Mitigating Hallucinations in Large Vision-Language Models by Adaptively Constraining Information Flow

Jiaqi Bai[1,2], **Hongcheng Guo**[3], **Zhongyuan Peng**[4], **Jian Yang**[3], **Zhoujun Li**[3], **Mohan Li**[1,2*], **Zhihong Tian**[1,2]

[1]Cyberspace Institute of Advanced Technology, Guangzhou University, China
[2]Huangpu Research School of Guangzhou University, China
[3]CCSE, Beihang University, China
[4]University of the Chinese Academy of Sciences, China
{jqbai, limohan, tianzhihong}@gzhu.edu.cn, {hongchengguo, jiaya, lizj}@buaa.edu.cn, zhongyuan.peng@cripac.ia.ac.cn

Abstract

Large vision-language models show tremendous potential in understanding visual information through human languages. However, they are prone to suffer from object hallucination, i.e., the generated image descriptions contain objects that do not exist in the image. In this paper, we reveal that object hallucination can be attributed to overconfidence in irrelevant visual features when soft visual tokens map to the LLM's word embedding space. Specifically, by figuring out the semantic similarity between visual tokens and LLM's word embedding, we observe that the smoothness of similarity distribution strongly correlates with the emergence of object hallucinations. To mitigate hallucinations, we propose using the Variational Information Bottleneck (VIB) to alleviate overconfidence by introducing stochastic noise, facilitating the constraining of irrelevant information. Furthermore, we propose an entropy-based noise-controlling strategy to enable the injected noise to be adaptively constrained regarding the smoothness of the similarity distribution. We adapt the proposed AdaVIB across distinct model architectures. Experimental results demonstrate that the proposed AdaVIB mitigates object hallucinations by effectively alleviating the overconfidence in irrelevant visual features, with consistent improvements on two object hallucination benchmarks.

Code — https://github.com/jiaqi5598/AdaVIB

Introduction

Large Vision-Language Models (LVLMs) (Zhu et al. 2023; Dai et al. 2023a; Bai et al. 2023) have recently attracted increasing attention. Due to the superiority of generating contextually relevant natural language descriptions grounded on visual patterns, LVLMs have shown impressive performance on various vision-language tasks, including image captioning (Deng et al. 2009; Lin et al. 2014), visual question answering (Antol et al. 2015; Li et al. 2023c) and multimodal machine translation (Yao and Wan 2020; Guo et al. 2022). Despite their success, LVLMs are prone to object hallucinations (Rohrbach et al. 2018; Biten, Gómez, and Karatzas 2022), where the generated image descriptions are inconsistent with the objects that appear in the grounded image. This inconsistency has significantly affected the reliability and applicability of LVLMs, especially in scenarios that demand precise judgment.

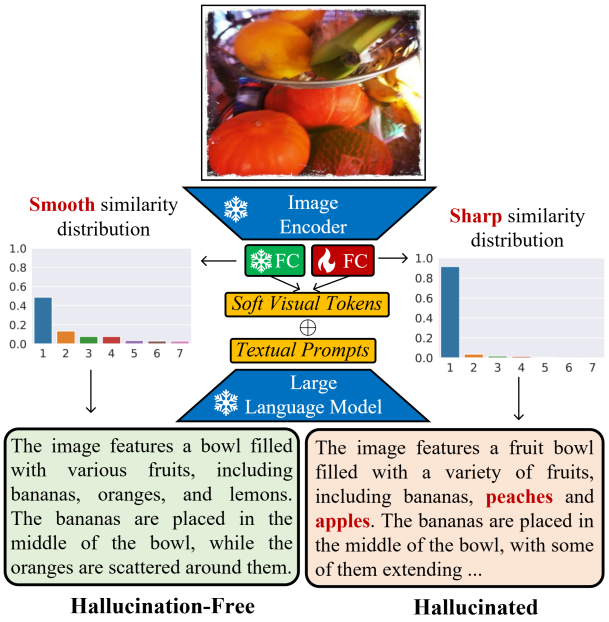

Figure 1: Impact on the smoothness of the similarity distribution correlates with the emergence of object hallucinations. We use the normalized dot product to measure the semantic similarity between soft visual tokens and LLM's word embedding. The y-axis of the similarity distribution denotes the similarity score. The x-axis is the top-ranked LLM's token sorting in descending order.

Most existing works focus on mitigating object hallucinations by reducing over-reliance on prior knowledge of LLMs through the decoding process (Huang et al. 2024; Favero et al. 2024). They devise various methods to penalize specific patterns that may induce hallucinations (Huang et al. 2024; Leng et al. 2024) or curate a training dataset to address statistical biases arising from co-occurrence and spuri-

*Corresponding author.

ous correlation among objects (Biten, Gómez, and Karatzas 2022; Zhou et al. 2024; Liu et al. 2023). Despite their efforts, most of these works rarely focus on narrowing the modality gap between visual patterns and textual descriptions. Recently, researchers have revealed that the vision-language connector plays a significant role in mitigating object hallucinations of LVLMs (Li et al. 2023a; Sun et al. 2023; Jiang et al. 2024). On the one hand, despite their impressive performance on various visual understanding tasks, existing vision encoding techniques are still challenging in expressing visual patterns precisely (Cho et al. 2022; Li et al. 2024; Jain, Yang, and Shi 2024). Hence, compressing irrelevant visual features encoded by the vision encoder is essential to generate hallucination-free descriptions. On the other hand, although previous work investigated various modality alignment approaches, such as Q-Formers (Li et al. 2023a; Dai et al. 2023b) and lightweight projectors (Liu et al. 2022; Alayrac et al. 2022; Gao et al. 2023; Chen et al. 2023; Liu et al. 2024a), the encoded soft visual tokens are still far from ideal in compressing irrelevant while preserving relevant visual information faithful to input images.

