NoiseBoost: Alleviating Hallucination with Noise Perturbation for Multimodal Large Language Models

Qingdong He¹

Kai WU^{1*} Boyuan Jiang^{1*} Zhengkai Jiang¹ Donghao Luo¹ Shengzhi Wang² Chengjie Wang¹ Tencent Youtu Lab Tongji University

Qingwen Liu^{2†}

Abstract

Multimodal large language models (MLLMs) contribute a powerful mechanism to understanding visual information building on large language models. However, MLLMs are notorious for suffering from hallucinations, especially when generating lengthy, detailed descriptions for images. Our analysis reveals that hallucinations stem from the inherent summarization mechanism of large language models, leading to excessive dependence on linguistic tokens while neglecting vision information. In this paper, we propose NoiseBoost, a broadly applicable and simple method for alleviating hallucinations for MLLMs through the integration of noise feature perturbations. Noise perturbation acts as a regularizer, facilitating a balanced distribution of attention weights among visual and linguistic tokens. Despite its simplicity, NoiseBoost consistently enhances the performance of MLLMs across common training strategies, including supervised fine-tuning and reinforcement learning. Further, NoiseBoost pioneerly enables semi-supervised learning for MLLMs, unleashing the power of unlabeled data. Comprehensive experiments demonstrate that NoiseBoost improves dense caption accuracy by 8.1% with human evaluation and achieves comparable results with 50% of the data by mining unlabeled data. Code and models are available at https://kaiwu5.github.io/noiseboost.

1 Introduction

Recent Large language models (LLMs) [1, 2, 3, 4] have demonstrated significant potential in approximating human intelligence and can serve as sophisticated assistants for intricate tasks. Building on the foundational LLMs, Multimodal Large Language Models (MLLMs) [5, 6, 7] are designed to transfer LLM's zero-shot understanding ability to vision, extending the advantages of LLMs to the realm of multi-modality comprehension. Despite the significant progress made in recent MLLM research, no MLLM method can be immune to hallucinations [8, 9] which limits their applicability in real-world applications.

Recent studies on mitigating hallucination predominantly concentrate on the development of a tailored decoder or the annotation of hallucination-specific data. OPERA [10], utilizing a discovery and re-decoding loop, implements a language over-trust penalty to discard hallucinated results and progressively regenerate them. By introducing visual contrastive decoding, [11] subtracts the decoded logits of hallucinated visual inputs from those of the original input. Despite these re-decoding-based methods do not require training, they achieve performance improvements by doubling or even tripling the inference time, rendering MLLMs challenging for deployment on personal devices. Conversely, HallDoctor [8, 12] establish a preference dataset annotation pipeline that distorts the ground truth answer with errors to form hallucinated pairs, which is later trained to align the model's honesty.

^{*}Equal contribution.

[†]Corresponding author.



Figure 1: MLLMs suffer from hallucinations due to the over-reliance on language priors. In (a), the hallucination tokens are overly dependent (0.27) on previous language tokens, and later tokens are all hallucinations. Meanwhile, in (b), NoiseBoost helps MLLMs distribute the attention weights evenly among visual and language tokens by noise perturbation, leading to honest results.

Without distorting the training response, Fine-grained PPO [13] annotates the model response on a word-by-word basis to train a reward model and align the model's generation with proximal policy optimization [14]. However, these manually curated reward datasets differ in distribution from real-world usage and cannot encompass all scenarios. In this paper, we aim to identify the fundamental reasons for hallucination and enhance MLLM's training without additional datasets or training costs.

Upon diving into the attention mechanism of MLLMs, we discovered that the occurrence of hallucination could be attributed to an excessive dependence on language priors. During the LLM response generation process, certain language tokens are automatically selected as anchors [15], causing subsequent generations to rely more heavily on the summarization of anchor token information, rather than on the comprehensive set of preceding visual and linguistic tokens in the context. As depicted in Fig. 1(a), MLLM's information flow is unevenly distributed from visual and language tokens. Furthermore, MLLM's visual and language tokens are from separately pre-trained visual encoder and LLMs [5, 6, 7], leading to a significant disparity in features even after training. Since MLLM's anchor token selection frequency is correlated with generation length, the hallucination phenomenon gets worse when generating long, detailed descriptions. Without appropriate training methodologies, the flow of information from visual tokens to linguistic tokens is hindered, leading to a neglect of visual information and an over-reliance on language priors.

In this paper, we propose NoiseBoost, a simple and widely applicable noise perturbation method designed to mitigate hallucination across various MLLM training stages. NoiseBoost disrupts the excessive dependence on language priors, facilitating a balanced distribution of the model's attention between visual and linguistic tokens. Specifically, we increase the hardship in MLLM's learning process by incorporating noise feature perturbation, achieved by injecting noise into visual tokens. This approach complicates visual understanding, necessitating more evenly distributed attention weights in LLM. Our extensive experiments demonstrate that the injection of Gaussian noise to projected visual tokens consistently enhances performance with negligible additional training costs. As depicted in Fig.1(b), token correlation is evenly distributed, significantly reducing overconfidence induced by summary tokens in LLMs. To further exhibit NoistBoost's generalizability, we conducted experiments across two MLLM training stages: supervised fine-tuning and reinforcement learning. NoiseBoost consistently improves performance in both training methods across hallucination and question-answer datasets, validating the efficacy of feature perturbation. To verify the results of long description generations, we evaluated 1k images by annotators for dense captions, which NoiseBoost shows an 8.1% improvement in accuracy.

