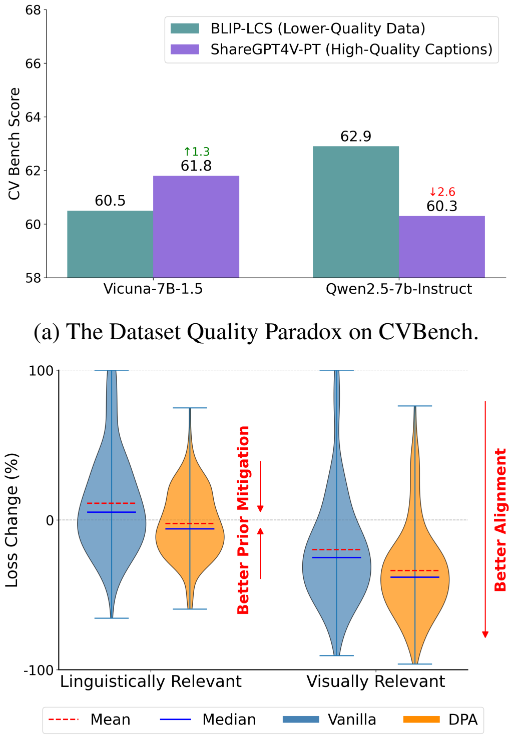
Decoupled Proxy Alignment: Mitigating Language Prior Conflict for Multimodal Alignment in MLLMs

Anonymous ACL submission

Abstract

Multimodal large language models (MLLMs) have gained significant attention due to their impressive ability to integrate vision and language modalities. Recent advancements in MLLMs have primarily focused on improving performance through high-quality datasets, novel architectures, and optimized training strategies. However, in this paper, we identify a previously overlooked issue, **language prior conflict**, a mismatch between the inherent language priors of large language models (LLMs) and the language priors in training datasets. This conflict leads to suboptimal vision-language alignment, as MLLMs are prone to adapting to the language style of training samples. To address this issue, we propose a novel training method called **Decoupled Proxy Alignment (DPA)**. DPA introduces two key innovations: (1) the use of a proxy LLM during pretraining to decouple the vision-language alignment process from language prior interference, and (2) dynamic loss adjustment based on visual relevance to strengthen optimization signals for visually relevant tokens. Extensive experiments demonstrate that DPA significantly mitigates the language prior conflict, achieving superior alignment performance across diverse datasets, model families, and scales. Our method not only improves the effectiveness of MLLM training but also shows exceptional generalization capabilities, making it a robust approach for vision-language alignment.



(a) The Dataset Quality Paradox on CVBench.

(b) Loss change (%) for linguistically relevant and visually relevant words before and after training.

Figure 1: In this paper, we identify the issue of language prior conflict. Figure 1a illustrates that datasets considered "high-quality" for one model may negatively affect another due to language prior conflict. Figure 1b shows that DPA enables models to focus more on vision-text alignment rather than overfitting to language priors in the training data. See Section 5.1 for more analysis.

1 Introduction

After the significant success of large language models (LLMs) (Dubey et al., 2024; Yang et al., 2024a), numerous efforts have been made to leverage the powerful language understanding capabilities of LLMs to construct multimodal large language models (MLLMs). Many recent studies are centered around enhancing the performance of MLLMs, which can be divided into three categories: (1) introducing high-quality datasets (Chen et al., 2024a; Deitke et al., 2024), (2) improving model architecture design (Tong et al., 2024a; Dubey et al., 2024), and (3) optimizing training strategies (Xiao et al., 2024; Chen et al., 2024b). These approaches are both distinctive and complementary, collectively driving the development of MLLMs and significantly improving their performance across diverse tasks.

Despite these efforts, our research has uncovered a critical issue during the training of MLLMs: a sig-

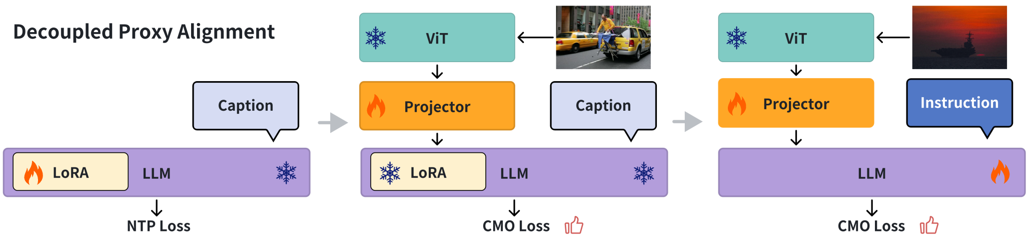


Figure 2: Illustration of Decoupled Proxy Alignment (DPA). From left to right: Proxy LLM Pretraining, Proxy MLLM Pretraining , and MLLM Instruction Tuning. See Section 4.3 for details.

nificant mismatch between the inherent language priors of LLMs and the language priors present in the training datasets. This mismatch causes MLLMs to adapt to the language style of the training samples, which compromises vision-language alignment and results in suboptimal performance. We term this phenomenon as **language prior conflict**, a challenge that has **not** been effectively addressed in existing methods. Consequently, there is an urgent need for a more effective multimodal training approach that mitigates the interference of language prior conflict and enhances the alignment between visual and language modalities.

To address this challenge, we propose a novel method called **Decoupled Proxy Alignment (DPA)**. The core idea of DPA is to decouple the vision-language alignment process from the interference caused by language prior conflicts. Specifically, DPA integrates two key components: (1) During the pretraining phase, a proxy LLM is introduced to mitigate the impact of language prior conflict, ensuring a less biased alignment process. (2) Throughout training, the loss weights are dynamically adjusted to strengthen the optimization signals for visually relevant tokens, rather than those related to linguistic style, further enhancing vision-language alignment.

Experimental results demonstrate that DPA effectively mitigates language prior conflicts and significantly outperforms baseline methods across various model families and training datasets. Furthermore, DPA exhibits strong generalization capabilities, achieving consistently superior performance across datasets and models of varying scales.

Our contributions can be summarized as follows:

- We are the first to define and investigate the issue of language prior conflict in MLLMs, experimentally verifying its negative impact on vision-language alignment.
- We introduce DPA, a three-stage training method that effectively mitigates language prior conflict and enhances vision-language alignment.
- Through extensive experiments, we validate the effectiveness of DPA, demonstrating significant improvements in alignment performance and outstanding generalization capabilities.

2 Related Work

Multimodal Large Language Models Multimodal large language models (MLLMs) have achieved significant advancements in visual understanding, progressing from basic image captioning to complex visual reasoning tasks (Li et al., 2024b; Team, 2025). These models typically combine a pretrained vision encoder (Radford et al., 2021; Zhai et al., 2023) with a pretrained language model (Touvron et al., 2023; Yang et al., 2024a), integrating the two modalities through connectors such as multi-layer perceptrons (MLPs) (Liu et al., 2024b,a) or cross-attention modules (Dai et al., 2023; Dubey et al., 2024).

A widely adopted training strategy for MLLMs is the two-stage visual instruction tuning framework, first proposed by Liu et al. (2024b). This methodology has been validated by subsequent studies (Agrawal et al., 2024; McKinzie et al., 2024; Tong et al., 2024a). While recent advancements, such as ShareGPT4V (Chen et al., 2023) and InternVL (Chen et al., 2024b), have introduced more sophisticated training protocols, they are fundamentally built upon the two-stage framework. Given its demonstrated efficacy, our study also adopts this two-stage training method.

