BiMax: Mitigating Object Hallucination in Large Multimodal Models via Bijective Maximum Likelihood Learning

Huu-Thien Tran[1], **Thanh-Dat Truong**[1], **Thai-Son Tran**[2], **Khoa Luu**[1]
[1]CVIU Lab, University of Arkansas, USA
[2]Vietnam National University, Ho Chi Minh City University of Science, Vietnam
`{ht035, tt032, khoaluu}@uark.edu` `ttson@fit.hcmus.edu.vn`
`https://uark-cviu.github.io/`

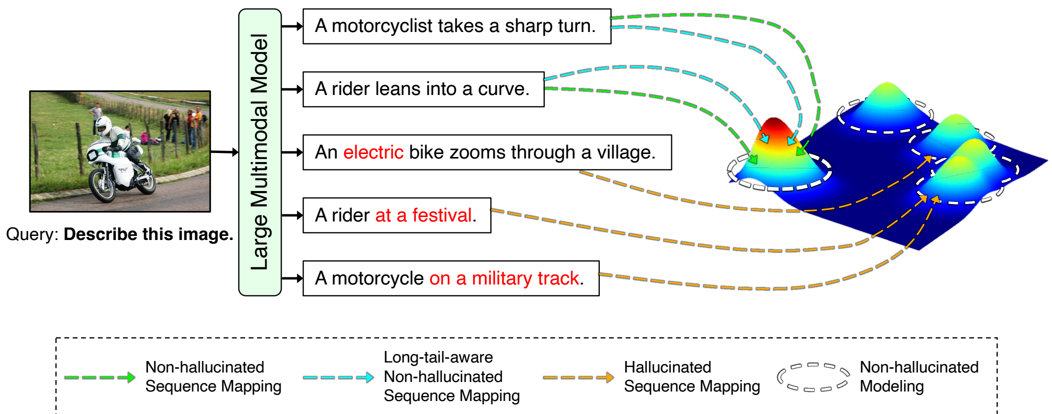

Figure 1: Overview of our proposed approach. Our **BiMax** employs a probabilistic model to address object hallucinations by proposing a bijective mapping, learning the distributional structure of model's generated sequences. It can be further enhanced by a long-tail-aware modeling, stemming from the inherent heavy-tailed effect in real-world training data. **Best viewed in color.**

Abstract

Large multimodal models have advanced rapidly and been widely integrated into interactive communication systems across various applications. Their versatility and extensive knowledge base have transformed the management of essential tasks, surpassing manual processes and facilitating unprecedented levels of efficiency. However, these models are inherently prone to hallucination, negatively impacting their reliability and trustworthiness. Specifically, object hallucination poses a critical challenge requiring attention in the context of vision-language applications. In this work, we tackle the object hallucination problem using a novel probabilistic Bijective Maximum Likelihood (**BiMax**) approach to analyze distributional structures of the output sequences. In addition, the proposed method is also able to manage the long-tail distribution observed in the prevalent training datasets. The experimental results demonstrate the performance improvements of the proposed **BiMax** methods in various settings, paving the way for alleviating object hallucination. This effort seeks to advance the safety and robustness of artificial intelligence models, both in the present era and in anticipated future developments.[1]

[1]The code and pretrained models of this work will be released after this work has been accepted.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Given the rapid developments of foundation models, particularly the recent progress of large language models (LLMs) [1–5], the emergence of high-quality large multimodal models (LMMs) has been witnessed [6–10]. LMMs enable the exploration of generalizability across diverse modalities, including textual data, visual inputs, or auditory signals. Furthermore, they facilitate interactive human-machine communications via multimodal prompts. The advancement of LMMs and the remarkable growth of computational infrastructure have increased accessibility to these models. However, it presents a significant challenge in employing these models, i.e., despite their prompt responses to users' inquiries, how can we ensure their trustworthiness and the accuracy of those responses?

During the process of enhancing their reliability, one of the most common problems is the occurrence of "hallucination". In LMMs, hallucinations occur when the model generates semantically coherent responses yet conflict with the input prompts, whether textual or visual. This phenomenon underscores the urgent need for an effective solution, as hallucinations can have catastrophic consequences resulting from erroneous decision-making and the dissemination of false information.

Problem Motivation. Hallucinations concerning the presence of objects [11–13] are challenging. Prior studies [14, 15] have developed various strategies to mitigate hallucinations, such as refining the decoding process in LMMs [16–18] or applying mechanistic interpretability techniques [19–21]. However, few have examined the underlying token sequence distribution. This study incorporates distributional structure modeling to fill that gap. As perceived in earlier research [22, 23], object frequencies follow a long-tail distribution, reflecting the world's inherent imbalance and common in prevalent large-scale training datasets. Tokenization of these objects reveals a pattern analogous to linguistic elements, which also follows Zipf's law [24, 25]. Addressing the probabilistic implications of this long-tail effect enables more accurate modeling, thereby reducing hallucination.

To this end, this work proposes a novel **Bi**jective **Max**imum Likelihood (**BiMax**) learning approach to solve the hallucination problem. **BiMax** represents the problem of object hallucination with a probabilistic model while alleviating the heavy-tailed dilemma. Extensive experiments across a wide range of benchmarks [11, 12, 26, 27] demonstrate that our proposed **BiMax** can mitigate object hallucinations effectively, thereby improving the reliability of LMMs.

Contributions of this Work. There are four key contributions in this work as follows. First, we study the object hallucination problem from a probabilistic perspective with the background of distributional structure modeling. Second, we detect and analyze the critical issue of long-tail distribution of object frequencies in training datasets, benefiting our probabilistic model framework. Third, we propose **BiMax**, which tackles object hallucinations via Bijective Maximum Likelihood learning, while simultaneously harmonizing token correlations to reduce the impact of heavy tails. Finally, the ablation studies and extensive empirical evaluations on standard benchmarks with State-of-the-Art (SOTA) baselines demonstrate the capability of **BiMax** in addressing these challenges, enhancing the model's reliability and trustworthiness.

2 Related Work

2.1 Hallucination in Large Multimodal Models

Recent studies have observed the emergence of large-scale foundation models capable of learning across diverse modalities, driven by early initiatives [28, 29], which investigate the alignment between visual and textual representations. In addition, the open-sourcing of LLMs [1, 2, 4, 30–32] has contributed enormously to this development. As a result, superior LMMs [6–8, 33–36] have significantly advanced the field by enabling users to engage in more immersive and interactive conversations with the models via visual and textual prompts. Contemporarily, LMMs utilize images and texts as input sources and typically follow a two-phase training procedure. The first phase, known as pre-training, aligns cross-modality representations. The second phase, instruction fine-tuning, teaches the model to converse with users. However, these LMMs are prone to a critical hallucination problem, which significantly impacts the credibility of their outputs.