In this paper, we reveal that object hallucinations can be attributed to *overconfidence* in irrelevant visual features when soft visual tokens project to the word embedding space of LLM. As Figure 1 shows, we present two distinct cases illustrating how the smoothness of the similarity distribution between soft visual tokens[1] and the LLM's word embedding correlates with the emergence of object hallucinations. Both frozen and fine-tuned fully connected (FC) layers, referred to as the *vision-language projector*, are introduced for comparison to project encoded visual features into soft visual tokens. By examining the normalized dot product similarity between soft visual tokens and the word embedding of LLM, we observe that the LVLMs equipped with a frozen FC layer (the green block on the left) generate a hallucination-free description with a smoother similarity distribution. In contrast, the variant equipped with a fine-tuned FC layer (the red block on the right) generates a hallucinated description with a sharper similarity distribution. This phenomenon indicates that the smoothness of the similarity distribution strongly correlates with the emergence of object hallucinations. A sharp similarity distribution indicates the occurrence of the *overconfidence* problem, which results from overfitting on statistically spurious correlations with the irrelevant visual features during training.

Therefore, alleviating the *overconfidence* on irrelevant features when soft visual tokens mapping to the LLM's word embedding is essential to mitigate object hallucinations, and it is critical to constrain information flow to soft visual tokens by elaborately devising the vision-language projector.

Motivated by the above analysis, we propose AdaVIB, a lightweight fine-tuning method that uses Variational Information Bottleneck (VIB) (Alemi et al. 2016) to mitigate object hallucinations. The proposed AdaVIB only requires fine-tuning weights of the vision-language projector with other modules frozen. It mitigates *overconfidence* problem by introducing a compression term to regularize the training of vision-language projector. Introducing the compression term can be regarded as adding stochastic noise to soft visual tokens during training, and increasing this noise constrains the information flow to them, which decreases the *overconfidence* in irrelevant visual features when visual tokens mapping to the LLM's word embedding space. To adaptively constrain the visual information flow to the soft visual tokens, we propose an entropy-based noise-controlling mechanism. The proposal adaptively constrains the injected noise regarding the smoothness of the similarity distribution between soft visual tokens and the LLM's word embedding, capturing the dynamic nature of a specific sample. We adapt AdaVIB on two distinct LVLM architectures, including MiniGPT4 (Zhu et al. 2023) and LLava-1.5 (Liu et al. 2024a). Experimental results demonstrate the effectiveness of AdaVIB in mitigating the overconfidence problem by effectively smoothing the similarity distribution between soft visual tokens and LLM's word embedding, with consistent improvements on two object hallucination benchmarks.

Our contributions are three-fold: **i):** We are the first to use VIB to mitigate object hallucinations, which decreases the overconfidence in irrelevant visual features when soft visual tokens map to the word embedding of LLM. **ii):** We propose AdaVIB, an entropy-based noise controlling strategy to adaptive constrain the information conveyed by visual tokens, regarding the smoothness of similarity distribution to LLM's word embedding. **iii):** Comprehensive experiments demonstrate the effectiveness of AdaVIB in mitigating object hallucinations. The proposed approach yields consistent improvements on two object hallucination benchmarks across different model architectures.

Related Work

Information Bottleneck

The Information Bottleneck (IB) principle (Tishby, Pereira, and Bialek 2000; Fischer 2020) is an excellent concept for regularizing internal representations to minimize the mutual information by compressing the original input representation. The compressed representation can further improve the model's generalization capability by ignoring irrelevant features in the original input. IB has been widely adopted for many machine learning tasks (Peng et al. 2018; Li and Eisner 2019; Belinkov, Henderson et al. 2020), such as image generation (Peng et al. 2018), explanation regeneration (Li et al. 2023b), retrieval-augmented generation (Zhu et al. 2024), and so on.

Based on the IB principle, Alemi et al. (2016) introduced variational information bottleneck (VIB), a variational approach that can be instantiated in deep neural networks, inspired by a similar approach in variational autoencoders (VAE) (Kingma and Welling 2013). It has been applied in the study of parsing (Li and Eisner 2019), natural language inference (Belinkov, Henderson et al. 2020), and graph structure learning (Sun et al. 2022). Compared to the above work, our work, to the best of our knowledge, is the first attempt to investigate VIB as a regularization technique to mitigate object hallucinations in LVLMs.

[1]We apply an average pooling over all visual tokens to obtain an overall representation.

Object Hallucinations in LVLMs

Mitigating object hallucinations (Rohrbach et al. 2018; Biten, Gómez, and Karatzas 2022) has been a long-standing challenge in realizing a trustworthy AI system. This issue can be attributed to several possible reasons, e.g., insufficient understanding of the real-world knowledge (Leng et al. 2024; Huang et al. 2024), statistical bias in training data (Tang et al. 2021; Biten, Gómez, and Karatzas 2022; Zhou et al. 2024) or uncertainty to the objects present in the image (Cho et al. 2022; Liu et al. 2022; Li et al. 2024).

To mitigate object hallucinations, existing studies typically involved methods such as contrastive decoding (Leng et al. 2024; Huang et al. 2024), balance co-occurrence patterns through data augmentation (Zhou et al. 2024; Liu et al. 2024a) and devise an aligner to narrow the modality gap between vision and languages (Zhu et al. 2023; Dai et al. 2023a). For example, Chuang et al. (2023) and Leng et al. (2024) introduced distinct decoding approaches to estimate output distributions by contrasting different sources. They effectively alleviate object hallucinations by reducing the over-reliance on the prior knowledge of a single source. Some studies (Biten, Gómez, and Karatzas 2022; Zhou et al. 2024) augmented training data by analyzing key factors underlying object hallucination, effectively reducing the statistical bias to alleviate object hallucinations. Compared to the above work, our work starts with an observation that the object hallucinations stem from the overconfidence problem. Based on this, we propose a VIB-based approach to mitigate object hallucinations by alleviating such a problem.

