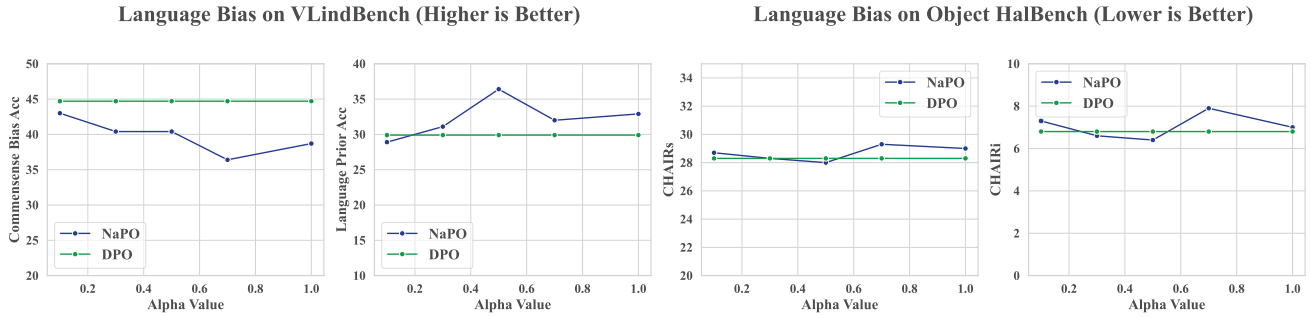

Figure 5. **Hyperparameter analysis on language-biasd data.** The chart illustrates the results of the model on VLindBench and Object HalBench when training on language-biased data with different α values in NaPO. We observed that the model achieves better performance across all four metrics when α is set to 0.5.

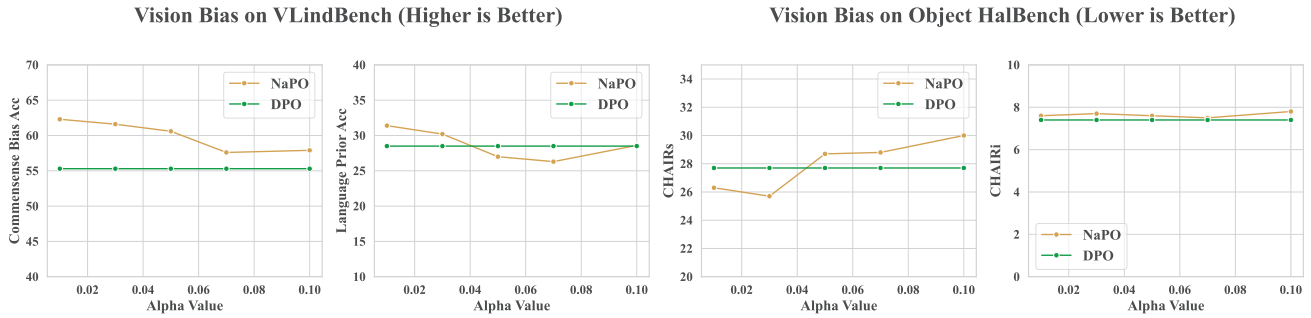

Figure 6. **Hyperparameter analysis on vision-biasd data.** The chart illustrates the results of the model on VLindBench and Object HalBench when training on vision-biased data with different α values in NaPO. We observed that the model's performance gradually decreases as the α value increases.

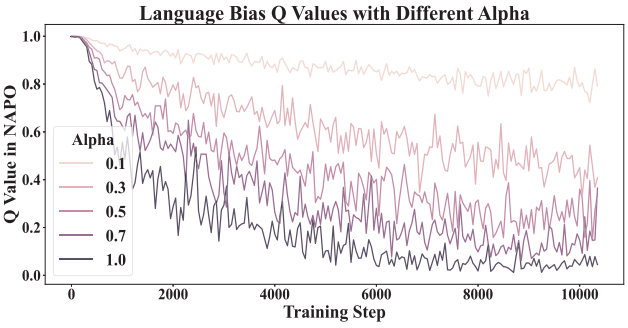

Figure 7. **The trend of q during training.** The figure illustrates the variation of q during training under different α values. As α increases, q exhibits larger fluctuations and decreases at a faster rate. In contrast, smaller α values result in more stable changes in q, with a slower and more consistent decline.

A. Additional Experiments

A.1. Hyperparameter Analysis of the NaPO

In this section, we detail the strategy for selecting the scaling parameter α and evaluate the model's performance under various α values across different datasets.