A comprehensive synthesis and outline of prior research addressing hallucinations in LMMs are presented in [14, 15]. The primary objective is to address object hallucination, misalignment with visual content, and over-reliance on language priors. Numerous methods [17, 18, 37] concentrate on the contrastive decoding technique, which utilizes perturbed inputs or internal descriptions as anchors to steer the generation of the next token. [38] subsequently employs the model's visual descriptions as contrastive anchors. In contrast, [39] revisits contrastive decoding with a multi-tentacled approach to determine the most suitable decoding method for each type of hallucination. Other findings about the negative impact of language priors—such as "text inertia" [40] or "anchor pattern" [16]—lead to the strategy of attention modulation. [41] improves grounding by merging global and local visual features, while [42] nullifies misleading priors using "HalluSpace" projection. Post-hoc correction approaches [43, 44] identify hallucinations by post-generation remedy and statistical insights. Some works focus on specialized techniques such as optimizing token relevance [45], length-aware mechanism [46], and pruning irrelevant visual tokens [47]. [48, 49] refine positional strategies, and [50–52] introduce training-based strategy as a preference-optimized problem.

2.2 Bijective Maximum Likelihood Learning

Flow-based generative models offer a principled approach to density estimation by transforming a simple base distribution into a complex target distribution through a sequence of invertible mappings, leveraging the change-of-variable formula and normalizing flow theory [53]. Among the seminal works in this domain, Dinh et al. [54] introduce RealNVP, a stable and tractable architecture that preserves invertibility and enables exact log-likelihood computation. Kingma et al. [55] later advanced this direction by proposing Glow, incorporating reversible 1×1 convolutions to generate high-resolution images efficiently. Other threads of development concentrate on integrating autoregressive mechanisms with normalizing flows. Germain et al. [56] introduce masked autoencoders to enforce autoregressive constraints within feedforward architectures, inspiring variants such as the inverse autoregressive flow [57] or masked autoregressive flow [58]. Applications of bijective maximum likelihood learning employing flow-based techniques extend beyond generative modeling. Duong et al. [59] explore a bijective metric learning approach for facial identity synthesis, demonstrating the capacity of these models for interpretable feature manipulation. Similarly, Truong et al. [60] introduce a bijective maximum likelihood framework tailored for unsupervised domain adaptation, reinforcing the versatility of flow-based formulations in diverse tasks.

2.3 Tokenizer, the Zipf's Law, and Fairness

While working with multimodal models, which accept inputs from various modalities, most models' outputs are primarily textual data. To transfer the knowledge of natural languages to computational utterances that machines can comprehend, these sentences are tokenized into the well-established fundamental element: a token. In prevalent LLM systems [2, 30, 32], tokens are considered imperative [61], defined by a set of building blocks that can either encrypt or decrypt any corpus of natural language texts. Tokenization also comprises the creation of a relevant vocabulary and the parsing, converting human text to a machine's vocabulary. Within the field of natural language, Zipf's law [24] is proposed for the rank-frequency distribution for words in a text or corpus, i.e., $\texttt{frequency} \propto 1/\texttt{rank}$ [25]. There are diverse tokenizers [62–66], and most of them are proven to create a vocabulary of tokens following Zipfian distribution [67]. Thus, there is an intrinsic ordering between tokens irrespective of the tokenizer employed. However, tokens that describe objects, e.g., person, car, bird, etc., should not be treated unequally since they hold syntagmatic information of the entity in the context. This problem of fairness among object labels is critical and well-studied in prior work, e.g., in domain adaptation [22, 23]. This research contributes to the ongoing study of fairness among tokens in LLMs by mitigating the heavy tail effect, enhancing the resilience of the models.

3 The Proposed BiMax Approach

In this section, we first provide an overview of the learning objective in autoregressive vision-language modeling (Sec. 3.1). Subsequently, we delve into a detailed analysis of this learning objective, considering the assumption of an ideal distribution for the distributional modeling of the output sequence (Sec. 3.2). This analysis culminates in the delineation of two primary components in the proposed **BiMax** approach, including the calibration loss (Sec. 3.3) and the balance loss (Sec. 3.4).

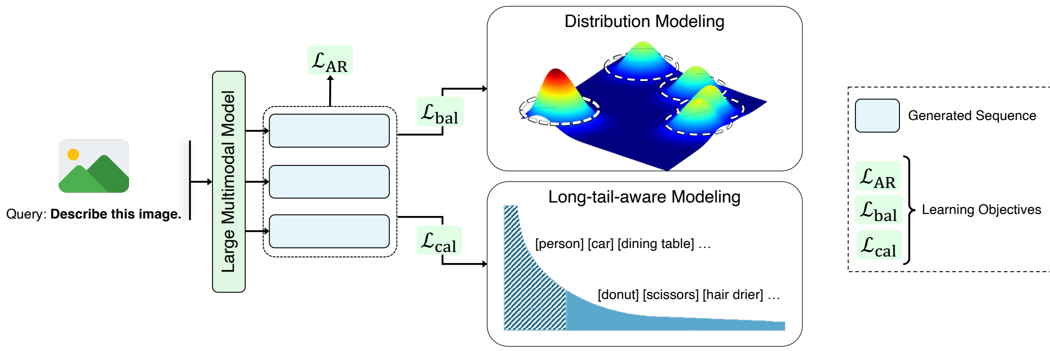

Figure 2: The overall framework of BiMax. **Best viewed in color.**

3.1 Preliminary to Autoregressive Vision-Language Models

The current advancements of multimodal learning of large-scale models result from the rapid development of autoregressive modeling [1–4, 7–9, 31, 32, 35, 36, 68]. Notably, autoregressive LLMs lead to many cutting-edge large-scale models capable of handling convoluted tasks. Autoregressive vision-language modeling is developed based on the growth of LLMs. Particularly, visual content is also incorporated in addition to receiving textual data, which typically encompasses the system context and the query as input.

Let $\mathbf{x}$ and $\mathbf{v}$ denote the textual and visual tokens, respectively. Next-token prediction follows a probabilistic distribution conditioned on $\mathbf{x}$ and $\mathbf{v}$. Precisely, given θ parameterizes the model, this procedure is formally written as in Eqn. (1).

$$p_\theta(\mathbf{y}|\mathbf{x},\mathbf{v}) \propto \prod_{i=0}^{L-1} p_\theta(y_i|\mathbf{x},\mathbf{v},y_{<i}), \tag{1}$$

where $\mathbf{y}$ is the pertinent response for the above derivation with the length of L; y_i is the i-th token with respect to the generation timestep, and $y_{<i} = \{y_j\}_{j=0}^{i-1}$ represents the preceding token sequence. The desired model θ^* can be attained by minimizing the negative log-likelihood as in Eqn. (2).

$$\theta^* = \operatorname*{argmin}_\theta \mathbb{E}\left[-\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{v})\right] = \operatorname*{argmin}_\theta \mathbb{E}\left[-\sum_{i=0}^{L-1} \log p_\theta(y_i|\mathbf{x},\mathbf{v},y_{<i})\right] \tag{2}$$

For further derivations pertaining to this objective, we define $\mathcal{L}_{\text{AR}}(\mathbf{y})$ as specified in Eqn. (3), which also depicts the subsequent substitution from Eqn. (2).

