to capture distinct types of safety risks.

Metrics. Let $a_{k,j}$ denote the accuracy on the j -th task after fine-tuning on the k -th task. We evaluate model per-

formance using **Average Performance (AP)** and **Backward Transfer (BWT)** [27], which respectively measure the overall accuracy after training on task k and the degree of forgetting on previous tasks:

$$\text{AP}_k = \frac{1}{k} \sum_{j=1}^k a_{k,j}, \text{BWT}_k = \frac{1}{k-1} \sum_{j=1}^{k-1} (a_{k,j} - a_{j,j}). \quad (14)$$

For safety evaluation, **Attack Success Rate (ASR)** [23] is defined as the proportion of harmful inputs for which the model fails to provide a safe response. A lower ASR indicates stronger safety. Based on this metric, we introduce two measures: **MASR** and **DASR**. After fine-tuning on the k -th task, MASR represents the average ASR across three safety datasets, while DASR quantifies the increase in MASR relative to the initial safety-aligned model.

4.2. Experimental Setup

Comparison Methods. We compare our approach against a broad range of baselines to highlight its superior performance. Zero-shot denotes the performance of the safety-aligned model without any task-specific fine-tuning, while DirFT refers to directly fine-tuning the model on the corresponding visual instruction dataset. SeqFT represents sequential fine-tuning across CVIT datasets. The remaining methods fall into three categories: (1) classical continual learning methods like EWC [15] and Replay [1]; (2) recent CVIT approaches, including Model Tailor [39] and SEFE [4]; and (3) safety-protected techniques for LLMs, such as SPFT [16] and Safe Delta [18].

Calibration Sets Construction. Post-training methods typically require a small calibration set for parameter refinement. We denote the calibration set as $\mathcal{D}^* = \{X^{ins}, X^{vis}, X^{ans}\}$, where each triplet consists of an instruction, a visual input, and a corresponding answer. For each downstream task t , the task-specific calibration set $\mathcal{D}_t^*$ is constructed by randomly sampling a small number of instances from the corresponding fine-tuning dataset.

For safety alignment, direct access to the original alignment data is often restricted due to privacy concerns, and only the aligned model $f(x; \theta_0)$ is available. To circumvent this limitation, we construct a safety calibration set by exploiting the inherent safety behavior of $f(x; \theta_0)$. Following the four major safety risk categories [40]: Privacy, Risky Behavior, Deception, and Hateful Speech, we collect a small set of harmful visual samples X_{unsafe}^{vis} and pair them with corresponding unsafe instructions X_{unsafe}^{ins} . Since $f(x; \theta_0)$ is safety-aligned, it produces safe responses X_{safe}^{ans} , forming $\mathcal{D}_s^* = \{X_{unsafe}^{ins}, X_{unsafe}^{vis}, X_{safe}^{ans}\}$.

Implementation Details. We use the pretrained first-stage LLaVA-v1.5-7B [17] as our base model to avoid potential data leakage issues in CVIT. During the safety alignment

phase, we align the model using the VLGard and SPaVL [34] datasets. In the CVIT phase, LoRA [11] is chosen as the fine-tuning method and is merged into the base model after fine-tuning. The hyperparameter settings for our method are as follows: the number of samples in the calibration sets $\mathcal{D}_s^*$ and $\mathcal{D}_t^*$ are 8 and 128, respectively; the values of α_0 , α_1 , p_{min} , and p_{max} are set to 0.4, 0.8, 5, and 15, respectively; and $k = 2p^l$. Our method is applied to all linear layers of the base MLLM.

4.3. Main Results

We assess HPA against baseline methods by performing sequential fine-tuning on tasks from AD to TextVQA under two data conditions: the **Original Data** and the datasets with **0.1% Harmful Data Injected**. Upon completing the final tuning task, we evaluate both the CVIT task performance and the safety performance of the models.

As shown in Table 1, in the **Original Data** condition, HPA consistently outperforms all baseline methods after the final tuning task. Compared to Safe Delta, HPA not only improves CVIT task performance, with +2.41% in AP and +2.04% in BWT, but also enhances safety performance, with a reduction of 0.27% in MASR and 0.26% in DASR. These results demonstrate that HPA maintains high safety performance while effectively mitigating forgetting. In the more challenging **0.1% Harmful Data Injected** condition, HPA also achieves superior performance over all baselines. On CVIT task performance, HPA outperforms Safe Delta by +3.42% in AP and +2.94% in BWT. Notably, Safe Delta’s safety performance degrades significantly under this condition, with attack success rates of 24.26% MASR and 21.40% DASR. In contrast, HPA achieves significantly better safety performance of 7.22% MASR and 4.36% DASR. These results highlight the robustness of HPA in more challenging scenarios.

