Mitigating Visual Knowledge Forgetting in MLLM Instruction-tuning via Modality-decoupled Gradient Descent

Junda Wu[1], **Yuxin Xiong**[1], **Xintong Li**[1], **Yu Xia**[1], **Ruoyu Wang**[2], **Yu Wang**[1]
Tong Yu[3], **Sungchul Kim**[3], **Ryan Rossi**[3], **Lina Yao**[2], **Jingbo Shang**[1], **Julian McAuley**[1]
[1]UC San Diego [2]University of New South Wales [3]Adobe Research
{juw069,y7xiong,xil240,yux078,yuw164,jshang,jmcauley}@ucsd.edu
{ruoyu.wang5,lina.yao}@unsw.edu.au {tyu,sukim,ryrossi}@adobe.com

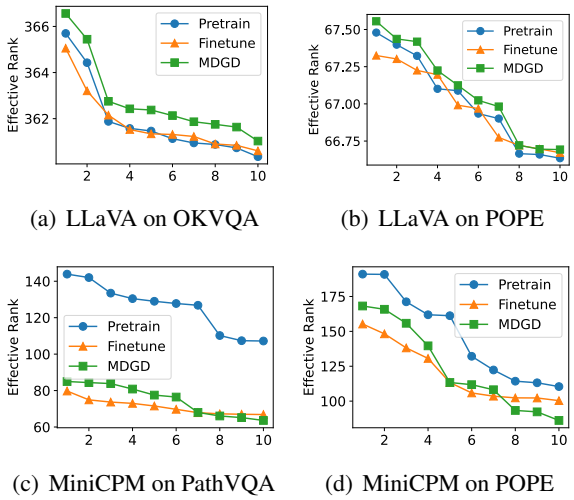
Abstract

Recent MLLMs have demonstrated strong visual understanding and reasoning after large-scale multimodal pre-training. However, instruction-tuning is typically text-driven with limited visual supervision, leading to significant visual forgetting and degradation of pre-trained visual knowledge. Existing fine-tuning and continual learning methods compress visual representations and emphasize task alignment over visual retention, failing to address this challenge. We present a novel perspective using effective rank to quantify the loss of visual representation richness, framing visual forgetting as excessive compression under the information bottleneck principle. To address this, we propose modality-decoupled gradient descent (MDGD), which regulates gradient updates to preserve the effective rank of visual features and explicitly disentangles visual learning from task-specific alignment. We further introduce a memory-efficient fine-tuning variant using gradient masking for parameter-efficient adaptation. Extensive experiments show that MDGD effectively mitigates visual forgetting across downstream tasks and models, maintaining pre-trained visual knowledge while supporting strong task adaptation.

1 Introduction

Multimodal large language models (MLLMs) enhanced visual understanding and reasoning by pre-training on large-scale multimodal datasets with comprehensive visual descriptions that integrate textual and visual knowledge (Liu et al., 2024b; Yao et al., 2024; Li et al., 2023b; Bai et al., 2023; Liu et al., 2024d; Wu et al., 2024d). These models achieve strong performance across various vision-language tasks, such as visual question answering (Jin et al., 2024; Wu et al., 2025b), multimodal reasoning (Zhang et al., 2024; Jiang et al., 2024b; Yan et al., 2024), multimodal recognition (Shenoy et al., 2024; Wu et al., 2024d), and personalized multimodality (Wu et al., 2024b, 2025a; Huang et al., 2025). However, adapting pre-trained MLLMs to downstream tasks via instruction-tuning (Wu et al., 2024a; Li et al., 2024, 2023a; Panagopoulou et al., 2023; Liu et al., 2024d) presents a critical challenge of visual forgetting. Unlike pre-training, where models receive rich visual-text alignment, instruction-tuning is often text-driven with limited direct visual supervision. This shift in training focus leads to the degradation of pre-trained visual encoding (Zhou et al., 2024; Niu et al., 2024; Wu et al., 2024a; Ko et al., 2023), negatively impacting model generalizability across downstream tasks that require strong visual knowledge (Bai et al., 2024; Huang et al., 2024). Addressing this challenge is essential for ensuring MLLMs retain their visual capabilities while aligning with new tasks efficiently.

While several approaches have attempted to mitigate catastrophic forgetting in neural networks



(a) LLaVA on OKVQA (b) LLaVA on POPE

(c) MiniCPM on PathVQA (d) MiniCPM on POPE

Figure 1: The top-10 image tokens with the highest effective ranks on OKVQA and POPE encoded by LLaVA, and PathVQA and POPE encoded by MiniCPM. We compare pretrained, finetuned, and MDGD-finetuned models. Effective rank (Wei et al., 2024) quantifies representation richness, which show that MDGD preserves higher effective rank, mitigating visual forgetting.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2282–2295
November 4–9, Suzhou, China, 2025

through direct fine-tuning and continual learning methods (Shi et al., 2024; Wu et al., 2024e; Zhu et al., 2024; Zheng et al., 2024), these methods often overlook the unique challenge of preserving visual knowledge in multimodal large language models (MLLMs). Directly fine-tuning MLLMs on new tasks often leads to overfitting to textual instructions while inadvertently suppressing visual representations (Zhai et al., 2023). Existing continual learning strategies, such as regularization and replay methods, tend to focus on retaining language-based knowledge, neglecting the trade-off between compressing visual representations and aligning them with task-specific instructions (Zhou et al., 2024; Niu et al., 2024; Wu et al., 2024a; Ko et al., 2023), leading to the degradation of pre-trained visual knowledge. Task-orthogonal gradient descent techniques have shown promise in disentangling gradients for multi-task optimization. However, their practical application in MLLMs poses unique challenges. MLLMs are pre-trained on vast and heterogeneous multimodal datasets (Liu et al., 2024b; Li et al., 2023b; Bai et al., 2023), where it is challenging to isolate task-specific gradients, causing the components critical for visual understanding to become entangled with other features.

To gain a fundamental view of the challenge of visual knowledge forgetting in MLLM instruction tuning, we adopt an information bottleneck (IB) perspective that characterizes the trade-off between retaining input information and ensuring output predictiveness (Tishby et al., 2000). To investigate the degradation of crucial pre-trained visual knowledge, we introduce a novel perspective that leverages effective rank to quantify the richness of the encoded visual representation from MLLMs. Specifically, we illustrate the visual forgetting problem in Figure 1, where we observe a consistent effective rank reduction problem caused by MLLM instruction tuning. Based on this view, we propose a modality-decoupled gradient descent (MDGD) method, which disentangles the optimization of visual understanding from task-specific alignment, MDGD regulates gradient updates to maintain the effective rank of visual representations compared with pre-trained MLLMs, while mitigating the over-compression effects described by the information bottleneck. Intuitively, visual forgetting occurs due to the shift from rich multimodal pre-training to instruction-tuning, where text-based supervision dominates without direct visual supervision. By explicitly decoupling the task-specific alignment with visual representation learning, MDGD preserves expressive and robust visual features. To further improve efficiency in instruction-tuning, we introduce a memory-efficient fine-tuning strategy using gradient masking, which selectively updates a subset of model parameters for parameter-efficient fine-tuning (PEFT). This approach reduces computational overhead while ensuring that crucial pre-trained visual representations are retained.