Language Prior The concept of Language prior refers to the unique linguistic characteristics that LLMs develop during training. These characteristics include distinct language patterns, styles, vocabularies, grammatical preferences, and implicit world knowledge. Language prior has already attracted significant attention in LLM research. Li et al. (2023); Wang et al. (2023) have demonstrated that models with similar language priors exhibit strong behavioral correlations in their predictions, enabling model tracing. Furthermore, Yang et al. (2024b); Wang et al. (2024) have shown that conflicts in language priors can lead to the forgetting of a model's original knowledge and capabilities.

In the context of MLLMs, language prior introduces two key challenges: **(1)** MLLMs are prone to capturing spurious correlations present in multimodal training data (Agarwal et al., 2020; Goyal et al., 2017). **(2)** MLLMs often rely disproportionately on textual prediction, which diminishes their dependence on the visual modality (Leng et al., 2024). These challenges have significant implications for the performance of multimodal models.

Image-text Modality Alignment Image-text modality alignment has long been regarded as a core challenge in multimodal understanding. Traditional approaches to image-text alignment often involve training multimodal models from scratch using strategies such as contrastive learning or autoregressive learning (Radford et al., 2021; Lin et al., 2024). In recent years, researchers have made significant strides by leveraging larger and higher-quality datasets, leading to notable advancements in cross-modal alignment (Chen et al., 2023; Deitke et al., 2024). However, these methods often come at the cost of substantial human and computational resources. More recently, Xiao et al. (2024) proposed CAL, which improves alignment by dynamically adjusting the importance of different tokens during the alignment process, achieving superior results. Despite these improvements, the underlying mechanisms driving these gains remain largely unexplored. In this study, we present a comprehensive analysis of the conflicts between the language priors in training data and those inherent to LLMs. Furthermore, we propose a novel method designed to effectively mitigate these conflicts.

3 Language Prior Conflict

In this section, we analyze how language prior conflict impedes the alignment training of multimodal models and ultimately degrades their performance. In Section 3.1, we start by defining language prior conflict, followed by an exploration of its causes and potential adverse effects. In Section 3.2, we present two quantitative experiments to comprehensively demonstrate the impact of language prior conflict on MLLMs.

3.1 Causes and Consequences

Language prior conflict refers to the mismatch between the inherent language priors of LLMs and those present in their training datasets. This phenomenon is particularly pronounced in the training of multimodal models. LLMs are typically trained on diverse, large-scale text corpora that cover a wide range of topics and styles. In contrast, image-caption datasets (Chen et al., 2024a; Deitke et al., 2024) primarily focus on objective descriptions of visual scenes, often generated by advanced models or through human annotation. These datasets exhibit linguistic distributions that differ significantly from the data used to pretrain LLMs.

During the pretraining phase of MLLMs, the model may prioritize minimizing training loss by adapting to the language style of the training samples rather than focusing on image-text alignment. This prioritization can even lead to severe conflicts, such as overfitting to the style and knowledge contained in the training dataset. In the following section, we present experimental evidence demonstrating the impact of language prior conflict on the performance of MLLMs.

3.2 Negative Impacts

3.2.1 Dataset Quality Paradox

A surprising discovery in MLLM training is that datasets regarded as "high-quality" for one model may negatively impact another due to conflicts in language priors. We refer to this phenomenon as **dataset quality paradox**. To explore this further, we conducted a comparative study using two LLM backbones and two image-caption datasets. More details can be seen in Appendix A.1.

The experimental results are presented in Figure 1a. Vicuna-7B-1.5, trained on the high-quality dataset (ShareGPT4V-PT), demonstrates superior performance compared to the model trained on BLIP-LCS. This aligns with the expectation that high-quality data enhances performance. However, for Qwen2.5-7B-Instruct, training on the high-quality dataset led to a performance decline.

This discrepancy is attributed to a significant conflict between the language priors of Qwen2.5-7B-Instruct and the ShareGPT4V-PT dataset. Specifically, Qwen2.5-7B-Instruct's advanced language capabilities may cause it to overly focus on textual content in the high-quality dataset, especially when processing lengthy captions, while underutilizing visual information. In contrast, Vicuna-7B-1.5 benefits from its extensive training on open-source GPT-4 distilled data, resulting in linguistic priors that are more compatible with the ShareGPT4V-PT dataset. Additionally, Vicuna's relatively weaker language capabilities reduce the risk of overfitting on complex captions, encouraging a greater reliance on visual features. This enables Vicuna-7B-1.5 to more effectively learn the correspondence between images and text.

3.2.2 Quantitative Analysis

To further investigate the underlying reasons for the performance drop observed when training Qwen2.5-7B-Instruct on ShareGPT4V-PT, we conduct a detailed quantitative analysis of word-level loss changes during training. Specifically, we examine how the model's loss on linguistically relevant and visually relevant words evolves before and after training on both BLIP-LCS and ShareGPT4V-PT. The detailed experimental setup is provided in Appendix A.2.

As shown in the figure 3, for BLIP-LCS, the loss change of linguistically relevant words centers around zero, indicating that the model does not overfit these words. In contrast, for ShareGPT4V-PT, the loss change for linguistically relevant words fluctuates dramatically due to overfitting of high-frequency words and increased loss for others. For visually relevant words, BLIP-LCS leads to a consistent loss decrease, reflecting effective visual-text alignment, while ShareGPT4V-PT shows an increase in loss, suggesting that language prior conflict hinders multimodal alignment. These results highlight the negative effect of language prior conflict on MLLM training with Qwen2.5-7B-Instruct.

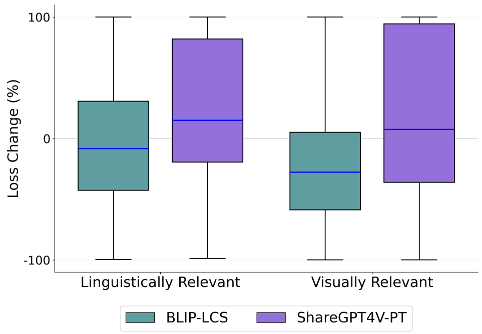


Figure 3: Word-level loss change for linguistically relevant and visually relevant words after training Qwen2.5-7B-Instruct on BLIP-LCS and ShareGPT4V-PT. The results highlight the negative impact of language prior conflict on multimodal alignment.

4 Methodology

In this section, we detail our three-stage training framework. First, we introduce Proxy Model Optimization (PMO) to mitigate language prior conflicts during pretraining via a proxy LLM. Next, we present Contrastive Modality Optimization (CMO), which enhances visual-language alignment through token reweighting. Finally, we combine these strategies into our **Decoupled Proxy Alignment (DPA)** method, effectively addressing the key challenges discussed at the end of Section 3.2.

4.1 Proxy Model Optimization

As described in Section 3.2, during the pretraining phase, when the language priors of LLMs conflict with those of an image-caption dataset, the model tends to focus on textual information while insufficiently utilizing visual information. To address this issue, we propose a novel approach called Proxy Model Optimization (PMO).