By integrating NoiseBoost, we pioneer the incorporation of semi-supervised learning (SSL) architecture for MLLM models. Current MLLM training relies on noisy web corpus, incurring substantial labeling costs for data cleaning without harnessing the potential of unlabeled data. The challenge is that MLLM does not have a mechanism for teacher-student learning with pseudo labels, which is a crucial element in traditional SSL architecture. We generate pseudo labels using original images and use NoiseBoost to be the noisy student, learning to produce consistent and robust results. Our experiments show that NoiseBoost can unleash the power of unlabeled data and achieve similar performance with 50% of labeled data. In summary, our contributions are as follows:

- With analyzing the cause of hallucination, We propose a simple and well-generalized method, NoiseBoost, which effectively alleviates hallucination for MLLM at negligible additional cost without introducing extra data.
- We are the pioneers in facilitating semi-supervised learning for MLLMs with NoiseBoost and reach the same performance with 50% of training data by mining the power of unlabeled data.
- Extensive experiments indicate the effectiveness of NoistBoost as a general training enhancement method, providing consistent performance improvement for MLLMs.

2 Related Work

2.1 Multimodal Large Language Foundation Models

Recent advancements in MLLMs research are primarily attributed to the evolution of large language models (LLMs). To integrate vision models with LLMs, existing MLLMs typically utilize lightweight layers such as QFormer [16] or linear projection [5]. Notably, LLaVA [5] integrates a vision encoder and an LLM to facilitate general-purpose visual and language understanding. This is achieved using multi-modal language-image instruction-following data, with the vision encoder designed to project image features into language token representations. MiniGPT-4 [17] incorporates a pretrained ViT and Q-Former and an LLM for multi-modal generation and understanding. Mini-gemini [18] enhances multi-modal reasoning capabilities through high-resolution visual tokens, employing an additional visual encoder for high-resolution refinement. However, directly bridging visual and language modalities causes hallucinations from over-reliance on the language priors. We propose NoiseBoost to redistribute attention weight to both visual and linguistic tokens by injecting feature perturbations to visual features.

2.2 Hallucinations in MLLMs

Hallucination in MLLMs has significantly impeded their usage in the real world, especially for tasks that rely on precise captions. Previous works focus on two perspectives: dataset construction and decoding schemes to alleviate the hallucination in MLLMs. For dataset construction, HallDocter [8] proposes a pipeline to annotate the hallucination dataset with the help of GPT4V. To enable reinforcement learning, [19, 20] propose object substitution using GPT4V and labor checking to create a hallucinated response pair. However, rectifying large models like MLLM with small curated data is contrary to the scaling law. With decoder scheme optimization, [21] proposes to achieve an un-hallucinated response by subtracting the hallucinated response decoded simultaneously using only language prompts. OPERA [10] proposes a penalty-based found and re-decoding method to reduce hallucinations. Although effective, decoding-based methods require iterative decoding, which incurs computational burden and impedes MLLM's deployment on personal devices. In this paper, we design a simple and well-generalized noise perturbation method for alleviating hallucinations without introducing additional datasets or inference costs.

3 Method

In this section, we first introduce the preliminaries of MLLM in Sec.3.1. Then we show how NoiseBoost is applied to different MLLM training methods, including Supervised Fine-tuning in Sec.3.2 and reinforcement Learning in Sec. 3.3. Finally, we incorporate Semi-Supervised Learning into MLLM by using NoiseBoost in Sec.3.4.

3.1 Preliminaries

Multimodal large language models (MLLMs) attain visual comprehension capabilities by integrating two well-established technologies—vision encoder and large language model (LLM). The process of using MLLM starts with an input image X_v and a question prompt X_q from multi-turn conversation data $(X_v^1, X_q^1, X_a^1, ..., X_v^n, X_q^n, X_a^n)$ where the turn number is *n* and X_a^i is the *i*-th turn's answer. Fig.2(a) illustrates a classic MLLM architecture [5], which employs a projection layer W_p to align the channel dimension of visual tokens extracted via a pre-trained vision encoder g_v to language embeddings, as



Figure 2: Framework of NoiseBoost. We add noise perturbation to visual tokens to mitigate the over-reliance on language tokens and thus fewer hallucinations. For SFT, we directly inject noise to visual features. We only inject perturbation to preferred response since that can make MLLMs harder to learn and achieve better results. For semi-supervised learning, we use freezed MLLM as a teacher to generate pseudo labels and NoiseBoost as students for consistency regularization.

follow:

$$z = (z_q, \mathbf{W}_p(g_v(X_v))) \tag{1}$$

,where **z** is the input embedding of MLLM, z_q is the language instruction embedding convert from X_q by a word to vector model and X_v is the input image.

