Journal of Electronics and Information Science (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2026.110108
ISSN 2371-9524 Vol. 11 Num. 1

DIVE: A Training-free Hallucination Mitigation Mechanism for Complex Scenes

Shuguo Jiang[1,a,*]

[1]*School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou, China*
[a]*1103976792jsg@gmail.com*
**Corresponding author*

Keywords: Language prior; Hallucination mitigation; Feature alignment

Abstract: When facing real-world scenes that are densely populated with objects or contain complex occlusions, Vision-Language Models are often constrained by the language prior in autoregressive decoding, producing severe hallucination phenomena. To address this pain point that limits the reliable deployment of large models, this paper proposes Dual-branch Inference for Visual-prior Elimination, a training- free hallucination mitigation mechanism for complex scenes. By constructing a dual-branch inference structure at the inference stage and introducing a dynamic visual-confidence penalty, this mechanism effectively quantifies and suppresses the overconfidence in the content generation process, forcing the model's output to be deeply aligned with the underlying visual features. Results on the object hallucination evaluation benchmark POPE show that, without consuming computing power for model fine-tuning, the proposed method reduces the model's hallucination rate when objects are dense or complex occlusions exist, and brings a slight improvement in the question-answering accuracy of the model on the MSCOCO and VG datasets.

1. Introduction

At present, the academic community has conducted extensive exploration to mitigate the hallucination problem of VLMs, which can be mainly divided into two routes: data-driven fine-tuning and external knowledge augmentation. Among them, data-driven fine-tuning corrects model weights by constructing large-scale, high-quality anti-hallucination fine-tuning datasets; external knowledge augmentation retrieves external knowledge graphs or invokes object detection tools before content generation. The above problems prompt scholars to consider a question: is it possible to achieve hallucination mitigation in complex scenes only by intervening in the decoding strategy at the inference stage, without consuming computing power to fine-tune the model weights and without introducing external tools?

The vanilla model describes a non-existent boat driven by the language prior, while DIVE faithfully captures the water birds standing on the withered branch (Figure 1). To address the above challenges, this paper proposes DIVE (Dual-branch Inference for Visual-prior Elimination). This mechanism focuses on the text generation stage of "image-to-text". By introducing a dynamic contrastive penalty term at each step of token generation, it mitigates the hallucination caused by the model's overconfident language prior and forces the model's output to be strictly aligned with the

actual visual features. This paper proposes a lightweight training-free hallucination mitigation mechanism for complex scenes. By constructing a dual-branch inference structure and introducing a dynamic visual-confidence penalty, it breaks the statistical regularities of the language prior and improves the trustworthiness of the model in real-world scenes. Moreover, experimental results prove that the proposed mechanism, without additional training cost, reduces the hallucination rate of the baseline model on benchmark datasets.

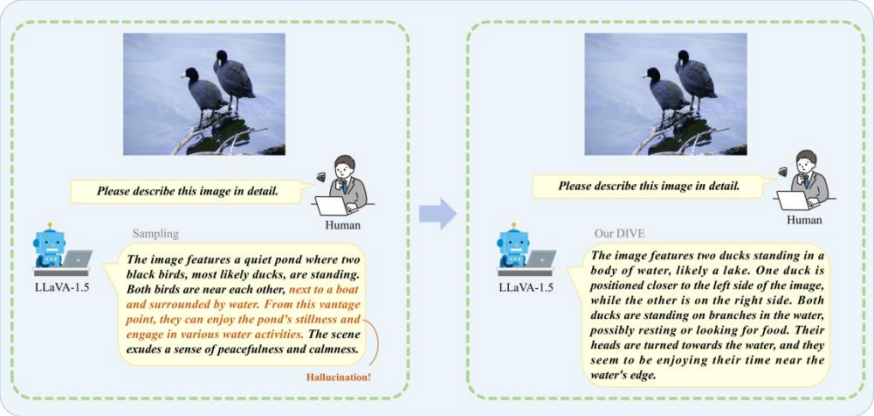

Figure 1. An illustrative example of DIVE reducing object hallucination.