$$\theta^* = \operatorname*{argmin}_\theta \mathbb{E}\left[\mathcal{L}_{\text{AR}}(\mathbf{y})\right] \quad \text{subject to} \;\; \mathcal{L}_{\text{AR}}(\mathbf{y}) \triangleq -\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{v}) \tag{3}$$

3.2 A Granular Analysis of the Autoregressive Learning Objective

Let $q(\mathbf{y})$ be the actual distribution of tokens within the given set of sequences. The following discussion assumes the existence of an ideal sequence distribution, denoted as $\tilde{q}(\mathbf{y})$, wherein every object-centric token is treated equitably. Under this assumption, our proposed approach can be devised, and afterwards, this assumption can be diminished, thereby eliminating any additional preconditions for achieving this ideal distribution. Formally, learning Eqn. (3) under the ideal data distribution across object-centric tokens can be formulated as in Eqn. (4).

$$\theta^* = \operatorname*{argmin}_\theta \mathbb{E}\left[\mathcal{L}_{\text{AR}}(\mathbf{y}) \cdot \frac{\tilde{q}(\mathbf{y})}{q(\mathbf{y})}\right], \tag{4}$$

where the quantity $\frac{\tilde{q}(\mathbf{y})}{q(\mathbf{y})}$ denotes the proportion between the ideal and actual data distributions. In other words, this can be expressed as the complementary component of the model, ensuring a reasonable difference between distributions. Considering the individual functionality of each token in the sequence with respect to the sequential ordering, we can derive Eqn. (4) as outlined in Eqn. (5).

$$\mathbb{E}\left[\mathcal{L}_{\text{AR}}(\mathbf{y}) \cdot \frac{\tilde{q}(\mathbf{y})}{q(\mathbf{y})}\right] = \mathbb{E}\left[\mathcal{L}_{\text{AR}}(\mathbf{y}) \cdot \sum_{k=1}^{L} \frac{\tilde{q}(y_k)\tilde{q}(\mathbf{y}_{<k})}{q(y_k)q(\mathbf{y}_{<k})}\right], \tag{5}$$

where y_k represents the k-th token in the sequence of length L; $\mathbf{y}_{<k}$ represents the predicted sequence of $\mathbf{y}$ up to the k-th token; $q(y_k)$ and $\tilde{q}(y_k)$ denote the likelihood of token y_k, while $q(\mathbf{y}_{<k})$ and $\tilde{q}(\mathbf{y}_{<k})$ denote the distributional structure of $\mathbf{y}_{<k}$ under the actual and ideal data distribution, respectively. At this point, the learning objective from Eqn. (3) is equipped with a greater granularity, contemplating the relationship between tokens in the output sequence. With further derivations, Eqn. (5) can be formally written as in Eqn. (6).

$$\mathbb{E}\left[\mathcal{L}_{\text{AR}}(\mathbf{y}) \cdot \frac{\tilde{q}(\mathbf{y})}{q(\mathbf{y})}\right] = \mathbb{E}\left[\mathcal{L}_{\text{AR}}(\mathbf{y})\right] + \underbrace{\frac{1}{L}\sum_{k=1}^{L}\mathbb{E}\left[\log\frac{\tilde{q}(y_k)}{q(y_k)}\right]}_{\text{balance loss}} + \underbrace{\frac{1}{L}\sum_{k=1}^{L}\mathbb{E}\left[\log\frac{\tilde{q}(\mathbf{y}_{<k})}{q(\mathbf{y}_{<k})}\right]}_{\text{calibration loss}}. \tag{6}$$

To sum up, the learning objective can be decomposed into three intuitive components: the autoregressive loss, the balance loss (denoted by $\mathcal{L}_{\text{bal}}$), and the calibration loss corresponding to how to overcome hallucinations (denoted by $\mathcal{L}_{\text{cal}}$). The autoregressive loss is already depicted in Eqn. (3), and the other terms are discussed in Sec. 3.3 and Sec. 3.4. The overall framework of our proposed **BiMax** is represented in Fig. 2.

3.3 The Probabilistic Lens of Object Hallucination

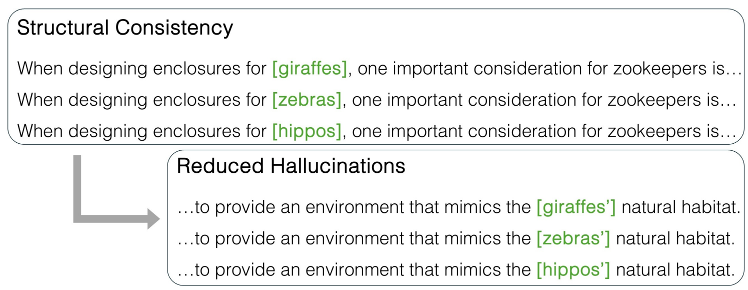

Figure 3: Illustration of the distributional structure surrounding object tokens of interest (highlighted by green texts). This illustration demonstrates the comparable structure enclosing object tokens with identical syntagmatic information to the context.

The calibration loss plays an important role in mitigating hallucinations happening among object tokens. Given that numerous research endeavors assume the distributional structure of natural language elements [69, 70] or pixel segmentation maps [22, 23], we posit that tokens possess a comparable characteristic, namely the distributional structure of tokens within a tokenized sequence. This structure of the tokenized sequence is depicted in Fig. 3.

In LLMs training, the pre-training stage focuses on developing the model's expertise whilst instruction fine-tuning entails refining the knowledge base of the model to enhance its ability to communicate effectively with users. Thus, it can be asserted that as long as the models' responses are favored by users, those responses should not be regarded as hallucinations. Furthermore, in the realm of multimodal models, the model's response should be coherent and cohesive and aligned with the cross-modality data, such as images. Therefore, we can interpret the instruction-tuning procedure as approximating the underlying distribution that underpins this behavior, thereby reducing the hallucination exhibited in multimodal models.

The Probabilistic Lens of Object Hallucination. To model this underlying distribution, let $\mathcal{X}$ be the considered dataset comprising an array of triplets: a query $\mathbf{x}$, an image $\mathbf{v}$, and a target response $\mathbf{y}$. In this research, these triplets exist under the token representation form.

Remark 1 (Desirable Sequence)**.** *While $\mathcal{X}$ is an instruction fine-tuning dataset, the model generates the output sequence $\hat{\mathbf{y}}$ for each triplet. If the model is well-trained and $\hat{\mathbf{y}} \approx \mathbf{y}$, there is a higher likelihood that $\hat{\mathbf{y}}$ will not be hallucinatory.*

In the context of instruction fine-tuning, the training model is instructed to produce responses most analogous to those in the fine-tuning dataset. It is reasonably assumed that the fine-tuning dataset, consisting of multiple multi-turn dialogues between humans and an intelligent assistant, is meticulously prepared and precisely reflects the system's behavior during conversations.