Methodology

Problem Formulation

We consider a general image-to-text generation problem with a dataset $\mathcal{D} = \{(x_i, q_i, y_i)\}_{i=1}^{N}$, where the i-th triple (x_i, q_i, y_i) consists of an image x_i, an textual prompt q_i and a image description y_i. The goal is to learn a probability distribution $p(y|x, q)$ with $\mathcal{D}$, and thus given a new pair (x, q) in an image-to-text generation, one can generate a response y token-by-token in terms of $p(y|x, q)$. The optimization of $p(y|x, q)$ can be formalized by maximizing the following conditional probability:

$$p(y|x, q) = \prod_{t=1}^{|\mathbf{y}|} p(y_t|y_{<t}, x, q) \tag{1}$$

Large Vision-Language Models

Given an input image x_j, a pre-trained visual encoder (e.g., CLIP (Radford et al. 2021)) first encodes x_j into a dense visual representation $\mathbf{x}_j$, which is then fed to a vision-language connector to obtain an intermediate representation $\mathbf{v}_j$, denoted as $\mathbf{v}_j = g_\theta(x_j)$, where θ is a set of parameters. The vision language connector is optional, and can be arbitrary architectures, such as a Q-Former (Li et al. 2023a; Dai et al. 2023b). After that, a vision-language projector receives $\mathbf{v}_j$ and maps it to the word embedding space of LLM, yielding a sequence of visual tokens $\mathbf{z}_j$, the vision-language projector can be a Multilayer Perceptron (i.e., MLP):

$$\mathbf{z}_j = \mathbf{W}_h \mathrm{GeLU}\left(\mathbf{W}_z \mathbf{v}_j\right) \tag{2}$$

where $\mathbf{W}_z$ and $\mathbf{W}_h$ are trainable parameters. Lastly, by embedding the textual prompt q into embedded representations $\mathbf{q}$, the gathered visual tokens $\mathbf{z}$ are concatenated with $\mathbf{q}$, yielding outputs y token-by-token. Suppose $p_\phi(y|\mathbf{v}, q)$ is an approximation of $p(y|x, q)$ parameterized by ϕ. The above process can be optimized by minimizing the following loss:

$$\min \mathcal{L}_{CE} = \mathbb{E}_{(x,q,y)\sim\mathcal{D},\mathbf{v}\sim g_\theta(x)}\left[-\log\left(p_\phi\left(y|\,\mathbf{v}, q\right)\right)\right] \tag{3}$$

In this paper, we focus on optimizing the vision-language projector with other modules frozen. Our approach can be easily extended to LVLMs with and without a vision-language connector (e.g., Q-Former) by simply replacing the vision-language projector with our component. For the convenience of description, we introduce our method upon the structure without the vision-language connector (e.g., LLaVa (Liu et al. 2024b,a)), and the inputs of the vision-language projector denote as $\mathbf{v}$ unless explicitly specified.

Adaptive Variational Information Bottleneck

Information Bottleneck The Information Bottleneck (IB) (Tishby and Zaslavsky 2015) has been demonstrated as a promising principle to find a compression $\mathbf{z}$ for the original input representation $\mathbf{v}$, that maximally compressing irrelevant information to $\mathbf{v}$ while maintaining relevant information to $\mathbf{y}$. The objective of IB is to minimize the combination of compression loss and prediction loss, which is formalized as follows:

$$\min \mathcal{L}_{\mathrm{IB}} = \underbrace{\beta I(\mathbf{v};\mathbf{z})}_{\text{Compression}} - \underbrace{I(\mathbf{z};\mathbf{y})}_{\text{Prediction}} \tag{4}$$

where $\beta \geqslant 0$ is the Lagrange multiplier for balancing the compression and prediction terms. $I(\cdot;\cdot)$ is the mutual information. The former term of Equation 4 improves the conciseness of the input signal $\mathbf{v}$ by minimizing the inclusion of irrelevant information, encouraging the network to concentrate more on useful content. The latter enables the network to selectively maintain useful information for supporting the predicted content $\mathbf{y}$ faithful to the input $\mathbf{v}$.

Variational Information Bottleneck Alemi et al. (2016) proposed Variational Information Bottleneck (VIB), a variational estimation of IB by approximating the probability distribution via a neural network:

$$\begin{aligned}\min \mathcal{L}_{\mathrm{VIB}} = \beta \mathop{\mathbb{E}}_{\mathbf{v}}[\mathrm{KL}(p_\theta(\mathbf{z}|\mathbf{v})\|r(\mathbf{z}))] \\ + \mathop{\mathbb{E}}_{\mathbf{z}\sim p_\theta(\mathbf{z}|\mathbf{v})}[-\log p_\phi(y|\mathbf{z}, q)]\end{aligned} \tag{5}$$

where $p_\phi(y|\mathbf{z}, q)$ is an estimation of response y parameterized by ϕ, given compressed input representation $\mathbf{z}$ and textual prompt q. $r(\mathbf{z})$ and $p_\theta(\mathbf{z}|\mathbf{v})$ are the estimation of prior and posterior probability to $\mathbf{z}$, respectively. The former part of Equation 5 serves as a compression term, which provides an explicit way to compress input representation $\mathbf{v}$ into $\mathbf{z}$. The compressed process can be regarded as introducing stochastic noise during training, where the noise can be injected by sampling $\mathbf{z}$ from $p_\theta(\mathbf{z}|\mathbf{v})$. Increasing this noise decreases the information conveyed by $\mathbf{z}$. The later part of Equation 5 is a prediction term, which preserves the useful