We summarize our contributions as follows:

- We analyze the visual knowledge forgetting problem in MLLM instruction tuning and frame the problem through the lens of effective rank and information bottleneck theory.
- We propose MDGD, which decouples visual optimization from task-specific alignment to preserve visual representations and introduces a PEFT variant MDGD-GM to reduce computational overhead through gradient masking.
- We conduct comprehensive experiments on various MLLMs and downstream tasks, demonstrating that MDGD effectively mitigates visual forgetting while enabling strong adaptation to new tasks.

1.1 Visual Knowledge Forgetting in MLLMs

Catastrophic forgetting, where a model loses previous knowledge while learning new tasks, is a major challenge in continual learning (Wang et al., 2023). This problem is now widely recognized in LLMs and MLLMs (Wu et al., 2024e; Luo et al., 2023; Zhai et al., 2023). Although various methods, including fine-tuning, task-orthogonal gradient descent, knowledge distillation, and replay, have been adapted to mitigate forgetting (Shi et al., 2024; Wu et al., 2024e; Zhu et al., 2024; Zheng et al., 2024), they often fail to preserve rich visual features. Fine-tuning on new tasks tends to overfit text and suppress visual information, while parameter-efficient methods like LoRA also suffer from forgetting (Fawi, 2024; Liu et al., 2024c). Model Tailor (Zhu et al., 2024) adapts the LLM backbone but does not address visual knowledge forgetting, which can lead to hallucination or degraded generalization (Zhai et al., 2023). In contrast, our approach synchronizes the training of the visual encoder and LLM, preserving pre-trained visual knowledge during instruction tuning.

1.2 Information Theory in LLMs

The Information Bottleneck (IB) principle (Tishby et al., 2000) has been used in LLMs to compress input while retaining task-relevant information (Delétang et al., 2023; Valmeekam et al., 2023; Wei et al., 2024; Wu et al., 2022). Prior works use IB to extract robust features (Zhang et al., 2022; Wu et al., 2024c) and enable feature attribution (Li et al., 2022; Jiang et al., 2020), but these focus on language models, not multimodal settings (Yang et al., 2025). Existing information-theoretic transfer learning methods (Tseng et al., 2024; Wu et al., 2024c; Ling et al., 2024) also do not address the unique challenges of MLLMs, where modalities are deeply entangled. Our method instead uses effective rank to measure and counteract visual representation compression. The proposed MDGD explicitly decouples visual learning from task alignment, going beyond previous IB-based approaches.

2 Preliminary

Task Definition. Given an MLLM π_θ and instruction-tuning dataset D, the image prompt $I \in \Omega$ is encoded by a visual encoder f into a sequence of M visual tokens $f(I) = X^v = (x_1^v, x_2^v, \ldots, x_M^v)$. During instruction tuning, the textual instructions $T \in D$ are tokenized as $X^l = (x_1^l, x_2^l, \ldots, x_N^l)$ using the tokenizer of the backbone LLM, which is querying the MLLM to generate textual responses conditioned on the multimodal inputs,

$$\hat{y}_k \sim \pi_\theta(\cdot \mid X^v, X^l, y_{<k}). \tag{1}$$

Therefore, the learning objective of visual instruction-tuning for K samples is to maximize the average log-likelihood of the ground truth answer tokens $y = (y_1, y_2, \ldots, y_T)$ of each sample,

$$\mathcal{L}_{vl}(\theta) = -\sum_{t=1}^{T} \log \pi_\theta \left(y_t \mid X^v, X^l, y_{<t} \right), \tag{2}$$

where multimodal instructions X^v and X^l both serve as generation conditions.

An Information Bottleneck Perspective on Visual Knowledge Forgetting. In multimodal models, the information bottleneck (Mai et al., 2022) (IB) framework provides a powerful lens to understand how representations are formed. In our setting, the IB principle seeks a representation Z that is maximally informative about the output y while discarding irrelevant details from the inputs. For an MLLM that processes visual inputs X^v and textual inputs X^l, a full IB objective might take the form:

$$\min_{\theta} \quad \mathcal{L}_{\text{IB}}^{\text{vision}}(\theta) = -I(y; Z) + \beta\, I(X^v; Z). \tag{3}$$

where $I(\cdot;\cdot)$ denotes mutual information and β controls the trade-off between predictive power and compression. This formulation explicitly highlights the risk of discarding visual details when the model is optimized primarily to predict y. Based on the information bottleneck view, we further analyze the visual forgetting problem in Appendix C.

Effective Rank as a Measure of Representation Richness. To quantify the information content retained in a representation, we use the effective rank metric (Roy and Vetterli, 2007). Given a representation matrix Z whose singular values are $\{\sigma_i\}$, the effective rank is defined as:

$$\text{erank}(Z) = \exp\left(-\sum_i p_i \log p_i \right), \tag{4}$$

where $p_i = \sigma_i / \sum_j \sigma_j$. This measure, based on the entropy of the singular value distribution, captures the "richness" or intrinsic dimensionality of Z. A higher effective rank indicates that the representation spans a larger subspace, whereas a lower effective rank implies that the representation has been overly compressed.

3 MDGD: Modality-Decoupled Gradient Regularization and Descent

Motivated by the visual forgetting problem caused by the degradation of multimodal encoding in Eq. (12), we introduce a modality-decoupling gradient regularization (**MDGD**) to approximate orthogonal gradients between visual understanding drift and downstream task optimization. Specifically, leveraging modality-decoupled gradients $\bar{g}_\theta$ and $\bar{g}_\phi$ derived from the current MLLM and a pre-trained MLLM respectively, we propose a gradient regularization term $\tilde{g}_\theta$ for more efficient multimodal instruction tuning, which promotes the alignment of downstream tasks while mitigating visual forgetting (Zhu et al., 2024). Since MDGD requires the estimation of parameter gradients, we could not directly apply parameter-efficient fine-tuning methods (*e.g.*, LoRA (Hu et al., 2021)). Thus, we alternatively formulate the regularization as a gradient mask $M_{\tilde{g}_\theta}$, which allows efficient fine-tuning only on a subset of masked model parameters.