2. Related Work

2.1. Vision-Language Models

The development of VLMs has fundamentally changed the paradigm of multimodal interaction. Early works represented by CLIP [1] and ALIGN [2] mainly focus on multimodal feature alignment through contrastive learning. With the explosion of LLMs [3], the research focus has shifted to how to endow LLMs with visual perception capabilities [4]. BLIP-2 [5] introduces a Q-Former structure to bridge the visual encoder and LLMs, ensuring zero-shot transfer performance while reducing training parameters and computational cost; models such as LLaVA [6] and Qwen- VL [7] adopt a more direct architecture, directly feeding visual features as special token sequences into the autoregressive language model through a simple linear projection or a visual adapter. As the model parameters expand and the instruction fine-tuning corpus increases, their hallucination phenomena in complex real-world scenes become increasingly prominent [8].

2.2. Multimodal Hallucination Mitigation Methods

2.2.1. Data-driven Methods Based on Training-time Intervention

Such methods attempt to mitigate hallucination by modifying model weights, such as M-HalDetect [9] and LRV-Instruction [10]. RLHF [11] optimizes the generation strategy by constructing a reward model and using a policy gradient algorithm to make it conform to human preferences for truthfulness; DPO [12] reduces the generation of model hallucination by directly maximizing the probability margin between the model's factual descriptions and hallucinated descriptions. Woodpecker [13] connects an external object detector and a VQA model after training and performs post-hoc correction on the generated text by constructing a visual knowledge base for the current image.

2.2.2. Decoding Strategy Optimization Based on Inference-time Intervention

To avoid the high training cost, recent studies have begun to explore mitigating hallucination by

modifying autoregressive decoding without updating model parameters [14][15]. In terms of contrastive decoding, the DoLa model [16] finds that, by contrasting the differences in logit values between the later layers of the model and the dynamically selected earlier layers after mapping to the vocabulary space, contrastive logits can be obtained at different layers of LLMs to reduce factual errors. The VCD model [17] reduces the dependence on the language prior by contrasting the distribution difference between the original input image and the perturbed input image. The CAD model [18] further enhances the model's faithfulness to the input context information by contrasting the output distributions with and without context conditions. In terms of mechanism optimization, the OPERA mechanism [19] is based on an over-confidence penalty and alleviates the attention collapse that easily occurs in the model during long-text generation through rollback decoding. The CoVe mechanism [20] guides the model to autonomously construct a verification chain for self-verification.

3. Algorithm Design

3.1. Overall Architecture and Motivation

Mainstream VLMs represented by LLaVA [6] and Qwen [7] generally adopt an autoregressive-based text generation paradigm. Given an input image and a text instruction, the model first maps the image into visual token vectors through a visual encoder represented by CLIP [1], then concatenates them with the text instruction and feeds them into the LLMs for step-by-step decoding. However, when facing densely populated objects or complex occlusions, this mechanism suffers from overconfidence in the language prior. As shown in Figure 2, DIVE constructs a dual-branch inference structure that can achieve hallucination mitigation without any parameter fine-tuning of the model.

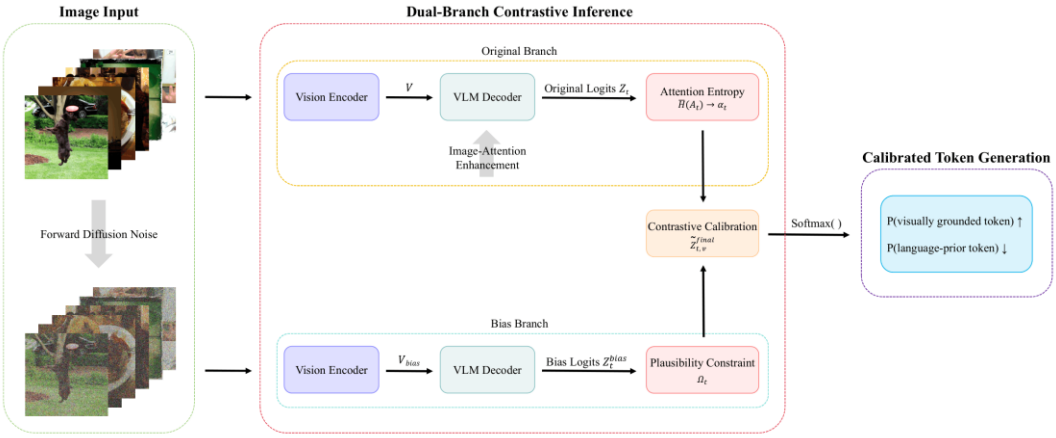

Figure 2. Overall framework of DIVE.

