VADE: Visual Attention Guided Hallucination Detection and Elimination

Vishnu Prabhakaran[1], **Purav Aggarwal**[1], **Vinay Kumar Verma**[1],
Gokul Swami[2], **Anoop Saladi**[1]
[1]Amazon, India, [2]Amazon, USA
{visprab,aggap,vkvermaa,swagokul,saladias}@amazon.com

Abstract

Vision Language Models (VLMs) have achieved significant advancements in complex visual understanding tasks. However, VLMs are prone to hallucinations—generating outputs that lack alignment with visual content. This paper addresses hallucination detection in VLMs by leveraging the visual grounding information encoded in transformer attention maps. We identify three primary challenges in this approach: the elective nature of visual grounding for certain tokens, the high-dimensional and noisy nature of attention maps, and the dynamic sequence length of attention on previous tokens. To address these, we propose VADE, a novel sequence modelling approach to effectively learn complex sequential patterns from high-dimensional and noisy attention maps for fine-grained hallucination detection and mitigation. VADE achieves an average PR-AUC of 80% in hallucination detection on M-HalDetect across four different model architectures and an $\sim$5% improvement in hallucination mitigation on MSCOCO.

1 Introduction

Large Vision-Language Models (VLMs) have ushered in a new era of image comprehension, achieving near-human performance on intricate tasks such as image captioning and visual question answering. This breakthrough was pioneered by the seminal work on CLIP (Radford et al., 2021), paving the way for several instruction-tuned VLM architectures (Alayrac et al., 2022; Touvron et al., 2023; Dai et al., 2023; Liu et al., 2024a) that seamlessly integrate visual and textual information. However, akin to their large language model (LLM) counterparts, VLMs exhibit a concerning tendency for hallucination, generating coherent yet unsubstantiated outputs that lack grounding in the actual visual data. Hallucinations in VLMs majorly occur due to: (a) Distributional Shift, when they encounter data distributions different from their training data (b) Data Biases, where they learn and reflect biases present in their training data (c) Lack of Grounding, where insufficient grounding to the visual input data leads to fabricated content, specifically when the LLM's language prior is stronger than the visual grounding.

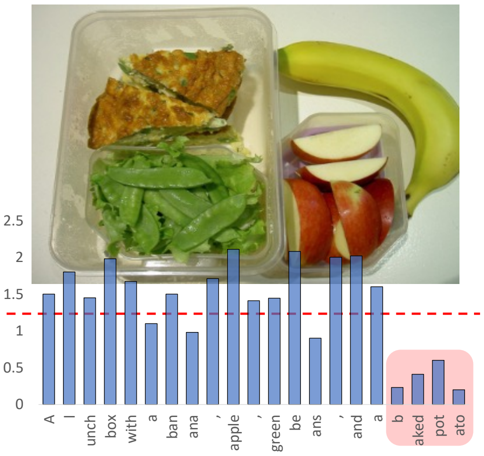
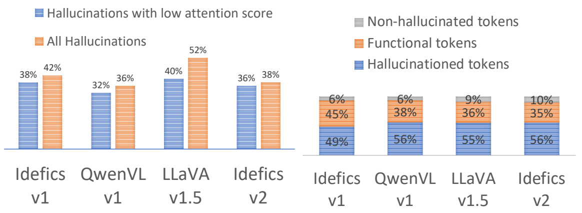


Figure 1: Illustration of the correlation between the aggregated visual attention weights and the token hallucinations.

This work focuses on detecting hallucinations in VLMs by analyzing the visual grounding of their outputs with respect to the input images. We leverage transformer's attention maps to assess visual grounding in VLMs, as they provide interpretability and a fine-grained view without requiring additional reference data, tools, or knowledge. Figure 1 illustrates the relationship between visual grounding signal and hallucinations for the generated image caption. The visual grounding signal used here is the aggregated attention weights pertaining to the input image tokens across all layers and heads for each generated text token. We clearly observe that the aggregated visual attention scores tend to be higher for accurately grounded tokens and drop significantly for hallucinated tokens. To empirically investigate this behavior, we conducted tests on four VLMs within the context of image describing tasks on 1000 samples from M-HalDetect dataset



(a) (b)

Figure 2: (a) Proportion of hallucinated tokens with low visual attention scores versus all hallucinated tokens in the dataset. (b) Distribution of tokens with low visual attention scores among true positives (hallucinations) and false positives (suffix tokens, punctuation, etc.).

(Gunjal et al., 2024) containing fine-grained hallucination labels. As shown in Figure 2a, we noticed a substantial proportion of hallucinated tokens had low visual attention scores (i.e., below the 20$^{\text{th}}$ percentile of all visual attention scores).

However, there are three major challenges in leveraging attention map based scores in detecting hallucinations:

(i) **Grounding is optional for some tokens**: In generated text, not all tokens require visual grounding, as VLMs learn to attend differently to various parts of speech. Content words like nouns, verbs, adjectives, and adverbs typically need grounding to convey semantic information about the depicted objects, actions, and attributes. However, functional and linguistic elements such as suffixes, prepositions, conjunctions, and punctuations may not necessitate direct visual grounding. Empirical evidence in Figure 2b shows that among tokens with low visual attention scores, approximately 30-40% belong to these non-grounded functional elements, potentially impacting the precision of hallucination detection methods using such measures.

(ii) **Noisy High-Dimensional Maze**: Raw attention maps from multiple heads and layers are intricate and cluttered with noise, a challenge further exacerbated by the presence of attention sinks (Xiao et al., 2024). Like CNN filters, attention heads specialize in capturing different input aspects, but only a subset correlates with visual grounding for accurate text generation in vision-language models. Directly aggregating these complex and noisy attention maps would result in information loss and suboptimal performance.

(iii) **Dynamic Sequence Length**: Attention maps are inherently dynamic in nature, with the sequence length expanding progressively as new tokens are generated during the decoding process. This continuous growth of the sequence length presents a challenge for conventional models, as they may struggle to effectively manage and process the evolving context. To address this, it is crucial to design a mechanism capable of efficiently handling sequences that increase in size over time.

To tackle these challenges, we introduce our approach called **V**isual **A**ttention Guided Hallucination **D**etection and **E**limination (VADE), a novel sequence learning paradigm to leverage the sequence of raw attention-maps corresponding to input/previous tokens for hallucination detection. Specifically, at each time step, VADE extracts attention weights from all layers and heads corresponding to the input image tokens, prompt tokens, and previously generated tokens. These attention weights are then used to construct a structured sequence of attention maps. Remarkably, with access to a small amount of annotated data, VADE can learn intricate patterns present in the attention maps, enabling it to distinguish between hallucinated and accurate segments. By formulating the task as a sequence learning problem, VADE can leverage fine-grained temporal information across the previous tokens from their attention maps without any lossy aggregation operations, thereby reducing the impact of noise and sinks. We can further integrate the VADE detector during decoding using rejection sampling strategy to mitigate hallucinations. The proposed method is architecture-independent, functioning across cross-attention and autoregressive architectures, applicable for diverse VLMs. Through extensive experiments on hallucination benchmarks, along with detailed ablation studies, we demonstrate that VADE achieves superior and generalized hallucination detection and mitigation performance across various VLMs.

2 Related Work

Previous approaches have focused on analyzing token probabilities, logits, and entropy (Guerreiro et al., 2023; Geng et al., 2024) or using probing classifiers with hidden states (Belinkov, 2021; Azaria and Mitchell, 2023). These methods often struggle to detect complex hallucinations. Recent studies (Gunjal et al., 2024; Yan et al., 2024) propose detection schemes through VLM fine-tuning with custom classification heads, but these lack cross-domain generalizability. Hallucination mitigation strategies involve integrating detection into VLM sampling (Gunjal et al., 2024; Chuang et al., 2024) or decoding processes (van der Poel et al., 2022; Favero et al., 2024; Leng et al., 2023) to improve candidate selection and probability calibration.