Remark 2 (Desirable Distribution)**.** *A target distribution π should be established that accurately represents the sequential representation of fine-tuning samples. It is certain that $\mathbf{y}$ conforms to π, i.e. $\mathbf{y} \sim \pi(\mathbf{y})$. Furthermore, the predicted $\hat{\mathbf{y}}$ is ideally anticipated to adhere to π.*

Subsequently, the problem of object hallucination should be alleviated if the model is trained with a clear objective to follow this structural design created by the dataset $\mathcal{X}$.

The Relaxation of Calibration Loss. As stated in the Eqn. (6), the calibration loss conditioned on the k-th element of the token sequence is defined as $\mathcal{L}_{\text{cal},k} \triangleq \mathbb{E}\left[\log \frac{\tilde{q}(\mathbf{y}_{<k})}{q(\mathbf{y}_{<k})}\right]$. Nevertheless, it poses a significant challenge when modeling the calibration loss, which is the ambiguity of the ideal conditional distribution $\tilde{q}(\mathbf{y}_{<k})$. Fortunately, $\mathcal{L}_{\text{cal},k}$ can be relaxed and closely approximated by optimizing the firm upper bound as in Eqn. (7).

$$\mathbb{E}\left[\log \frac{\tilde{q}(\mathbf{y}_{<k})}{q(\mathbf{y}_{<k})}\right] \leq \mathbb{E}\left[-\log q(\mathbf{y}_{<k})\right]. \tag{7}$$

Eqn. (7) remains valid regardless of the definition of the ideal distribution $\tilde{q}(\mathbf{y})$ employed based on the definition of a probabilistic density function. Therefore, by calculating the optimal negative log-likelihood of the R.H.S of Eqn. (7), the L.H.S can also be optimized. It is also noteworthy that this relaxation disentangles the requirement of the ambiguous ideal distribution during training.

The following is about how to accurately approximate the upper bound in Eqn. (7). The problem can be solved through a lens of the maximum likelihood estimation problem. In this work, we discuss using bijective maximum likelihood learning based on the theoretical foundation of normalizing flow.

Preliminary to Normalizing Flow. The core idea behind density estimation with normalizing flows [53] is to construct a bijective and differentiable mapping $f : \mathbb{R}^d \to \mathbb{R}^d$, such that its inverse g satisfies $g \circ f(z) = z$, where $\circ$ denotes function composition. Given a base variable $z_0 \sim p_0(z_0)$ drawn from a known distribution, a sequence of transformations is defined as $z_i = f_i(z_{i-1})$ for $i \in \{1, 2, \ldots, K\}$, with each f_i being invertible to preserve the bijective property throughout the chain. After applying this sequence of transformations, the final variable z_K follows the target distribution $p_K(z_K)$, computed as in Eqn. (8).

$$z_K = f_K \circ \cdots \circ f_2 \circ f_1(z_0). \tag{8}$$

This structure enables tractable and exact density computation by changing variables formula, leveraging the base distribution p_0 and the Jacobian determinants of each f_i. The result is a flexible yet mathematically grounded approach to modeling complex distributions.

Effective Modeling for Calibration Loss. The goal now is to find the bijective transformation between the actual data distribution q and the target distribution π as denoted in Remark 2. Here, we consider mapping the structural distribution $q(\mathbf{y}_{<k})$ to the target distribution $\pi(\bar{\mathbf{y}})$, where $\bar{\mathbf{y}}$ represents the correct distributional structure for non-hallucinated sequences constructed across data. Thus, given the bijective mapping function denoted by $\mathcal{F}$, the objective transformation can be formalized as in Eqn. (9) by the change of variables theorem.

$$\log q(\mathbf{y}_{<k}) = \log \pi(\bar{\mathbf{y}}) + \log\left(\left|\det\left(\frac{\partial \mathcal{F}}{\partial \mathbf{y}_{<k}}\right)\right|\right), \tag{9}$$

where $\det\left(\frac{\partial \mathcal{F}}{\partial \mathbf{y}_{<k}}\right)$ denotes the Jacobian determinant at $\mathbf{y}_{<k}$.

3.4 Long-tail Distribution of Object Frequencies

The Zipfian distribution, also known as a heavy-tail or long-tail distribution, presents a classical yet intricate modeling challenge in the realm of frequency distribution analysis. It is studied and applied in various aspects, such as natural language modeling [24, 25, 67], information theory [71], and economics [72]. This work extends the study of long-tail distribution with respect to the frequency of objects of interest in the problem of object hallucination.

As illustrated in Fig. 4, the supervised fine-tuning and the evaluation sets exhibit long-tail distributions across MSCOCO objects. In other words, the frequency of major categories (e.g. person, car, or dining table) substantially exceeds that of the minority group (e.g. scissors, or hair drier). From a Zipfian distribution perspective, the x-axis represents the order of token rank, while the y-axis represents the token frequency. Hence, due to the disproportionately high frequency towards major categories, the rank is thereby decreased. This phenomenon warrants remediation, as a well-performing model is

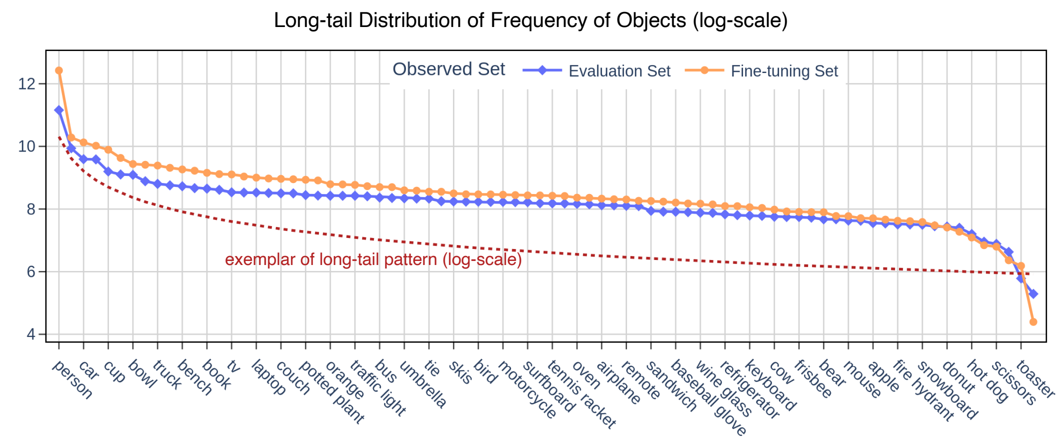

Figure 4: Illustrations of long-tail distributions across MSCOCO objects in the observed sets. The y-axis represents the log-scale of frequency of occurrence of each category on images.

expected to avoid favoring a specific group of classes of tokens, particularly when they all convey the same syntagmatic information.

For the model, to correctly learn how to respond the questions about world knowledge, the distribution of entities in the data must be met. However, the fine-tuning data is curbed by the impact of the heavy tail of lower-frequency objects. By one way or another, the established rank-frequency correlation between objects in the fine-tuning set considerably impacts the responses from a proficient model. Therefore, the distribution of objects among questions answered by a high-quality model still has a heavy tail towards rarer categories.