3.2. Dual-branch Logits Contrast and Adaptive Calibration

3.2.1. Dual-branch Contrastive Decoding

Suppose that at the t-th step of the standard visual-feature decoding process, the image features processed by the visual encoder are V, and the context sequence concatenated from the user instruction and the historically generated tokens is $Y_{<t}$. During decoding, the language model of the VLMs outputs the hidden state in the vocabulary space and maps it into the original logits Z_t , which represents the model's preliminary prediction trend, as shown in Equation (1):

$$Z_t = \mathrm{LLMs}(V, Y_{<t}) \in R^{|V_{vocab}|} \quad (1)$$

where $|V_{vocab}|$ denotes the size of the model's vocabulary.

DIVE innovatively introduces the logic of bias reasoning. While keeping the text context sequence completely unchanged, a visually-corrupted image is input to the model, which represents the visually-corrupted features V_{bias} [21] obtained by re-encoding the original image after applying forward diffusion noise. At this point, let the bias logits output by the model at the t-th decoding step be Z_t^{bias}, as shown in Equation (2):

$$Z_t^{bias}=LLMs(V_{bias},Y_{<t}) \in R^{|V_{vocab}|} \tag{2}$$

Since the effective physical information in V_{bias} is significantly weakened or masked, the tokens that still have high scores at this point reflect the model's over-reliance on the text context in the absence of reliable visual guidance.

After obtaining the original logits Z_t and the bias logits Z_t^{bias}, DIVE reconstructs the output distribution through a contrastive mechanism. Its goal is to enhance the tokens that have significant scores in Z_t but perform poorly in Z_t^{bias}, it strictly penalizes the tokens that also have high scores in Z_t^{bias}, i.e., the bias-driven hallucinated tokens. Specifically, the logits Z_t^{final} obtained through basic calibration are shown in Equation (3):

$$Z_t^{final}=Z_t+\alpha(Z_t-Z_t^{bias}) \tag{3}$$

where $\alpha \geq 0$ denotes the contrast penalty coefficient. And $Z_t-Z_t^{bias}$ denotes the marginal contribution of visual information to generating the token.

3.2.2. Attention-entropy-based Adaptive Penalty

DIVE introduces an image-attention amplification mechanism in the forward propagation of the standard original branch to strengthen visual alignment from the source of attention. Specifically, for the l-th decoding layer ($l \geq l_0$), before Softmax normalization, DIVE applies a positive bias λ to the key-value columns corresponding to all visual tokens in the attention scores, as shown in Equation (4):

$$A^{(l)}=Softmax\left(\frac{QK^T}{\sqrt{d}}+M+\lambda \cdot 1_{vis}\right) \tag{4}$$

where M is the original attention mask, 1_{vis} is an indicator vector that takes the value 1 only on the key-value columns of visual tokens, λ is the attention amplification strength, and l_0 is the starting layer where the intervention is applied.

DIVE does not only read the last layer, but averages the attention distribution of the current step over the visual tokens across the deepest L decoding layers, obtaining the aggregated attention distribution $A_t=\{a_t^1,a_t^2,...,a_t^N\}$, whose attention entropy $H(A_t)$ is shown in Equation (5):

$$H(A_t)=-\sum_{j=1}^{N} a_t^j \log(a_t^j) \tag{5}$$

where N denotes the total number of visual tokens. When the attention is completely uniformly distributed over all visual patches, the entropy reaches its maximum value $H_{max}(A_t)=\log(N)$. To eliminate the scale differences caused by different input resolutions, DIVE normalizes the entropy, as shown in Equation (6):

$$\bar{H}(A_t)=\frac{H(A_t)}{\log(N)} \in [0,1] \tag{6}$$

The smaller the value of $\bar{H}(A_t)$, the more the model highly focuses on certain local visual features and the better the current step is visually anchored. Accordingly, let the adaptive penalty coefficient at the t-th step be α_t, as shown in Equation (7):

$$\alpha_t=\alpha_{min}+(\alpha_{max}-\alpha_{min})\cdot\bar{H}(A_t)^{\beta} \tag{7}$$

where α_{min} and α_{max} are the lower and upper bounds of the penalty strength respectively, and β is the attention-entropy exponent. α_{min} provides a non-zero guaranteed strength for the contrast penalty, when the attention is scattered, α_t climbs toward the upper bound α_{max} , applying stronger suppression to the potential hallucinated tokens; when the attention is focused, α_t falls back to α_{min} , protecting the recall of small objects with a gentler intervention.