Some recent studies (Chuang et al., 2024; Huang et al., 2024; Liu et al., 2024b; Zhu et al., 2024) have explored leveraging attention weights for hallucination detection and mitigation. Lookback Lens (Chuang et al., 2024) handcraft a simple feature vector from aggregated attention weights and train a logistic regression classifier to detect contextual hallucinations for text summarization task in LLMs. OPERA (Huang et al., 2024), proposes to detect the knowledge aggregation patterns in VLM's self-attention maps (i.e. attention sinks) to detect the onset of hallucinated continuations. PAI (Liu et al., 2024b) and IBD (Zhu et al., 2024), propose to amplifying the attention weights assigned to image tokens to enhance visual grounding during inference for hallucination mitigation. These methods depending upon handcrafted features and manually identified emerging patterns have limited effectiveness and generalizability to detect complex and intricate cases. Differing from existing works, our proposed method (VADE), leverages the attention maps in their raw form to learn complex sequential patterns to determine visual grounding for VLMs using small annotated data. VADE, proves effective in distinguishing visual grounding signal from noise in attention maps caused by the presence outliers, sinks and null attentions.

3 Method

3.1 Task Definition

Recent generative VLMs typically consist of a vision encoder for visual perception, a large language model for causal text generation and a modality connector to glue the two modalities. Depending on the choice of the modality connector, VLMs can belong to: i) *fully autoregressive architecture* (Dai et al., 2023; Liu et al., 2023; Laurençon et al., 2024), that utilizes a learnable projector that maps the vision hidden space to the text hidden space, post which the visual tokens and text tokens are concatenated as input to the language model, ii) *cross-attention architecture* (Alayrac et al., 2022; Laurençon et al., 2023; Bai et al., 2023b), uses learnable cross-attention blocks that are interleaved within the language model, allowing the text tokens to cross-attend to the image tokens.

Let a VLM M, containing L decoder layers with H attention heads each, take input visual image (V) and a input text prompt (P) to generate text (y). The visual input tokens can be denoted as $x^V = \{x_1, \ldots, x_v, \ldots, x_V\}$, prompt tokens as $x^P = \{x_1, \ldots, x_p, \ldots, x_P\}$ and generated tokens as $y = \{y_1, \ldots, y_t, \ldots, y_T\}$. VLM takes the input embeddings and uses attention mechanisms to process and understand the contextual relationships within the inputs. It then generates output tokens autoregressively, predicting the most probable next token by leveraging the model's learned representations and the previously generated tokens as context:

$$\begin{aligned} \mathrm{z}^L &= M(x^V, x^P) \\ P(y_t|y_{<t}, x^V, x^P) &= \text{SoftMax}\left[h(\mathrm{z}^L)\right]_{\mathcal{V}} \end{aligned} \tag{1}$$

where z^L is the output hidden states of the last layer, h is a linear head that projects the hidden states z^L to obtain logits over the text vocabulary $\mathcal{V}$ and $y_{<t}$ represents the generated token sequence up to time step $t-1$. For hallucination detection, the task is to predict whether the generated text (y) is hallucinated or not. Formally, given a multi-modal hallucination detection dataset $D = \{(v_i, p_i, y_i, y_i^h)\}_{i=1}^N$, containing N image-prompt-response-label quadruplets, where $y_i^h \in \{0, 1\}$ is the ground-truth hallucination label (1 hallucinated, 0 not), the task is to predict the hallucination likelihood $(\hat{y}_i^h)$ for the generated text. The granularity of the detection can be at token, phrase/segment or sentence levels. For hallucination mitigation, given an input text prompt and image, the task is to generate accurate continuations grounded to the visual information and object hallucinations are measured with respect to ground truth objects in the image.

3.2 VADE

Formally, the attention mechanism is defined by the attention equation, which computes the attention scores between a query (Q) and key (K) :

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

In *cross-attention architecture*, Q represents the previous set of text tokens, and K represents the visual tokens. The interaction between these two components determines how the model assigns importance to image tokens while generating text tokens. While in *fully autoregressive architecture*, both K and Q are represented by concatenating the visual and text tokens and the self attention takes care of computing the dependency between

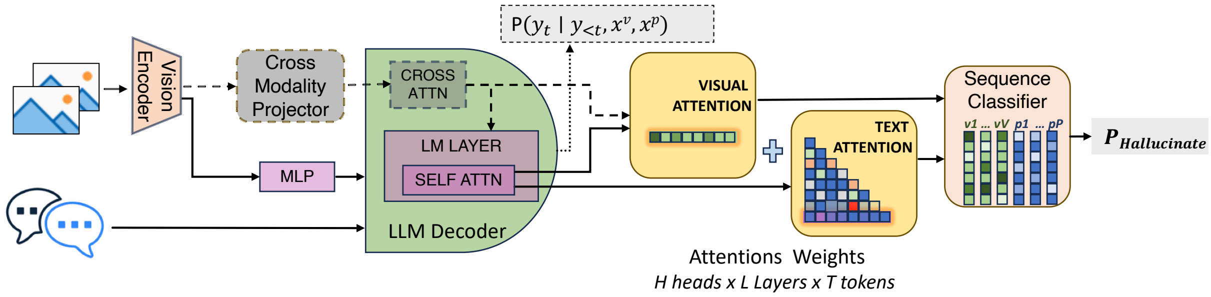


Figure 3: The architecture diagram leveraging a pre-trained VLM alongside the proposed Hallucination Detection (VADE) Components. The visual and text attention vectors are extracted separately and concatenated across all layers and heads. The visual attention is computed from the cross attention module in case of the *cross-attention* architecture and from the attention on image tokens for the *fully autoregressive* architecture.

the two modalities. We represent the two architectures in Figure 3 and explicitly demarcate the overall attention into visual and textual based on the weights emphasised on the respective tokens within the attention mechanism. We highlight how both the visual and text attention are separately extracted from all heads and layers and concatenated to create a comprehensive representation.

Specifically, at a particular time step t, as illustrated in Figure 4, we concatenate the attention weights on a particular image token x_v and text token x_p across all layers and heads and represent them as A_v and A_p.

$$A_v = \left[A_v^{(l_1,h_1)}, \dots, A_v^{(l_L,h_H)}\right] \in \mathbb{R}^{1\times LH}, \quad (3)$$

$$A_p = \left[A_p^{(l_1,h_1)}, \dots, A_p^{(l_L,h_H)}\right] \in \mathbb{R}^{1\times LH}, \quad (4)$$

where $A_v^{(l,h)}$ and $A_p^{(l,h)}$ represent a scalar attention weight of the current token on the image token x_v and text token x_p respectively for each head h in layer l at the time step t. Note that, in VLMs that use interleaved cross-attention blocks (Alayrac et al., 2022), we mask the values of A_v for the remaining layers to denote missing features.

Then, we construct a structured sequence of attention weight vectors pertaining to all input visual tokens and previous text tokens at time step t:

$$\mathbf{A}(t) = \left[A_{v_1}, .., A_{v_V}, A_{p_1}, .., A_{p_P}\right] \in \mathbb{R}^{(V+P)\times LH} \quad (5)$$

where V is the number of visual tokens and P is the number of text tokens (i.e. $P = t - 1$). $\mathbf{A}(t)$ encapsulates the nuanced interaction between the modalities at each decision point.

The length of the sequence (i.e., $V + P$) grows dynamically with each newly generated token. Consequently, converting these sequences into fixed-length feature vectors suitable for feature-based classifiers becomes challenging without employing techniques such as aggregators, padding, truncation, or fixed local windows. To address this limitation, we propose a sequence modelling approach to effectively capturing long-range dependencies and temporal relationships within the input sequences.