With that goal, the proposed balance objective $\mathcal{L}_{\text{bal}}$ between tokens of interest amplifies the object-centric fairness and enhance the significance of the tail group in the overall improvement of the model's robustness. If all object tokens are treated equally, one presumption is that they must be distributed uniformly. Therefore, we provide a relaxation of the ideal distribution $\tilde{q}$ by modeling it as a uniform distribution, i.e., $\tilde{q}(y_k) \triangleq \frac{1}{|\mathcal{V}|}, \forall k : 1 \leq k \leq L$; where $|\mathcal{V}|$ is the size of model's vocabulary.

4 Experimental Results

4.1 Dataset and Baseline

Dataset. The instruction fine-tuning dataset[2] released by LLaVA v1.5 [33] is adopted for empirical evaluation. In particular, the `llava_v1_5_mix665k` variant is used. It comprises over 665,000 multi-turn dialogues between users and an intelligent assistant, covering a broad spectrum of instruction-following tasks. The utilization of this dataset also ensures the alignment with prior studies and enables a fair empirical comparison.

Table 1: Evaluation results of our proposed **BiMax** compared to prior hallucination mitigation approaches on the CHAIR benchmark [11]. LLaVA v1.5 [33] serves as the baseline. Best values are in **bold** and underlined.

Method	$\text{CHAIR}_\text{S}\downarrow$	$\text{CHAIR}_\text{I}\downarrow$
Baseline	48.8	14.2
OPERA [16]	44.6	12.8
ICD [37]	47.4	13.9
VCD [17]	46.8	13.2
SID [18]	44.2	<u>12.2</u>
ProjectAway [19]	<u>42.0</u>	<u>12.2</u>
BiMax (ours)	**38.2**	**11.5**

Baseline. LLaVA-v1.5 [33] is a widely adopted open-source vision-language model, frequently used in hallucination-related research. It integrates a visual encoder with a large language model via a projection layer that maps visual features into the language model's embedding space, enabling robust multimodal alignment. Building on the framework introduced in [6], LLaVA-v1.5 [33] incorporates the `AnyRes` mechanism to support higher-resolution image processing. LLaVA-v1.5 can be utilized with various settings of LLM backbones. The Vicuna-7B [30] and Qwen2-7B [2] are

[2] `https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K`

Table 2: Evaluation results of our proposed **BiMax** in comparison with prior hallucination mitigators on the POPE MSCOCO benchmark [12]. The baseline LLaVA v1.5 [33] is employed in the experiment. Best values are in **bold** and <u>underlined</u>.

Method	Random		Popular		Adversarial	
	Accuracy↑	F1 Score↑	Accuracy↑	F1 Score↑	Accuracy↑	F1 Score↑
Baseline	84.77	82.28	79.98	79.34	76.03	76.26
OPERA [16]	88.85	88.67	82.77	83.40	79.16	80.93
ICD [37]	87.97	87.84	84.03	84.22	80.21	80.97
VCD [17]	87.02	86.96	83.53	84.56	78.12	80.16
SID [18]	<u>89.46</u>	**89.62**	85.13	<u>85.94</u>	<u>83.24</u>	<u>82.21</u>
AGLA [41]	88.54	87.71	<u>85.14</u>	84.68	81.13	81.36
CCA-LLaVA [49]	88.03	86.65	86.87	85.54	85.67	84.42
BiMax (ours)	**89.60**	<u>88.95</u>	**88.73**	**87.72**	**87.47**	**86.52**

suitable for fair comparison with prior works, while Qwen2-0.5B-Instruct [2] is a lightweight model, well-suited for rapid prototyping in ablation studies in this research.

4.2 Experimental Results across Benchmarks

CHAIR. The Caption Hallucination Assessment with Image Relevance (CHAIR) [11] is a rule-based metric for measuring object hallucination in image captions, specifically how often captions mention objects not in the ground-truth. It includes two variants, i.e., $\mathrm{CHAIR_S}$ (sentence-level) and $\mathrm{CHAIR_I}$ (image-level), offering complementary views on caption accuracy. Tab. 1 shows that the proposed **BiMax** surpasses prior methods with a large margin across evaluation metrics. Particularly, **BiMax** reaches 38.2% and 11.5% on $\mathrm{CHAIR_S}$ and $\mathrm{CHAIR_I}$, respectively, showcasing its remarkable improvement on this image captioning benchmark.

Table 3: Evaluation results of our proposed **BiMax** compared to SOTA models on the AMBER benchmark [26]. Best values are in **bold** and <u>underlined</u>.

Model	Vision	LLM	AMBER Score↑
mPLUG-Owl [73]	ViT-L14	LLaMA-2-7B	48.7
mPLUG-Owl2 [74]	ViT-L14	LLaMA-2-7B	84.0
LLaVA [6]	ViT-L14	Vicuna-7B	60.6
MiniGPT-4 [35]	ViT-L14	Vicuna-7B	75.6
CogVLM [75]	ViT-L14	Vicuna-7B	83.4
LLaVA-v1.5 [33]	ViT-L14	Vicuna-7B	83.5
InstructBLIP [34]	ViT-G14	Vicuna-7B	<u>86.5</u>
BiMax (ours)	ViT-L14	Qwen2-7B	**89.0**

POPE. The Polling-based Object Probing Evaluation (POPE) [12] is a benchmark for detecting object hallucination in LMMs. It asks models yes/no questions like "Is there a <object> in the image?" using a balanced mix of present and absent objects. POPE has three splits based on how negative samples are chosen: random (uniformly sampled), popular (common objects to test over-generalization), and adversarial (contextually tricky objects to test bias). As shown in Tab. 2, **BiMax** shows superiority across POPE's settings. Notably, it reaches an accuracy of 89.60% $(+0.14\%)$, 88.73% $(+3.60\%)$, 87.47% $(+4.23\%)$ on random, popular, and adversarial splits, respectively. **BiMax** also improves F1 score on popular and adversarial splits by 1.78 and 4.31.

AMBER. The LLM-free multi-dimensional benchmark AMBER [26] facilitates the evaluation of object hallucination on both generative and discriminative tasks, with the AMBER score as the

Table 4: Evaluation results of our proposed **BiMax** compared to prior LMMs on the AMBER discriminative benchmark [26], comprising Existence, Attribute, and Relation hallucination settings (details in [26]). Best values are in **bold** and <u>underlined</u>.