The PMO methodology involves a two-stage training process. (1) We train the LLM exclusively on the textual portion of the image-caption dataset. This aligns the LLM's language priors with the linguistic style and characteristics of the dataset, resulting in a dataset-adapted LLM, referred to as the Proxy LLM. (2) We construct a Proxy MLLM by integrating the Proxy LLM, which is kept frozen during subsequent training on the full image-caption dataset. Since the Proxy LLM has already captured the language priors of the training data, the model can focus more effectively on vision-language alignment rather than relearning language priors. This two-stage training process effectively resolves the issue of language prior conflicts during the pretraining phase.

To further optimize the training of the Proxy LLM, we introduce Low-Rank Adaptation (LoRA), a method that offers significant advantages in mitigating catastrophic forgetting and reducing computational overhead. On the one hand, directly training the Proxy LLM on the dataset's textual data may lead to catastrophic forgetting of the original LLM's pretrained knowledge and general language capabilities. LoRA addresses this by freezing the weights of the original LLM and training only a small number of low-rank matrices, enabling the Proxy LLM to adapt to the language priors of the dataset while retaining the core knowledge of the base LLM. On the other hand, compared to full fine-tuning, LoRA saves a significant number of trainable parameters, drastically lowering the computation costs and the training time.

4.2 Contrastive Modality Optimization

To further enhance the alignment between the visual and language modalities during training stage, we propose contrastive modality optimization (CMO). The primary motivation behind CMO lies in the varying degrees of visual relevance among tokens in the captions. For example, compared to nouns, verbs, and adjectives that directly describe visual attributes, function words, discourse markers, and expansive descriptions may contribute less to visual alignment. Treating all tokens equally in loss computation may therefore not be optimal. The design goal of CMO is to dynamically adjust the loss weights to strengthen the optimization signals for visually relevant tokens, rather than those related to linguistic style.

CMO achieves this dynamic token weighting through a contrastive method. During training, for each token in the caption, CMO estimates its visual relevance by comparing the predicted probabilities of the token in two scenarios: (1) when the multimodal context with the input image is provided and (2) when only the textual context is provided. Notably, we directly adjust the weights based on probabilities, which is simpler and more effective compared to previous works such as CAL (Xiao et al., 2024) that rely on logits. Through this contrastive approach, CMO can further decouple the influence of language priors, thus re-evaluating the relevance of the current token to the visual input. Intuitively, tokens with higher visual relevance will exhibit greater differences in predicted probabilities between the scenarios with and without image input. As a result, CMO can effectively capture and amplify the visual alignment signals of these tokens. The detailed algorithm is depicted in Algorithm 1.

Algorithm 1 Detail Procedure of $\mathcal{L}_{CMO}$

```
Input: Visual input $V_i$, Textual sequence $S = \{s_1, s_2, \ldots, s_m\}$, Model distribution $D_\phi$
  Extract probability vectors:
      $\mathbf{r}^{i,j} = D_\phi(V_i, S^{i,<j}), \mathbf{q}^{i,j} = D_\phi(S^{i,<j})$
  Calculate differential score $\delta_s[s_j]^{i,j}$:
      $\delta_s[s_j]^{i,j} = \mathbf{r}[s_j]^{i,j} - \mathbf{q}[s_j]^{i,j}$
  Transform scores to weights through normalization:
      $\omega'_{i,s_j} = \text{aggregate}_\Omega(\text{clip}(\delta_s[s_j]^{i,j}, \alpha, \beta))$
      $\omega_{i,s_j} = \frac{\omega'_{i,s_j}}{\sum_{k=1}^{m} \omega'_{i,s_k}}$
  Formulate final loss through token weighting:
      $\mathcal{L}_{CMO} = -\sum_{i=1}^{N} \sum_{j=1}^{m} \omega_{i,s_j} \log D_\phi(s_j | V_i, S^{i,<j})$
Output: Optimized model parameters $\phi^*$
```

4.3 Decoupled Proxy Alignment

The overall methodology comprises three stages:

- **Proxy LLM Pretraining**: We train the LLM solely on the text portion of the image-caption dataset to obtain a Proxy LLM that is adapted to the language priors of the dataset.
- **Proxy MLLM Pretraining**: The Proxy LLM is integrated with vision encoder and visual connector layer to construct Proxy MLLM. Then, the complete image-caption dataset is used to train the Proxy MLLM with CMO. In this stage, only the connector layer is trainable.
- **MLLM Instruction Tuning**: The final MLLM is constructed by combining the original LLM, the pretrained connector layer, and the vision encoder. It then undergoes instruction tuning using CMO, during which both the connector layer and the LLM are trainable.

This approach aims to decouple the visual-language alignment process from the potential interference of the language prior conflicts. Additionally, a proxy model is introduced in Stage 1 to enhance the alignment process. Therefore, this method is referred to as **Decoupled Proxy Alignment**.

5 Experiments

In this section, we evaluate the performance of our models through comparative analysis on a variety of visual benchmarks, demonstrating the advantages of our approach.

Please refer to Appendix A.5 for detailed experimental setup, including datasets, evaluation metrics, baselines, and implementation details.

5.1 Main Results

Compared to Vanilla As shown in Table 1, after incorporating DPA, MLLMs trained on diverse pretraining data consistently demonstrated significant performance improvements compared to Vanilla, indicating that DPA effectively mitigates the prevalent language prior conflicts existing between different pretraining datasets and different LLMs. **Notably**, the MLLM trained with DPA on Llama-3.1-8B-Instruct using PixMo-Cap (a high-diversity dataset with multiple expert annotations) achieved an average improvement of 2.8 points compared to Vanilla. This **highlights** that even with high-quality, well-aligned annotation data, language prior conflicts still exists. By decoupling language priors and modality alignment processes, DPA effectively