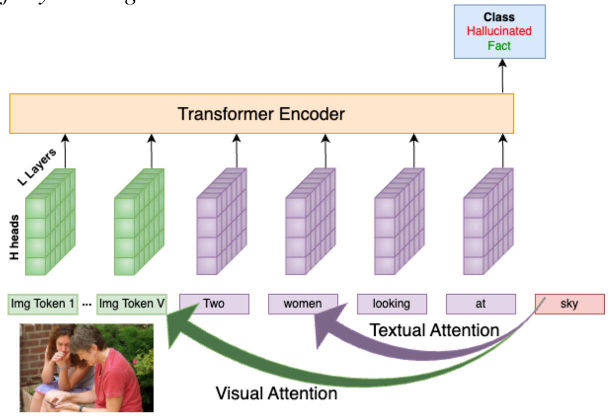


Figure 4: Illustration of how the visual and text attention vector are computed and aggregated before being processed by the sequence classifier to make the prediction for hallucination at step t.

We input the sequence $\mathbf{A}(t)$ into a sequence classifier f_θ (parameterized by θ) to learn patterns that distinguish grounded representations from hallucinated ones, enabling the classifier to predict hallucination likelihood $\hat{y}^h(t)$ for each generated token at each time step t.

$$\hat{y}^h(t) = f_\theta(\mathbf{A}(t)) \quad (6)$$

The overall training objective would be to minimize a loss function $\mathcal{L}$, such as binary cross-entropy if hallucination detection is framed as a binary classification task:

$$\mathcal{L} = -\sum_{t} \Big(y^h(t) \log(\hat{y}^h(t)) + (1 - y^h(t)) \log(1 - \hat{y}^h(t)) \Big) \tag{7}$$

where $y^h(t)$ represents the ground-truth label for hallucination at time step t. This setup allows the classifier to learn attention patterns that are indicative of hallucinations, thus enabling reliable predictions at each time step.

The innovative approach we propose for handling the structured sequence of raw attention maps addresses the challenges outlined (in Section 1) and overcomes limitations in existing formulations (detailed in Appendix C). By employing a sequence classifier, VADE can seamlessly process fine-grained raw attention information across preceding tokens without resorting to any lossy aggregation techniques, enabling it to differentiate between visual grounding signals and noise caused by outliers and sinks. This design allows VADE to identify common and complex patterns within attention maps, facilitating the distinction between hallucinated and accurate segments using minimal annotated data while achieving generalizable performance on out-of-distribution (OOD) datasets.

For hallucination mitigation task, we employ our hallucination detector (VADE) trained at segment level into the decoding process to select candidate segments with less likely hallucinations. Let $\{S_1, ..., S_k\}$ be the k candidate segments of a fixed length (t_s) generated at time step t, we select the candidate with least hallucination likelihood (S^*):

$$S^* = \arg\min_{s \in \{S_1,...,S_k\}} f_\theta(\mathbf{A}(t + t_s)) \tag{8}$$

4 Experiment

4.1 Setup

Models and Datasets: We conduct our experiments on four VLMs: Idefics-v1-9b-instruct (Laurençon et al., 2023), QwenVL-9b (Bai et al., 2023b), LLAVA-v1.5-7b (Liu et al., 2024a) and Idefics-v2-8b (Laurençon et al., 2024). Idefics-v1 and QwenVL adopt the cross-attention architecture, while LLAVA-v1.5 and Idefics-v2 leverage the fully-autoregressive architecture. We evaluate our proposed methods on two multi-modal hallucination detection datasets (M-HalDetect (Gunjal et al., 2024) and ViGoR (Yan et al., 2024)) with fine-grained annotations for hallucinated segments in detailed image descriptions. For learning based methods, we train the models on M-HalDetect's train set and evaluate on its validation set. For cross-dataset transfer evaluation, we evaluate the models on ViGoR validation set. We measure the area under the Precision-Recall (PR) curve (PR-AUC) to evaluate hallucination detection methods. For more details on the datasets and models, refer to Appendix A and Appendix B.

Baselines: As baseline methods for hallucination detection, we consider proven methods from uncertainty and hallucination literature from CV and NLP domains. We introduce them briefly below and explain them in detail in Appendix C.

Token Probability / Entropy (Guerreiro et al., 2023; Geng et al., 2024; Huang et al., 2023), methods based on the hypothesis that VLMs often hallucinate when showing high uncertainty in generation. We implement mean and maximum calculations of token probabilities and entropy as indicators of hallucination risk.

Aggregated Attention Scores, measure visual grounding by averaging attention weights across image tokens, layers, and heads. Similarly, textual attention scores are computed based on attention weights across previous text tokens. These aggregated scores help quantify visual-textual alignment in generated text.

Attention Sinks (Xiao et al., 2024)(Huang et al., 2024), represent patterns where intermediate tokens aggregate knowledge from prior tokens and are often linked to hallucination. We implement a metric identifying these patterns within local windows to evaluate knowledge aggregation.

Probing Classifiers (Gunjal et al., 2024; Yan et al., 2024; Azaria and Mitchell, 2023), involve learning a classifier on model intermediate representations to predict hallucination likelihood. Features from hidden states across layers inform this classifier, enabling layer-wise detection.

Attention Map-based Detectors (Chuang et al., 2024), train a classifier on multi-head attention maps to identify patterns linked to hallucination. Averaged attention weights from input image, prompt, and previous tokens are concatenated to form a feature vector. Detectors focus on image tokens alone (*VisAttnDet*) or a combination of all tokens (*VisTexAttnDet*).

Implementation Details: We use the attention weights before the softmax operation, as we found it to be more effective for hallucination detection than the softmaxed version. In VADE, we use a single layer transformer based encoder with 8 attention heads for the sequence model, followed by a linear classification head (details in Appendix D). For segment-level task, for each sample we set target labels at segment-end tokens while masking other indices, and optimize using cross-entropy loss. For remaining classifiers, we use a single layer neural network with a hidden dimension of 1024. All classifiers are trained with a batch size of 8 for 2 epochs. We use Adam optimizer with a learning rate of 2×10^{-4} and weight decay of 10^{-2}. We performed all experiments on NVIDIA Tesla A10 GPUs.

4.2 Main Results

Table 1 reports the hallucination detection performance for baselines and VADE on M-HalDetect and ViGoR validation sets. Entropy-based methods (MeanEnt & MaxEnt) perform slightly better than probability-based scores (MeanProb & MaxProb). Aggregated attention scores (VisAttnAgg & VisTexAttnAgg) show a 10-15% improvement over the best uncertainty-based metric, indicating that visual grounding information from naive attention weight aggregation helps in hallucination detection. Attention sink-based detectors with local window set to current segment or current + previous segment perform poorly, often worse than random, for Idefics-v1 and Qwen-VL, while showing limited detection ability for LLaVA-v1 and Idefics-v2. In cross-attention models, sink occurrence in self-attention maps may not limit the model's focus on vision tokens due to the cross-attention mechanism allowing text tokens to attend to image tokens.

Probing Classifiers significantly outperform the uncertainty and attention based metrics, across all models. Specifically, probe classifier with Idefics-v2 achieves 90% gain over VisTexAttnAgg metric, highlighting the effectiveness of learning based methods. Among the probing classifier variants, we observe across models that use hidden states from multiple layers or all layers did not bring considerable gains over using only the last layer. Attention map based detectors (VisAttnDet & VisTexAttnDet) further boost the detection performance and achieve a relative improvement of about 15% over probing classifiers. The VisTexAttnDet model outperforms the VisAttnDet model, highlighting the significance of leveraging both visual and textual attention mechanisms to detect fabricated segments. Our proposed method VADE, shows superior performance amongst attention map based detectors, proving the potential of learning complex sequential patterns from high-dimensional and noisy attention maps for effecting hallucination detection. We show qualitative results in Appendix E.

Transferability Performance. From Table 1, probing classifiers leveraging hidden states suffer a significant 35% performance drop when transferred from M-HalDetect (training) to ViGoR (out-of-distribution). This supports the hypothesis that hidden states optimized for next-token prediction tasks are less effective for hallucination detection on out-of-distribution data, being overly specialized to the training dataset and limiting their utility in unseen contexts. Attention map-based detectors (VisAttnDet, VisTexAttnDet and VADE), on the other hand, show a relatively smaller performance decline (14%) in the transfer setting. The ability of these models to leverage both visual and textual attention maps makes them more adaptable to unseen data, as they are less sensitive to the idiosyncrasies of specific datasets. Within these detectors, VADE exhibits the best absolute performance and the least degradation (11%) in the transfer setting. VADE's ability to process sequence of raw attention maps without any lossy operators needed to convert to fixed size vectors, allows it effectively discern visual grounding signal from noisy attention maps and learn transferable patterns for hallucination detection across different datasets.