Model	Existence			Attribute		Relation	
	Precision↑	Recall↑	F1 Score↑	Accuracy↑	F1 Score↑	Accuracy↑	F1 Score↑
mPLUG-Owl [73]	99.7	9.4	17.2	55.7	22.9	59.6	6.2
mPLUG-Owl2 [74]	**100**	<u>80.4</u>	<u>89.1</u>	<u>76.6</u>	72.4	58.6	54.3
LLaVA [6]	<u>99.9</u>	4.4	8.4	62.9	48.6	63.8	58.1
LLaVA-1.5 [33]	**100**	71.5	83.3	72.0	64.6	**73.9**	65.6
MiniGPT-4 [35]	<u>99.9</u>	66.7	80.0	61.7	43.7	63.4	52.7
CogVLM [75]	**100**	73.3	84.5	66.8	57.4	66.7	59.8
InstructBLIP [34]	**100**	80.2	89.0	76.1	<u>76.3</u>	66.8	<u>67.6</u>
BiMax (ours)	**100**	**92.7**	**96.2**	**83.8**	**83.2**	<u>68.1</u>	**70.7**

evaluation metric. Particularly, the discriminative task has three settings considering the object's interrelationship, namely existence, attribute, and relation. Tab. 3 demonstrates that our proposed **BiMax** outperforms SOTA models while evaluated on this benchmark, scoring 89.0% ($+2.5\%$). Furthermore, as observed in Tab. 4, our **BiMax** achieves remarkable performance on discriminative sub-tasks. Remarkably, **BiMax** increases the F1 score of these three sub-tasks by 3.1–7.1%.

PhD. The ChatGPT-Prompted Visual Hallucination Evaluation Dataset (PhD) [27] provides a comprehensive benchmark with four distinct variants: base (normal VQA), sec (specious context), icc (incorrect context), and ccs (counter common sense artificial images). The metric employed in this benchmark is PhD-Index, computed based on the Yes/No recall ratio. Tab. 5 exhibits the superior performance of **BiMax** on this benchmark, surpassing prior SOTA approaches by a considerable amount. Specifically, **BiMax** achieves the PhD-Index of 50.8%, demonstrating **BiMax**'s benefit in combating hallucinations.

Table 5: Evaluation results of our proposed **BiMax** compared to prior hallucination mitigation approaches on the PhD benchmark [27]. Best values are in **bold** and <u>underlined</u>.

Model	Vision	LLM	PhD-Index↑
InstructBLIP-L [34]	ViT-G14	Vicuna-13B	27.8
InstructBLIP [34]	ViT-G14	Vicuna-7B	30.5
mPLUG-Owl2 [74]	ViT-L14	LLaMA-2-7B	32.0
MiniGPT-v2 [76]	ViT-G14	LLaMA-2-7B	39.0
LLaVA [6]	ViT-L14	Vicuna-7B	13.5
LLaVA-1.5 [33]	ViT-L14	Vicuna-7B	26.5
LLaVA-1.5-L [33]	ViT-L14	Vicuna-13B	27.0
LLaVA-1.6 [77]	ViT-L14	Vicuna-7B	37.3
LLaVA-1.6-L [77]	ViT-L14	Vicuna-13B	42.3
Qwen-VL [78]	ViT-bigG/14	Qwen-7B	<u>48.8</u>
BiMax (ours)	ViT-L14	Qwen2-7B	**50.8**

4.2.1 Ablation Study

Effectiveness of $\mathcal{L}_{\text{bal}}$ and $\mathcal{L}_{\text{cal}}$. Tab. 6 demonstrates the performance of the model on CHAIR evaluation with three different settings: without $\mathcal{L}_{\text{bal}}$ and $\mathcal{L}_{\text{cal}}$, without $\mathcal{L}_{\text{cal}}$, and the full form of the proposed objective loss. The illustrated results show that with the incorporation of $\mathcal{L}_{\text{bal}}$, the performance of the model on CHAIR benchmark is enhanced, with 0.5% decrease on CHAIR_S and 0.5% decrease on CHAIR_I. Moreover, with the addition of $\mathcal{L}_{\text{cal}}$, the model achieves a greater improvement (2.6% and 0.2%, respectively). It concludes the impact of our proposed learning objectives.

Effectiveness of bijective mapping in sequence modeling. This ablation study focuses on the impact of sequence modeling incorporating different density estimators. As illustrated in Tab. 7, modeling distributional structure using flow-based model as in the proposed framework achieves an improvement of 1.6% and 1.0% on CHAIR_S and CHAIR_I, respectively. This culminates in the effectiveness of the proposed bijective maximum likelihood learning in modeling the distributional structure of output sequences, thereby reducing object hallucinations.

5 Conclusion

In this work, we confront the persistent challenge of object hallucination in LMMs by adopting a probabilistic perspective on sequence distribution modeling. Our analysis also reveals a long-tail distribution pattern in training data that exacerbates hallucination and distorts the underlying token correlations during generation. To address this, we introduce **BiMax**, which leverages bijective maximum likelihood learning to model sequence distributions more faithfully while mitigating heavy-tail effects. Extensive experiments and ablation studies demonstrate **BiMax**'s impact in reducing object hallucination, enhancing the reliability and robustness of vision-language models.

Table 6: Ablation study on the effectiveness of each learning objective in the performance of the LLaVA-v1.5/Qwen2-0.5B-Instruct model on CHAIR benchmark.

Objective			Metric	
$\mathcal{L}_{\text{AR}}$	$\mathcal{L}_{\text{bal}}$	$\mathcal{L}_{\text{cal}}$	$\text{CHAIR}_\text{S}\downarrow$	$\text{CHAIR}_\text{I}\downarrow$
✓	-	-	70.5	24.5
✓	✓	-	69.0	24.0
✓	✓	✓	**66.4**	**23.8**

Table 7: Ablation study on the effectiveness of bijective maximum likelihood estimator in modeling distributional structure of output sequences. The LLaVA-v1.5/Qwen2-0.5B-Instruct model is evaluated on CHAIR benchmark.

Method	Metric	
	$\text{CHAIR}_\text{S}\downarrow$	$\text{CHAIR}_I\downarrow$
Gaussian Mixture Modeling	68.0	24.8
Bijective Mapping	**66.4**	**23.8**

Acknowledgments and Disclosure of Funding

Our work is supported by the National Science Foundation under Award No. OIA-1946391 (NSF DART) and NSF SBIR Phase II under Award No. 2247237. This research is also supported by the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission.

References

[1] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

[5] Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic, 2024.

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[7] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv*, 2024.

[8] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv*, 2025.

[9] 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*, 2023.

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[11] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

[12] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

[13] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143, 2024.

[14] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

[15] Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *arXiv*, 2024.

[16] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.

[17] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[18] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-Introspective Decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[19] Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025.

[20] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*, 2024.

[21] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024.

[22] Thanh-Dat Truong, Ngan Le, Bhiksha Raj, Jackson Cothren, and Khoa Luu. Fredom: Fairness domain adaptation approach to semantic scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19988–19997, 2023.

[23] Thanh-Dat Truong, Hoang-Quan Nguyen, Bhiksha Raj, and Khoa Luu. Fairness continual learning approach to semantic scene understanding in open-world environments. *Advances in Neural Information Processing Systems*, 36:65456–65467, 2023.

[24] George Kingsley Zipf. *Selected studies of the principle of relative frequency in language*. Harvard university press, 1932.

[25] Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130, 2014.