Dataset	Method	General				Vision-centric				Avg.
		MMB.	OCRB.	DocVQA	AI2D	CV-2D	CV-3D	MMVP	NaB.	
					Qwen2.5-7B-Instruct					
BLIP-LCS	Vanilla	74.5	341	33.3	72.8	59.7	66.1	**40.0**	18.8	49.9
	CAL	74.5	**356**	34.2	**74.0**	58.4	67.1	32.0	18.4	49.3
	DPA	**75.8**	345	**34.7**	72.2	**63.5**	**68.7**	38.7	**20.5**	**51.1**
ShareGPT4V-PT	Vanilla	73.1	365	34.7	72.9	58.5	62.1	36.0	18.5	49.0
	CAL	75.7	367	35.3	**74.7**	60.0	**70.2**	38.0	17.8	51.0
	DPA	**76.1**	**368**	**36.0**	74.1	**61.1**	68.1	**40.7**	**20.4**	**51.7**
PixMo-Cap	Vanilla	76.4	377	37.7	74.7	61.5	**70.4**	41.3	19.8	52.4
	CAL	76.4	392	38.0	**75.0**	61.8	69.9	42.7	19.9	52.9
	DPA	**77.0**	**404**	**38.2**	**75.0**	**67.9**	65.4	**45.3**	**20.7**	**53.7**
					Llama-3.1-8B-Instruct					
BLIP-LCS	Vanilla	70.4	319	30.1	**65.4**	58.5	59.2	28.0	13.7	44.6
	CAL	70.4	329	**30.6**	65.1	58.9	67.6	28.0	**16.8**	46.3
	DPA	**70.9**	**336**	**30.6**	65.3	**61.4**	**68.5**	**28.7**	15.4	**46.8**
ShareGPT4V-PT	Vanilla	69.4	**350**	**32.3**	65.8	62.2	62.2	33.3	15.6	47.0
	CAL	68.3	344	31.7	66.7	57.4	65.5	32.7	10.9	46.0
	DPA	**71.5**	346	32.2	**66.9**	**66.1**	**71.8**	**39.3**	**16.6**	**49.9**
PixMo-Cap	Vanilla	68.4	347	33.1	66.9	60.3	60.6	37.3	16.1	47.2
	CAL	71.1	**361**	**34.7**	**67.4**	62.3	66.8	**38.0**	17.8	49.3
	DPA	**72.3**	349	34.4	67.2	**63.7**	**71.4**	37.3	**18.6**	**50.0**
					Gemma-2-9B-it					
BLIP-LCS	Vanilla	72.5	334	31.0	67.6	59.4	60.9	**28.7**	15.3	46.1
	CAL	71.9	**340**	31.2	67.6	58.3	**62.8**	**28.7**	15.7	46.3
	DPA	**72.9**	335	**31.6**	**67.7**	**62.7**	61.5	25.3	**15.8**	**46.4**
ShareGPT4V-PT	Vanilla	72.9	354	34.2	68.4	64.0	60.5	33.3	16.7	48.2
	CAL	71.7	362	34.3	67.6	60.4	**61.8**	30.0	17.4	47.4
	DPA	**74.3**	**377**	**35.5**	**69.6**	**64.8**	58.7	**38.7**	**21.1**	**50.0**
PixMo-Cap	Vanilla	74.3	364	37.0	70.7	64.1	65.2	**45.3**	19.9	51.6
	CAL	**75.3**	367	37.4	**71.0**	65.0	66.8	34.7	19.8	50.8
	DPA	74.7	**383**	**37.5**	70.4	**65.6**	**70.8**	42.7	**21.4**	**52.7**

Table 1: Evaluation results of baselines and DPA. The best performances within each setting are **bolded**. Abbreviations: MMB.(MMBench), OCRB.(OCRBench), CV-2D(CVBench-2D), CV-3D(CVBench-3D), NaB.(NaturalBench).

suppresses the language-dominated overfitting tendency. This fully validates the effectiveness of our method in multimodal alignment.

Compared to CAL When compared with the CAL method, DPA also consistently achieves better performance across all mainstream models and datasets. For instance, using the Llama-3.1-8B-Instruct with ShareGPT4V-PT, the DPA method achieves a score of 16.6 on the NaturalBench benchmark, surpassing CAL by 5.7 points. Furthermore, DPA's average performance across metrics exceeds CAL's by 3.9 points.

Notably, CAL exhibits **inferior** performance compared to even the Vanilla method in several configurations (e.g., Llama-3.1-8B-Instruct with ShareGPT4V-PT). This phenomenon reveals that simply adjusting the loss weight of visually-related tokens is **insufficient** to decouple language priors from the modality alignment process. Interference from language priors disrupts the optimization trajectory of modality alignment, ultimately leading to performance degradation. These results highlight the unique advantages of DPA in harmonizing language priors with multimodal alignment.

Word-level Loss Analysis To further illustrate how DPA alleviates language prior conflict, we conduct a word-level loss analysis by tracking the loss changes for "Linguistically Relevant" and "Visually Relevant" words before and after training with either the Vanilla or DPA method (see Appendix A.3 for details). As illustrated in Figure 1b, DPA substantially reduces loss fluctuations for linguistically relevant tokens compared to Vanilla, indicating a lower tendency to overfit linguistic styles. At the same time, DPA achieves greater loss reductions for visually relevant tokens, signifying improved visual-text alignment. These results demonstrate that DPA effectively re-prioritizes optimization, suppressing language prior overfitting and enhancing multimodal alignment. This also

Method	General Avg.	Vision. Avg.
Vanilla	54.3	43.8
+ PMO	55.4	46.1
+ CMO	55.7	47.2
DPA	**55.8**	**47.6**

Table 2: Ablation study on DPA's components.

explains why DPA excels even when training on high-quality datasets, overcoming the dataset quality paradox observed with conventional methods.

Conclusion In summary, DPA significantly outperforms both Vanilla and CAL in alleviating language prior conflicts and improving overall multimodal performance. The word-level loss analysis further demonstrates that DPA re-prioritizes optimization, effectively suppressing overfitting to linguistic priors while enhancing visual-text alignment. These results validate the generalizability and robustness of DPA, providing a superior and principled solution for multimodal alignment.

6 Analysis

In this section, we first verify the effectiveness of each component of DPA in multimodal alignment. We then evaluate DPA across different model scales and data sizes. Finally, we analyze the impact of various reweighted loss strategies on multimodal alignment. The detailed experimental setup is provided in Appendix A.4.

6.1 Ablations Studies

Component Analysis To evaluate the effectiveness of the core components in the DPA framework, we systematically ablated PMO and CMO to train different models. As shown in Table 2, combining either PMO or CMO with the Vanilla model improves performance on both general benchmarks and vision-centric benchmarks. Notably, CMO achieves greater improvements (from 54.3 to 55.7 on General benchmarks, and from 43.8 to 47.2 on vision-centric benchmarks) compared to PMO (from 54.3 to 55.4 on General benchmarks, and from 43.8 to 46.1 on vision-centric benchmarks), as CMO enhances modality alignment during both pretraining and instruction tuning, whereas PMO only impacts pretraining. Furthermore, combining both PMO and CMO yields additional performance gains. Specifically, as shown in the detailed tables in the appendix B.6, DPA significantly outperforms models with only CMO or PMO on benchmarks like MMVP and MMBench.

Stages of conducting CMO within MLLM Training CMO can be integrated into both the Pre-Training (PT) stage and the Instruction Tuning (IT) stage in existing MLLMs. In this section, we investigate which stage benefits the most from CMO in Table 11. Our experimental analysis reveals distinct advantages of CMO integration across different training phases: The instruction tuning (IT) stage contributes the majority of performance gains across all evaluated benchmarks, while pretraining (PT) stage integration further enhances model capabilities, particularly demonstrating marked improvements on vision-centric benchmarks.

We present more ablation experiments in the Appendix B.1 These include the necessity of LoRA in PMO, the selection of its rank, whether the third stage of DPA continues to use the proxy LLM, as well as ablation studies on the hyper-parameters $[\alpha, \beta]$ in the clamping operation.

6.2 Results Across Different Model Scales

Scales	Method	General Avg.	Vision. Avg.
1.5B	Vanilla	48.2	36.5
	DPA	49.8	42.0
3B	Vanilla	51.5	42.4
	DPA	53.3	43.3
14B	Vanilla	55.6	46.4
	DPA	55.2	49.4
32B	Vanilla	58.2	54.6
	DPA	58.0	56.3

Table 3: Performance across different LLM scales.

Scales	Method	General Avg.	Vision. Avg.
5%	Vanilla	1.5	13.7
	DPA	0.4	10.0
10%	Vanilla	1.6	14.3
	DPA	0.1	13.5
25%	Vanilla	32.6	32.6
	DPA	39.5	38.0
50%	Vanilla	38.9	41.8
	DPA	42.3	44.1
100%	Vanilla	39.2	46.7
	DPA	42.5	49.1

Table 4: Performance across different data size.