Firstly, as shown in Figure 4, the numerical range of the logp margin is typically larger compared to the avg logp margin. To ensure effective value scaling, we leverage the sigmoid function, which is most sensitive to changes within the range of $-2 \sim 2$. To achieve this, we scale the margin values using the parameter α. Specifically, for logp margins, α is selected within the range $[0.01, 0.1]$, while for avg logp margins, α is chosen from the range $[0.1, 1.0]$.

Secondly, Figures 5 and 6 illustrate the model's performance on VLindBench and Object HalBench when trained on language-biased and vision-biased data, respectively. The results demonstrate that selecting an appropriate α allows NaPO to estimate the noise-robust coefficient q more effectively, leading to improved model performance. In contrast, using an unsuitable α value can result in suboptimal q estimation and degrade performance.

Finally, in Figure 7, we analyze the variation of q during training under different α settings. We observe that higher α values amplify the fluctuations of q during training and cause q to decrease more rapidly. Conversely, smaller α values stabilize q, keeping it at relatively higher values throughout training. This analysis underscores the importance of

Loss and Data	VLindBench		Object HalBench	
	CB ↑	LP ↑	CHAIRs ↓	CHAIRi ↓
$\mathcal{L}_{\text{DPO}}$ with RLAIF-V	39.4	**25.4**	32.0	8.5
$\mathcal{L}_{\text{MDPO}}$ with RLAIF-V	0.3	0.4	35.3	10.5
$\mathcal{L}_{\text{NaPO}}$ with RLAIF-V	**48.3**	22.3	**26.7**	**7.5**

Table 5. **NaPO with RLAIF-V.** We tested the effectiveness of NaPO on the RLAIF-V and found that the results of NaPO outperform those of DPO on the RLAIF-V.

Loss	VLindBench		Object HalBench	
	CB ↑	LP ↑	CHAIRs ↓	CHAIRi ↓
$\mathcal{L}_{\gamma}$	**58.9**	44.0	**25.7**	**6.2**
w/o γ_i + repl. $\mathcal{L}_{\text{NaPO}}$	54.0	**47.8**	27.3	7.0
repl. $\mathcal{L}_{\text{NaPO}}$	21.9	21.1	36.7	9.2

Table 6. **Replace with NaPO.** We found that replacing DPO in $\mathcal{L}_{\gamma}$ with NaPO leads to a certain degree of performance degradation. Moreover, utilizing γ_i to balance the loss weights in this case causes the model performance to decline sharply. Therefore, dynamic weight balancing may not be suitable for all scenarios.

carefully selecting α to balance robustness and adaptability during training.

A.2. NaPO with RLAIF-V

To evaluate the effectiveness of NaPO on the original dataset, we conducted experiments using the same default settings as the main experiments. Specifically, we used $\log p$ to estimate the noise-robust coefficient q, and, consistent with the main experiments, we set $\alpha = 0.01$. The experimental results are shown in Table 5. From Table 5, we observed that NaPO outperforms DPO and MDPO across most metrics. This observation prompted the question: *would replacing DPO with NaPO in $\mathcal{L}_{\gamma}$ lead to further performance improvements*? To explore this, we replaced DPO with NaPO in $\mathcal{L}_{\gamma}$, and the results are presented in Table 6. Surprisingly, this replacement resulted in a performance drop. Moreover, under these conditions, employing dynamic weight balancing with γ_i caused the model to collapse. This indicates that dynamic weight balancing with γ_i may not be suitable for all scenarios. We leave the detail discussion of these issues for future work.

B. Data Construction and Analysis

B.1. Data Construction

We use the LLaVA-v1.5-7B model to construct our dataset, employing inference hyperparameters of temperature `t=0`, `num_beams=3`, and `max_new_tokens=1024`. No additional prompts are utilized to guide the model. For language-biased responses, we retain only the textual information from the questions to prompt the model's answers, whereas for vision-biased responses, we exclusively preserve visual information from the questions to generate the model's replies. Additionally, we do not employ any explicit data-filtering strategies. Our NaPO approach can instead be viewed as a soft data-selection method, dynamically adjusting the optimization strength by adaptively controlling the noise robustness coefficient q.

B.2. Data Analysis

From the observation of Figure 4, we note that in language-biased responses, noise-free (biased) data exhibit a higher avg LogP margin compared to noisy (unbiased) data. Similarly, in vision-biased responses, noise-free (biased) data show a higher LogP margin than noisy (unbiased) data.