[26] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. AMBER: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

[27] Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. PhD: A ChatGPT-Prompted Visual hallucination Evaluation Dataset. *arXiv preprint arXiv:2403.11116*, 2024.

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

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

[30] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, March 2023.

[31] 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.

[32] 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.

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

[34] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[35] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[36] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[37] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *ACL (Findings)*, pages 15840–15853, 2024.

[38] Junho Kim, Hyunjun Kim, Kim Yeonju, and Yong Man Ro. CODE: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 37:133571–133599, 2024.

[39] Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating Hallucination via Dynamic Contrastive Decoding. *arXiv preprint arXiv:2503.00361*, 2025.

[40] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer, 2024.

[41] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. AGLA: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*, 2024.

[42] Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. *arXiv preprint arXiv:2412.13817*, 2024.

[43] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67, 2024.

[44] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[45] Xinyu Lyu, Beitao Chen, Lianli Gao, Hengtao Shen, and Jingkuan Song. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *Advances in Neural Information Processing Systems*, 37:122811–122832, 2024.

[46] Hongliang Wei, Xingtao Wang, Xianqi Zhang, Xiaopeng Fan, and Debin Zhao. Toward a stable, fair, and comprehensive evaluation of object hallucination in large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[47] Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. VASparse: Towards efficient visual hallucination mitigation for large vision-language model via visual-aware sparsification. *arXiv preprint arXiv:2501.06553*, 2025.

[48] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418, 2024.

[49] Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. *Advances in Neural Information Processing Systems*, 37:92012–92035, 2024.

[50] Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088, November 2024.

[51] 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.

[52] Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via DPO: On-policy data hold the key. *arXiv preprint arXiv:2501.09695*, 2025.

[53] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[54] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

[55] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2018.

[56] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pages 881–889. PMLR, 2015.

[57] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 2016.

[58] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

[59] Chi Nhan Duong, Thanh-Dat Truong, Khoa Luu, Kha Gia Quach, Hung Bui, and Kaushik Roy. Vec2face: Unveil human faces from their blackbox features in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6132–6141, 2020.

[60] Thanh-Dat Truong, Chi Nhan Duong, Ngan Le, Son Lam Phung, Chase Rainwater, and Khoa Luu. Bimal: Bijective maximum likelihood approach to domain adaptation in semantic scene segmentation. In *International Conference on Computer Vision*, 2021.

[61] Julia Witte Zimmerman, Denis Hudon, Kathryn Cramer, Alejandro J Ruiz, Calla Beauregard, Ashley Fehr, Mikaela Irene Fudolig, Bradford Demarest, Yoshi Meke Bird, Milo Z Trujillo, et al. Tokens, the oft-overlooked appetizer: Large language models, the distributional hypothesis, and meaning. *arXiv preprint arXiv:2412.10924*, 2024.

[62] Philip Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

[63] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018.

[64] Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. Multilingual language processing from bytes. *arXiv preprint arXiv:1512.00103*, 2015.

[65] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[66] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *EMNLP 2018*. Association for Computational Linguistics, 2018.

[67] Elizaveta Zhemchuzhina, Nikolai Filippov, and Ivan P Yamshchikov. Pragmatic constraint on distributional semantics. *arXiv preprint arXiv:2211.11041*, 2022.

[68] 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.

[69] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[70] Magnus Sahlgren. The distributional hypothesis. *Italian Journal of linguistics*, 20:33–53, 2008.

[71] Peter Harremoës and F Topsoe. Zipf's law, hyperbolic distributions and entropy loss. *General Theory of Information Transfer and Combinatorics*, pages 788–792, 2006.

[72] Gerrit de Wit. Zipf's law in economics. *Scales research reports*, 2005.

[73] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[74] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration (2023). *arXiv preprint arXiv:2311.04257*, 2023.

[75] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023.

[76] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[77] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[78] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

A Further Discussion

A.1 Future Work and Broader Impact

Future Work. This research proposes object hallucination via bijection-based distributional structure modeling to enhance the robustness of LLaVA-v1.5. Further extension to this approach can be built by incorporating diverse LMMs to assess its universality in improving system performance against object hallucination. Additionally, we investigate the frequency of tokens representing objects within a specific training dataset to identify the intrinsic long-tail distribution. Future research can explore this phenomenon in a larger dataset.

Broader Impact. Our research can benefit the community by mitigating the hallucinations found in prevalent LMMs. Additionally, our method can be incorporated seamlessly with various off-the-shelf models to help them become more robust against hallucinations.

A.2 Limitations

Our work models hallucinations using a probabilistic model and enhances the robustness of existing LMMs. However, our model requires the training procedure of the language model component. Several methods [17, 18, 41] enable training-free paradigms. This can be left as the potential future work for a training-free or lightweight probabilistic model that addresses this challenge.

B Mathematical Derivations

B.1 Proof of Eqn. (7)

Given $\tilde{q}(\mathbf{y}_{<k})$ the probabilistic density function, we have:

$$0 \leq \mathbb{E}[\tilde{q}(\mathbf{y}_{<k})] \leq 1 \tag{10}$$

$$\Rightarrow \mathbb{E}[\log \tilde{q}(\mathbf{y}_{<k})] \leq \log\left(\mathbb{E}[\tilde{q}(\mathbf{y}_{<k})]\right) \leq 0 \tag{11}$$

$$\Rightarrow \mathbb{E}\left[\log \frac{\tilde{q}(\mathbf{y}_{<k})}{q(\mathbf{y}_{<k})}\right] = \mathbb{E}\left[\log \tilde{q}(\mathbf{y}_{<k})\right] - \mathbb{E}\left[\log q(\mathbf{y}_{<k})\right] \leq -\mathbb{E}\left[\log q(\mathbf{y}_{<k})\right] \tag{12}$$

$$\Rightarrow \mathbb{E}\left[\log \frac{\tilde{q}(\mathbf{y}_{<k})}{q(\mathbf{y}_{<k})}\right] \leq \mathbb{E}\left[-\log q(\mathbf{y}_{<k})\right] \tag{7}$$

Thus, the proof for Eqn. (7) is completed. □

B.2 Derivation for Eqn. (9)

Given $\mathcal{F}$ the bijective transformation matrix mapping the structural distribution $q(\mathbf{y}_{<k})$ to the target distribution $\pi(\bar{\mathbf{y}})$, we have the following derivation using the change of variables theorem, given the following conditions:

$$\mathbf{y}_{<k} \sim q(\mathbf{y}_{<k}), \bar{\mathbf{y}} \sim \pi(\bar{\mathbf{y}}), \bar{\mathbf{y}} = \mathcal{F}(\mathbf{y}_{<k}) \tag{13}$$

$$q(\mathbf{y}_{<k}) = \pi(\bar{\mathbf{y}}) \left|\det\left(\frac{\partial \mathcal{F}}{\partial \mathbf{y}_{<k}}\right)\right| \tag{14}$$

$$\Rightarrow \log q(\mathbf{y}_{<k}) = \log \pi(\bar{\mathbf{y}}) + \log\left(\left|\det\left(\frac{\partial \mathcal{F}}{\partial \mathbf{y}_{<k}}\right)\right|\right) \tag{9}$$

Thus, the derivation for Eqn. (9) is completed. □

C Implementation Details

In our experiments, we adopted the framework from LLaVA-v1.5 [33] as the baseline model. For the vision encoder, we utilized the CLIP-ViT-L14 (336px) variant. For the language model, we employed

Vicuna-7B v1.5 [30] and Qwen2-7B [2] in the main experiments, while Qwen2-0.5B-Instruct [2] served as the ablation study model. For the projection between the vision encoder and language model, we employed a two-layer perceptron. In terms of distributional structure modeling, we utilized the RealNVP [54] model, with the specified number of scales (4) and blocks (8). All models were trained for a single epoch.