4.3 Ablation Study

Impact of Aggregation Operator. For attention score aggregation, we found that computing *median* over the attention values across layers and heads performs marginally better than *mean* aggregator due to the presence of outlier values as reported in Table 2. While using *minimum*, *quartile* and *maximum* operators resulted in very poor performance. We conjecture that presence of extremum values in uninformative heads (Vig and Belinkov, 2019) and knowledge aggregation patters (Xiao et al., 2024; Huang et al., 2024) causes this sensitivity.

Effect of Attention Head Selection. The relevance of a particular head in a layer can be quantitatively assessed by studying the distribution of attention weights on image tokens, $A_v^{(l,h)}$, for each output token y_t in an unsupervised manner on a small set of hold out images and prompts. Given the image and the input instruction prompt (e.g. *Describe the image briefly*), we generate captions

	M-HalDetect (Source)				**ViGoR** (Target)			
Method	Idefics v1	QwenVL v1	LLaVA v1.5	Idefics v2	Idefics v1	QwenVL v1	LLaVA v1.5	Idefics v2
Random	23.7	23.7	23.7	23.7	25.2	25.2	25.2	25.2
Token Probability / Entropy								
MeanProb	21.2	26.7	27.2	30.1	21.5	24.2	25.5	27.2
MaxProb	22.4	26.9	28.4	25.6	23.7	25.6	26.9	26.5
MeanEnt	25.1	32.1	30.2	29.5	26.8	28.9	29.4	25.6
MaxEnt	27.9	31.6	29.4	27.4	27.6	30.7	28.1	28.4
Aggregated Attention Scores								
VisAttnAgg	27.4	31.4	32.8	33.1	35.5	33.6	33.4	35.2
VisTexAttnAgg	30.1	37.6	35.4	34.9	38.4	39.7	36.7	39.8
Attention Sink								
AttnSink (*k=curr*)	18.4	21.4	27.6	28.7	20.1	24.4	34.9	33.8
AttnSink (*k=prev+curr*)	19.1	19.6	30.1	30.4	22.3	25.6	32.8	35.4
*__Probing Classifiers__								
Last Layer	63.5	67.1	66.4	64.8	44.2	49.4	50.4	52.4
Last Four Layers	62.1	62.8	68.2	68.2	41.9	47.5	49.6	54.2
All Layers	60.9	69.4	70.4	66.4	43.1	47.2	45.8	52.8
*__Attention Map based Detectors__								
VisAttnDet	67.5	70.2	75.1	74.4	58.1	60.4	62.1	66.4
VisTexAttnDet	70.1	70.9	77.5	79.7	60.7	64.1	63.8	67.5
VADE	**74.9**	**73.7**	**81.5**	**84.7**	**66.8**	**69.4**	**67.1**	**72.5**

Table 1: PR-AUC for Hallucination Detection on M-HalDetect and ViGoR datasets. *For learning based methods we operate in a cross dataset setup where we train on M-HalDetect and evaluate on ViGoR.

Method	Setting	Idefics-v1	LLaVA-v1.5
VisAttnAgg	*Mean*	27.4	32.8
VisAttnAgg	*Median*	**28.5**	**33.3**
VisAttnAgg	*Min*	24.1	27.4
VisAttnAgg	*q(75)*	25.5	28.1
VisAttnAgg	*Max*	21.6	22.8
VisAttnAgg	*k=10*	**33.2**	**35.2**
VisAttnAgg	*k=50*	30.5	34.3
VisAttnAgg	*k=100*	27.1	33.5
VisAttnAgg	*k=All*	27.4	33.1
VADE	*k=10*	55.5	63.2
VADE	*k=50*	62.1	64.7
VADE	*k=100*	65.9	69.9
VADE	*k=All*	**74.9**	**81.5**

Table 2: Ablation Study on different settings and hyperparameters.

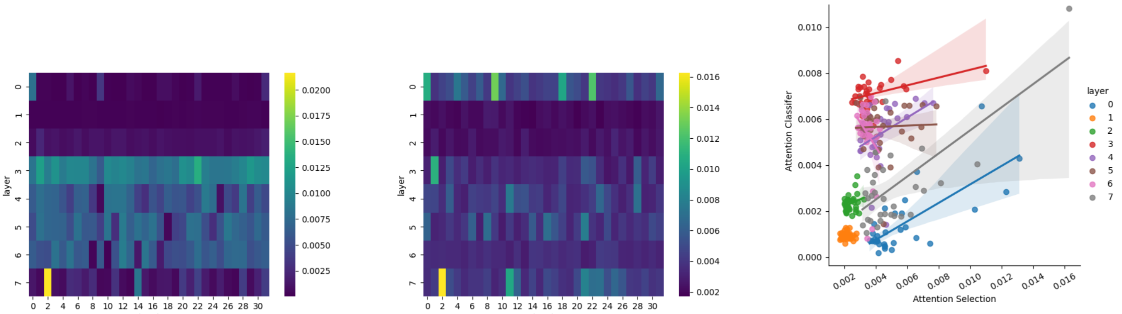
and aim to analyze the token-level visual attention weights for each layer and head. We introduce the concept of **Attention Head Selection**, where the variance in softmax-normalized attention weights on image tokens is used as a criterion for selecting the most impactful layers and heads. We define the importance score $S^{l,h}$ for a given layer l and head h as:

$$S^{(l,h)} = \text{Var}_{\alpha}^{(l,h)} = \frac{1}{N}\sum_{i=1}^{N}\left({\alpha_i}^{(l,h)} - \mu_{\alpha}^{(l,h)}\right)^2,$$

where $\alpha_i^{(l,h)}$ is the softmax-normalization of the visual attention weights $A_i^{(l,h)}$ and $\mu_{\alpha}^{(l,h)}$ is the mean of $\alpha_i^{(l,h)}$ across N generated tokens.

Figure 5a illustrates the heat map of layer-head importance scores extracted from above discussed attention head selection logic and Figure 5b shows the importance scores derived from coefficients of the classifier used in VisAttnDet, with both methods identifying layer 7 and head 2 as the most important. The strong correlation between the importance scores (Figure 5c) qualitatively emphasizes the reliability of statistical head selection method, especially in data-scarce scenarios. We futher observe that several layers and heads attend to image tokens with similar weights (low variance) across all tokens, thereby not contributing to the model's ability to selectively capture visual features.

By focusing on heads with higher variance, we can identify components that effectively ground generated text in visual content. To validate this, we selected top-k layer-heads using attention head selection logic discussed above on 1000 random images from M-HalDetect train set and show the results for *VisAttnAgg* and *VADE* methods for different k in Table 2. We notice that for *VisAttnAgg*, the detection performance drops as more layer-head combinations are selected, likely due to noise from less informative layers and heads. On the contrary, in *VADE*, adding more heads steadily improves the performance. The analysis reveals that atten-



(a) Attention Head Selection

(b) VisAttnDet Classifier

(c) Correlation b/w AS and AC importance scores

Figure 5: *(a)(b)* Visualization of Layer-Head importance scores derived from Attention Head Selection logic and VisAttnDet Classifier coefficients. We see a strong correlation between both methods in identifying informative layers and heads. The former found layers=$\{0, 7, 5, 4\}$ as Top-4 important layers, while the later identified them to be $\{3, 5, 4, 6\}$. *(c)* Layer-Head importance scores from Attention Head Selection against VisAttnDet Classifier coefficients. We see that for all layers (except layer 6), there is a positive correlation. Pearson's correlation coefficient between the importance scores from two methods was found to be $+0.3$.

tion layer heads exhibit distinct receptive fields, with learning-based methods more adept at detecting complex patterns than statistical approaches. However, this advantage is tempered by limited access to high-quality hallucination labels and challenges in cross-dataset transferability, underscoring the need for large-scale multi-modal hallucination datasets. For additional results, ablations and discussions on ternary classification setting, sentence-level detection, sequence model architecture and computational efficiency refer to Appendix E.