3.1 Modality Decoupling

Based on the information bottleneck objective in Eq. (3), the objective encourages the model to maximize $I(y; Z)$ while compressing $I(X^v; Z)$ (Tishby et al., 2000; Alemi et al., 2016). In practice, this compression may discard useful visual details, leading to visual forgetting. To mitigate such compression and preserve the pre-trained visual knowledge, we follow the KL divergence loss $D_{\text{KL}}\Big(\mu_\phi(X^v) \Big\| \pi_\theta(X^v)\Big)$ to constrain the current model's visual representation $\pi_\theta(X^v)$ to remain close to the pre-trained distribution $\mu_\phi(X^v)$, thereby preserving the mutual information $I(X^v; Z)$ that would otherwise be reduced by the compression (Hinton, 2015; Lopez et al., 2018). However, since MLLMs cannot directly track the distributions of image tokens, we instead introduce an auxiliary loss function

$$\mathcal{L}_v(\phi, \theta) = \|\mu(X^v|\phi) - \pi(X^v|\theta)\|_1, \quad (5)$$

which approximates the KL divergence loss (Zhu et al., 2022b, 2017) by penalizing discrepancies between the pre-trained visual representation and that obtained during instruction tuning.

In the MLLM instruction tuning, the visual output tokens (e.g., $\{z_k^{vl}\}_{k=1}^M$) are encoded as latent representations. Such visual encoding cannot be directly supervised by any learning objective but is learned through textual gradient propagation of the negative log-likelihood loss in downstream tasks. To approximate the visual optimization direction, we derive the gradients of $\mathcal{L}_v(\phi, \theta)$ for both the pre-trained MLLM π_ϕ and the current MLLM π_θ:

$$h_\phi = \nabla_\phi \mathcal{L}_v(\phi) = \boldsymbol{\lambda}(\phi, \theta) \cdot \nabla_\phi \mu(X^v|\phi),$$
$$h_\theta = \nabla_\theta \mathcal{L}_v(\theta) = -\boldsymbol{\lambda}(\phi, \theta) \cdot \nabla_\theta \pi(X^v|\theta),$$

where $\boldsymbol{\lambda}(\phi, \theta) = \text{sign}\left(\mu(X^v|\phi) - \pi(X^v|\theta)\right)$. Intuitively, when the MLLM's visual understanding drift causes visual forgetting, we further derive the orthogonal task gradients $\bar{g}_\phi$ and $\bar{g}_\theta$:

$$\bar{g}_\phi = \nabla_\phi \mathcal{L}_{vl}(\phi) - \frac{\nabla_\phi \mathcal{L}_{vl}(\phi)^\top h_\phi}{\|h_\phi\|^2} \cdot h_\phi, \quad (6)$$

$$\bar{g}_\theta = \nabla_\theta \mathcal{L}_{vl}(\theta) - \frac{\nabla_\theta \mathcal{L}_{vl}(\theta)^\top h_\theta}{\|h_\theta\|^2} \cdot h_\theta, \quad (7)$$

which enables **modality decoupling** of the downstream task loss gradient in Eq.(2) orthogonal to the visual understanding drift for the pretrained MLLM $\bar{g}_\phi \perp h_\phi$ and current MLLM $\bar{g}_\theta \perp h_\theta$.

Algorithm 1 MDGD: Modality Decoupled Gradients Descent

Inputs: Pre-trained MLLM μ_ϕ, current MLLM π_θ, instruction-tuning dataset D, and learning rate η
Outputs: The optimized model weights of π_θ
Initialize $\pi_\theta \leftarrow \mu_\phi$
for Receive minibatch $D_i \subset D$ **do**
  Calculate $\mathcal{L}_{vl}(\phi)$ of μ_ϕ, based on Eq.(2);
  Calculate $\mathcal{L}_{vl}(\theta)$ of π_θ, based on Eq.(2);
  Extract visual encodings of $\mu(X^v|\phi)$;
  Extract visual encodings of $\pi(X^v|\theta)$;
  Calculate $\mathcal{L}_v(\phi, \theta)$, based on Eq.(5);
  Derive orthogonal task gradients $\bar{g}_\phi$ and $\bar{g}_\theta$, according to Eq.(6);
  if Parameter-efficient fine-tuning **then**
    Calculate $M_{\tilde{g}_\theta}$,based on Eq.(10);
    Update the model following Eq.(11).
  else
    Calculate $\tilde{g}_\theta$, based on Eq.(8);
    Update the model following Eq.(9).
  end if
end for

3.2 Regularized Gradient Descent

The auxiliary loss in Eq. (5) preserves the visual representation at a distribution level via the feature alignment auxiliary loss in Eq. (5). However, the information bottleneck framework indicates that the gradient component compressing $I(X^v; Z)$ (*i.e.*, $\nabla_\theta I(X^v; Z)$), can harm visual preservation by reducing the effective rank of the features (Achille and Soatto, 2018; Lee et al., 2021).

To address this compression-induced drift, we incorporate an orthogonal gradient as a regularize. Motivated by multi-task orthogonal gradient optimization (Yu et al., 2020; Zhu et al., 2022a; Dong et al., 2022), we leverage the gradient $\bar{g}_\phi$ from the pre-trained model μ_ϕ, which reflects the accumulated visual drift and approximates a global orthogonal learning effect in the downstream task. We then project the current model's gradient onto this direction:

$$\tilde{g}_\theta = \frac{\bar{g}_\theta^\top \bar{g}_\phi}{\|\bar{g}_\phi\|^2} \cdot \bar{g}_\phi. \quad (8)$$

In addition, to prevent discrepancies between the regularization and task gradients, we include the feature alignment auxiliary loss (Eq. (5)) in the overall objective. The final parameter update is:

$$\pi_\theta \leftarrow \pi_\theta - \nabla_\theta \mathcal{L}_{vl}(\theta) - \nabla_\theta \mathcal{L}_v(\theta) - \tilde{g}_\theta. \quad (9)$$

3.3 Enabling Parameter-efficient Fine-tuning of MDGD via Gradient Masking

Parameter-efficient fine-tuning (PEFT) methods, such as adapters (Houlsby et al., 2019) and LoRA (Hu et al., 2021), aim to reduce the computational cost and memory usage when fine-tuning models on downstream tasks under practical constraints (Han et al., 2024). However, due to the requirement of directly estimating gradient directions on the pre-trained model parameters, MDGD cannot be directly applied to these PEFT methods, which introduce additional model parameters whose gradients are separate from the original model weights.