3.2.3. Rationality Constraint

If the original logits Z_t and the bias logits Z_t^{bias} are subtracted without restriction, semantic collapse may be triggered. DIVE innovatively introduces a rationality constraint mechanism. Let $v\in V_{vocab}$ be any token in the vocabulary, and $Z_{t,v}$ denote the logit component corresponding to token v in the vector Z_t. To ensure the logical coherence of generation, DIVE introduces a rationality constr aint set Ω_t. DIVE retains the candidate tokens in the original logits Z_t whose probability is no lower than τ times the peak probability, and takes the top-k ranked tokens among them as the set tokens, as shown in Equation (8):

$$\Omega_t=\left\{v\in V_{vocab} \mid Z_{t,v}\geq \log\tau+\max_{v'} Z_{t,v'}\right\}\cap \mathrm{TopK}(Z_t;k) \tag{8}$$

where $\tau\in(0,1]$ is the rationality threshold and k is the truncation size of the candidate set. The final calibrated logits are shown in Equation (9):

$$\tilde{Z}_{t,v}^{final}=\begin{cases} Z_{t,v}+\alpha_t\left(Z_{t,v}-Z_{t,v}^{bias}\right), & \text{if } v\in\Omega_t \\ -\infty, & \text{otherwise} \end{cases} \tag{9}$$

4. Experiments

4.1. Experimental Setup

This paper adopts the mainstream object-probing evaluation benchmark POPE [21]. The experiments use both the MSCOCO dataset, and the VG dataset. This paper uses POPE to construct three sampling subsets for each dataset, namely Random, Popular, and Adversarial. This paper additionally introduces the ImageNet dataset.

LLaVA adopts the typical architecture of "visual encoder + large language model", maps the image into a visual token sequence and concatenates it with the text instruction, and then gradually generates the answer or description in an autoregressive manner. This paper selects the following representative baseline methods for comparison experiments: 1) the original LLaVA; 2) VCD. By comparing with these methods, this paper will verify the superiority of DIVE in scenes that are densely populated with objects or contain complex occlusions.

4.1.1. Evaluation Metrics

This paper uses accuracy, recall, and F1 score to evaluate the experimental performance. For the evaluation of the generation quality of open-ended long descriptions in complex environments, this paper draws on the LLM-as-a-judge paradigm [22] and introduces a Claude-aided scoring mechanism: the original image and the descriptions generated by each method are provided to Claude together,

which scores them separately along six mutually complementary dimensions, namely Fluency (FLU), Detail Richness (DET), Scene Coverage (COV), Hallucination Mitigation (HAL), Logical Clarity (LOG), and Salience and Focus (SAL).

4.1.2. Implementation Details

This paper sets the lower bound of the penalty strength to α_{min}=1.0 and the upper bound to α_{max}=2.0, the attention-entropy exponent to β=1.0, the rationality threshold to τ=0.25, the truncation size of the candidate set to k=50, and the number of deep layers for entropy aggregation to L=8; the image-attention amplification strength is set to λ=1.5 and the starting layer where the intervention is applied is set to l_0=2. In the dual-branch inference, the visually-corrupted features V_{bias} used by the bias branch are obtained by applying forward diffusion noise to the original image, and the noise step adopted by DIVE is t_{noise}=800, so that the bias branch approximates a purer language prior and thereby fully exposes the hallucination driven by co-occurrence statistics.

4.2. Results

4.2.1. POPE Object Existence Evaluation Results

On the two datasets MSCOCO and VG, adopts the three sampling strategies Random, Popular, and Adversarial under the POPE [22], transforms the image captioning task into the binary-classification question, and computes accuracy, recall, and F1, with the results shown in Table 1. Under all six combinations of datasets and sampling strategies, the F1 score that measures the comprehensive balanced performance of DIVE is the best among the three methods, the recall is also fully leading, and the accuracy is overall dominant, being only slightly lower than VCD in a few settings.

Table 1. POPE Object Existence Evaluation Results.

Strategy	Method	MSCOCO			VG		
		Accuracy	Recall	F1 Score	Accuracy	Recall	F1 Score
Random	LLaVA	0.8437	0.7600	0.8294	0.7683	0.7360	0.7606
	VCD	0.8743	0.8073	0.8653	0.7937	0.7773	0.7902
	DIVE	0.8923	0.8733	0.8902	0.7803	0.8500	0.7946
Popular	LLaVA	0.8237	0.7600	0.8117	0.6937	0.7360	0.7061
	VCD	0.8510	0.8073	0.8442	0.7043	0.7773	0.7244
	DIVE	0.8577	0.8733	0.8599	0.7180	0.8500	0.7509
Adversarial	LLaVA	0.7970	0.7627	0.7898	0.6633	0.7367	0.6863
	VCD	0.8260	0.8073	0.8227	0.6723	0.7773	0.7035
	DIVE	0.8140	0.8733	0.8244	0.6723	0.8500	0.7218