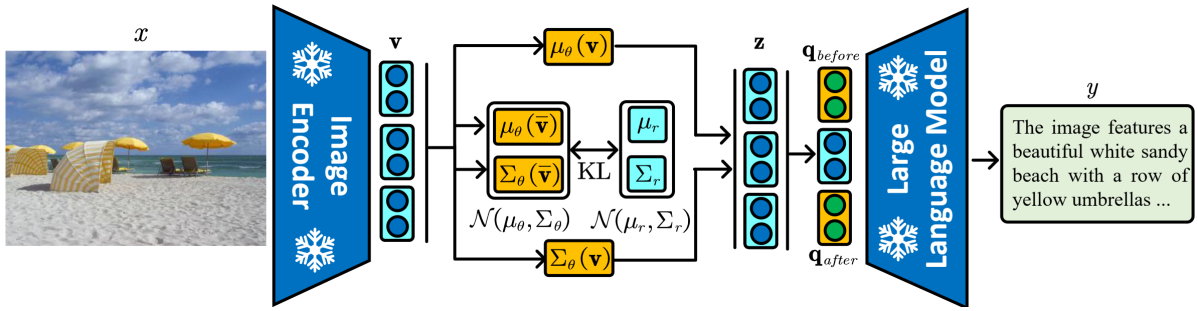

Figure 2: The model architecture of AdaVIB. AdaVIB compresses the input representations $\mathbf{v}$ into soft visual tokens $\mathbf{z}$ with mean $\mu_\theta(\mathbf{v})$ and constrain the irrelevant information by injecting the Gaussian noise with variance $\Sigma_\theta(\mathbf{v})$.

information for the prediction of y. When $\beta = 0$, there is no incentive to inject a noise perturbation; thereby, the VIB degrades to deterministic dimensionality reduction with an MLP, similar to the implementation of Equation 3. During testing, we use the expected value of $\mathbf{z}$ to predict answer y through $p_\phi(y|\mathbf{z}, q)$.

In our experiment, the prior distribution $r(\mathbf{z})$ and posterior distribution $p_\theta(\mathbf{z}|\mathbf{v})$ are modeled by parametric Gaussian distributions, formalized by:

$$r(\mathbf{z}) = \mathcal{N}(\mu_r, \Sigma_r) \tag{6}$$

$$p_\theta(\mathbf{z}|\mathbf{v}) = \mathcal{N}(\mu_\theta(\bar{\mathbf{v}}), \Sigma_\theta(\bar{\mathbf{v}})) \tag{7}$$

where μ_r and μ_θ are mean vectors, Σ_r and Σ_θ are diagonal covariance matrices. $\bar{\mathbf{v}}$ is the average-pooling of $\mathbf{v}$ over all tokens. Because each dimension of these variables is independent and identically distributed, the Kullback-Leibler (KL) divergence of the multivariate Gaussian distribution can be estimated as follows:

$$\begin{aligned}\mathrm{KL}(\mathcal{N}_r||\mathcal{N}_\theta) = \frac{1}{2}[\log\frac{\det(\Sigma_\theta)}{\det(\Sigma_r)} - d_\mathbf{z} + \mathrm{tr}(\Sigma_\theta^{-1}\Sigma_r) \\ + (\mu_\theta - \mu_r)^\top \Sigma_\theta^{-1}(\mu_\theta - \mu_r)]\end{aligned} \tag{8}$$

where $d_\mathbf{z}$ is the dimensionality of $\mathbf{z}$. Similar to Li and Eisner (2019), we apply the reparameterization strategy (Kingma and Welling 2013) to approximate gradients backpropagate to $\mathbf{v}$, formalized by:

$$\mathbf{z} = \mu_\theta(\mathbf{v}) + \Sigma_\theta(\mathbf{v}) \odot \epsilon \tag{9}$$

where ϵ is modeled with a standard Gaussian distribution. In practice, the compressed input representation $p_\theta(\mathbf{z}|\mathbf{v})$ is modeled by two distinct linear layers, each projects input to the same dimensionality with $\mathbf{z}$ for computing $\mu_\theta(\mathbf{v})$ and $\Sigma_\theta(\mathbf{v})$, where $\mu_\theta(\mathbf{v})$ is initialized from the pre-trained weights of the LVLM's vision-language projector, $\Sigma_\theta(\mathbf{v})$ is randomly initialized and apply a softplus transform to ensure outputs non-negativity.

Adaptive β Previous studies (Peng et al. 2018; Li and Eisner 2019) have shown that balancing the compression and prediction terms using a Lagrange multiplier β is crucial for removing irrelevant information while preserving information that is predictive of the model output. While setting the β to be fixed may prevent the constrained procedure from capturing the dynamic nature of the specific sample.