During training, we employed 32 NVIDIA A100 GPUs, each equipped with 48GB of VRAM. The 7B models underwent training for an average duration of 3 to 5 days, while the 0.5B model required approximately 5 hours of training. During evaluation, all models can be tested on a single GPU with a minimum VRAM capacity of 24GB.

For evaluations, publicly available repositories were utilized for the implementation (CHAIR[3], POPE[4], AMBER[5], PhD[6]).

D Impact of Rank-Frequency of Fine-tuning data on That of Model's Response

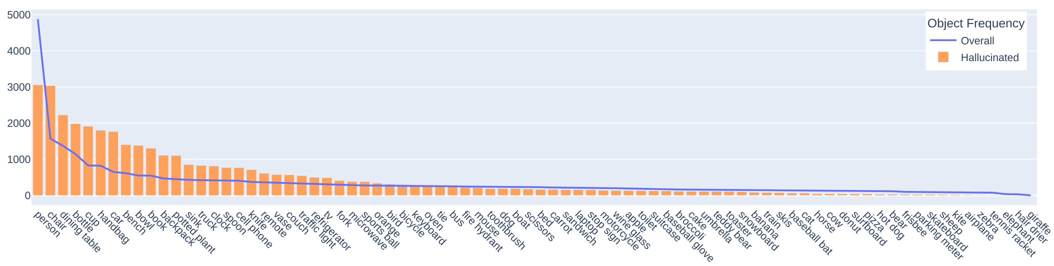

Figure 5: Illustration of the frequency distribution of objects across two settings: (1) the entire evaluation dataset and (2) the hallucinatory captions generated by LMMs. Both settings exhibit a long-tail distribution pattern.

In order for a model to effectively learn how to answer questions that requires knowledge of surrounding world, it is crucial that the training/fine-tuning data accurately reflects the underlying distribution of entities present in the real world. However, training/fine-tuning data often suffers from an imbalanced distribution, where a small number of high-frequency entities dominate, and a long tail of lower-frequency or rare entities is significantly underrepresented. This skewed distribution presents a substantial challenge during training and combating object hallucination. By one way or another, the rank-frequency relationship among entities in the fine-tuning set inevitably influences the model's learning representation. In other words, the statistical structure of entity occurrences in the data tends to imprint itself onto the model's output behavior. As a result, as depicted in Fig. 5, the proficient model like LLaVA-v1.5 [33] tend to inherit this long tail bias. Its responses still disproportionately favor frequently occurring objects, while obscure entities remain underrepresented, thereby preserving the heavy-tailed distribution in the model's answers.

E Some Qualitative Results

Tabs. 8 and 9 show several qualitative results while evaluating our proposed **BiMax** in comparison with prior work [33]. As observed in the results, our model can effectively reduce object hallucinations.

[3]https://github.com/Maxlinn/CHAIR-metric-standalone
[4]https://github.com/EvolvingLMMs-Lab/lmms-eval
[5]https://github.com/junyangwang0410/AMBER
[6]https://github.com/jiazhen-code/PhD

	
User	Please describe this image in detail.
LLaVA-v1.5 [33]	The image features a **man** standing in a field, flying a yellow **kite** high in the sky. The **kite** is visible in the upper part of the scene, soaring above the **man**. The field is surrounded by a few cars parked nearby, with some closer to the foreground and others further in the background. There are several other **people** in the field, possibly enjoying the outdoor activity or watching the man fly the **kite**.
BiMax	In the image, a **man** is standing on grass in an open field flying a yellow and black **kite**. The sky appears to be cloudy but not dark or stormy at this moment. Several other **people** are scattered around the area; some of them may also be engaged with their own activities or simply enjoying each other's company. In addition to the **people**, there are multiple **tents** set up throughout the scene.

	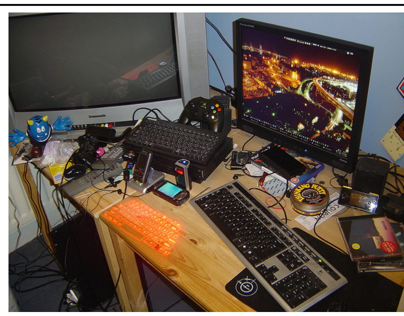
User	Please describe this image in detail.
LLaVA-v1.5 [33]	The image features a **wooden desk** with a **computer setup**, including a **keyboard** and a **monitor**. The **monitor** is positioned above the **keyboard**, and a **mouse** can be seen on the **desk** as well. Additionally, there are two books on the **desk**, one located near the **keyboard** and the other near the **monitor**. The **desk** appears to be a busy workspace with various electronic devices and items.
BiMax	The image features a **computer desk** with various **electronic devices** and **gaming peripherals** on it. A large **monitor** takes up much of the space, covering most of the top area of the **table**. To its right is an additional **screen**, which might be used for multiple purposes or connected to different equipment like **computers**.

Table 8: Some qualitative results on MSCOCO compared with prior work. **Bold text** indicates the correct objects. Underlined red text indicates hallucinated objects.

	
User	Please describe this image in detail.
LLaVA-v1.5 [33]	The image features a **white dog** lying on the floor, with its head inside **a bowl of food**. The **dog** appears to be enjoying its meal, which includes a variety of food items. The **bowl** is placed on a dining table, and there is a **spoon** nearby, possibly used for serving the food.
BiMax	The image features a **white dog** laying on the floor near its food and **water bowls**. Both **dishes** are placed in close proximity, with one located closer to the right side of the room and the other slightly more towards the left. The **dog** is sitting comfortably next to both **bowls** while eating from them.

	
User	Please describe this image in detail.
LLaVA-v1.5 [33]	The image features three vases of varying sizes and shapes, each with a unique design. The first vase is located on the left side of the image, while the second vase is positioned in the middle, and the third vase is on the right side.
BiMax	The image displays three small, decorative **plates** sitting next to each other on a white surface. These ceramic **dishes** feature different patterns and designs, showcasing various styles of artistic expression. They are positioned close together, creating an appealing display for the viewer.

Table 9: Some qualitative results on MSCOCO compared with prior work. **Bold text** indicates the correct objects. Underlined red text indicates hallucinated objects.