4.4 Performance on Hallucination Mitigation

Method	CHAIRi ↓	CHAIRs ↓	Recall ↑
Greedy Decoding	14.5	44.0	53.1
Beam Search	13.9	48.8	**55.4**
PMI	7.6	28.4	50.2
M3ID	8.1	29.2	52.7
OPERA	12.8	42.6	55.1
PAI	6.6	20.7	49.4
VisTexAttnAgg	7.8	28.5	51.1
VADE	**6.3**	**19.9**	50.2

Table 3: Performance comparison of hallucination mitigation methods on MS COCO validation set with LLaVA-v1.5.

We evaluate VADE's hallucination reduction for detailed image description on MS COCO 2014 val set using rejection sampling (as in (Gunjal et al., 2024; Chuang et al., 2024)), selecting candidates with least hallucination likelihood in a best-of-K setting via Beam Search. We assess using CHAIR metrics (Rohrbach et al., 2018) on 500 random images, prompting models with "*Please describe this image in detail.*" with `max_new_tokens` = 512. From Table 3, mutual information-based decoding techniques like PMI (van der Poel et al., 2022) and M3ID (Favero et al., 2024) demonstrate substantial hallucination reduction. Among existing methods that leverage attention map-based, OPERA (Huang et al., 2024) shows modest improvements, while PAI (Liu et al., 2024b) achieves significant hallucination mitigation. Compared with baseline methods, we can observe our VADE attains the highest performance among these mitigation strategies, albeit with marginal gains over PAI. Refer to Appendix F for evaluation details and more results.

5 Conclusion

In this work, we conducted a comprehensive analysis of the properties of multi-head and multi-layered visual attention patterns, evaluating their effectiveness in detecting hallucinations in VLMs. Our findings revealed that statistical measures, aggregation schemes, and feature selection strategies used to derive visual grounding signals from attention maps are suboptimal for detecting hallucinations. Unlike conventional methods that rely on handcrafted features or attention aggregation, our proposed method VADE learns complex sequential patterns from high-dimensional and noisy attention maps using a limited amount of annotated data. Our extensive experiments across multiple VLMs and datasets demonstrate the superiority of VADE over existing techniques, achieving state-of-the-art performance in both hallucination detection and mitigation tasks. Furthermore, VADE exhibits strong generalization capabilities, with minimal performance degradation when transferred to out-of-distribution datasets, highlighting its potential for real-world applications.

Limitations

Despite its promising performance, VADE has certain limitations, including its reliance on high-quality hallucination annotations, lack of explicit cross-modal modeling, computational overhead in processing attention maps, and potential performance degradation on domain-shifted or adversarial data. Additionally, VADE's hallucination mitigation strategy is limited by the samples generated by the base VLM and cannot correct the hallucinations in the samples. Addressing these challenges through advanced modeling techniques, efficient computation strategies, and robust training methods could further enhance the reliability and scalability of VADE for real-world applications.

Ethics Statement

In this work, we used publicly available datasets and open-sourced models in accordance to their intended usage terms and licenses. Our method discusses and addresses the problem of hallucinations in Vision-Language Models (VLMs) and aims to reduce the inherent data bias and factual errors, enhancing the reliability and trustworthiness of VLMs in real-world applications. When implemented, our method still inherits the challenges associated with visual language models. Consequently, there remains a potential risk that the VLM may generate output that is biased, harmful, or offensive in nature.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *Preprint*, arXiv:2304.13734.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and advances. *Preprint*, arXiv:2102.12452.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *Preprint*, arXiv:2407.07071.

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

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14303–14312.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. *Preprint*, arXiv:2311.08298.

Nuno M. Guerreiro, Elena Voita, and André F. T. Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. *Preprint*, arXiv:2208.05309.

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

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *Preprint*, arXiv:2311.17911.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *Preprint*, arXiv:2307.10236.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. *Preprint*, arXiv:2103.03206.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Preprint*, arXiv:2306.16527.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024b. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *Preprint*, arXiv:2407.21771.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *Empirical Methods in Natural Language Processing (EMNLP)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *Preprint*, arXiv:1906.04284.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. *Preprint*, arXiv:2309.17453.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. *Preprint*, arXiv:2402.06118.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *Preprint*, arXiv:2402.18476.

A Datasets

M-HalDetect The dataset comprises of images from MS COCO 2014 validation set (Lin et al., 2015), with 3200 images for training and 800 for validation. Each image is accompanied by four diverse input prompts and corresponding detailed responses comprising multiple segments of varying length, totaling 12800 training and 3200 validation samples. The text segments within the responses were labelled by expert annotators into four categories: i) Accurate ii) Analysis iii) Inaccurate and iv) Unsure. Approximately 25% of the segments exhibit hallucinations, highlighting the challenge in detailed image description tasks with VLMs. Following the previous work (Gunjal et al., 2024), we first merge the "Unsure" category into "Inaccurate" class and report the dataset stats in Table 4. We experiment with two settings, namely, binary classification (Accurate, Inaccurate), by further merging Analysis category into Accurate category and ternary classification (Accurate, Analysis, Inaccurate).

Splits	Samples		Segments	
		Accurate	Analysis	Inaccurate
Train	12800	40209	17289	18129
Val	3200	11770	5207	5115

Table 4: M-HalDetect dataset statistics

ViGoR The dataset comprises of 7200 images from MS COCO 2017 train set (Yan et al., 2024). We consider the entire set for our validation in transfer setting. Each image is queried with diverse input prompts related to image description and corresponding detailed responses comprising multiple sentences are collected, totaling 15400 samples. Unlike M-HalDetect, expert annotations are only available at sentence level and has following label form {Accurate: Yes/No, Creative: Yes/No} per sentence. We transform these labels with the label definition used in M-HalDetect for consistency using the rubric in Table 5 and report the dataset statistics in Table 6.

Accurate	Creative	Final Label
Yes	No	Accurate
Yes	Yes	Analysis
No	Yes/No	Inaccurate

Table 5: Rubric for converting ViGoR labels into M-HalDetect's label definition

Splits	Samples		Sentences	
		Accurate	Analysis	Inaccurate
Val	7200	36253	15306	20362

Table 6: ViGoR dataset statistics

B Models

IDEFICS-v1 (Laurençon et al., 2023) is an open-source state-of-the-art VLM based on the Flamingo architecture (Alayrac et al., 2022). It combines a CLIP ViT H14 (Radford et al., 2021) vision encoder and an LLaMA-7B (Touvron et al., 2023) language model. It uses a Perceiver Resampler (Jaegle et al., 2021) to convert variable image features to 64 visual tokens, connecting the vision encoder to the frozen language model. Learnable cross-attention transformer blocks (Vaswani et al., 2023) are interleaved with the pre-trained LM layers for visual conditioning during text generation. The 9B instruction-finetuned model has 8 cross-attention layers with 32 heads each and 32 self-attention blocks with 32 heads each. Visual attention maps are from the cross-attention layers, and textual attention maps are from the self-attention heads.

QwenVL (Bai et al., 2023b), uses the Qwen-7B (Bai et al., 2023a) as the LLM backbone and Openclip's ViT-bigG (Radford et al., 2021) as the vision encoder. Similar to IDEFICS-v1, they use a Perceiver Resampler (Jaegle et al., 2021) to connect the vision encoder to the frozen language model that compresses the visual feature sequence to a fixed length of 256. The model has 32 layers, the embedding dimension is 4096, and the number of attention heads is 32. We use the 9B instruction fine-tuned version for our experiments.