Inspired by our observation that the smoothness of the similarity distribution between soft visual tokens and the LLM's word embedding strongly correlates with the emergence of object hallucinations, we propose Adaptive Variational Information Bottleneck (AdaVIB). The proposal adaptively constrains the injected noise regarding the smoothness of the similarity distribution for capturing the dynamic nature per sample. Specifically, we use the entropy (Shannon 1948; Zhai et al. 2023; Farquhar et al. 2024) of the similarity distribution as an indicator to reflect the degree of soft visual tokens suffering from the overconfidence problem. A sample with a low entropy level refers to a sharp similarity distribution, indicating that the soft visual tokens are prone to be overconfident (Pereyra et al. 2017) when mapping to the LLM's embedding space. In contrast, a sample with a high entropy level indicates a smooth similarity distribution, corresponding to the uncertainty of the probability distribution (Alemi, Fischer, and Dillon 2018). Therefore, ensuring the entropy within a proper value is essential to effectively delivering essential information. With this intuition, our method computes the smoothing indicator on the fly during the forward propagation in training, relying on the entropy of the similarity distribution per sample, which is formalized as follows:

$$H = -\sum_{i=1}^{|V|} p(E_{LLM}^{(i)}|\mathbf{z}) \log p(E_{LLM}^{(i)}|\mathbf{z}) \tag{10}$$

$$p(E_{LLM}|\mathbf{z}) = \texttt{softmax}(\bar{\mathbf{z}} \cdot E_{LLM}^\top) \tag{11}$$

where $\bar{\mathbf{z}}$ is the average pooling of $\mathbf{z}$ for all soft visual tokens, E_{LLM} is the word embedding of the LLM, $|V|$ is the vocabulary size of the LLM. Since entropy H has an unfixed range, we normalize H between 0 and 1 with its maximum value $\log(|V|)$. The adaptive β updates as follows:

$$\beta \leftarrow -\beta \cdot \log(\frac{H}{\log(|V|)}) \tag{12}$$

With this mechanism, a sample with low entropy is trained with high β, in which the compression term in Equation 5 dominates the optimization process, thereby constraining the information conveyed by input through adding a significant noise perturbation. Conversely, a sample with high entropy corresponds to low β, where the prediction term dominates, preventing the learning process from converging to a

worse performance. Note that the gradient computation for β is excluded from the computation graph, thereby the gradient does not flow through adaptive β.

Experiments

Datasets and Evaluation Metrics

MSCOCO (Lin et al. 2014) The Microsoft Common Objects in Context (MSCOCO) stands as a comprehensive dataset for evaluating various visual tasks, including image recognition, segmentation, and captioning. It has more than 300k images with annotations for over 80 object categories. In this paper, we employ COCO2014 to assess object hallucinations on the image captioning task. To train the vision-language projector, we randomly select 5000 image-text pairs from LLaVA-150k (Liu et al. 2024b), which is a set of GPT-generated multi-modal instruction-following data grounded on the images from COCO2014. Following Zhou et al. (2024), we additionally select 5000 unique images from the training dataset of COCO2014 to evaluate object hallucinations, ensuring that the selected images do not overlap with those used in training.

We use Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al. 2018) metric to evaluate object hallucinations in the image captioning task. CHAIR assesses object hallucinations by counter objects mentioned in the predicted caption but not present in the grounded image. It has two commonly used variants, CHAIR_S and CHAIR_I, where the former assesses object hallucinations at the sentence level, and the latter assesses at the instance level.

POPE (Li et al. 2023c) The Polling-based Object Probing Evaluation (POPE) is a widely adopted benchmark for assessing object hallucination on the Visual Question Answering (VQA) task. We conduct the evaluation on this benchmark to verify the generalization capability of AdaVIB on different tasks. POPE designs a binary question format, requiring the LVLM to deliver a binary-like answer to discriminate whether the mentioned objects are within the grounded image. We use the official question sets introduced in Li et al. (2023c). The dataset comprises three splits: In **Random** split, the absent objects are randomly selected from the whole dataset. In **Popular** split, the absent objects are chosen from the most frequently appeared objects in the dataset. In **Adversarial** split, the absent objects are selected from those frequently co-occurred with ground-truth objects. Each split is composed of 3000 questions on images taken from the validation set of COCO2014.

We use Accuracy and F1 scores to evaluate model performance, where Accuracy reflects the proportion of samples that correctly predict the golden answer. F1 score is the harmonic average of precision and recall. Following Li et al. (2023c), we use it as the major metric for evaluation.

Baselines

We employ the following competitive baselines to compare with our approach: **Chain-of-Thought (CoT)**: Wei et al. (2022) decomposed a hard problem by generating intermediate steps for the final answer. Following Zhou et al. (2024), we leverage CoT by asking the model to first list the identified objects and then describe the image in terms of these objects. **Decoding by Contrasting Layers (DOLA)**: Chuang et al. (2023) mitigated LLM's hallucinations by contrasting the differences between output logits from the later and earlier layers of LLMs. **Teacher**: Saha, Hase, and Bansal (2024) integrated several short descriptions into a long-form version. Following Zhou et al. (2024), we leverage BLIP-2 (Li et al. 2023a) to generate short descriptions as contextual guidance, facilitating a long-form description using the caption generator. **Visual Contrastive Decoding (VCD)**: Leng et al. (2024) alleviated hallucinations by introducing a visual contrastive decoding strategy, effectively decreasing the over-reliance on statistical bias and unimodal priors. **LVLM Hallucination Revisor (LURE)**: Zhou et al. (2024) introduced a post-hoc rectification method to train a hallucination revisor for rectifying initially generated descriptions.

Apart from the above baselines, we employ two distinct LVLMs as backbone frameworks, i.e., MiniGPT-4 and LLaVa-1.5. We employ both original and fine-tuned (FT) version as baselines. Additionally, we investigate the effectiveness of the variant without adaptive β (w/o Ada β) in Equation 12, and the one without reparameterization strategy (w/o Repara.) in Equation 9 for AdaVIB.

Implementation Details

We employ Vicuna-7B as the caption generator for both MiniGPT4 and LLaVa-1.5. We train all models in one epoch to avoid overfitting. All hyperparameters of baselines are selected via cross-validation on the training dataset of MSCOCO. Specifically, the Lagrange multiplier β is set to $\beta = 1e^{-7}$ unless explicitly specified. We set the batch size to 2 with gradient accumulation steps to 8. The maximum sequence length during training is set to 512. We use greedy decoding with a maximum decoding length of 256 during inference. We set the learning rate to $3e^{-5}$ with a weight decay of 0.05, and use a linear warm-up schedule for the first 1/10 optimization steps, followed by a polynomial decay. We use an A100-PCIE-40G GPU for training, which takes approximately 20 minutes for MiniGPT4 and 40 minutes for LLaVa-1.5. There are around 6.3M and 25.2M trainable parameters for MiniGPT4 and LLaVa-1.5, respectively.