To address this challenge, we propose a variant, MDGD-GM, by formulating the gradient regularization term in Eq. (8) as gradient masking that selects model weights with efficient gradient directions. Specifically, we define it as

$$M_{\bar{g}_\theta} = \mathbf{1}\left\{\frac{\bar{g}_\theta^\top \bar{g}_\phi}{\|\bar{g}_\phi\|\|\bar{g}_\theta\|} \geq T_\alpha\right\}, \qquad (10)$$

where T_α is determined by a percentile α of trainable parameters with the highest similarity scores between $\bar{g}_\theta$ and $\bar{g}_\phi$. Consequently, the optimization in Eq. (9) is reformulated as

$$\pi_\theta \leftarrow \pi_\theta - M_{\bar{g}_\theta} \cdot (\nabla_\theta \mathcal{L}_{vl}(\theta) + \nabla_\theta \mathcal{L}_v(\theta)). \quad (11)$$

We summarize and illustrate the optimization process of MDGD and MDGD-GM in Algorithm 1.

4 Experiments

Datasets To evaluate catastrophic forgetting, we follow the setup of Zhu et al. (2024) and consider two models: LLaVA-1.5 (7B) and MiniCPM-V-2.0 (2.8B). For each model, we use a set of pre-trained tasks (including VQAv2, GQA, VizWiz, TextVQA, POPE, and MM-Bench) and fine-tuning tasks on previously unseen datasets (such as Flickr30k, OKVQA, TextCaps, and PathVQA). Full details on dataset composition are provided in the appendix.

Baselines We compare our approach against several baselines: standard fine-tuning following Zhu et al. (2024), LoRA-based fine-tuning (Hu et al., 2021), and Model Tailor (Zhu et al., 2024). For Model Tailor, we report the original results from their paper for comparison.

Implementation Details All experiments use the official Huggingface implementations of LLaVA-1.5 and MiniCPM-V-2.0, with LoRA adapters where applicable. Models are fine-tuned using BFloat16 precision on 2 NVIDIA A100 80GB GPUs. We include fine-grained implementation details in Appendix B.

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

4.1 Comparison Results

LLaVA-1.5 adapts better to downstream tasks but is more prone to visual forgetting. We study the visual forgetting problem on the LLaVA-1.5 MLLM, and report performance comparison results in Table 1. We observe that the pre-trained LLaVA enables efficient instruction tuning on target tasks, where the zero-shot performance is near zero. When the model is fine-tuned on the image caption task, Flickr30K, which largely differs from the pre-trained tasks of visual question-answering, the model can learn a degraded multimodal representation, which causes visual forgetting in its projected visual representation space (in Section C). Fine-tuning on visual question-answering task OKVQA, which is similar to the pre-trained tasks, can also lead to MLLM's visual understanding drift, due to the limited image-text pairs existing in the downstream task.

MiniCPM-V-2.0 also experiences visual forgetting while limited in downstream task improvements. To validate the observation on a smaller MLLM, we report the comparison results of MiniCPM-V-2.0 with 2.8B model parameters in Table 2. We observe that compared with the LLaVA MLLM, MiniCPM suffers from less prominent visual forgetting. We attribute this observation to MiniCPM learning a more compact and constrained visual representation space during pre-training, causing the visual representations of target task images to be less aligned with those of the pre-trained MLLM. Consequently, MiniCPM exhibits limited improvement in downstream tasks, as its restricted ability to acquire additional visual knowledge leads to ineffective instruction tuning.

MDGD prevents visual forgetting while maintaining downstream task improvements. By employing MDGD in MLLM instruction tuning, we observe the LLaVA's average performance drops on pre-trained tasks when fine-tuned on OKVQA and also improves when fine-tuned on Flickr30K, which demonstrates the efficiency of MDGD in mitigating visual forgetting. For the smaller MLLM, MiniCPM, MDGD achieves comparable fine-tuning improvements with direct fine-tuning, while completely eliminating visual for-

Method	#Params	Pre-trained tasks						Target task	Metrics	
		GQA	**VizWiz**	**SQA**	**TextVQA**	**POPE**	**MMBench**	**Flickr30k**	**Avg**	**Hscore**
Zero-shot	–	61.94	50.00	66.80	58.27	85.90	64.30	3.5	55.82	59.86
Fine-tune	1.2B	56.26	44.45	28.34	38.98	38.40	50.56	**78.82**	47.97	45.26
LoRA	29M	17.74	40.63	5.38	30.48	2.40	9.55	64.18	24.33	20.49
Model Tailor	273M	52.49	42.28	67.15	43.89	82.88	63.40	75.40	61.07	59.85
MDGD	1.2B	67.71	48.18	**69.05**	57.32	**85.12**	65.43	73.47	**66.61**	**66.03**
w/o visual align	1.2B	57.64	36.95	53.96	32.84	30.43	56.66	65.58	47.72	46.19
MDGD-GM	124M	**69.89**	**51.22**	65.87	**58.18**	84.39	**66.42**	64.18	65.74	65.86

Method	#Params	Pre-trained tasks						Target task	Metrics	
		GQA	**VizWiz**	**SQA**	**TextVQA**	**POPE**	**MMBench**	**OKVQA**	**Avg**	**Hscore**
Zero-shot	–	61.94	50.00	66.80	58.27	85.90	64.30	0.14	55.34	59.58
Fine-tune	1.2B	62.98	40.59	59.84	48.38	71.42	51.98	69.10	57.76	56.79
LoRA	29M	63.44	41.61	51.29	48.02	75.27	37.31	**71.46**	55.49	54.12
Model Tailor	273M	60.39	**46.49**	**69.51**	**54.88**	**85.44**	63.32	38.10	59.73	61.48
MDGD	1.2B	**66.55**	42.72	64.60	52.54	85.17	61.73	62.29	62.23	62.22
w/o visual align	1.2B	66.39	39.89	60.19	52.40	84.92	62.97	62.39	61.31	61.22
MDGD-GM	124M	66.02	43.97	67.91	52.80	84.70	**63.97**	61.04	**62.92**	**63.07**

Table 1: Performance on various pre-trained tasks of LLaVA-1.5 models fine-tuned on Flickr30K and OKVQA. We report the best performance for each task in a **bold font** while the second best performance underlined.

getting in the pre-trained tasks. MDGD and its variants consistently achieve the best average performance for both MLLMs, demonstrating its great potential for incremental learning on individual downstream tasks.

Comparison with baseline methods. Table 1 shows that MDGD consistently outperforms both LoRA fine-tuning and Model Tailor (Zhu et al., 2024) on LLaVA-1.5. LoRA suffers from visual forgetting due to projecting multimodal features into lower-rank spaces, especially on Flickr30K and OKVQA. Model Tailor, while effective for anti-forgetting in LLMs, is less robust for MLLMs and remains sensitive to the target dataset, performing better on Flickr30K than OKVQA. In contrast, MDGD achieves higher average scores and H-scores across datasets. In Table 2, MDGD improves average performance on MiniCPM tasks, reducing visual forgetting by 2.43% and 1.83% on PathVQA and TextCaps, respectively.