To verify that DPA is applicable to LLMs of different scales, we further trained MLLMs based on LLMs with various parameter scales and evaluated them on multimodal benchmarks. As shown in Table 3, models of all scales exhibit a stable performance improvement trend on vision-centric benchmarks. For smaller-scale models (e.g., 1.5B), performance improved from 36.5 to 42.0. For larger-scale models (e.g., 14B), performance increased from 46.6 to 49.4. For 32B models, performance rose from 54.6 to 56.3. Considering that unified training hyperparameters were used in the experiments, further adjustments could lead to additional improvements. This phenomenon strongly demonstrates that DPA has impressive adaptability to LLM scales, and its optimization effect is not significantly affected by the number of LLM parameters.

On general benchmarks, performance showed a slight decline as the LLM scale increased. This may be due to the fact that as the parameter count grows, LLMs are more likely to overfit to the linguistic priors from caption data, causing interference from these linguistic priors during inference. In contrast, all the vision-centric tasks are multiple-choice questions, where inference is less affected by such interference, resulting in no performance decline in the metrics.

6.3 Results Across Different Data Size

To investigate the impact of data scale on the performance of MLLMs trained with the DPA method, we conducted experiments by adjusting the data volumes for pretraining (PT) and instruction fine-tuning (IT). Specifically, we trained the models using different subsets of the ShareGPT4V-PT and Cambrian-1 datasets.

Table 4 indicate that data scale is critical to the effectiveness of the DPA method. When the data size is relatively small ($\leq$10%), the performance of the DPA method is lower than that of the baseline model. This is primarily due to insufficient data, which hinders the Proxy LLM from decoupling language priors and limits the MLLM's ability to assess visual relevance. However, when the data volume reaches 25% or more, the performance of the baseline model improves significantly. The DPA method further enhances modality alignment, leading to additional performance improvements. As the data scale continues to increase, the DPA method provides even greater improvements in modality alignment and overall performance.

In summary, while the DPA method is limited in effectiveness with small data scales, it demonstrates significant advantages with larger data scales, making it highly valuable for improving the performance of multimodal models.

6.4 Resutls Across Different Reweighted Loss

To further validate the effectiveness and generalization of our proposed CMO, we conduct a direct comparison with CAL under the same training strategy (PMO) across multiple model backbones. As shown in Table 5, CMO consistently achieves the best performance on all models, while CAL sometimes even underperforms the baseline. This suggests that CAL's performance is highly sensitive to the underlying model, likely due to its reliance on model-specific logits distributions. In contrast, CMO demonstrates strong robustness and generalization, benefiting from its probability-based design. These results highlight the practical advantage of CMO for multimodal model training across diverse architectures.

Method	General. Avg.	Vision. Avg.	Avg.
Qwen2.5-7B-Instruct			
PMO	55.4	46.0	50.7
PMO + CAL	**56.0**	46.9	51.5
PMO + CMO (Ours)	55.8	**47.6**	**51.7**
Llama-3.1-8B-Instruct			
PMO	49.2	42.4	45.8
PMO + CAL	50.3	44.1	47.2
PMO + CMO (Ours)	**51.3**	**48.5**	**49.9**
Gemma-2-9B-it			
PMO	53.7	44.7	49.2
PMO + CAL	53.5	44.3	48.9
PMO + CMO (Ours)	**54.3**	**45.8**	**50.0**

Table 5: Comparison between our proposed CMO loss and CAL loss when combined with PMO. All models are trained on ShareGPT4V-PT dataset. The best performances within each setting are **bolded**.

7 Conclusion

In this paper, we introduced the concept of language prior conflict and proposed a novel method, Decoupled Proxy Alignment (DPA), to effectively address this challenge and enhance the alignment between visual and language modalities. Extensive experiments demonstrate that DPA significantly reduces the negative impact of language prior conflict, achieving superior alignment performance across a wide range of datasets, model families, and scales. It not only enhances the training efficiency of MLLMs but also shows exceptional generalization capabilities, making it a robust approach for vision-language alignment.

Limitations

While our proposed DPA demonstrates significant improvements in mitigating language prior conflicts and enhancing vision-language alignment, certain limitations remain. Specifically, the selection of lower and upper bounds in CMO process is currently determined empirically, which could be extended to more adaptive settings in further explorations.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. 2021. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688.

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.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024a. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024b. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Linyang Li, Pengyu Wang, Ke Ren, Tianxiang Sun, and Xipeng Qiu. 2023. Origin tracing and detecting of llms. *arXiv preprint arXiv:2304.14072*.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. 2024. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*.

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.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. 2023. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. 2024. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Babak Saleh and Ahmed Elgammal. 2015. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2025. Qwen2.5-vl.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

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

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. Seqxgpt: Sentence-level ai-generated text detection. *arXiv preprint arXiv:2310.08903*.

Rui Wang, Fei Mi, Yi Chen, Boyang Xue, Hongru Wang, Qi Zhu, Kam-Fai Wong, and Ruifeng Xu. 2024. Role prompting guided domain adaptation with general capability preserve for large language models. *arXiv preprint arXiv:2403.02756*.

Xin Xiao, Bohong Wu, Jiacong Wang, Chunyuan Li, Xun Zhou, and Haoyuan Guo. 2024. Seeing the image: Prioritizing visual correlation by contrastive alignment. *arXiv preprint arXiv:2405.17871*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024b. Self-distillation bridges distribution gap in language model fine-tuning. *arXiv preprint arXiv:2402.13669*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

A Experimental Details

A.1 Dataset Quality Paradox

To explore the dataset quality paradox, we conducted a comparative study using two LLM backbones:

- **Vicuna-7B-1.5** A relatively weaker model in text generation.
- **Qwen2.5-7B-Instruct** A model with strong text generation capabilities.

and two image-caption datasets:

- **BLIP-LCS** A noisier dataset with shorter captions, commonly used in LLaVA-1.5 pretraining.
- **ShareGPT4V-PT** A high-quality dataset featuring longer, more detailed captions generated by GPT-4.

Both models were fine-tuned based on the LLaVA-1.5 architecture under consistent experimental settings and hyperparameters (Liu et al., 2024b). The Cambrian-1 dataset was utilized as the instruction-tuning dataset. Performance evaluation was conducted using CVBench, a vision-centric benchmark specifically designed to account for sensitivity to language priors.

A.2 Analysis of Word-Level Loss

To further investigate the impact of language prior conflict on MLLM training, we conducted a word-level loss analysis based on the fine-tuning experiments of Qwen2.5-7B-Instruct on BLIP-LCS and ShareGPT4V-PT.

Specifically, we randomly sampled 100 examples from each dataset. Each example was tokenized at the word level, and GPT-4.1 was used to classify each word as either *Language Prior* or *Visually Relevant*. For each word, we computed its loss before and after fine-tuning. If a word was split into multiple tokens, we used the loss of the first token as the word-level loss. The percentage change in loss for each category was then calculated to analyze the model's tendency to fit language priors versus visually grounded content.

All other training settings were kept consistent with the main experiments described above.