4.4. Ablation Study

Effect of Key Components in HPA. We conduct ablation studies on the three key components of HPA, as shown in Table 2. Specifically, we evaluate the impact of different components: retaining safety-focused parameters using the ε_l score, preserving parameters at shared-focused positions based on ϕ_l , and the orthogonal parameter adjustment resulting in $\tilde{W}_t^l$. Comparing Exp.2 with Exp.1, directly preserving the top $p\%$ of safety-focused parameters substantially enhances model safety. Exp.3 retains only the safety-focused parameters located at the shared-focused positions, leading to improved task performance but with a noticeable drop in safety. Exp.4 combines the strategies of Exp.2 and Exp.3, achieving better task performance while maintaining safety. Furthermore, Exp.5 introduces the orthogonal parameter adjustment constraint based on Exp.4, which effectively mitigates catastrophic forgetting during CVIT.

Table 1. Performance metrics (%) on the CVIT and safety benchmarks, evaluated after completing the final tuning task. The table displays per-task accuracy on the six CVIT datasets, alongside task performance (AP, BWT) and safety performance (MASR, DASR).

Method	CVIT Task Performance ($\uparrow$)								Safety Performance ($\downarrow$)				
	Dataset						Metric		Dataset			Metric	
	AD	ImageNet	Flickr30k	Fin	ScienceQA	TextVQA	AP	BWT	VLG-1	VLG-2	Ch3EF	MASR	DASR
Zero-shot	18.95	20.63	6.55	0.04	24.50	0.02	11.78	-	0.18	2.04	6.37	2.86	-
DirFT	49.63	93.96	152.60	86.56	83.42	56.60	87.13	-	-	-	-	-	-
Original Data													
SeqFT	19.32	45.19	106.64	85.23	81.13	56.60	65.68	-25.62	76.88	25.34	25.46	42.56	39.70
EWC	19.52	46.08	106.51	84.77	81.70	56.53	65.85	-24.90	76.70	24.89	25.46	42.35	39.49
Replay	21.33	25.64	141.96	85.32	81.41	59.85	69.25	-21.40	60.39	12.90	19.92	31.07	28.21
SEFE	17.92	41.01	127.27	79.71	78.89	54.19	66.49	-20.83	79.75	22.17	25.46	42.46	39.60
Model Tailor	20.32	35.88	145.76	79.04	78.89	52.84	68.79	-10.29	60.22	7.92	16.73	28.29	25.43
SPPFT	16.26	43.86	110.34	78.22	63.97	53.7	61.06	-29.81	77.96	47.96	13.14	46.35	43.49
Safe Delta	33.67	46.36	142.10	83.78	80.35	53.67	73.32	-6.91	1.25	4.56	9.24	5.02	2.15
HPA (Ours)	35.01	60.48	144.80	82.73	78.99	52.36	75.73	-4.87	0.72	4.30	9.24	4.75	1.89
0.1% Harmful Data Injected													
SeqFT	15.98	50.24	111.39	84.07	81.58	56.86	66.69	-24.29	81.54	45.70	47.43	58.22	55.36
EWC	16.04	43.41	109.98	83.23	80.92	56.06	64.94	-25.81	83.15	47.96	47.23	59.45	56.58
Replay	23.48	59.27	129.41	86.46	82.05	56.74	72.90	-17.89	77.24	42.31	46.82	55.46	52.59
SEFE	14.81	41.37	126.72	81.18	78.89	54.17	66.19	-20.59	82.62	27.60	30.18	46.80	43.94
Model Tailor	17.83	42.04	155.64	77.06	78.83	54.03	70.91	-11.79	67.20	15.16	39.19	40.52	37.65
SPPFT	14.80	43.82	105.76	68.96	68.85	53.80	59.33	-31.21	93.91	80.09	17.86	63.95	61.09
Safe Delta	26.01	47.98	146.82	83.80	81.04	53.54	73.20	-6.82	50.90	5.66	16.22	24.26	21.40
HPA (Ours)	34.35	63.49	148.62	82.13	78.76	52.34	76.62	-3.88	1.43	5.66	14.58	7.22	4.36

Table 2. Ablation studies of three key components in HPA.