LLaVA-v1.5 (Liu et al., 2024a) utilizes the Vicuna-7B (Bai et al., 2023a) as its LLM backbone, and the CLIP-ViT-L-336px (Radford et al., 2021) as the vision encoder. An MLP cross-modal connector is employed to project the visual tokens (576 in number) into the word embedding space of the language model. The input sequence fed into the language model is the concatenation of these visual tokens and the text tokens. The model architecture comprises 32 layers, 32 attention heads, and a hidden dimension of 4096.

IDEFICS-v2 (Laurençon et al., 2024) employs Mistral-7B (Jiang et al., 2023) as the language model backbone and SigLIP-SO400M (Zhai et al., 2023) from Google as the vision encoder backbone. It utilizes a fully-autoregressive architecture where the vision encoder's output is concatenated with text embeddings, and the entire sequence is fed into the language model. The visual token sequence is pooled to a shorter length (64) for computational efficiency. The 8B model has 32 layers, 32 attention heads, and uses 4096-dimensional embeddings.

C Baselines: Implementation Details

C.1 Token Probability / Entropy

These methods are motivated by the hypothesis that VLMs often tend to hallucinate when they exhibit high uncertainty during generation. Previous works for LLMs (Guerreiro et al., 2023; Geng et al., 2024; Huang et al., 2023) used aggregated output token probabilities or entropy as a measure for hallucination detection. We implement two most common methods calculating the mean and maximum of these scores on the generated text segment:

$$\text{MeanProb} = \frac{1}{S}\sum_{t=1}^{S} P(y_t|y_{<t}, x^v, x^p) \tag{9}$$

$$\text{MaxProb} = \max_{t\in S} P(y_t|y_{<t}, x^v, x^p) \tag{10}$$

$$\text{MeanEnt} = \frac{1}{S}\sum_{t=1}^{S} \mathcal{H}_t \tag{11}$$

$$\text{MaxEnt} = \max_{t\in S} \mathcal{H}_t \tag{12}$$

where S is the length of the generated text segment and $\mathcal{H}_t$ is the entropy of the generative distribution at time step t.

C.2 Aggregated Attention Scores

These methods are based on the hypothesis that low visual grounding can lead to hallucinated content and leverage the transformer attention weights to measure the *visual groundedness* of the generated text with respect to the input image. The aggregated visual attention metric $A^V_{avg}(t)$ for the token at t is computed as the average attention weight across the V image tokens, L layers, and H heads:

$$A^V_{avg}(t) = \frac{1}{L \times H \times V}\sum_{l=1}^{L}\sum_{h=1}^{H}\sum_{v=1}^{V} A_v^{(l,h)} \tag{13}$$

where $A_v^{(l,h)}$ is the attention weight of the token at t on the v^{th} image token in head h and layer l.

Similarly, the aggregated textual attention metric $A^P_{avg}(t)$ for the token at t is computed as the average attention weight across the $t-1$ previous text tokens, L layers, and H heads:

$$A^P_{avg}(t) = \frac{1}{L \times H \times t-1}\sum_{l=1}^{L}\sum_{h=1}^{H}\sum_{t=1}^{t-1} A_p^{(l,h)} \tag{14}$$

Finally, for the text segment of interest, the above aggregated attention values are averaged:

$$\text{VisAttnAgg} = \frac{1}{S}\sum_{t=s}^{S} A^V_{avg}(t) \tag{15}$$

$$\text{VisTexAttnAgg} = \frac{1}{S}\sum_{t=s}^{S} \frac{A^V_{avg}(t)}{A^P_{avg}(t)} \tag{16}$$

C.3 Attention Sinks

Attention sinks are (Xiao et al., 2024) are columnar attention patterns in self-attention maps that aggregate knowledge from previous tokens. Prior studies (Huang et al., 2024) show hallucinated content often follows tokens with this pattern. We implement a metric to detect knowledge aggregation patterns within a local window as detailed in (Huang et al., 2024) and evaluate its effectiveness for hallucination detection. Let $W \in \mathbb{R}^{k^2}$ be the local window self-attention map of the last layer with max pooled attention values across multi-heads, where k is the window size. We conduct column-wise multiplication on the lower triangle of the attention matrix and obtain a vector of column-wise scores. The maximum value of this vector is the characteristic of knowledge aggregation patterns. Formally,

$$W^k_{t-1} = \{w^i\}^{t-1}_{i=t-k}, \quad \text{s.t. } w^i = \{\omega_{i,j}\}^{i}_{j=t-k}, \tag{17}$$

$$\phi(\omega_{<t}) = \prod_{i=c}^{t-1} \sigma\omega_{i,c}, \quad \text{s.t. } c = \mathop{\arg\max}_{t-k\leq j\leq t-1} \prod_{i=j}^{t-1} \sigma\omega_{i,j}. \tag{18}$$

where $\omega_{i,j}$ means the attention weight assigned by the j^{th} token to the i^{th} token and σ is a configurable scaling factor. For the token concluding each segment, we take the previous and current segment span as the local window, to detect the presence of hallucinated content in current segment.

C.4 Probing Classifiers

These methods resort to learning a classifier on model's intermediate representations (z^l_t) from a specific VLM layer l for the generated token t to predict the likelihood of hallucination (Gunjal et al., 2024; Yan et al., 2024; Azaria and Mitchell, 2023). Similar to previous works, we implement a classification head with input features from the hidden states of the VLM's decoder from last, multiple or all layers. For instance, the classifier that uses the hidden states from last four layers can be formalized as:

$$\bar{z}_t = \left[z_t^{l_{L-3}}, z_t^{l_{L-2}}, z_t^{l_{L-1}}, z_t^{l_L}\right] \tag{19}$$

$$\hat{y}^h = f_\theta\left(\bar{z}_t\right) \tag{20}$$

In segment-level task, for each sample we set target labels at segment-end tokens while masking other indices, and optimize using cross-entropy loss. Given the hidden dimension of 4096 for the models under consideration, concatenating the features from the last 4 layers would yield a high-dimensional 16384-dim feature vector. While utilizing the hidden states from all layers was also explored, concatenating them would lead to an extremely large feature vector, significantly slowing down the classifier training process. Consequently, to obtain input feature vectors of a manageable size (4096-dim) for the classifier, an average pooling operation was performed across the features from different layers.

C.5 Attention Map based Detectors

These detectors involve training a classifier on model's multi-head attention maps from several layers for the read or generated token t to learn and detect attention patterns leading to hallucination (Chuang et al., 2024). For each time step t, averaged attention weights pertaining to input image tokens and previously read/generated text tokens are fetched for every head h in layer l and their absolute values/ratios are concatenated to form single vector $\bar{\mathbb{A}}_t$.

$$A^V_{(l,h)} = \frac{1}{V}\sum_{v=1}^{V} A_v^{(l,h)} \tag{21}$$

$$A^P_{(l,h)} = \frac{1}{t-1}\sum_{p=1}^{t-1} A_p^{(l,h)} \tag{22}$$

$$\begin{aligned} A(t) = \big[&A^V_{(l_1,h_1)}, \ldots, A^V_{(l_L,h_H)}, \\ &A^P_{(l_1,h_1)}, \ldots, A^P_{(l_L,h_H)}\big] \end{aligned} \tag{23}$$

$$\begin{aligned} \bar{\mathbb{A}}(t) &= \frac{1}{S}\sum_{t=s}^{S} A(t) \\ \hat{y}^h(t) &= f_\theta\left(\bar{\mathbb{A}}(t)\right) \end{aligned} \tag{24}$$

We refer the detector that takes only image tokens $\left[A^V\right]$ as inputs as *VisAttnDet* and a combination of all tokens $\left[A^V, A^P\right]$ as *VisTexAttnDet* respectively.

D VADE: Implementation Details

In VADE, we employ a single-layer transformer encoder with 8 attention heads for the sequence model, followed by a linear classification head for binary/ternary classification tasks. The model's embedding dimension is set to 1024 (i.e. $L \times H$ from Equation 5). In the position-wise feed-forward sub-layer, we use an intermediate size of 2048 and the SiLU non-linear activation function. Layer normalization is performed using RMSNorm with $\epsilon = 1e^{-6}$. A dropout probability of 0.2 is applied to all dropout layers for regularization. Additionally, we incorporate rotatory positional embeddings to encode positional information in the input embeddings.