A.3 Analysis of Word-Level Loss for Main Results

The experimental setup for the word-level loss analysis in Section 5.1 closely follows the procedure described in Section A.2. Specifically, this analysis is based on the fine-tuning results of Qwen2.5-7B-Instruct on the BLIP-LCS dataset.

The only difference from Section A.2 is that, to facilitate clearer visualization, we excluded words with a frequency less than 3 in the sampled data.

All other experimental settings remain consistent with those described above.

A.4 Ablation Study and Analysis

Unless otherwise specified, all experiments in Section 6 are conducted using the Qwen2.5-Instruct series models trained on the BLIP-LCS dataset. If the model size is not explicitly mentioned, Qwen2.5-7B-Instruct is used by default.

A.5 Detailed Experimental Setup

Datasets During the pretraining stage, we selected three representative multi-modal pretraining datasets for comparative analysis: **BLIP-LCS** [1](Li et al., 2022): Used as the pretraining dataset for LLaVA-1.5(Liu et al., 2024a). It is a filtered subset of LAION(Schuhmann et al., 2021), CC(Sharma et al., 2018), and SBU(Saleh and Elgammal, 2015), with a more balanced distribution of concept coverage. **ShareGPT4V-PT**(Chen et al., 2024a): Used as the pretraining dataset for ShareGPT4V(Chen et al., 2024a). It utilizes high-quality image-text descriptions generated by the GPT-4, offering significantly richer semantics and contextual coherence compared to BLIP-LCS. **PixMo-Cap**(Deitke et al., 2024): Used as the pretraining dataset for Molmo(Deitke et al., 2024), a state-of-the-art open-source MLLM. This dataset is constructed by expert annotations, featuring precise visual attribute labeling and complex scene descriptions.

During the instruction tuning stage, we follow Tong et al. (2024a) and use the Cambrian-1[2] dataset as our training data. This dataset builds upon LLaVA-665k(Liu et al., 2024a), systematically expanding the model's understanding of structured visual information by incorporating OCR data and chart data.

Evaluation metrics We employs a dual-dimensional evaluation system: **General benchmarks** and **Vision-centric benchmarks**. General benchmarks include MMBench(Liu et al., 2024c) (commonsense reasoning), AI2D(Hiippala et al., 2021) (diagram parsing), DocVQA(Hudson and Manning, 2019) (document understanding) and OCRBench(Liu et al., 2023) (OCR capability),

[1] *LCS* abbreviates the LAION, CC, and SBU datasets

[2] https://huggingface.co/datasets/nyu-visionx/Cambrian-10M/blob/main/jsons/Cambrian737k.jsonl

covering the assessment of fundamental cognitive abilities. Vision-centric benchmarks focus on evaluating core visual capabilities, comprising three specialized test sets: CVBench(Tong et al., 2024a) examines structured visual understanding through 2D/3D spatial relationship analysis, MMVP(Tong et al., 2024b) emphasizes fine-grained feature recognition, and NaturalBench(Li et al., 2024a) tests comprehensive visual perception capabilities through challenging tasks such as understanding attribute bindings and reasoning about object relationships.

We use VLMEvalKit(Duan et al., 2024) for systematic evaluation. Specifically, multiple choice questions (AI2D / MMBench / CVBench / MMVP) primarily use the accuracy of the options as the core metric. Document parsing (DocVQA) uses Normalized Levenshtein Distance. OCR recognition (OCRBench) is based on the hit rate of detections contained in the ground-truth answers. Cross-combination evaluation (NaturalBench) sets four fine-grained metrics, including single-question accuracy, group accuracy (requiring all four combined questions to be correct), Question Accuracy (correct rate for the same question on both images) and image accuracy (correct rate for the same image on both questions).

Baselines We compare DPA with two representative training paradigms: **(1) Vanilla** is the classic two-stage alignment method (Liu et al., 2024b), where only the MLP projection layer is unfrozen during pretraining, while both the MLP and LLM are unfrozen during fine-tuning, **(2) CAL**(Xiao et al., 2024) introduces a dynamic weight adjustment mechanism on top of Vanilla, optimizing the loss weights of different tokens through logits differences to enhance alignment of key semantics.

Implementation Details The model architecture employs CLIP-pretrained ViT-L as the visual encoder, a two-layer MLP cross-modal connection layer with GeLU activation, and three different LLMs: Qwen2.5-7B-Instruct(Yang et al., 2024a), Llama-3.1-8B-Instruct(Dubey et al., 2024), and Gemma-2-9B-it(Team et al., 2024). The training parameters were optimized through grid search, with learning rates set to 2e-3 and 4e-5 for the pretraining and fine-tuning stages, respectively. The learning rate for the Proxy LLM pretraining phase was set to 4e-5. The batch size was fixed at 256. The LoRA configuration uses rank=256, alpha=512, and weight boundaries α=0.05, β=0.5, with a pooling layer window size of 3. Experiments were conducted on 8 NVIDIA H100 GPUs.

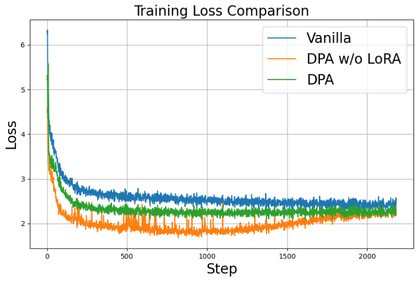


Figure 4: Comparison of pretraining loss function behavior across three training strategies: Vanilla, DPA without LoRA, and DPA with LoRA.

B More Experimental Results

B.1 Ablation Studies

Necessity of LoRA To systematically validate the effectiveness of the LoRA-enhanced training strategy, this study investigates the pretraining loss function behavior of three training strategies: Vanilla, DPA w/o LoRA, and DPA (with LoRA). As shown in Figure 4, the DPA w/o LoRA strategy exhibits a clear non-monotonic convergence pattern: the training loss initially decreases rapidly but then unexpectedly increases, a significant deviation from the typical training curve. This phenomenon indicates that DPA w/o LoRA overfits to a subset of text descriptions during the Proxy Model Optimization phase—the model's loss decreases abnormally on specific samples while its generalization performance degrades significantly on others.

This overfitting phenomenon has a dual negative impact. First, the LLM focuses excessively on local features within the text data rather than learning the overall prior language distribution. Second, in the subsequent visual modality alignment phase, the model struggles to disentangle the already solidified text feature representations, leading to semantic mismatches during cross-modal alignment. Notably, the introduction of LoRA results in the expected monotonic convergence characteristic, validating its effectiveness in suppressing overfitting.

Rank for LoRA Table 6 using the Qwen2.5-7B-Instruct model on the BLIP-LCS dataset demonstrates a non-linear relationship between LoRA rank and model performance. As rank increases, metrics on both general and vision-centric benchmarks initially improve, then decline. This phenomenon can be explained as follows: Initially, in-

Method	LoRA rank	General Avg.	Vision. Avg.
DPA w/o LoRA	-	**54.2**	44.5
DPA with LoRA	128	53.3	43.5
	256	**54.2**	**45.7**
	512	53.8	44.4

Table 6: Performance difference when applying different rank for LoRA. The LoRA alpha is set to twice the LoRA rank.

creasing the rank appropriately increases the number of trainable parameters, enhancing the language model's ability to fit the textual descriptions. This, in turn, allows the model to focus more on semantic matching with the visual modality during cross-modal alignment (DPA). However, when the rank exceeds a certain threshold, the excessive degrees of freedom lead to the model overfitting the textual descriptions, ultimately weakening the visual-language modality alignment.