4.2.2. AI Evaluation Results on Open-ended Long Descriptions

This paper randomly samples 50 images on each dataset, and provides the original image together with the descriptions of each method under the same image to Claude, which scores them separately along the six dimensions FLU, DET, COV, HAL, LOG, and SAL. After each image obtains a raw score of 0–100 on each dimension, the scores are first averaged over all images for the same method and the same dimension, and then normalized to [0, 1], the final results shown in Figure 3. Combining the results of POPE and the AI evaluation, DIVE demonstrates a more consistent trustworthiness improvement at both the object-level judgment and the open-ended long-description levels.

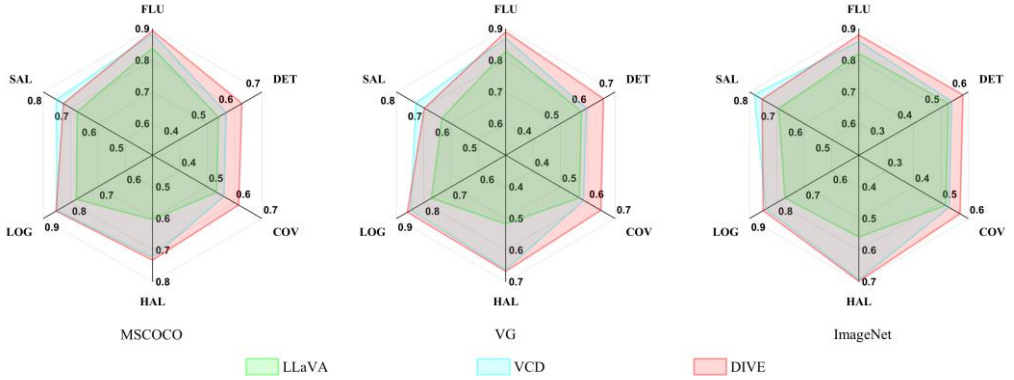

Figure 3. Claude-assisted evaluation results on MSCOCO, VG, and ImageNet datasets.

4.2.3. Ablation Experiment

This paper conducts an ablation experiment on the MSCOCO dataset under the Random sampling strategy, and measures the contribution of each module by accuracy, recall, and F1 score, with the experimental results shown in Figure 4. The performance improvement of DIVE mainly comes from the explicit modeling of the language prior by the visually-corrupted branch; the dual branches can accurately identify vision-driven and language-prior contributions, and can effectively suppress object hallucination.

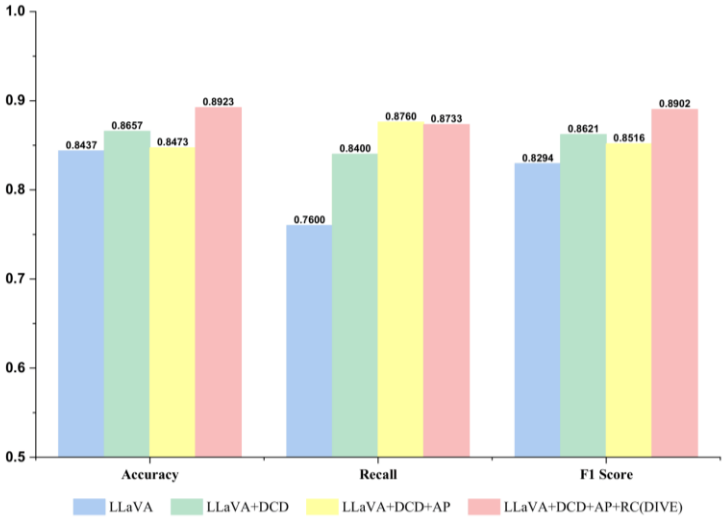

Figure 4. Results of the ablation study of DIVE - with accuracy, recall and F1 score.

5. Conclusion

Different from existing methods that rely on a fixed contrast penalty and inference-time intervention, DIVE, this paper proposed, introduces a dual-branch contrastive decoding architecture to explicitly estimate the language-prior bias and dynamically calibrate the token probability during generation. In addition, this paper constructs a parallel bias branch for language-prediction driving, while the original visual branch maintains the visual basis through image-attention amplification, and proposes an attention-entropy-based adaptive penalty mechanism that dynamically adjusts the intervention strength according to the model's visual confidence. The experimental results on the MSCOCO and VG datasets show that DIVE performs excellently in accuracy, recall, and F1 score, and the ablation experiment further verifies the importance of each sub- module of DIVE.