4.2 Ablation Study

Ablation study on visual alignment. We compare MDGD with its two variants, MDGD w/o visual align and MDGD-GM. MDGD w/o visual align enables MDGD without including visual representation loss $\mathcal{L}_v(\phi, \theta)$ Eq.(5), to understand the effect of directly optimizing to reduce the visual representation discrepancy between the current model and pre-trained model. We observe that MDGD w/o visual align maintains relatively comparable performance to MDGD on OKVQA and PathQA, due to the reduced need for visual representation adaptation in such visual question-answering tasks. In contrast, tasks like image captioning on Flickr30K and TextCaps benefit from feature alignment regularization, which directly mitigates visual understanding drift in the MLLM.

Ablation study on gradient masking. The other variant, MDGD-GM, leverages gradient masking to enable parameter-efficient fine-tuning (PEFT). We observe the PEFT variant of MDGD consistently achieves comparable performance across all tasks and backbone MLLMs, which only fine-tunes a subset of 10% original MLLM parameters used for direct fine-tuning and original MDGD. Different from conventional PEFT methods such as adapters, MDGD and its variants do not introduce additional parameters to the original model architecture, enabling incremental learning in an online setting (Maltoni and Lomonaco, 2019; Gao et al., 2023).

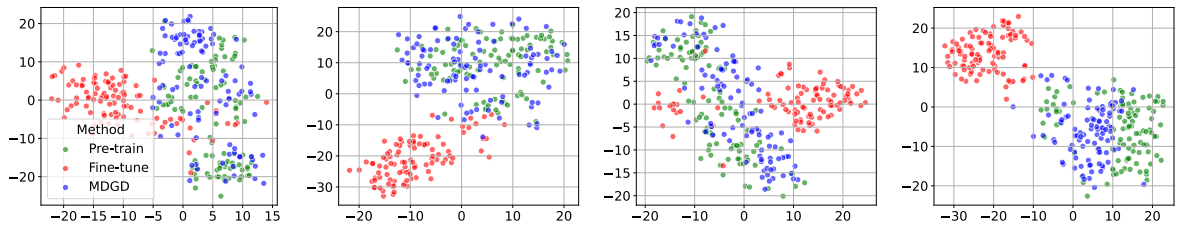
4.3 Representation Learning Analysis

T-SNE Analysis on Visual Representation To analyze the learning of visual and multimodal representation distributions in MLLMs, we create T-SNE (Van der Maaten and Hinton, 2008) plots to visualize the feature distributions extracted from pre-trained MLLMs, as well as MLLMs after standard fine-tuning and MDGD We illustrate the distributions of the multimodal features z^{vl} extracted from the last token of the multimodal instruction

Method	#Params	Pre-trained tasks							Target task	Metrics	
		VizWiz	A-OKVQA	OKVQA	TextVQA	IconQA	POPE	MMBench	PathVQA	Avg	Hscore
Zero-shot	-	55.27	79.39	64.86	77.98	79.01	88.93	70.98	5.44	65.23	10.04
Fine-tune	517M	52.91	76.94	59.06	58.34	76.96	**89.60**	70.16	11.04	61.88	18.74
LoRA	35M	52.95	76.24	**64.45**	77.18	77.80	88.08	67.47	**15.03**	64.90	**24.41**
MDGD	517M	**55.73**	78.25	64.33	77.54	**79.45**	89.19	**71.94**	9.09	**65.69**	15.97
w/o visual align	517M	54.92	78.52	64.17	77.42	79.37	89.10	70.96	8.49	65.37	15.03
MDGD-GM	52M	55.04	**78.78**	64.31	**77.78**	79.10	88.76	70.98	5.72	65.06	10.52

Method	#Params	Pre-trained tasks							Target task	Metrics	
		VizWiz	A-OKVQA	OKVQA	TextVQA	IconQA	POPE	MMBench	TextCaps	Avg	Hscore
Zero-shot	-	55.27	79.39	64.86	77.98	79.01	88.93	70.98	15.77	66.52	25.50
Fine-tune	517M	52.03	77.73	59.16	67.24	78.67	88.20	71.42	**33.85**	66.04	**44.76**
LoRA	35M	53.30	78.17	63.99	77.68	78.28	87.31	69.23	32.41	67.55	43.80
MDGD	517M	**55.17**	78.17	63.67	76.08	79.40	**89.11**	71.58	28.90	67.76	40.52
w/o visual align	517M	51.35	78.08	63.06	76.48	78.99	88.98	71.30	25.93	66.77	37.35
MDGD-GM	52M	55.04	**78.43**	**65.26**	**78.08**	**79.65**	88.93	**71.88**	29.14	**68.30**	40.85

Table 2: Performance on various pre-trained tasks of MiniCPM-V2.5 models fine-tuned on PathVQA and TextCaps. We report the best performance for each task in a **bold font** while the second best performance underlined.



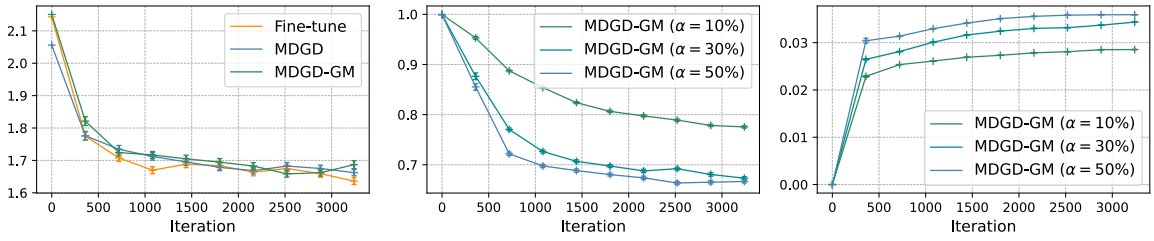
(a) z^{vl} on **PathVQA** (b) $\pi(X^v)$ on **PathVQA** (c) z^{vl} on **TextCaps** (d) $\pi(X^v)$ on **TextCaps**

Figure 2: T-SNE plots of the distribution of extracted visual $\pi(X^v)$ and multimodal z^{vl} representations from pre-trained MiniCPM, and models with direct fine-tuning and MDGD on PathVQA and TextCaps.

tokens, and the visual features $\pi_\theta(X^v)$ extracted from the last token of the input image tokens. We observe a consistent visual understanding drift in the MLLMs' visual representation spaces after standard fine-tuning on PathVQA and TextCaps with MiniCPM (Figure 2b and 2d). By employing MDGD to mitigate visual forgetting, we observe that visual understanding drift is effectively reduced, allowing the fine-tuned MLLM to retain pre-trained visual capabilities.