Initialization	General Avg.	Vision. Avg.	Avg.
Proxy LLM	54.7	46.8	50.7
Target LLM	54.3	47.9	51.1

Table 7: Performance comparison between models initialized with the proxy LLM and the target LLM for instruction fine-tuning.

Proxy LLM vs. Target LLM in Instruction Fine-tuning To investigate the effect of restoring the proxy LLM to the target LLM in stage 3, we conducted comparative experiments analyzing the performance differences when using different LLMs as starting points for the instruction-tuning phase. Table 7 shows that using the target LLM as the starting point significantly improves the model performance on vision-centric benchmarks, while there is a slight decrease on general benchmarks. The experimental results indicate that restoring the proxy LLM to the target LLM in stage 3 is more beneficial for vision-text modality alignment.

Hyper-parameters for $[\alpha, \beta]$ in clamping We further conduct an ablation study on α and β to study the effect of the hyperparameters in our clamping operation. First, we plot the δ distribution on MLLMs in Figure 7. Tokens whose δ lower than 0.5 constitute approximately 96% of the total label sequences. Based on this observation, we set 0.5 as the upper bound β. To prevent language style-related tokens from being completely ignored, we set the lower bound α to 0.05. This is because a δ of 0 implies that the weights of these tokens are zero, meaning they will not be optimized. We then extended both the lower and upper bounds to their extreme values, i.e., 0 and 1. The results are presented in Table 12. (1) When the upper bound β is set to 1, the model's performance degrades significantly. This indicates that allowing a few visually correlated tokens to dominate the importance weights across all label tokens negatively impacts the model. A possible explanation is that these tokens become over-optimized, while other tokens are ignored. (2) When the lower bound α is set to 0, the model also shows a performance drop. This suggests that focusing solely on optimizing visually correlated tokens is harmful. Instead, the optimization process should cover all tokens while emphasizing visually correlated tokens.

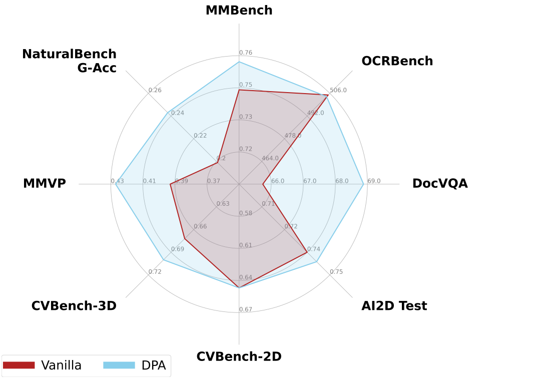


Figure 5: Anyres performance when using DPA or not.

B.2 Dynamic Resolution Input

Supporting dynamic resolution input is a trend in MLLMs. Based on the Qwen2.5-7B-Instruct model architecture, we experimented with the Anyres training strategy of LLaVA 1.5 on the BLIP-LCS dataset. As shown in Figure 5, the experimental results demonstrate that this method can effectively accommodate dynamic resolution input schemes.

B.3 Training on Multi-Datasets

To address the challenge of training on multiple datasets with conflicting language priors, we conducted an additional experiment by mixing 200K samples each from BLIP-LCS (web-crawled), ShareGPT4V-PT (GPT-generated), and PixMo-Cap (expert-annotated) into a composite dataset containing 600K samples. This mixed dataset naturally introduces diverse and potentially conflicting language priors. We trained Qwen2.5-7B-Instruct on this dataset and compared the performance of DPA against the Vanilla baseline. As

Method	MMB.	OCRB.	DocVQA	AI2D	CV-2D	CV-3D	MMVP	NaB.	Avg.
Vanilla	0.754	372	36.898	0.738	0.597	0.687	0.327	0.188	51.52
DPA	0.755	394	37.229	0.736	0.619	0.695	0.453	0.210	54.34

Table 8: Performance comparison of Vanilla and DPA on a mixed dataset (BLIP-LCS, ShareGPT4V-PT, and PixMo-Cap, 600K samples total). DPA demonstrates superior average performance and significant gains in vision-centric tasks, indicating effective adaptation to multiple conflicting language priors.

shown in Table 8, DPA outperformed Vanilla by a notable margin (average score: 54.34 vs. 51.52), with especially significant improvements in vision-centric tasks such as MMVP (0.453 vs. 0.327). These results indicate that proxy LLM pretraining with LoRA can effectively adapt to diverse language priors while preserving the base LLM's knowledge, thereby mitigating interference and reducing the risk of catastrophic forgetting.

B.4 Computational Overhead

Our proposed PMO requires an additional round of pre-training on the dataset in conjunction with LoRA. For CMO, each iteration involves two forward passes of text tokens. Table 9 presents the training time of Llama-3.1-8B-Instruct on BLIP-LCS using 8 H100 GPUs. DPA introduces approximately 33% additional training time. Memory usage increased by approximately 20% due to LoRA and CMO computations. Given the performance improvements achieved by DPA (e.g., +2.8 on PixMo-Cap, as shown in Table 1), this trade-off is considered justified.

Method	Other	Pretraining	Instruction Tuning	Overall
Vanilla	-	1.37h	6h	6.37h
DPA	0.5h	1.85h	6.12h	8.47h

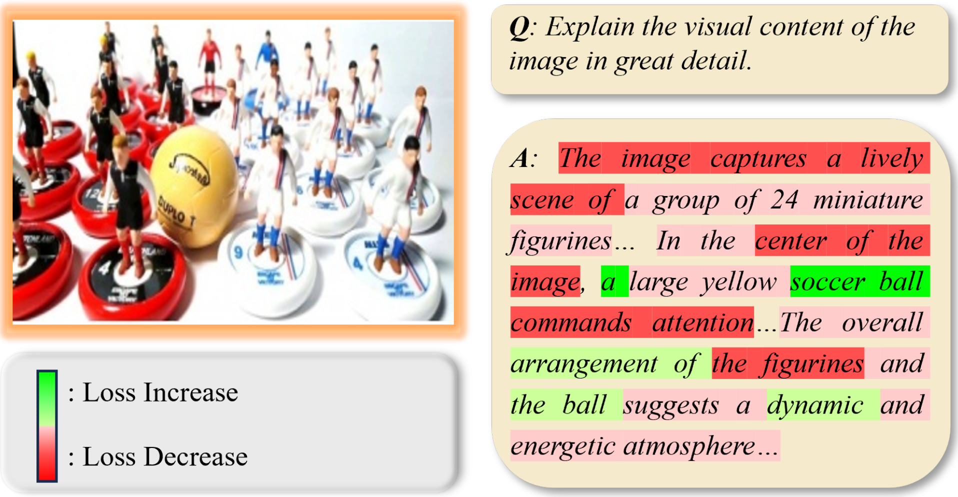
Table 9: Training time of different methods (Vanilla and DPA) for pretraining and instruction tuning of Llama-3.1-8B-Instruct on BLIP-LCS using 8 H100 GPUs.