References

[1] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever I. (2021) Learning transferable visual models from natural language supervision. In 2021 International conference on machine learning, PMLR, pp. 8748–8763.

[2] Jia, C., Yang, Y.F., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z. and Duerig, T. (2021) Scaling up visual and vision-language representation learning with noisy text super-vision. In 2021 International conference on machine learning, PMLR, pp. 4904–4916.

[3] Naveed, H., Ullah Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. and Mian, A. (2025) A comprehensive overview of large language models. ACM Trans. Intell. Syst. Technol., 16(5), 1–72.

[4] Zhang, J., Huang, J., Jin, S. and Lu, S. (2024) Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence, 46(8), 5625–5644.

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

[6] Liu, H., Li, C., Wu, Q. and Lee, Y.J. (2023) Visual instruction tuning. Advances in neural information processing systems, 36, 34892–34916.

[7] Bai, J.Z., Bai, S., Yang, S.S., Wang, S.J., Tan, S.N., Wang, P., Lin, J.Y., Zhou, C. and Zhou, J.R. (2023) Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv:2308.12966 [cs.CV].

[8] Huang,, L., Yu, W.J., Ma, W.T., Zhong, W.H., Feng, Z.Y., Wang, H.T. Chen, Q.L., Peng, W.H., Feng, X.C., Qin, B. and Liu, T. (2025) A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), 1–55.

[9] Gunjal, A., Yin, J. and Bas, E. (2024) Detecting and preventing hallucinations in large vision language models. In Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence, 38, 18135– 18143.

[10] Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y. and Wang, L. (2024) Mitigating hallucination in large multimodal models via robust instruction tuning. In International Conference on Learning Representations 2024, 57689–57733.

[11] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. and Lowe R. (2022) Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730–27744.

[12] Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S. and Finn, C. (2023) Direct preference optimization: Your language model is secretly a reward model. Adv. Neural Inf. Process. Syst., 36, 53728–53741.

[13] Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X. & Chen, E. (2024) Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences, 67(12), 220105.

[14] Lee. N., Ping, W., Xu, P., Patwary, M., Fung, P.N., ShoeybI, M. and Catanzaro, B. (2022) Factuality enhanced language models for open-ended text generation. Adv. Neural Inf. Process. Syst., 35, 34586–34599.

[15] Li, K., Patel, O., Viégas, F., Pfister, H. and Wattenberg, M. (2023) Inference-time intervention: Eliciting truthful answers from a language model. Adv. Neural Inf. Process. Syst., 36, 41451–41530.

[16] Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J.R. and He, P. (2024) Dola: Decoding by contrasting layers improves factuality in large language models. In International Conference on Learning Representations 2024, 54158–54183.

[17] Leng, S.C., Zhang, H., Chen, G.Z., Li, X., Lu, S.J., Miao, C.Y. and Bing, L.D. (2024) Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In Proceedings of the 2024 IEEE/CVF CVPR, 13872–13882.

[18] Shi, W., Han, X., Lewis, M., Tsvetkov, Y., Zettlemoyer, L. and Yih, W.T. (2024) Trusting your evidence: Hallucinate less with context-aware decoding. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2 (Short Papers), 783–791.

[19] Huang, Q., Dong, X.Y, Zhang, P., Wang, B., He, C.H., Wang, J.Q., Lin, D.H., Zhang, W.M. and Yu N.H. (2024) Opera: Alleviating hallucination in multi-modal large language models via overtrust penalty and retrospection-allocation. In Proceedings of the 2024 IEEE/CVF - CVPR, 13418–13427.

[20] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A. and Weston J. (2024) Chain-of-verification reduces hallucination in large language models. In Findings of the association for computational linguistics: ACL 2024, 3563–3578.

[21] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X. and Wen, J.R. (2023) Evaluating object hallucination in large vision-language models. In Proceedings of the 2023 conference on empirical methods in natural language processing, 292–305.

[22] Zheng, L.M., Chiang, W.L., Sheng, Y., Zhuang, S.Y., Wu, Z.H., Zhuang, Y.H., Lin, Z., Li, Z.H., Li, D.C., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I. (2023) Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36, 46 595–46 623.