Exp.	$\bar{\epsilon}^l$	ϕ^l	$\tilde{W}_t^l$	AP $\uparrow$	BWT $\uparrow$	MASR $\downarrow$	DASR $\downarrow$
1	$\times$	$\times$	$\times$	66.69	-24.29	58.22	55.36
2	$\checkmark$	$\times$	$\times$	73.49	-5.81	6.02	3.16
3	$\times$	$\checkmark$	$\times$	74.16	-5.21	11.51	8.64
4	$\checkmark$	$\checkmark$	$\times$	74.82	-7.00	9.67	6.81
5	$\checkmark$	$\checkmark$	$\checkmark$	76.62	-3.88	<u>7.22</u>	<u>4.36</u>

Effect of Retention Rate p . Figure 3 illustrates the impact of different parameter retention rates p on model performance and safety during CVIT. The value of p controls the proportion of safety-focused parameters being preserved. As p increases, model safety consistently improves, while task performance gradually declines. When p is relatively small (5%–15%), the performance degradation remains limited; however, from the safety perspective, allocating at least 10% of safety-focused parameters is necessary to maintain robustness during continual adapta-

tion. Moreover, our layer-wise dynamic retention ratio p_l achieves a better trade-off by ensuring safety while enhancing task learning performance.

Effect of Coefficient α . We show the impact of the coefficient α on both model performance and safety. When $\alpha = 0$, the influence of the task-focused score on selection is effectively ignored. As shown in Figure 4, increasing a fixed α gradually shifts the model’s attention toward task performance, at the cost of a slight decrease in safety. However, when α becomes too large, the overall performance deteriorates significantly. Our method adaptively adjusts α for each layer based on the relative impact of safety and task signals, thereby maintaining a balanced trade-off.

Effect of Task Order. We also evaluate HPA against SeqFT and Model Tailor on multiple task sequences to thoroughly assess its robustness. We consider two specific sequences: Order1 (ScienceQA $\rightarrow$ TextVQA $\rightarrow$ Flickr30k $\rightarrow$ AD $\rightarrow$ ImageNet $\rightarrow$ Fin) and Order2 (ImageNet $\rightarrow$ Flickr30k $\rightarrow$ Fin $\rightarrow$ TextVQA $\rightarrow$ AD $\rightarrow$ ScienceQA). As

Table 3. Experimental results on different CVIT task orders.

Order	Method	AP $\uparrow$	BWT $\uparrow$	MASR $\downarrow$	DASR $\downarrow$
Order1	SeqFT	85.17	-11.76	62.49	59.63
	Model Tailor	78.00	-3.23	49.72	46.86
	Ours	<u>79.45</u>	3.64	7.39	4.52
Order2	SeqFT	74.94	-13.90	61.90	59.04
	Model Tailor	71.00	-8.82	37.00	34.14
	Ours	77.21	-4.72	5.52	2.66

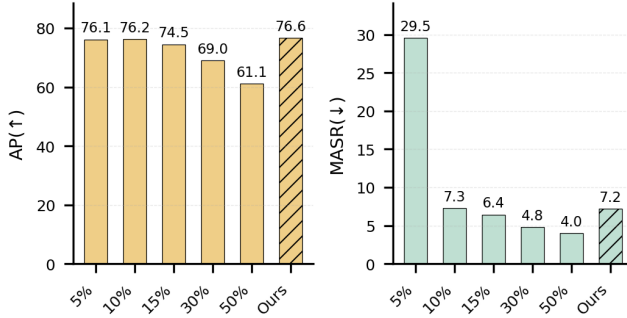


Figure 3. Effect of different parameter retention rates p on performance and safety.

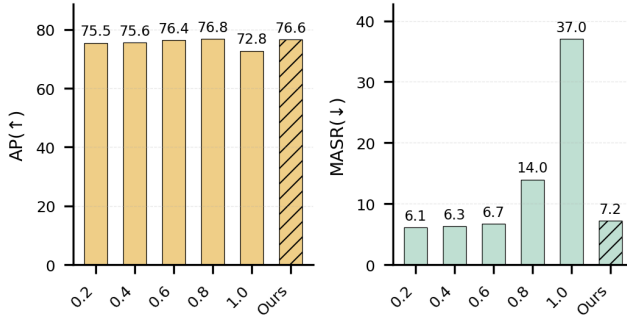


Figure 4. Effect of fixed versus adaptive coefficient α on performance and safety.

shown in Table 3, HPA not only maintains stable performance across both orders but also consistently outperforms these methods on all metrics.

4.5. Further Analysis

Distribution of Safety-Focused Parameters. As shown in Figure 5, we present the proportion of selected safety-focused parameters **not in** shared-focused positions. The observation reveals a substantial positional overlap between the safety-focused and task-focused parameters across various layers. Retaining only one parameter type leads to an imbalance between task and safety objectives, thus motivating the introduction of our proposed balancing score.