E Additional Results

E.1 Ternary Classification Setting

While the Table 1 in the main paper reports the results for binary classification settting, Table 7 presents the performance of the methods in the ternary classification task. For learning based methods, we train the models on M-HalDetect's train set for the ternary classification task and evaluate on its validation set. For cross-dataset transfer evaluation, we evaluate the models on ViGoR validation set. For each individual class, the precision-recall curve is computed, and subsequently, these curves are averaged to yield a single metric that encapsulates the model's overall performance across the three classes. The task is relatively difficult, since the separated "Analysis" class contains portions of text that are subjective and not visually grounded, but are not inaccurate. From the table, we observe similar performance trends both among baselines and VADE shows improvement across multiple models in both source and target validation datasets.

E.2 Sentence-level Hallucination Detection

We consolidate M-HalDetect's labeled sub-sentence segments into sentence-level segments, labeling each sentence as inaccurate if any sub-segment is inaccurate, as analysis if any sub-segment is analysis (and none are inaccurate), and as accurate if any sub-segment is accurate (and none are inaccurate or analysis). ViGoR already contains sentence-level labels and hence requires no change. We train the models on the sentence-level labels on M-HalDetect train set and evaluate on validation sets of M-HalDetect and ViGoR as before. Table 8 reports the results of the detection in binary and ternary setting for sentence level labels. We see that, transfer performance of probing classifier is much better when trained on sentence level labels, as the source and target tasks are similar. The drop in performance between source and target validation sets has narrowed to about 15% compared to 35% in Table 1. Attention based detectors, specifically VADE, shows robust performance in this setting and also maintain their transfer capa-

	M-HalDetect (Source)				**ViGoR** (Target)			
Method	Idefics v1	QwenVL v1	LLaVA v1.5	Idefics v2	Idefics v1	QwenVL v1	LLaVA v1.5	Idefics v2
Random	21.3	21.3	21.3	21.3	22.7	22.7	22.7	22.7
Token Probability / Entropy								
MeanProb	19.1	24.0	24.5	27.1	19.4	21.8	23.0	24.5
MaxProb	20.2	24.2	25.6	23.0	21.3	23.0	24.2	23.9
MeanEnt	25.6	28.9	27.2	26.6	24.1	28.0	26.5	23.0
MaxEnt	25.1	28.4	26.5	24.7	24.8	27.6	25.3	25.6
Aggregated Attention Scores								
VisAttnAgg	24.7	28.3	29.5	29.8	32.0	30.2	30.1	31.7
VisTexAttnAgg	27.1	33.8	31.9	31.4	34.6	35.7	33.0	35.8
Attention Sink								
AttnSink (*k=curr*)	16.6	19.3	24.8	25.8	18.1	22.0	31.4	30.4
AttnSink (*k=prev+curr*)	17.2	17.6	27.1	27.4	20.1	23.0	29.5	31.9
*__Probing Classifiers__								
Last Layer	57.2	60.4	59.8	58.3	39.8	44.5	45.4	49.2
Last Four Layers	59.9	56.5	61.4	61.4	37.7	42.8	44.6	48.8
All Layers	54.8	62.5	63.4	62.8	38.8	42.5	41.2	47.5
*__Attention Map based Detectors__								
VisAttnDet	60.8	63.2	67.6	67.0	52.3	54.4	55.9	59.8
VisTexAttnDet	63.1	63.8	**74.4**	71.7	54.6	57.7	59.4	60.8
VADE	**67.4**	**66.3**	73.4	**76.2**	**60.1**	**62.5**	**60.4**	**65.3**

Table 7: PR-AUC for hallucination detection on M-HalDetect and ViGoR datasets in ternary classification setting. *For learning based methods, we operate in a cross dataset setup where we train on M-HalDetect and evaluate on ViGoR.

bility across datasets.

E.3 Ablation on Sequence Model Architecture

Table 9 reports the effect of adding more attention blocks in the sequence encoder. We see that as we increase the network depth, the model tends to overfit to the training distribution and shows deterioration on the validation sets. This is clearly evident when we look at the performance on the transfer validation dataset (ViGoR), where the model with $L = 1$ and $H = 8$ shows superior results. Table 10 compares the model parameter counts and FLOPS per token between the baseline methods and VADE using the method introduced in previous work (Kaplan et al., 2020). VADE requires slightly higher computational resources than baseline methods but delivers better hallucination detection results.

E.4 Qualitative Analysis

We show qualitative example from M-HalDetect in Table 11 to illustrate how VADE improves hallucination detection performance over competitive baselines. Firstly, the aggregated attention scores in *VisTexAttnAgg* method fails to detect much of the hallucinated segments. Probe Classifier (using the hidden states from the last layer), identifies all of the hallucinated segments but also contains many false positive detections. VADE, with the sequential classifier over the attention maps, effectively detects most hallucinated segments with better precision.

F Hallucination Mitigation

Metrics. We measure object hallucinations using the CHAIR (Captioning Hallucination Assessment with Image Relevance) metrics (Rohrbach et al., 2018), that use ground truth annotations to calculate the proportion of objects that appear in the model generated text but absent in the image. CHAIRi measures the fraction of hallucinated objects among the generated text and CHAIRs computes the fraction of captions with at least one object hallucination. Additionally, Recall measures the fraction of correctly predicted objects compared to ground truth.

$$\text{CHAIRi} = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{generated objects}\}|}$$

$$\text{CHAIRs} = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}$$

$$\text{Recall} = \frac{|\{\text{correctly predicted objects}\}|}{|\{\text{ground truth objects}\}|}$$

	M-HalDetect (Source)				**ViGoR** (Target)			
Method	Idefics v1	QwenVL v1	LLaVA v1.5	Idefics v2	Idefics v1	QwenVL v1	LLaVA v1.5	Idefics v2
Binary Classification Setting								
Probing Classifiers (Last Layer)	55.8	59.4	60.8	54.1	47.2	46.5	49.1	48.2
VisAttnDet	58.2	61.7	62.5	56.9	52.3	54.4	55.9	59.8
VisTexAttnDet	60.3	62.8	63.1	59.2	54.6	57.7	59.4	60.8
VADE	**65.1**	**64.3**	**63.9**	**76.2**	**59.8**	**63.1**	**64.7**	**63.9**
Ternary Classification Setting								
Probing Classifiers (Last Layer)	47.4	50.5	51.7	46.0	40.1	39.5	41.7	41.0
VisAttnDet	49.5	52.4	53.1	48.4	44.5	46.2	47.5	50.8
VisTexAttnDet	51.3	**55.4**	**54.3**	50.3	46.4	49.0	50.5	51.7
VADE	**55.3**	53.7	**54.3**	**54.8**	**50.8**	**53.6**	**52.0**	**54.3**

Table 8: PR-AUC for hallucination detection at sentence level on M-HalDetect and ViGoR datasets.

			M-HalDetect (Source)		**ViGoR** (Target)	
Method	*Layers*	*Heads*	Idefics-v1	LLaVA-1.5	Idefics-v1	LLaVA-1.5
VADE	1	1	72.1	79.4	65.1	65.9
VADE	1	8	74.9	81.5	66.8	67.1
VADE	3	8	75.5	82.1	66.2	66.8
VADE	5	8	73.9	80.8	63.1	59.2

Table 9: Ablation on different model sizes for the sequence model.

Method	#Parameter (M)	#FLOPS (M)
Probing Classifiers		
Last Layer	4.2	8.3
Last Four Layers	16.7	33.5
All Layers	4.2	8.4
Attention Map based Detectors		
VisAttnDet	1.1	2.1
VisTexAttnDet	2.1	4.2
VADE (L=1, H=8)	6.2	12.7

Table 10: Comparison of model parameter counts and FLOPS (Floating-Point Operations) per token.