We further observe a distributional discrepancy in the multimodal representation z^{vl} of LLaVA (Figures 5a and 5c) between MDGD and the pre-trained MLLM. This discrepancy arises from the alignment of the MLLM to the target task through multimodal instructions, demonstrating effective adaptation to the downstream task of the LLaVA model. In addition, we also observe such multimodal distribution discrepancy reduces in a smaller MLLM, MiniCPM. This observation aligns with our findings on MiniCPM in Section 4.1, where we noted limited effects in model adaptation to downstream tasks. However, applying MDGD to MiniCPM mitigates visual forgetting by preventing degradation of both image and multimodal encodings into lower-rank representation spaces.

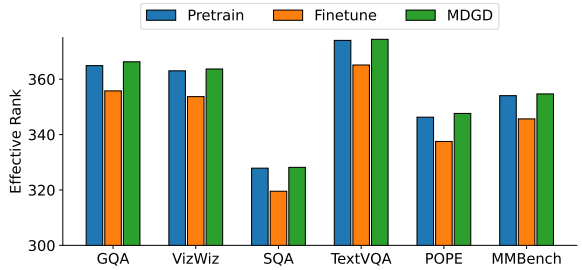
Effective Rank Analysis on Visual Representation To quantitatively analyze the visual forgetting problem (in Section C) described in Eq. (12), we calculate effective ranks of the visual representations extracted from the last hidden layer on the position of image tokens in individual MLLMs. We show the comparison results of LLaVA models in Figure 4(a) and MiniCPM models in Figure 4(b). We observe that with both the backbone models of LLaVA and MiniCPM, directly fine-tuning the pre-trained models on downstream tasks can lead to a consistent reduction of effective ranks in visual representations. Such observations validate the hypothesis in Section C regarding the potential visual forgetting problem in MLLM instruction tuning. In addition, we can observe that MDGD achieves consistent improvements in effective ranks compared with the standard fine-tuning method for both backbone MLLMs across various pre-trained tasks. In Figure 4(a), we observe that MDGD achieves com-



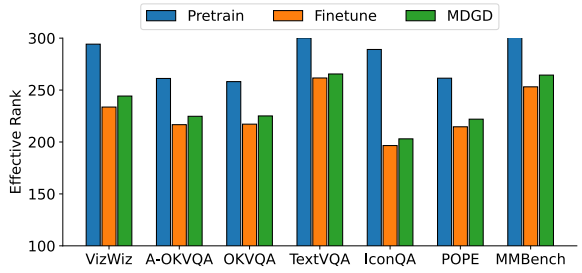
(a) Task Loss $\mathcal{L}_{vl}(\phi, \theta)$ (b) Cosine Similarity $\bar{g}_\theta^\top \bar{g}_\phi / \|\bar{g}_\phi\| \|\bar{g}_\theta\|$ (c) Representation Loss $\mathcal{L}_v(\phi, \theta)$

Figure 3: Illustration of (a) the learning process of three methods based on task loss $\mathcal{L}_{vl}(\phi, \theta)$, (b) the average regularized cosine similarity $\frac{\bar{g}_\theta^\top \bar{g}_\phi}{\|\bar{g}_\phi\| \|\bar{g}_\theta\|}$ in Eq.(10) for gradient masking at varying ratios, and (c) the visual representation loss $\mathcal{L}_v(\phi, \theta)$ in Eq.(5) for gradient masking at varying ratios α.

parable or even better effective ranks on pre-trained tasks, compared with the pre-trained LLaVA model. However, MDGD on MiniCPM in Figure 4(b) also suffers from the visual representation degradation problem, while MDGD consistently alleviates the problem. Such observation suggests a higher risk of visual forgetting in smaller-scale MLLMs.



(a) LLaVA models pretrained, finetuned, and fine-tuned with MDGD



(b) MiniCPM models pretrained, finetuned, and fine-tuned with MDGD

Figure 4: The effective rank comparison on individual downstream fine-tuning datasets.

4.4 Sensitivity Study

We evaluate the learning curves of MDGD and MDGD-GM compared with standard fine-tuning in Figure 3(a), where we observe that MDGD and MDGD-GM achieve comparable training efficiency compared with the standard fine-tuning method. We also investigate the sensitivity of gradient cosine similarity between $\bar{g}_\theta$ and $\bar{g}_\phi$ in Figure 3(b) and the representation loss in Figure 3(c), with respect to the gradient masking ratio in MDGD-GM. In Figure 3(b), we observe that MDGD-GM with lower gradient masking ratios can better align the modality-decoupled learning gradients between the target model and the pre-trained model, while MDGD-GM maintains over 70% alignment with 50% gradient masking. In Figure 3(c), we show that MDGD-GM with 50% gradient masking still effectively alleviates the visual representation degradation problem by reducing the visual representation discrepancy $\mathcal{L}_v$, while learning with a more active gradient can achieve better alignment.

5 Conclusion

In this work, we addressed the challenge of visual forgetting in MLLMs during instruction tuning by introducing a novel modality-decoupled gradient descent (MDGD) approach. MDGD disentangles the gradient updates for visual representation learning from task-specific alignment, thereby preserving the effective rank of pre-trained visual features and mitigating the over-compression effects highlighted by the information bottleneck perspective. This decoupling enables MLLMs to retain rich visual knowledge while adapting robustly to new downstream tasks. Furthermore, our gradient masking variant, MDGD-GM, enhances memory efficiency and optimizes parameter usage, making fine-tuning both practical and scalable. Extensive experiments across various downstream tasks and backbone models demonstrate that MDGD not only effectively prevents visual forgetting but also outperforms existing strategies in achieving balanced multimodal representation learning and task adaptation. Our findings underscore the importance of preserving visual representations during instruction-tuning and offer a viable solution for efficient and effective multimodal learning in real-world scenarios.

6 Limitation

In this work, we focus on MLLMs that process multimodal instructions consisting solely of visual and textual inputs. Given the limited availability of MLLMs across other modalities, our primary goal is to mitigate visual forgetting. However, our modality-decoupling approach is generalizable to other input modalities. Consistent with standard practices, we limit the instructions to two input modalities, though extending this to more diverse, free-form multimodal inputs remains an avenue for future research.

Acknowledgment

This work is partially supported by NSF IIS-2432486.

References

Alessandro Achille and Stefano Soatto. 2018. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2897–2905.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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*.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, and 1 others. 2023. Language modeling is compression. *arXiv preprint arXiv:2309.10668*.