It is easy to notice that the vision encoder and LLM are pre-trained separately, with the projector being the only newly introduced component. Consequently, MLLMs take shortcuts to be excessively dependent on language priors, neglecting the visual aspect because of the disparity in features. NoiseBoost introduces noise perturbation to visual tokens, thereby complicating the visual understanding process and compelling the MLLM to allocate more attention to the visual aspect, reducing its reliance on language priors.

3.2 Supervised Noise Boosting Fine-tuning

Supervised Fine-tuning(SFT) is a widely used training technology, in Fig.2(a). Given an image and a language prompt, the MLLM model directly predicts the linguistic results autoregressively following LLM's convention. Without specialized for multi-modality training, MLLM inherited the characteristics of LLM with over-reliance on language priors, as depicted in the token correlations matrix in Fig.4. To help MLLM redistribute attention evenly, we propose a noise feature perturbation method ϕ_v to disturb pre-trained visual features. Based in Eq. 1, the noise feature perturbation can be represented as $z = z_q + \phi_v(W_p(g_v(X_v)))$, where z is the perturbed visual tokens. So the SFT loss is as follows:

$$\mathcal{L}_{sft} = -H(y_w | \phi_v(z), X_q), \tag{2}$$

where *H* represents the cross-entropy operation used in SFT. The vision feature disturbance makes it hard for MLLM to discern the visual information and pay more attention to image understanding. We found that adding Gaussian Noise perturbation as ϕ_v to the projected vision tokens can effectively reduce the overreliance on language tokens.

3.3 Reinforcement Noise Boosting Learning

Reinforcement learning has emerged as an essential technology with the rise in popularity of LLMs [14]. However, the direct application of reinforcement learning techniques from LLMs to MLLMs without adaptation has proven to be unstable due to data limitations, as also observed in [19]. To address this, we propose the integration of noise feature perturbation for visual tokens to augment visual understanding and SFT into reinforcement learning to enhance training stability as illustrated in Fig.2(b). The noise feature perturbation is added directly to visual tokens but in a different training corpus. We employ the DPO [22], a classical reinforcement learning algorithm, to illustrate the loss equation:

$$\mathcal{L}_{dpo} = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[log\sigma \left(\beta log \frac{\pi_{\theta}(y_w | \phi_v(X_V), X_q)}{\pi_{ref}(y_w | X_V, X_q)} - \beta log \frac{\pi_{\theta}(y_l | \phi_v(X_V), X_q)}{\pi_{ref}(y_l | X_v, X_q)} \right) \right].$$
(3)

In this equation, the random feature perturbation function ϕ_v is incorporated into the projected visual tokens $\phi_v(X_v)$. The variables y_w and y_l represent the preferred and less preferred outputs, respectively. The model's objective is to maximize the probability of the preferred output and minimize that of the less preferred one. The function π_{θ} denotes the model's policy, while π_{ref} signifies the reference policy. The sigmoid function σ , compresses its input into the range (0, 1), and β is a temperature parameter controlling the distribution's sharpness. The final reinforcement loss is defined as $\mathcal{L}_{rl} = \mathcal{L}_{sft} + \mathcal{L}_{dpo}$. In experiments, we observed that a larger noise perturbation on y_w and less on y_l resulted in superior performance. This aligns with our intuition that a challenging visual feature enables MLLM to learn a better attention weight distribution, which the less preferred output y_l does not need.

3.4 Semi-Supervised Noise Boosting Learning

Semi-supervised learning is a mature technology for mining the ability of unlabeled data but has not been applied to MLLMs. The reason is that MLLM's training strategy prevents it from creating a weak augmentation for pseudo labels generation and a strong augmentation for consistency regularization. Particularly, MLLMs have no visual augmentation methods, with the assumption that pixel-level image disturbance can mislead understanding of the image content. Thanks to NoiseBoost, we can incorporate noise feature perturbations to provide weak and strong distortions for MLLMs without affecting visual understanding and comply with semi-supervised mechanism at the same time. The unlabeled loss for semi-supervised learning is

$$\mathcal{L}_{\mu} = \frac{1}{\mu B} \sum_{i=1}^{\mu B} \mathbb{1}(\max(\pi_{ref}(X_{\nu}, X_q)) \ge t) H(\hat{q}_i, \pi_{\theta}(\phi_{\nu}(X_{\nu}), X_q))$$
(4)

where the reference model π_{ref} generate pseudo labels without feature distortion and π_{θ} keep training on noise distorted data for consistency regularization. *t* is the threshold for filtering noisy pseudo labels, μ is the ratio of label and unlabeled data, *B* is the batch size and $\hat{q}_i = \arg \max(\pi_{ref}(X_v, X_q))$ for artificial labels generation. The final semi-supervised loss is $\mathcal{L}_{semi} = \mathcal{L}_u + \mathcal{L}_{sft}$. With NoiseBoost, we can mine the unlabeled data power without waiting for labor-intensive trillions of data cleaning.