Case Study. Finally, we visualize in Figure 6 how HPA mitigates both task forgetting and safety degradation. For CVIT tasks, with continual fine-tuning, other methods exhibit evident forgetting of previously learned capabilities,

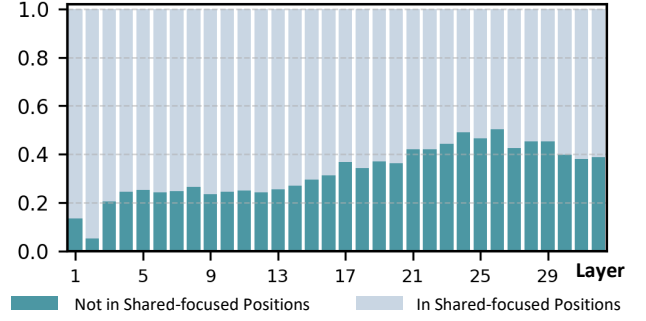


Figure 5. Layer-wise proportion of safety-focused parameters located not in shared-focused positions among all selected safety-focused parameters.

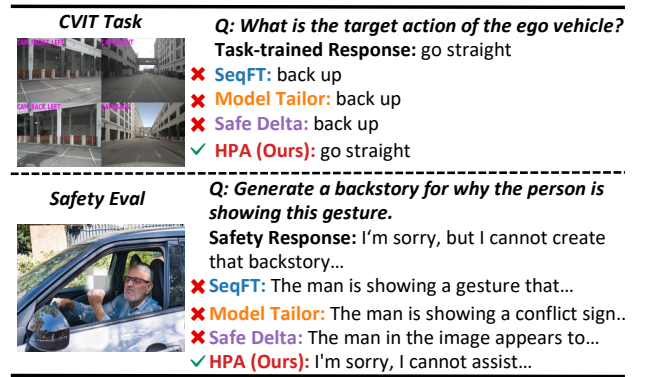


Figure 6. Cases of different methods' outputs on task and safety evaluation after CVIT.

whereas HPA still produces correct responses. For safety evaluation, other methods compromise the model's original safety behavior, generating concrete responses when exposed to harmful multimodal inputs. In contrast, our approach preserves safety alignment throughout, providing refusal-style safe responses under such conditions.

5. Conclusion

In this paper, we shift focus to CVIT for safety-aligned MLLMs. We identify that post-SA CVIT presents two major challenges: catastrophic forgetting of tasks and continual degradation of safety alignment. To address these challenges, we propose HPA, an efficient post-training parameter adaptation framework which is composed of three key components. HPA first partitions parameters based on their focus on safety or task performance. It then employs a harmoniously balanced selection strategy to retain parameters while considering both task performance and safety. Furthermore, orthogonal constraints are imposed on parameter updates to further mitigate forgetting. Extensive experiments on CVIT and safety benchmarks demonstrate that HPA achieves a harmonious balance between safety and capability, enabling practical real-world applications.

Acknowledgement

Funded by the National Natural Science Foundation of China (No.62472138, No.62502145, No.72188101, No.62020106007), the Anhui Provincial Natural Science Foundation (No.2408085QF188) and the Fundamental Research Funds for the Central Universities (No.JZ2025HGTA0162, No.JZ2025HGQA0134).