Xin Dong, Ruize Wu, Chao Xiong, Hai Li, Lei Cheng, Yong He, Shiyou Qian, Jian Cao, and Linjian Mo. 2022. Gdod: Effective gradient descent using orthogonal decomposition for multi-task learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 386–395.

Muhammad Fawi. 2024. Curlora: Stable llm continual fine-tuning and catastrophic forgetting mitigation. *arXiv preprint arXiv:2408.14572*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, and 1 others. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, and 1 others. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

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*.

Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Tong Yu, Ruiyi Zhang, Ryan A Rossi, Branislav Kveton, Dongruo Zhou, and 1 others. 2025. Towards agentic recommender systems in the era of multimodal large language models. *arXiv preprint arXiv:2503.16734*.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multimodal large language models. *arXiv preprint arXiv:2402.14683*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Jingjing Jiang, Ziyi Liu, and Nanning Zheng. 2024a. Correlation information bottleneck: Towards adapting pretrained multimodal models for robust visual question answering. *International Journal of Computer Vision*, 132(1):185–207.

Wenyuan Jiang, Wenwei Wu, Le Zhang, Zixuan Yuan, Jian Xiang, Jingbo Zhou, and Hui Xiong. 2024b. Killing two birds with one stone: Cross-modal reinforced prompting for graph and language tasks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1301–1312.

Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. 2020. Inserting information bottlenecks for attribution in transformers. *arXiv preprint arXiv:2012.13838*.

Congyun Jin, Ming Zhang, Weixiao Ma, Yujiao Li, Yingbo Wang, Yabo Jia, Yuliang Du, Tao Sun, Haowen Wang, Cong Fan, and 1 others. 2024. Rjuameddqa: A multimodal benchmark for medical document question answering and clinical reasoning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5218–5229.

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*.

Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. 2021. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34:19538–19552.

Chen Li, Yixiao Ge, Dian Li, and Ying Shan. 2024. Vision-language instruction tuning: A review and analysis. *Transactions on Machine Learning Research*.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2023a. Finetuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*.

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

Qintong Li, Zhiyong Wu, Lingpeng Kong, and Wei Bi. 2022. Explanation regeneration via information bottleneck. *arXiv preprint arXiv:2212.09603*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zhenqing Ling, Daoyuan Chen, Liuyi Yao, Yaliang Li, and Ying Shen. 2024. On the convergence of zeroth-order federated tuning for large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1827–1838.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Jialin Liu, Jianhua Wu, Jie Liu, and Yutai Duan. 2024c. Learning attentional mixture of loras for language model continual learning. *arXiv preprint arXiv:2409.19611*.

Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. 2024d. Multimodal pretraining, adaptation, and generation for recommendation: A survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6566–6576.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Romain Lopez, Jeffrey Regier, Michael I Jordan, and Nir Yosef. 2018. Information constraints on auto-encoding variational bayes. *Advances in neural information processing systems*, 31.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

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. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134.

Davide Maltoni and Vincenzo Lomonaco. 2019. Continuous learning in single-incremental-task scenarios. *Neural Networks*, 116:56–73.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Junyu Liu, and Benji Peng. 2024. From text to multimodality: Exploring the evolution and impact of large language models in medical practice. *arXiv preprint arXiv:2410.01812*.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2023. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE.

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.

Ashish Shenoy, Yichao Lu, Srihari Jayakumar, Debojeet Chatterjee, Mohsen Moslehpour, Pierce Chuang, Abhay Harpale, Vikas Bhardwaj, Di Xu, Shicong Zhao, and 1 others. 2024. Lumos: Empowering multimodal llms with scene text recognition. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5690–5700.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Yu-Hsiang Tseng, Pin-Er Chen, Da-Chen Lian, and Shu-Kai Hsieh. 2024. The semantic relations in llms: An information-theoretic compression approach. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge)@ LREC-COLING-2024*, pages 8–21.

Chandra Shekhara Kaushik Valmeekam, Krishna Narayanan, Dileep Kalathil, Jean-Francois Chamberland, and Srinivas Shakkottai. 2023. Llmzip: Lossless text compression using large language models. *arXiv preprint arXiv:2306.04050*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023. Orthogonal subspace learning for language model continual learning. *arXiv preprint arXiv:2310.14152*.

Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. 2024. Diff-erank: A novel rank-based metric for evaluating large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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, Xintong Li, Tong Yu, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Jingbo Shang, and Julian McAuley. 2024a. Commit: Coordinated instruction tuning for multimodal large language models. *arXiv preprint arXiv:2407.20454*.

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

Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao, and Ani Nenkova. 2022. Context-aware information-theoretic causal de-biasing for interactive sequence labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3436–3448.

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, 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*.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024e. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.

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*.

Zhou Yang, Zhengyu Qi, Zhaochun Ren, Zhikai Jia, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2025. Exploring information processing in large language models: Insights from information bottleneck theory. *arXiv preprint arXiv:2501.00999*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Cenyuan Zhang, Xiang Zhou, Yixin Wan, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. 2022. Improving the adversarial robustness of nlp models by information bottleneck. *arXiv preprint arXiv:2206.05511*.

Yipeng Zhang, Xin Wang, Hong Chen, Jiapei Fan, Weigao Wen, Hui Xue, Hong Mei, and Wenwu Zhu. 2024. Large language model with curriculum reasoning for visual concept recognition. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6269–6280.

Junhao Zheng, Qianli Ma, Zhen Liu, Binquan Wu, and Huawen Feng. 2024. Beyond anti-forgetting: Multimodal continual instruction tuning with positive forward transfer. *arXiv preprint arXiv:2401.09181*.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2024. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*.

Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Kun Kuang, and Chao Wu. 2024. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. *arXiv preprint arXiv:2402.12048*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Shijie Zhu, Hui Zhao, Pengjie Wang, Hongbo Deng, Jian Xu, and Bo Zheng. 2022a. Gradient deconfliction via orthogonal projections onto subspaces for multi-task learning.

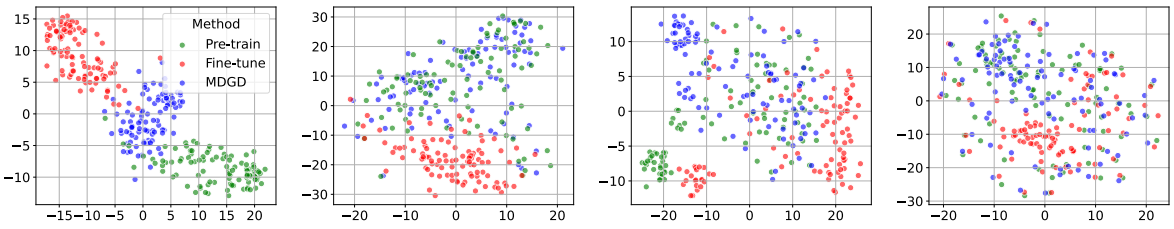
Xianchao Zhu, Tianyi Huang, Ruiyuan Zhang, and William Zhu. 2022b. Wdibs: Wasserstein deterministic information bottleneck for state abstraction to balance state-compression and performance. *Applied Intelligence*, pages 1–14.