4 Experiments

4.1 Setup

Baseline and Data. We have chosen LLaVA-1.5 [5] and QwenVL [7], two recently released stateof-the-art MLLMs, as our baseline. In addition to the data incorporated in LLaVA-1.5, we have supplemented our dataset with COCO captions and ShareGPT4v [23] to reach a total of 800K entries, thereby matching the performance of LLaVA-1.5 because thousands of images could not be downloaded from the original 665K meticulously curated dataset [6]. The data serves as a comparable baseline for LLaVA-1.5-7B, but it falls short for QwenVL-Chat [7]. Given that QwenVL [7] has not released its data for retraining, we only evaluate QwenVL for human evaluation to establish a relatively fair comparison. Reinforcement learning datasets for MLLM are limited, we use HA-DPO dataset [19] which contains 18k images. Although the dataset is limited in scale, NoiseBoost can also achieve performance gain compared to previous methods. Semi-supervised learning datasets are constructed by splitting the SFT data into 30% and 50% with others used for unlabeled data learning. In summary, with the data managed to maintain a fair comparison, we tested NoiseBoost on LLaVA-1.5 using Llama7B and Llama13B to assess backbone generalizability and on QwenVL to evaluate performance across different LLM styles.

Implementation Details. For supervised fine-tuning, We use a batch size of 192 with accumulation steps setting to 2 for training, similar to [6], with 24 V100 training for 16 hours, about 384 GPU hours. For reinforcement learning and semi-supervised learning 30% setting, the training only needs around 90 GPU hours and 150 GPU hours to finish because of not much data. We set the learning rate to 2e-5 in for SFT and semi-supervised learning. Reinforcement learning uses a small learning rate 2e-6 because of data deficiency resulting in unstable training. The weight decay and warmup ratio are set to 0.0 and 0.03 respectively. The model length is 2048, the same as [6] but 1024 for QwenVL [7] for fewer training hours. All of our experiments are conducted on float16 with deepspeed due to GPU memory limitations. For noise perturbation, we set the noise scale to 0.5 with a 50% chance of triggering perturbation for all experiments if not specified without tuning the parameters.

Table 1: Supervised Fine-tuning results. NoiseBoost consistently improves the performance of MLLM on hallucinated and question-answer datasets. For the caption dataset Flickr30k, we achieve comparable performance since the traditional caption dataset cannot manifest the long, dense captions generated by NoiseBoost. * means trained on our collected dataset.

Method	Backbone	POPE	GQA	VixWiz	Text- VQA	MME	SEED Bench	Flickr30K
Existing Methods								
Flamingo	80B	-	-	31.6	-	-	-	67.2
VLIP-2	Vicuna-13B	-	32.3	19.6	-	-	-	71.6
InstructBLIP	Vicuna-13B	-	49.5	33.4	-	-	-	82.8
mPLUG-Owl2	-	-	56.1	54.5	-	-	-	85.1
Qwen-VL	Qwen-7B	-	59.3	35.2	-	-	-	85.8
Qwen-VL-Chat	Qwen-7B	-	57.5	38.9	-	-	-	81.0
LLaVA-1.5	Llama-13B	87.1	63.3	56.6	48.69	1523	68.2	79.5
LLaVA-1.5	Llama-7B	86.9	61.9	54.3	46.07	1507	66.2	74.9
NoiseBoost								
LLaVA-1.5*	Llama-13B	88.3	64.0	59.8	49.5	1540	69.2	81.2
+ NoiseBoost	Llama-13B	88.4	64.2	61.5	49.8	1580	69.1	80.8
LLaVA-1.5*	Llama-7B	87.2	62.3	54.6	47.18	1501	66.9	73.3
+ NoiseBoost	Llama-7B	88.4	63.4	57.1	47.8	1531	67.7	73.8



Figure 3: Human evaluation of dense captions for 1k images with prompt "Describe the image and its style in detail". The correct category means the totally accurate image with other categories are errors identified by human annotators. NoiseBoost consistently reduce error on nearly all categories.

Evaluations. We conduct an evaluation of various datasets, including the hallucination dataset POPE [24], question-answer datasets [25, 26, 27, 28], and the caption dataset [29]. Given the inherent difficulty in assessing long captions using automated tools, we supplement our study with a collection of 1,000 images for human evaluation. For automated evaluations, we utilized [30], a publicly available tool designed to facilitate the evaluation of all MLLM datasets. However, we observed suboptimal performance when assessing QwenVL-Chat using [30], attributable to a minor modification in the evaluation prompt can lead to significant discrepancies in the MLLM results. To ensure a fair evaluation, we maintained uniformity in all evaluations, refraining from prompt tuning for a single model and only evaluating QwenVL on human evaluation. For human-labeled captions, we ask annotators to select reasons for captioning results other than the binary right or wrong.