Evaluation Results

Results on MSCOCO Table 1 reports the results on MSCOCO, we have following observations: First, the proposed AdaVIB yields consistent improvements across both MiniGPT4 and LLaVa-1.5 model architectures. Specifically, it outperforms one of the strongest baseline LURE_{13B} by around 14.5% under both CHAIR_S and CHAIR_I metrics, especially under the CHAIR_S metric, it surpasses LURE by around 24.9% across two distinct model architectures. Second, AdaVIB substantially outperforms the vanilla fine-tuning method. The improvement of AdaVIB over fine-tuning under CHAIR_S and CHAIR_I metrics are 23.6% and 21.2%, respectively. This indicates that the vanilla fully connected layer may not be strong enough to model the sophisticated mapping relations between vision and language. In contrast, our approach can improve the vision-language

Model	*MiniGPT-4*		*LLaVa*	
	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
Teacher$_{13B}$	24.0	5.7	49.9	9.3
CoT$_{13B}$	31.6	9.4	47.6	9.0
DOLA$_{7B}^{\dagger}$	26.7	7.9	31.9	8.3
VCD$_{7B}^{\dagger}$	24.4	7.5	26.2	7.8
LURE$_{13B}$	19.7	**4.9**	27.1	6.4
Original$_{7B}$	27.1	7.8	47.8	10.8
FT$_{7B}$	19.3	6.4	26.7	7.2
AdaVIB$_{7B}$	**16.2**	5.2	**18.4**	**5.5**
w/o Ada β	17.1	5.6	19.2	5.7
w/o Repara.	18.5	6.2	22.3	6.4

Table 1: Experimental results on MSCOCO. The evaluation is performed using CHAIR$_S$ (C_S) and CHAIR$_I$ (C_I), with smaller values indicating fewer object hallucinations. † means rerun using their released code. The best and the second-best are marked in bold and underlined, respectively.

alignment by better compressing irrelevant visual information while maintaining relevant information for generating hallucination-free descriptions. Third, the ablation results indicate the effectiveness of each component. Specifically, by removing adaptive β in equation 12, the results drop over 5.0% on average across two model architectures. Notably, by removing the reparameterization strategy in equation 9, the results drop over 15.0% on average, indicating the significance of this component in mitigating object hallucinations on the MSCOCO dataset.

Results on POPE Table 2 reports the results on POPE. We observe that the improvements of AdaVIB on Popular are clearer than on Random and Adversarial. Concretely, compared with one of the strongest baselines LURE on MiniGPT4 backbone, the improvements of AdaVIB on Popular (8.1%) are more significant than on Random (4.8%) and Adversarial (3.9%). Similar observations can be found when using the LLaVa backbone, AdaVIB achieves 7.5% improvements on Popular, while 2.5% improvements on Random and Adversarial by average. These observations indicate the advantages of AdaVIB in effectively mitigating the statistical bias arising from the most frequently appeared objects during training, reducing the over-reliance on irrelevant features, thereby alleviating object hallucinations. Additionally, the ablation results on POPE indicate the effectiveness of AdaVIB. Concretely, removing adaptive β decreases AdaVIB by 1.5% and 1.4% on MiniGPT4 and LLaVA-1.5 by average, respectively. While removing the reparameterization strategy results in performance degradation by 2.1% and 2.2% on MiniGPT4 and LLaVA-1.5 by average, respectively. This observation indicates that both adaptive β and reparameterization strategy is important to improve the performance of AdaVIB on the POPE dataset.

Discussions

Does the AdaVIB mitigate object hallucinations by alleviating the *overconfidence* problem? To address this, we figure out the proportion of hallucinated samples in a specific range of the max similarity score to overall hallucinated samples. Figure 3 presents the results that employ the MiniGPT4 as the model backbone on the MSCOCO dataset. We have the following observations: First, for the fine-tuned (FT) variant, the similarity score ranging $[0.9, 1.0)$ dominates the hallucinated samples, with an average similarity entropy of 0.64. A low entropy denotes a sharp similarity distribution, indicating that the fine-tuned variant is prone to overconfidence when visual tokens map to the LLM's word embedding space. Second, both VIB and AdaVIB have a smoother distribution than the fine-tuned variant, with a consistent improvement on alleviating object hallucinations from the results reported in Table 1 and Table 2. This observation indicates that mitigating the *overconfidence* problem is one of the effective solutions to mitigate object hallucinations. Third, AdaVIB has fewer hallucinated samples with a large max similarity score than VIB. It has a smoother distribution than VIB, with a larger average similarity entropy (1.30 v.s. 0.97). This result indicates that the proposed adaptive β mitigates object hallucinations by effectively controlling the smoothness of the similarity distribution.