A T-SNE Analysis on LLaVA-1.5 Model

In addition to Section 4.3, we further include the T-SNE analysis on LLaVA-1.5 model. We observe a consistent visual understanding drift in the MLLMs' visual representation spaces after standard fine-tuning on Flickr30K and OKVQA with LLaVA (Figure 5b and 5d).

B Implementation Details

Datasets To evaluate the effectiveness of MDGD in mitigating catastrophic forgetting, we used two models of different sizes. Our experimental design follows the settings from the work of Zhu et al. (2024). For each model, datasets were categorized into two types: **pre-trained tasks**, which assess the model's ability to retain inherent knowledge after fine-tuning, and **fine-tuning tasks**, consisting of unseen datasets used to test adaptability. After fine-tuning, we evaluated performance on both task



(a) z^{vl} on **OKVQA** (b) $\pi(X^v)$ on **OKVQA** (c) z^{vl} on **Fllickr30K** (d) $\pi(X^v)$ on **Fllickr30K**

Figure 5: T-SNE plots of the distribution of extracted visual $\pi(X^v)$ and multimodal z^{vl} representations from pre-trained LLaVA-1.5, and models with direct fine-tuning and MDGD on OKVQA and Flickr30K.

types to measure forgetting and generalization. Below, we detail the datasets used for each model. **LLaVA-1.5 (Vicuna-7B) (Liu et al., 2024a)**: This model has 7 billion parameters. In line with Liu et al. (2024a), we used the following datasets:

- **Pre-trained Tasks**: VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), VizWiz (Gurari et al., 2018), SQA (Lu et al., 2022), TextVQA (Singh et al., 2019), POPE (Li et al., 2023c), and MM-Bench (Liu et al., 2023).
- **Fine-tuning**: Flickr30k (Young et al., 2014) and OKVQA (Marino et al., 2019), which were not encountered in the pre-training stage.

MiniCPM-V-2.0 (Yao et al., 2024): This model has 2.8 billion parameters. We evaluated its performance on:

- **Pre-trained Tasks**: VizWiz, OKVQA, A-OKVQA (Schwenk et al., 2022), Text-VQA, IconQA (Lu et al., 2021), POPE, and MM-Bench.
- **Fine-tuning**: TextCaps (Sidorov et al., 2020) and PathVQA (He et al., 2020), which were not part of its pre-training exposure.

Baselines We compare our approach against several baselines:

- **Standard Fine-Tuning.** For a fair comparison, we follow the setting of Model-Tailor (Zhu et al., 2024), where LLaVA-1.5 is fine-tuned on the last 6 layers and its feature adapter, with a total of 1.2B parameters. MiniCPM is fine-tuned on the last 8 layers and its feature resampler, with 517M parameters.
- **LoRA-based Fine-Tuning (Hu et al., 2021).** LoRA introduces low-rank matrices to update only a small subset of parameters, reducing memory consumption and computational cost. In our experiments, LLaVA-1.5 and MiniCPM are fine-tuned by modifying the query and key projection layers within the attention mechanism.
- **Model Tailor (Zhu et al., 2024).** This baseline employs a hybrid strategy that mitigates catastrophic forgetting by identifying and adjusting the most critical parameters for adaptation. It has been evaluated through experiments on multimodal large language models (MLLMs). As the method is not open source, we report only the original results of the LLaVA-1.5 experiments provided in the original paper as a baseline.

Implementation Details We use the official Huggingface implementations of the LLaVA-1.5 and the MiniCPM-V-2.0 models and their LoRA adapters. For model fine-tuning, we use BFloat16 precision for memory-efficient training. Experiments are conducted using 2 NVIDIA A100-SXM4-80GB GPUs.

C Visual Forgetting in MLLM Instruction-tuning

Building on the IB objective Eq. (3) introduced in Section 2, we examine how instruction tuning affects the richness of visual representations. Let the pre-trained MLLM induce a latent representation,

$$Z \sim p(\cdot \mid X^v, X^l),$$

where Z is is decomposed into modality-specific components, $Z = (Z^v, Z^l)$ with Z^v captures the visual features extracted from X^v, and Z^l encapsulates the textual features from X^l. Define the *pre-trained* visual representation space as,

$$\mathcal{Z}_0^v = \left\{ Z_\phi^v : Z \sim p_\phi(\cdot \mid X^v), \quad X^v \in \Omega \right\}.$$

During instruction tuning, the model is optimized primarily to predict the target y. As described in Eq. (3), the IB objective introduces a trade-off between retaining visual information $I(X^v; Z)$ and ensuring that Z remains predictive of y via $I(y; Z)$ (Jiang et al., 2024a). In practice, however, instruction-tuning datasets are predominantly text-driven; thus, the learned visual representation Z^v receives only indirect and often weaker supervision (Wang et al., 2024).

Let the tuned model's latent representation be $Z_\theta \sim p_\theta(\cdot \mid X^v, X^l)$, and denote the corresponding visual representation space by,

$$\mathcal{Z}_\theta^v = \left\{ Z_\theta^v : Z \sim p_\theta(\cdot \mid X^v, X^l), (X^v, X^l) \in D \right\},$$

where D is the instruction-tuning dataset. To measure the richness of the visual representation, we employ the effective rank metric from Eq. (4). A higher effective rank indicates that the representation spans a broader subspace, whereas a lower effective rank signals more aggressive compression.

The Visual Forgetting Problem. During instruction tuning, the visual representation undergoes significant compression as the model prioritizes textual supervision. This reduction occurs because the model effectively sacrifices part of $I(X^v; Z)$ to focus on $I(y; Z)$, thereby reducing the effective dimensionality of the visual features. As a result, the model progressively loses its ability to retain and utilize rich visual information, leading to a phenomenon we define as ***visual forgetting***. Empirically, in Figure 1 we observe,

$$\text{erank}(\mathcal{Z}_\theta^v) < \text{erank}(\mathcal{Z}_0^v). \tag{12}$$

This indicates that the tuned visual representation is compressed relative to the pre-trained space, making it harder for the model to leverage visual information effectively. In RQ3 (Section 4.3), we validate such empirical observations and demonstrate that our method helps to preserve effective ranks in the visual representation learning of MLLMs.