4.2 Quantative Experiments

In this section, we analyze NoiseBoost's performance gain in SFT, reinforcement learning and semi-supervised learning.

Supervised Fine-tuning. We conduct experiments on LLaVA-1.5 7B/13B and QwenVL to test variations in backbone and architecture. As demonstrated in Tab.1, NoiseBoost consistently enhances performance across nearly all datasets, with gains exceeding 1% over most datasets, no matter whether the datasets are hallucination-based or question-answer-based. For the LLaVA-1.5 with the

Table 2: Reinforcement learning result. NoiseBoost consistently improves over all datasets.

Model	POPE	GQA	VizWiz	MME	SeedBench	ScienceQA
LLaVA-1.5 DPO	86.3	60.1	53.9	1516	66.3	66.9
+ NoiseBoost	87.2	61.8	54.7	1528	66.5	70.3

Table 3: Semi-supervised learning experiments, NoiseBooost enables MLLM mining unlabeled data and achieve similar performance with 50% data.

	POPE	GQA	VizWiz	MME	Seedbench	ScienceQA
30% Data	86.0	60.3	44.1	1426	67.0	67.9
+ NoiseBoost	87.4	62.5	54.9	1509	67.2	69.1
50% Data	86.9	62.4	54.3	1490	66.8	70.0
+ NoiseBoost	88.0	62.5	55.2	1553	67.0	71.0

Llama 13B model, the MME [27] reached 1580, 40 points higher than the original model. Notably, we only achieve performance comparable to the baseline on Flickr30K [29] because NoistBoost is more likely to generate rich caption data, which differs from traditional caption evaluation datasets. However, most MLLM automatic evaluations are notorious for not aligning with human feelings.

We further assessed the dense caption performance using human labeling, as depicted in Fig. 3. NoiseBoost achieves an accuracy of 540/1000, which is 8.1% higher than the QwenVL-Chat baseline. With human labeling, error categories are classified. A detailed explanation of each category can be found in the supplementary material Sup. A.1. Our improvements primarily stem from object error and hallucinations, which refer to the description of erroneous objects or non-existent objects, respectively. The results indicate that noise feature perturbation can redistribute the attention weights of the MLLM, leading to more pronounced improvements in object-related information in the image.

Reinforcement Learning. To align the behaviour of MLLM with actual human responses, DPO [22] serves as a prevalent reinforcement learning technique that requires only paired data for training. However, the DPO is first proposed in LLM and has no adaptation for MLLMs. We inject the noise perturbation to the preferred visual features with the assumption that harder consistency learning achieves better results. Upon testing with HA-DPO [19] as shown in Table 2, NoiseBoost improve ScienceQA with 3.4% and consistently enhances performance by approximately 1% on both the hallucination dataset [24] and various question-answer datasets [25, 26, 27, 28, 31]. However,



Figure 4: Analysis of the cause of hallucination is the over-reliance on language tokens circled in read which NoiseBoost doesn't have.

it is noteworthy that the degree of improvement is relatively less in comparison to SFT. This can be attributed to two primary factors: the limited scale of HA-DPO, which restricts the full potential of NoiseBoost due to fewer training steps, and the lack of proper tuning of the noise scale injected into the features, which prevents a fair comparison.

Semi-Supervised Learning. To unleash the power of unlabeled data, we incorporate NoiseBoost to create teacher-student architecture as in MeanTeacher [32], which is a classic semi-supervised learning technique. The teacher produces pseudo labels, and the student learns with strong augmented images for consistency regularization. With noise perturbation, we inject Gaussian noise during student learning and keep the original model frozen as a teacher. The experiments in Tab.3 show that LLaVA-1.5 can reach similar performance with only 50% of the data, which sheds light on mining the power of unlabeled data.



Figure 5: Qualitative evaluation shows that NoiseBoost can generate long detailed captions without hallucinations.

Table 4: Different noise probability and noise scale. With an increase in noise prob and scale, the MLLM's performance is robust, but too much noise may affect the learning process.

Noise Prob	Noise Scale	POPE	GQA	Viz Wiz	Text VQA	Seed bench	MME	Text Caps	Flickr 30K
0	0	87.2	62.3	54.6	47.6	66.9	1501	96.9	73.3
0.1	0.1	88.1	63.4	56.4	47.9	67.2	1506	98.4	73.1
0.5	0.1	88.2	63.4	54.4	47.5	66.9	1504	97.2	73.2
0.5	0.5	88.4	63.4	57.1	47.8	67.7	1531	100.6	73.8
0.5	0.9	87.9	63.0	55.2	47.0	66.6	1517	98.6	73.1
0.7	0.5	87.8	63.0	54.0	47.7	66.6	1532	96.8	72.3
0.9	0.5	87.9	62.9	55.8	47.1	66.8	1522	96.8	72.7

4.3 Qalitative Experiments

We conduct a series of experiments using a street image, which is prone to cause hallucinations of "people" due to the common association of people walking in streets in language. As illustrated in Fig. 5, the original model tends to hallucinate during response generation, primarily due to an over-reliance on language priors. To substantiate our hypothesis, we visualized the token correlation map, using the final layer attention maps from LLM's last token generation. The column attention, highlighted in red in Fig. 4 (a), indicates that the subsequently generated tokens are overly dependent on a specific language anchor token, leading to a neglect of visual tokens. The column phenomenon emerges in the middle, coinciding with the occurrence of hallucination. The tendency for hallucination becomes severe during the generation of longer responses. NoiseBoost, however, disrupts the overconfidence in specific hallucination tokens and is capable of generating extended captions without errors. After the introduction of noise feature perturbation, the LLM redistributes attention weights more evenly, as shown in Fig. 4 (b).