Model	Random		Popular		Adversarial	
	$ACC \uparrow$	$F1 \uparrow$	$ACC \uparrow$	$F1 \uparrow$	$ACC \uparrow$	$F1 \uparrow$
MiniGPT4						
DOLA$_{7B}^{\dagger}$	76.6	77.9	65.8	69.3	63.3	68.5
VCD$_{7B}^{\dagger}$	78.4	80.2	68.5	72.0	64.4	70.3
LURE$_{13B}^{\dagger}$	78.0	79.2	66.0	70.3	63.9	71.1
FT$_{7B}$	79.0	80.8	70.0	72.9	65.0	71.2
AdaVIB$_{7B}$	**81.2**	**83.6**	**71.1**	**76.3**	**66.7**	**73.6**
w/o Ada β	80.0	82.2	69.8	75.6	65.3	73.1
w/o Repara.	79.4	82.0	69.5	74.0	65.6	72.5
LLaVa						
DOLA$_{7B}^{\dagger}$	84.0	84.4	79.5	82.2	76.6	77.2
VCD$_{7B}^{\dagger}$	**86.8**	86.2	83.2	84.3	**80.1**	79.6
LURE$_{13B}$	86.3	85.8	80.3	80.7	77.2	78.4
FT$_{7B}$	84.3	85.8	84.9	84.0	76.5	77.7
AdaVIB$_{7B}$	86.3	**87.0**	**86.3**	**86.8**	78.0	**80.4**
w/o Ada β	85.4	86.0	85.0	85.6	76.9	79.0
w/o Repara.	84.9	85.3	84.5	85.1	76.2	78.1

Table 2: Experimental results on POPE. Larger values indicate less hallucinations.

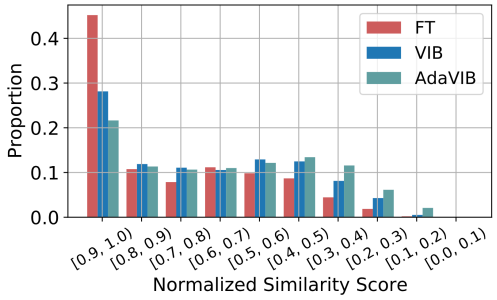

Figure 3: Distribution of the max similarity score between soft visual tokens and LLM's word embedding. We use MiniGPT4 as the model backbone. The x-axis denotes the normalized similarity score, ranging from 0-1. The y-axis denotes the proportion of hallucinated samples in a specific range to overall hallucinated samples.

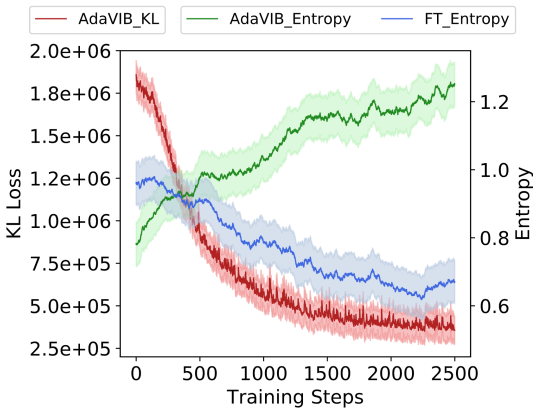

Figure 4: Correlation between the KL loss (Equation 8) and the similarity entropy (Equation 10) over the course of training. All curves are smoothed by exponential moving average for better understanding the tendency.

Does the AdaVIB alleviate the *overconfidence* issue by effectively constraining irrelevant information? In addressing this question, we present the learning curves of the KL loss for AdaVIB, as well as the curves of similarity entropy for both AdaVIB and fine-tuning. From Figure 4, we have two main observations: First, with the training progress, the KL loss of AdaVIB (red curve) is decreased, accompanied by the increase of its similarity entropy (green curve). This phenomenon indicates that the compression term in Equation 5 effectively compresses the irrelevant information by adding stochastic noise, thereby progressively reducing the overconfidence in irrelevant visual features between the visual tokens and the LLM's word embedding. Second, compared with the curves of similarity entropy between AdaVIB and fine-tuning, the entropy curve of the fine-tuning variant severely decreases with the training process going on, indicating the emergence of the overconfidence problem. In contrast, AdaVIB improves the performance by progressively smoothing the similarity distribution, avoiding overconfidence in irrelevant visual features.

How does different β impact the model performance on mitigating object hallucinations? To answer this question, we employ both MiniGPT4 and LLaVa-1.5 as the backbone, adjusting β and analyzing the change of the CHAIR score on MSCOCO evaluation dataset. From Figure 5, we observe that the performance of both CHAIR_S and CHAIR_I score improves along with the reduction of β from $\beta = 1e^{-1}$ to $\beta = 1e^{-7}$, followed by an obvious performance degradation with $\beta = 1e^{-9}$. This observation indicates that setting the β to a proper value is essential to effectively constrain the information flow while preserving the useful information as much as possible. A large β introduces a big noise during the model optimization, resulting in the model scarcely capturing the essentials relevant to the golden answer, thereby converging to a worse performance. When the β is too small, the model tends to preserve relevant information rather than compress the irrelevant features present in the original input, delivering a sub-optimal performance in mitigating object hallucinations. Moreover, setting β to a fixed value, regardless of the dynamic nature per sample, cannot achieve the optimal performance in mitigating hallucinations. This conclusion is consistent with the results reported in Table 1, where removing the adaptive β results in a worse performance compared to the variant with the adaptive one.

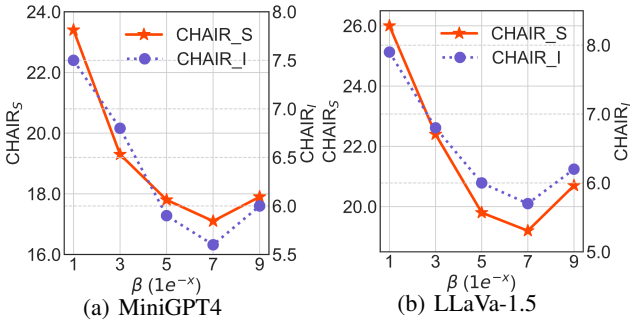

Figure 5: Impact of different β on object hallucinations. Figure 5(a) and Figure 5(b) present the results that leverages MiniGPT4 and LLaVa-1.5 as backbone, respectively.