To better understand this phenomenon, we must clarify the main distinction between LogP and avg LogP. LogP is highly sensitive to response length, meaning that variations in length can significantly influence its value. In contrast, avg LogP normalizes by response length, reducing the impact of length differences and making it more reflective of semantic consistency.

To explore this further, we computed the average length margin of different types of sampled data, as shown in Table 7. The results indicate that the length margin between noisy and noise-free data in language-biased responses is minimal, suggesting that avg LogP is more effective in distinguishing noise from noise-free samples in this case. Conversely, the substantial length margin between noisy and noise-free data in vision-biased responses amplifies the influence of length sensitivity, making LogP a better metric for identifying noise in vision-biased responses.

Mean of Length Margin	Language Bias	Vision Bias
Biased (noise-free)	60.83	-199.20
Unbiased (noisy)	65.57	199.61

Table 7. **Analysis of noise and length margin.** We observe that vision-biased responses have significant length margin differences between noise and non-noise data, while language-biased responses show minimal length margin differences between them.

The differences in length margins between noisy and noise-free data can be attributed to the vision-biased generation process. In this process, textual instructions are masked, prompting the model to rely primarily on image content. This often results in image captions, which are classified as noisy data when responding to descriptive prompts like "Describe the objects in the image in detail," as they provide general rather than detailed descriptions. Conversely, for question-answering prompts such as "What is the person in the image wearing? " the responses are typically concise and specific, relying directly on visual information. These are categorized as noise-free data. Therefore, the length margin for noise-free data is typically very small, whereas that for noisy data is significantly larger.

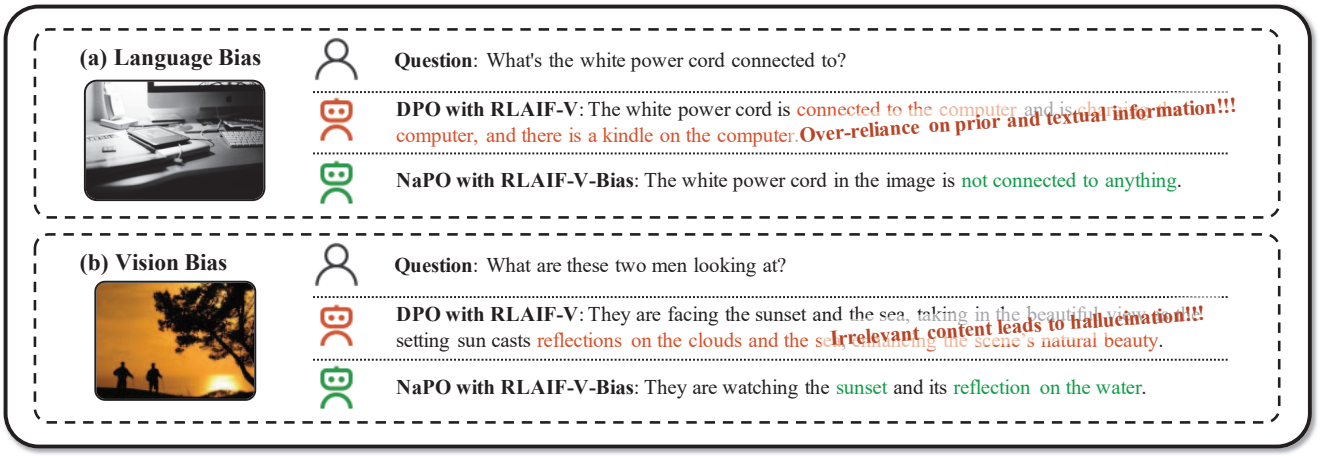

Figure 8. **Case study.** Models trained on general datasets over-rely on prior knowledge and specific modalities. In example (a), the model assumes the power cord is connected to a computer, though it's visibly disconnected. In example (b), irrelevant details lead to hallucinations of clouds and the sea, despite neither being visible.

C. Case Study

By observing the comparison in Figure 8, it becomes evident that models trained on general datasets often rely too heavily on prior knowledge and specific modalities. In example (a), the model assumes that the power cord should be connected to a computer, whereas a closer inspection shows that it isn't connected to anything in the image. In example (b), the model includes excessive, irrelevant details, resulting in hallucinations of objects like clouds and the sea, even though no clear clouds or ocean are visible in the image.