5 Ablations

Different Feature Perturbation Scale. We choose Gaussian noise with upper and lower bound [0, 1] in our paper. To test the robustness of feature perturbation on different scales and probabilities, we conduct extensive experiments. As in Tab.4, NoiseBoost is robust to the scale of noise-injected with not much variation with changing the hyperparameters, duo to page limits see full table in Sup.A1. An interesting phenomenon is that with the increasing of noise scale, the performance first increase and then decrease, which can be explained by the fact that the MLLM training process needs noise to break the language reliance, but too much noise can harm the learning process.

Other Perturbation Methods. Perturbation methods can be broadly classified into pixel-level and feature-level categories. In the case of pixel-level perturbations, we evaluate the efficacy of conventional image augmentations. For feature-level perturbations, we opt for dropout as a comparative measure, given the absence of alternative feature augmentation techniques.

As demonstrated in Table 5, pixel-level distortions such as RandomCrop and GaussianNoise induce more hallucinations in POPE [24], as these distortions impact the appearance or even the existence

1		1			1		
Perturbation Methods	POPE	GQA	Scien ceQA	Viz Wiz	MME	Seed bench	Flickr 30K
ColorJitter	88.6	62.5	69.9	58.7	1485	67.3	70.5
RandomCrop	86.4	63.1	69.5	56.6	1516	66.7	74.6
GaussianNoise	86.0	62.4	69.5	55.4	1507	65.4	70.7
Dropout	88.0	63.3	69.7	55.8	1489	67.3	71.4
Dropout + NoiseBoost	88.1	63.3	69.5	57.9	1491	67.4	74.3
Ours	88.4	63.4	69.9	57.1	1531	67.7	73.8

Table 5: Comparision with pixel level and feature level perturbations.

Table 6: Comparision with noise perturbation on language tokens.

	POPE	GQA	ScienceQA	VizWiz	TextVQA	MME	Seedbench
+ Lan	87.9	61.1	59.7	46.6	45.1	1477	65.1
+ Lan Vis	88.1	62.7	66.8	56.3	46.4	1507	67.1
Ours	88.4	63.4	69.9	57.1	47.8	1531	67.7

of the object. ColorJitter, which solely alters the image's colour, does not increase incorrect object hallucinations in POPE, but it does degrade performance in visual understanding datasets like MME [27] due to the disparity in colour. Pixel-level distortions, therefore, either crop images, induce object hallucinations, or disrupt the colour space, thereby affecting visual comprehension.

Since the inception of the Deep Learning era, Dropout has been a widely used feature perturbation method. We incorporate dropout into visual features post-projection, akin to NoiseBoost, and adhere to the convention by setting the dropout rate at 0.1. As per Table 5, Dropout only yields performance comparable to the baseline, which can be attributed to the fact that MLLM training methods already employ this technology for backbone training. The performance can be enhanced with NoiseBoost, thereby validating the effectiveness of our method.

Feature Perturbation on Language. NoistBoost only adds feature perturbation to visual features to align with LLM feature space and break the over-reliance on language priors. We also study whether the perturbation is effective vice versa. By adding noise to language embeddings before LLM backbone, we found a performance degradation among nearly all benchmarks. From Tab.6, we can conclude that the foundation LLM has a strong pre-trained knowledge, which should not be affected during training. However, with visual token noise perturbation in NioseBoost, the performance can also be enhanced.

5.1 Limitations and Societal Impacts

NoiseBoost serves as a fundamental method capable of mitigating the hallucination phenomenon in MLLM throughout all stages of training. The feature perturbation technique employed by NoiseBoost is a rudimentary training strategy that not only avoids negative societal impacts but also propels the advancement of multi-modal AI assistants. However, it is important to note that while NoiseBoost does not necessitate any additional costs or modifications to the MLLM structure, it doesn't change the existing methods. Presently, MLLM incorporates large language models without any module resembling the human brain, which should be developed at the architecture level.