B.5 Qualitative Analysis of Word-level loss change

To further illustrate the impact of language prior conflict, we present a qualitative analysis of word-level loss changes based on a sample from the ShareGPT4V-PT dataset, as shown in Figure 6. In the figure, words highlighted in red indicate a decrease in loss after fine-tuning, while those in green indicate an increase in loss.

It can be observed that many words with decreased loss are primarily related to language style, such as “captures” and “a lively scene of.” In contrast, some words with increased loss are highly relevant to visual content, such as “soccer” and “ball.” This case further supports our quantitative findings: language prior conflict leads to suboptimal performance when training Qwen2.5-7B-Instruct with ShareGPT4V-PT, as the model tends to fit language priors at the expense of visually grounded content.

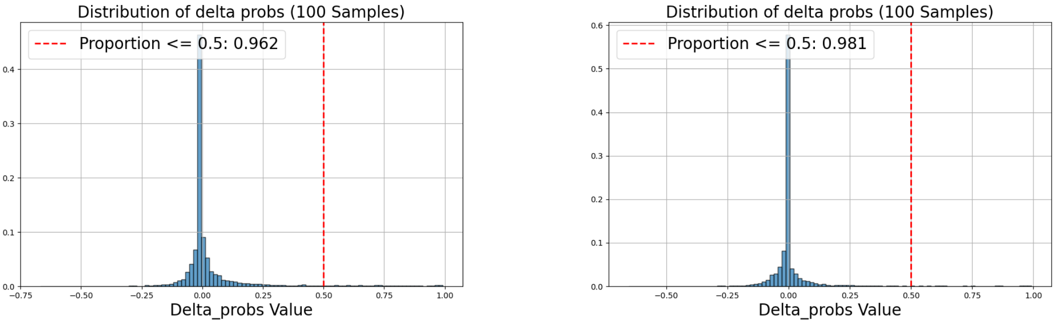
B.6 Detailed Results



Q: *Explain the visual content of the image in great detail.*

A: *The image captures a lively scene of a group of 24 miniature figurines... In the center of the image, a large yellow soccer ball commands attention...The overall arrangement of the figurines and the ball suggests a dynamic and energetic atmosphere...*

Figure 6: Qualitative Analysis of Word-level loss change.



(a) Qwen2.5-7B-Instruct

(b) Llama-3.1-8B-Instruct

Figure 7: $\Delta\mathbf{p}$ distribution for models on 100 random sampled cases.

Method	**General**				**Vision-centric**			
	MMB.	OCRB.	DocVQA	AI2D	CV-2D	CV-3D	MMVP	NaB.
Vanilla	73.1	365	34.7	72.9	58.5	62.1	36.0	18.5
DPA w/o PMO	75.2	385	35.9	73.1	60.6	69.1	39.3	19.8
DPA w/o CMO	74.9	380	35.2	73.4	59.1	65.3	42.0	17.8
DPA	76.1	368	36.0	74.1	61.1	68.1	40.7	20.4

Table 10: Ablation study on DPA's components.

PT	IT	**General**				**Vision-centric**			
		MMB.	OCRB.	DocVQA	AI2D	CV-2D	CV-3D	MMVP	NaB.
		72.4	352	33.9	73.7	60.1	67.5	37.3	17.4
✓		73.2	357	34.1	72.8	57.5	67.7	40.7	18.5
	✓	74.5	368	35.0	72.9	61.5	69.2	39.3	18.6
✓	✓	75.8	345	34.7	72.2	63.5	68.7	38.7	20.5

Table 11: Performance difference when *CMO* is applied at different training stages.

$[\alpha, \beta]$	**General**				**Vision-centric**			
	MMB.	OCRB.	DocVQA	AI2D	CV-2D	CV-3D	MMVP	NaB.
$[0, 1]$	65.0	335	30.8	61.5	60.7	53.8	25.3	10.8
$[0, 0.5]$	67.8	349	32.5	62.5	62.1	65.4	22.7	9.7
$[0.05, 1]$	74.6	355	34.1	72.4	59.9	62.3	41.3	18.3
$[0.05, 0.5]$	75.8	345	34.7	72.2	63.5	68.7	38.7	20.5

Table 12: Performance difference when applying different weights $[\alpha, \beta]$ for clamping.

Scales	Method	General				Vision-centric			
		MMB.	OCRB.	DocVQA	AI2D	CV-2D	CV-3D	MMVP	NaB.
1.5B	Vanilla	68.8	300	29.4	64.8	55.9	54.5	22.7	12.8
	DPA	70.7	321	30.6	65.6	57.9	61.1	34.7	14.3
3B	Vanilla	71.0	328	31.9	70.2	57.5	66.5	31.3	14.4
	DPA	72.6	354	34.9	70.3	56.8	67.9	31.3	17.1
14B	Vanilla	76.6	351	35.9	75.0	63.3	64.9	38.7	18.5
	DPA	78.4	351	33.9	73.5	67.0	69.8	39.3	21.3
32B	Vanilla	79.8	369	38.1	78.0	71.5	75.8	48.2	22.7
	DPA	81.0	366	37.5	76.7	71.8	75.9	54.4	22.9

Table 13: Performance difference across different LLM parameter scales.

Data size	Method	General				Vision-centric			
		MMB.	OCRB.	DocVQA	AI2D	CV-2D	CV-3D	MMVP	NaB.
5%	Vanilla	44.2	27	0.4	56.6	4.6	0.6	0.7	1.2
	DPA	19.0	22	0.3	57.3	1.6	0.0	0.0	0.0
10%	Vanilla	47.9	59	0.4	53.8	6.3	0.0	0.0	2.1
	DPA	46.7	74	0.3	52.8	0.4	0.0	0.0	2.3
25%	Vanilla	53.8	158	11.4	49.7	53.8	55.8	20.7	10.9
	DPA	52.4	276	19.0	46.9	57.9	61.2	38.7	16.7
50%	Vanilla	71.8	255	16.8	64.8	56.0	68.2	31.3	16.4
	DPA	75.4	256	18.0	64.9	61.2	70.8	37.3	17.5
100%	Vanilla	73.1	365	34.7	72.9	58.5	62.1	36.0	18.5
	DPA	76.1	368	36.0	74.1	61.1	68.1	40.7	20.4

Table 14: Performance difference across different data size.

Method	LoRA rank	General				Vision-centric			
		MMB.	OCRB.	DocVQA	AI2D	CV-2D	CV-3D	MMVP	NaB.
DPA w/o LoRA	-	73.7	360	34.6	72.4	59.3	67.2	34.0	17.6
DPA with LoRA	128	73.0	343	32.7	73.2	60.6	67.1	28.7	17.6
	256	75.0	350	33.6	73.2	60.1	67.5	37.3	18.0
	512	74.1	348	33.8	72.5	58.6	65.4	35.3	18.1

Table 15: Performance difference when applying different rank for LoRA. The LoRA alpha is set to twice the LoRA rank.

Initialization	General				Vision-centric			
	MMB.	OCRB.	DocVQA	AI2D	CV-2D	CV-3D	MMVP	NaB.
Proxy LLM	75.1	363	34.0	73.2	61.9	67.9	38.7	18.6
Target LLM	75.8	345	34.7	72.2	63.5	68.7	38.7	20.5

Table 16: Performance comparison between models initialized with the proxy LLM and the target LLM for instruction fine-tuning.