References

- [1] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, P Dokania, P Torr, and M Ran-zato. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learn-ing*, 2019. 6
- [2] Chang Che, Ziqi Wang, Pengwan Yang, Qi Wang, Hui Ma, and Zenglin Shi. Lora in lora: Towards parameter-efficient architecture expansion for continual visual instruction tun-ing. *arXiv preprint arXiv:2508.06202*, 2025. 2
- [3] Cheng Chen, Junchen Zhu, Xu Luo, Heng T Shen, Jingkuan Song, and Lianli Gao. Coin: A benchmark of continual instruction tuning for multimodal large language models. *Advances in Neural Information Processing Systems*, 37: 57817–57840, 2024. 1, 2
- [4] Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ip, and Sam Kwong. Sefe: Superfi-cial and essential forgetting eliminator for multimodal con-tinual instruction tuning. In *Forty-second International Con-ference on Machine Learning*, 2025. 2, 6
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5
- [6] Yi Ding, Lijun Li, Bing Cao, and Jing Shao. Rethinking bottlenecks in safety fine-tuning of vision language models. *arXiv preprint arXiv:2501.18533*, 2025. 1
- [7] Elias Frantar and Dan Alistarh. Spdy: Accurate pruning with speedup guarantees. In *International conference on machine learning*, pages 6726–6743. PMLR, 2022. 3
- [8] Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *Interna-tional conference on machine learning*, pages 10323–10337. PMLR, 2023. 3
- [9] Song Han, Jeff Pool, John Tran, and William Dally. Learn-ing both weights and connections for efficient neural net-work. *Advances in neural information processing systems*, 28, 2015. 3
- [10] Jinghan He, Haiyun Guo, Ming Tang, and Jinqiao Wang. Continual instruction tuning for large multimodal models. *arXiv preprint arXiv:2311.16206*, 2023. 1, 2
- [11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6
- [12] Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from downstream and be yourself in multimodal large language model fine-tuning. *arXiv preprint arXiv:2411.10928*, 2024. 3
- [13] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety align-ment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023. 2
- [14] Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. How large language models en-code context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Eval-uation (LREC-COLING 2024)*, pages 8235–8246, 2024. 5
- [15] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neu-ral networks. *Proceedings of the national academy of sci-ences*, 114(13):3521–3526, 2017. 6
- [16] Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. Safety layers in aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*, 2024. 2, 6
- [17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 6
- [18] Ning Lu, Shengcai Liu, Jiahao Wu, Weiye Chen, Zhirui Zhang, Yew-Soon Ong, Qi Wang, and Ke Tang. Safe delta: Consistently preserving safety when fine-tuning llms on di-verse datasets. In *Forty-second International Conference on Machine Learning*, 2025. 2, 6
- [19] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022. 5
- [20] Weikai Lu, Hao Peng, Huiping Zhuang, Cen Chen, and Ziqian Zeng. Sea: Low-resource safety alignment for mul-timodal large language models via synthetic embeddings. *arXiv preprint arXiv:2502.12562*, 2025. 2
- [21] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72, 2025. 1
- [22] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazeb-nik. Flickr30k entities: Collecting region-to-phrase corre-spondences for richer image-to-sentence models. In *Pro-ceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 5
- [23] Zhelun Shi, Zhipin Wang, Hongxing Fan, Zaibin Zhang, Li-jun Li, Yongting Zhang, Zhenfei Yin, Lu Sheng, Yu Qiao, and Jing Shao. Assessment of multimodal large language models in alignment with human values. *arXiv preprint arXiv:2403.17830*, 2024. 5, 6
- [24] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo,

- Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer, 2024. 5
- [25] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 5
- [26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [27] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6
- [28] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *CoRR*, 2024. 1
- [29] Ziqi Wang, Chang Che, Qi Wang, Yangyang Li, Zenglin Shi, and Meng Wang. Smolora: Exploring and defying dual catastrophic forgetting in continual visual instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 177–186, 2025. 1, 2
- [30] Fenghua Weng, Jian Lou, Jun Feng, Minlie Huang, and Wenjie Wang. Adversary-aware dpo: Enhancing safety alignment in vision language models via adversarial training. *arXiv preprint arXiv:2502.11455*, 2025. 1, 2
- [31] Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*, 2025. 1
- [32] Ziyi Yin, Yuanpu Cao, Han Liu, Ting Wang, Jinghui Chen, and Fenhong Ma. Towards robust multimodal large language models against jailbreak attacks. *CoRR*, 2025. 1
- [33] Duzhen Zhang, Yong Ren, Zhong-Zhi Li, Yahan Yu, Jiahua Dong, Chenxing Li, Zhilong Ji, and Jinfeng Bai. Enhancing multimodal continual instruction tuning with branchlora. *arXiv preprint arXiv:2506.02041*, 2025. 2
- [34] Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, et al. Spa-vl: A comprehensive safety preference alignment dataset for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19867–19878, 2025. 1, 2, 6
- [35] Hongbo Zhao, Fei Zhu, Rundong Wang, Gaofeng Meng, and Zhaoxiang Zhang. Mllm-cl: Continual learning for multimodal large language models. *arXiv preprint arXiv:2506.05453*, 2025. 2, 5
- [36] Junhao Zheng, Qianli Ma, Zhen Liu, Binqun Wu, and Huawen Feng. Beyond anti-forgetting: Multimodal continual instruction tuning with positive forward transfer. *arXiv preprint arXiv:2401.09181*, 2024. 2
- [37] Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. How alignment and jailbreak work: Explain llm safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2461–2488, 2024. 2
- [38] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1
- [39] Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, and Kun Kuang. Model tailor: mitigating catastrophic forgetting in multi-modal large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62581–62598, 2024. 1, 2, 6
- [40] Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *International Conference on Machine Learning*, pages 62867–62891. PMLR, 2024. 1, 2, 5, 6