6 Conclusions

Recent advancements in MLLM have been swift, yet these models can induce hallucinations, thereby limiting their practical applications. This paper introduces a simple, broadly applicable method, termed NoiseBoost, designed to enhance visual comprehension and mitigate hallucinations in MLLM without incurring additional costs. Specifically, NoiseBoost incorporates Gaussian noise into visual tokens to diminish the excessive reliance on language priors, a characteristic inherited from LLMs. Through comprehensive experimentation, we demonstrate that feature perturbation can augment MLLM performance without extra expenditure, and that NoiseBoost currently stands as the most efficacious feature perturbation technique. Moreover, we equip MLLM with semi-supervised learning

capabilities by employing NoiseBoost to establish teacher-student networks. Collectively, we posit that NoiseBoost can serve as a fundamental method for training MLLMs and illuminate the path towards exploiting unlabeled data for large language models.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- [2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.
- [3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.
- [4] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295, 2024.
- [5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [6] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint*, 2023.
- [8] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. arXiv preprint arXiv:2311.13614, 2023.
- [9] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023.
- [10] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. arXiv preprint arXiv:2311.17911, 2023.
- [11] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. arXiv preprint arXiv:2311.16922, 2023.
- [12] Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. *arXiv preprint arXiv:2403.08730*, 2024.
- [13] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2024.
- [14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [15] Jianhui Pang, Fanghua Ye, Derek F Wong, and Longyue Wang. Anchor-based large language models. arXiv preprint arXiv:2402.07616, 2024.
- [16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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, 2023.

- [17] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [18] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. arXiv preprint arXiv:2403.18814, 2024.
- [19] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization, 2023.
- [20] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:2311.16839, 2023.
- [21] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. arXiv preprint arXiv:2312.06968, 2023.
- [22] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.
- [23] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793, 2023.
- [24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.
- [25] Drew A Hudson and Christopher D Manning. 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, 2019.
- [26] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.
- [27] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023.
- [28] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seedbench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- [29] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer* vision, pages 2641–2649, 2015.
- [30] Bo Li*, Peiyuan Zhang*, Kaichen Zhang*, Fanyi Pu*, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024.
- [31] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.
- [32] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

A Appendix

A.1 Human Evaluation Guidance for Dense Captions.

To align the MLLM's evaluation with human preferences, we ask human annotators to evaluate the dense captions with detailed error category labeling, including errors about object, number, name, posture, position, color, relation, hallucination, style, text, and others.

Object Error. Errors of object descriptions. For example, describing a phone when it's actually an iPad, or describing short hair as long hair, short sleeves as padded jackets

Number Error. Errors of number. For example, if there are two people dancing in the picture, the MLLM says three people.

Name Error. Error for proper noun. Such as incorrect descriptions of a person's name, place of interest, or idiom.

Posture Error. Errors of posture or movement. The object is not doing the described action.

Position Error. Errors of object position. The object is not in the described position, such as in the picture's top, bottom, left, or right.

Color Error. Errors of color descriptions.

Relation Error. Errors for relations among subjects. such as the description of "two people, one on the other's shoulder," but in the image is a left-right or front-back relationship.

Hallucination Error. Error of hallucinations. The object described in the picture does not exist.

Style Error. The described style is wrong, such as light black and white, the actual picture is colorful and heavy, etc.

Text Error. Error of the text descriptions in the image.

Other Errors. Errors not listed in the above such as repetition of MLLM response.

Correct. The descriptions not labeled into any categories of errors are considered correct.

A.2 Noise Perturbation Scale and Probability Variations

The full table shows changing the scale and probability of noise feature perturbation for NioseBoost. We found that NoiseBoost is robust with the variation of hyperparameters.

s performance is robust, but too much noise may affect the rearining process.									
Noise Prob	Noise Scale	POPE	GQA	Viz Wiz	Text VQA	Seed bench	MME	Text Caps	Flickr 30K
0	0	87.2	62.3	54.6	47.6	66.9	1501	96.9	73.3
0.1	0.1	88.1	63.4	56.4	47.9	67.2	1506	98.4	73.1
0.3	0.5	88.0	63.1	54.5	47.4	66.8	1517	98.9	72.8
0.5	0.1	88.2	63.4	54.4	47.5	66.9	1504	97.2	73.2
0.5	0.3	88.2	63.2	54.0	47.5	67.0	1522	97.9	72.8
0.5	0.5	88.4	63.4	57.1	47.8	67.7	1531	100.6	73.8
0.5	0.7	88.1	63.2	55.4	47.2	66.9	1525	98.4	73.2
0.5	0.9	87.9	63.0	55.2	47.0	66.6	1517	98.6	73.1
0.7	0.5	87.8	63.0	54.0	47.7	66.6	1532	96.8	72.3
0.9	0.1	88.3	63.2	57.1	47.7	67.1	1524	97.8	72.2
0.9	0.5	87.9	62.9	55.8	47.1	66.8	1522	96.8	72.7

Table A1: Different noise probability and noise scale. With an increase in noise prob and scale, the MLLM's performance is robust, but too much noise may affect the learning process.

A.3 Hard Cases Qualitative Examples

We test NoiseBoost on hard captioning cases with various categories. NoiseBoost shows honest performance on test casts.