Baselines. Greedy Decoding, selects the token with the highest probability from the output distribution. Beam Search Decoding, maintains multiple candidate hypotheses or "beams" to explore a broader range of possibilities and ultimately select the best hypothesis from the set.

Mutual Information-based Decoding (PMI (van der Poel et al., 2022), M3ID (Favero et al., 2024)) aim to maximize the mutual information between generated tokens and the input image, formulated as:

$$y_t = \arg\max_{y \in \mathcal{V}} \big(P(y_t|y_{<t}, x^v, x^p) \\ - \mu \mathbb{1}\left[\mathcal{H}_t \geq \tau\right] \log P(y_t|y_{<t}, x^p) \big) \tag{25}$$

where, μ is the calibration strength factor, $\mathcal{H}_t$ denotes the Shannon entropy, τ is entropy threshold and $P(y_t|y_{<t}, x^p)$ is the predictive probability distribution unconditioned on the input image at time t. The above formulation ensures the predictive distribution is calibrated when the model exhibits uncertainty to reduce the language bias.

OPERA (Huang et al., 2024) employs a penalty term on the model logits during beam search decoding to address over-trust on the summary tokens, along with a rollback strategy that checks for summary tokens in previously generated tokens and re-allocates token selection if needed. PAI (Liu et al., 2024b) adaptively adjusts and amplifies the attention weights assigned to image tokens, giving greater prominence to visual elements and leveraged mutual information-based decoding to eliminate the language prior.

VADE. We employ the our hallucination detector (VADE) trained at segment level into the decoding process to select candidate segments with less likely hallucinations. Let $\{S_1, ..., S_k\}$ be the k candidate segments of a fixed length (t_s) generated at time step t, we select the candidate with least hallucination likelihood (S^*):

$$S^* = \arg\min_{s \in \{S_1, \ldots, S_k\}} f_\theta(\mathbf{A}(t + t_s)) \tag{26}$$

We used a segment size of 20 tokens with 8 candidates for each step. The hyperparameters for baselines and VADE are selected using standard grid search on a held out set.

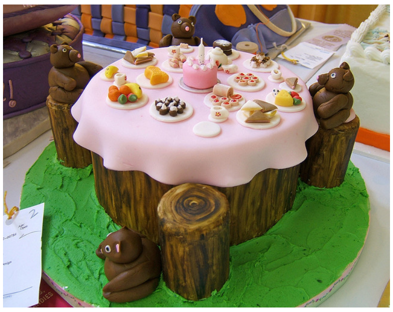
Hallucination Detection Example	
	
Prompt	Describe the following image in detail.
Groundtruth	The image features a large cake decorated to resemble a tree stump with bears sitting around it. The cake is placed on a table, and there are several teddy bears scattered around the scene. The bears are positioned in different areas of the table, some closer to the center of the cake, while others are more towards the edges. There are two handbags placed on the table, one near the left edge and the other near the right edge.
VisTexAttnAgg	The image features a large cake decorated to resemble a tree stump with bears sitting around it. The cake is placed on a table, and there are several teddy bears scattered around the scene. The bears are positioned in different areas of the table, some closer to the center of the cake, while others are more towards the edges. There are two handbags placed on the table, one near the left edge and the other near the right edge.
Probe Classifier	The image features a large cake decorated to resemble a tree stump with bears sitting around it. The cake is placed on a table, and there are several teddy bears scattered around the scene. The bears are positioned in different areas of the table, some closer to the center of the cake, while others are more towards the edges. There are two handbags placed on the table, one near the left edge and the other near the right edge.
VADE	The image features a large cake decorated to resemble a tree stump with bears sitting around it. The cake is placed on a table, and there are several teddy bears scattered around the scene. The bears are positioned in different areas of the table, some closer to the center of the cake, while others are more towards the edges. There are two handbags placed on the table, one near the left edge and the other near the right edge.

Table 11: Illustration of hallucination detection methods on M-HalDetect validation sample (ID: 260808). Groundtruth and predicted hallucination segments are highlighted in red.

	Idefics-v1			**QwenVL-v1**			**Idefics-v2**		
Method	CHAIRi ↓	CHAIRs ↓	Recall ↑	CHAIRi ↓	CHAIRs ↓	Recall ↑	CHAIRi ↓	CHAIRs ↓	Recall ↑
Greedy Decoding	16.7	50.6	45.1	13.6	41.1	56.8	12.8	37.4	60.1
Beam Search	15.2	56.1	**47.1**	13.0	45.8	**59.2**	11.8	43.9	**63.7**
PMI	8.7	32.7	42.7	7.1	26.2	53.7	6.8	24.1	57.7
M3ID	9.3	33.6	44.8	7.5	27.5	56.4	6.9	26.3	60.6
OPERA	14.7	49.0	46.8	11.8	39.2	58.9	11.5	36.2	63.4
PAI	**7.1**	**21.8**	42.0	6.2	19.5	52.7	**5.6**	18.6	56.8
VisTexAttnAgg	9.0	32.8	43.4	7.3	26.7	54.5	6.6	25.7	58.8
VADE	7.2	22.9	42.7	**5.9**	**18.7**	53.7	**5.6**	**17.9**	57.7

Table 12: Performance comparison of hallucination mitigation methods on MS COCO (Lin et al., 2015) validation set.

Results. Table 12 reports the extended results to other VLM models on 500 randomly sampled images from MS COCO dataset. VADE performs significantly better than most of the baselines including VisTexAttnAgg across models, while achieving marginal improvement over PAI. Table 13 shows a case where VADE proves its ability to eliminate hallucinations.

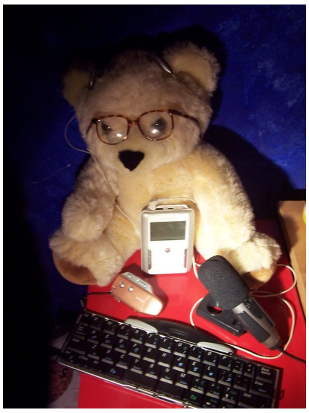
Hallucination Mitigation Example	
	
Prompt	Please describe this image in detail.
BeamSearch	The image features a white teddy bear sitting on a red desk, surrounded by various items. The teddy bear is wearing glasses and appears to be holding a cell phone. A computer keyboard is placed on the desk, along with a mouse nearby. There is also a remote control on the desk, and a book can be seen close to the teddy bear. The scene creates a playful and cozy atmosphere, as if the teddy bear is a part of a home office setup.
VADE	The image features a white teddy bear sitting on a red desk, surrounded by various items. The teddy bear is wearing glasses and is listening to an ipod. A computer keyboard and a microphone are placed on the desk. The scene creates a playful and cozy atmosphere, as if the teddy bear is a part of a home office setup.

Table 13: VADE's performance on reducing hallucinations (ID: 42889). Hallucination segments are highlighted in red.

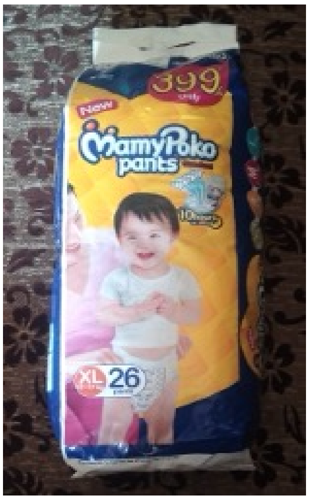
Hallucination Mitigation Example	
	
Prompt	Please describe this image in detail.
BeamSearch	The image shows a pack of Huggies baby wipes. The packaging is yellow and red with a picture of a baby on it. The wipes are in a roll and are folded. The wipes are white and have a blue stripe on them. The wipes are made by Kimberly-Clark.
VADE	The image shows a MamyPoko pants-style diapers package containing 26 diapers. The packaging is yellow and red with a picture of a baby on it. The package is placed on a brown table.

Table 14: VADE's performance on reducing hallucinations (ID: 42889). Hallucination segments are highlighted in red.