How does AdaVIB perform compared to other regularization strategies? In addition to the fine-tuning baseline that applies the weight decay (Loshchilov and Hutter 2017) as a regularizer. We also deploy dropout (Srivastava et al. 2014) to the input (DRP_{in}) and output (DRP_{out}) of the vision-language projector upon the fine-tuning method, with the dropout rate of 0.1. Table 3 presents results. We observe that variants equipped with dropout effectively reduce object hallucinations compared to the FT, while they still have a performance gap between AdaVIB. This indicates the superiority of AdaVIB in the trade-off between compressing irrelevant while preserving relevant information to represent visual tokens precisely.

Model	*MiniGPT-4*		*LLaVa*	
	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
FT	19.3	6.4	26.7	7.2
FT+DRP_{in}	18.5	6.1	23.4	6.4
FT+DRP_{out}	18.6	5.9	22.7	6.2
AdaVIB	**16.2**	**5.2**	**18.4**	**5.5**

Table 3: Impact on different regularization strategies. "DRP" is the abbreviation of the "dropout".

Conclusions

In this paper, we propose using VIB to mitigate object hallucinations. The proposal is based on our observation that the object hallucination can be attributed to the overconfidence in irrelevant visual features when soft visual tokens project to the word embedding space of LLM. Motivated by our observation, we propose AdaVIB to adaptively constrain irrelevant information regarding the smoothness of similarity distribution. Experimental results and comprehensive analysis demonstrate the effectiveness of our approach with consistent improvements over competitive baselines.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62372126, 62372129, U2436208, 62272119, 62276017, 62072130), the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515030142), the Key Technologies R&D Program of Guangdong Province (No. 2024B0101010002), the Strategic Research and Consulting Project of the Chinese Academy of Engineering (No. 2023-JB-13), and the State Key Laboratory of Complex & Critical Software Environment (No. SKLCCSE-2024ZX-18).

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Alemi, A. A.; Fischer, I.; and Dillon, J. V. 2018. Uncertainty in the variational information bottleneck. *arXiv preprint arXiv:1807.00906*.

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Belinkov, Y.; Henderson, J.; et al. 2020. Variational information bottleneck for effective low-resource fine-tuning. In *International Conference on Learning Representations*.

Biten, A. F.; Gómez, L.; and Karatzas, D. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1381–1390.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Cho, J.; Yoon, S.; Kale, A.; Dernoncourt, F.; Bui, T.; and Bansal, M. 2022. Fine-grained Image Captioning with CLIP Reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 517–527.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J. R.; and He, P. 2023. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023a. InstructBLIP: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 49250–49267.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B. A.; Fung, P.; and Hoi, S. C. H. 2023b. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *ArXiv*, abs/2305.06500.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Favero, A.; Zancato, L.; Trager, M.; Choudhary, S.; Perera, P.; Achille, A.; Swaminathan, A.; and Soatto, S. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14303–14312.

Fischer, I. 2020. The conditional entropy bottleneck. *Entropy*, 22(9): 999.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Guo, H.; Liu, J.; Huang, H.; Yang, J.; Li, Z.; Zhang, D.; and Cui, Z. 2022. LVP-M3: Language-aware Visual Prompt for Multilingual Multimodal Machine Translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2862–2872.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. 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*, 13418–13427.

Jain, J.; Yang, J.; and Shi, H. 2024. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27992–28002.

Jiang, C.; Xu, H.; Dong, M.; Chen, J.; Ye, W.; Yan, M.; Ye, Q.; Zhang, J.; Huang, F.; and Zhang, S. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27036–27046.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13872–13882.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Q.; Wu, Z.; Kong, L.; and Bi, W. 2023b. Explanation Regeneration via Information Bottleneck. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Li, X. L.; and Eisner, J. 2019. Specializing Word Embeddings (for Parsing) by Information Bottleneck. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023c. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 292–305.

Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; and Bai, X. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26763–26773.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, R.; Li, Y.; Tao, L.; Liang, D.; and Zheng, H. 2022. Are we ready for a new paradigm shift? A survey on visual deep MLP. *Patterns*, 3(7): 100520.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Peng, X. B.; Kanazawa, A.; Toyer, S.; Abbeel, P.; and Levine, S. 2018. Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow. In *International Conference on Learning Representations*.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045.

Saha, S.; Hase, P.; and Bansal, M. 2024. Can language models teach? teacher explanations improve student performance via personalization. *Advances in Neural Information Processing Systems*, 36.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Sun, Q.; Li, J.; Peng, H.; Wu, J.; Fu, X.; Ji, C.; and Philip, S. Y. 2022. Graph structure learning with variational information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4165–4174.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Tang, K.; Miao, D.; Peng, W.; Wu, J.; Shi, Y.; Gu, Z.; Tian, Z.; and Wang, W. 2021. Codes: Chamfer out-of-distribution examples against overconfidence issue. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1153–1162.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, 1–5. IEEE.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yao, S.; and Wan, X. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 4346–4350.

Zhai, S.; Likhomanenko, T.; Littwin, E.; Busbridge, D.; Ramapuram, J.; Zhang, Y.; Gu, J.; and Susskind, J. M. 2023. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, 40770–40803. PMLR.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2024. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. In *The Twelfth International Conference on Learning Representations*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, K.; Feng, X.; Du, X.; Gu, Y.; Yu, W.; Wang, H.; Chen, Q.; Chu, Z.; Chen, J.; and Qin, B. 2024. An Information Bottleneck Perspective for Effective Noise Filtering on Retrieval-Augmented Generation. *arXiv preprint arXiv:2406.01549*.