In the center of the image, a small dog with a black mask is comfortably seated on a brown couch. The dog's eyes are wide open, and it's looking directly at the camera, giving the impression of a playful and curious demeanor. The mask and hat add a touch of whimsy to the scene. The couch, which is the dog's chosen spot, is brown and has a patterned fabric that adds texture to the image. The dog's position on the couch suggests it's at ease in its environment. In the dog's paws, there's a silver knife with a black handle. The knife is open, and its position suggests it's not in use at the moment. The presence of the knife adds an element of intrigue to the image, making one wonder about its purpose. Overall, the image captures a unique and interesting moment, with the small dog in its mask as the

main focus. The knife adds an unexpected twist to the scene, making it all the more captivating.



In the heart of a serene landscape, a majestic bald eagle. The eagle, facing towards the right side of the image, seems to be in a state of rest, its gaze fixed on something unseen. The branch it's perched on is adorned with a few leaves, their colors a mix of orange and yellow, perhaps a sign of the changing seasons. The background is a blur of green and brown hues, suggesting a lush forest or park. The eagle's position on the branch and its focused gaze give it an air of quiet determination. This image captures a moment of tranquility in nature, with the bald eagle as its central figure.

Figure A1: Animal Captioning



The image captures a scene from a video game, where a futuristic robot stands in the foreground. The robot, with its metallic body and angular design, is the central figure in the image. It's positioned on a gray concrete surface, facing towards the right side of the image. The background is a blend of urban and industrial elements, featuring a tall building and a bridge. The colors in the image are predominantly gray and blue, with some orange and green accents adding depth to the scene. The robot's location in the foreground and the cityscape in the background create a sense of depth and perspective. The image does not contain any discernible text. The relative positions of the objects suggest a futuristic city setting, with the robot possibly being a character in the game.



The image captures a moment of tranquility and simplicity. A hand, appearing from the bottom right corner, holds a small, vibrant orange ball. The hand is positioned in such a way that the orange ball is held up against a backdrop of a solid blue color. The orange ball, with its bright color, stands out against the blue background, drawing the viewer's attention. The hand, with its fingers slightly curled, holds the ball gently, suggesting a sense of care and attention. The image does not contain any text or other discernible objects. The relative position of the hand and the ball is such that the ball is in the center of the image, with the hand appearing to the right of the ball.



In the image, a large brown teddy bear is seated in the driver's seat of a black car. The teddy bear, donning a pair of sunglasses, appears to be in the midst of a journey, as suggested by the steering wheel in its hands. The car is parked in a garage, with a large window serving as the backdrop. The teddy bear's position in the driver's seat, along with the steering wheel, gives the impression that it is in control of the car. The sunglasses on the teddy bear add a touch of whimsy to the scene.

Figure A2: Generated Image Captioning



In the image, a young woman is the central figure, seated on a bus. She is dressed in a white dress adorned with ruffles, and her feet are clad in white sneakers. Her hair, styled in loose waves, adds a touch of elegance to her overall appearance. She is looking off to the side, her gaze directed towards something outside the frame of the image. The bus she is seated on is characterized by a blue seat, providing a vibrant contrast to her white attire. The window of the bus offers a glimpse of the outside world, where palm trees sway gently, suggesting a warm and tropical location. The image captures a moment of tranquility amidst the hustle and bustle of a bus journey.



In the image, a group of people are gathered in front of a white building, their hands raised in a wave. The building is characterized by a green awning and a white window. The people, dressed in blue and green clothing, are the main focus of the image. Their smiles and the raised hands suggest a sense of camaraderie and celebration. The image is taken from a low angle, which gives a unique perspective, making the people appear larger and more prominent. The overall composition of the image suggests a joyful gathering, possibly a celebration or a farewell, captured in a moment of shared happiness.



In the heart of a field of white flowers, a man stands as the central figure. His attire is a dark coat and a scarf, providing a stark contrast to the surrounding whiteness. His gaze is directed off to the side, a serious expression etched on his face, suggesting deep contemplation or perhaps distraction. The field of flowers, white and dense, stretches out behind him, their petals untouched and pristine.

Figure A3: Human Captioning



The image captures a breathtaking aerial view of a desert landscape. Dominating the center of the frame is a large rock formation, its reddish-brown hues contrasting with the surrounding terrain. This formation is composed of multiple layers, each one stacked atop the other, creating a sense of depth and scale. The desert itself is a vast expanse of sand, its light beige color stretching out to meet the horizon. Scattered throughout this sea of sand are smaller rock formations, their presence adding texture and variety to the landscape. The sky above is a clear blue, its tranquility undisturbed by clouds. The horizon line is clearly visible in the distance, where the blue of the sky meets the beige of the desert.



The image captures the majestic Machu Picchu, an ancient Incan citadel nestled high in the Andes Mountains of Peru. The ruins, constructed from stone, are perched on a steep mountain ridge, offering a breathtaking view of the surrounding landscape. The perspective of the image is from a high angle, providing a comprehensive view of the ruins and the mountainous terrain. The colors in the image are predominantly earth tones, reflecting the natural environment, with the stone ruins providing a stark contrast.

Figure A4: Scenary Captioning