AMEND: ADAPTIVE MODULATION VIA ENHANCED NAVIGATION DECODING FOR MITIGATING HALLUCINATIONS IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) have demonstrated extraordinary capabilities in cross-modal tasks. However, they frequently suffer from hallucinations, generating descriptions that are inconsistent with the visual inputs. Existing training-free attention intervention methods lack dynamic adjustment, resulting in poor visual grounding or instruction neglect, while indiscriminately enhancing all visual tokens amplifies background noise. To address this, we propose AMEND, a novel training-free approach that mitigates hallucinations by dynamically recalibrating visual attention. Specifically, this adjustment is performed from two aspects: (1) Uncertainty-Aware Adaptive Scaling, which leverages the entropy of output token logits to estimate generation uncertainty and adaptively adjusts the visual attention enhancement coefficient; and (2) Expert-Layer Attention-Driven Token Extraction, which analyzes attention weights across vision-dominant expert layers to selectively identify the most critical and question-relevant visual tokens, rather than indiscriminately amplifying all image tokens. Finally, using the dynamically enhancement coefficient, we amplify attention to the selected key visual tokens while suppressing attention to irrelevant visual tokens and system prompts. Extensive experiments on standard benchmarks demonstrate that AMEND significantly alleviates hallucinations and outperforms existing baselines. The code is available at `https://anonymous.4open.science/r/AMEND-7576/`

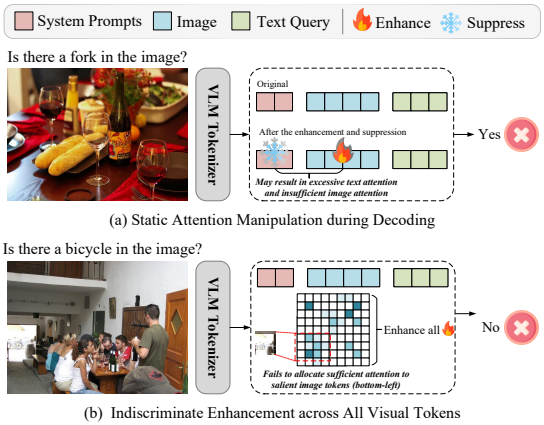
1 INTRODUCTION

By integrating visual and textual modalities with the knowledge and generative capabilities of foundational Large Language Models (LLMs), LVLMs have demonstrated strong performance across a wide range of downstream tasks, including image captioning, visual question answering (VQA), and intricate multi-turn visual dialogues (Yin et al., 2024; Liang et al., 2024; Geng et al., 2024; Li et al., 2024; Zhang et al., 2024). However, despite their impressive capabilities, LVLMs still suffer from **hallucination** issue (Liu et al., 2024a; Bai et al., 2024). Specifically, hallucinations commonly appear as the fabrication of non-existent objects or the incorrect description of object attributes and relationships, such as color and position (Rohrbach et al., 2018; Guan et al., 2024). These issues undermine the reliability of LVLMs and hinder their deployment in safety-critical applications, such as autonomous driving and medical visual analysis (Cui et al., 2024; Khan et al., 2025).

To mitigate hallucinations in LVLMs, researchers have proposed a variety of strategies, which can be broadly categorized into training-based and training-free paradigms. Training-based methods primarily focus on data construction, fine-tuning, and reinforcement learning (Liu et al., 2023a; Yu et al., 2024; Sun et al., 2024; Zhao et al., 2025; Dai et al., 2025; Yang et al., 2025; Wu et al., 2025). However, these methods rely on large amounts of high-quality training data, the collection and annotation of which are often costly and time-consuming. Conversely, training-free methods provide a lightweight alternative (Wang et al., 2024b; Huang et al., 2024; Cao et al., 2025). Among them, contrastive decoding has emerged as an effective strategy (Ruan et al., 2025), mitigating hallucinations by contrasting token distributions under different visual or decoding conditions and adjusting token probabilities based on their discrepancies (Leng et al., 2024; Chen et al., 2026a; Yu et al., 2026). However, its effectiveness heavily depends on the quality of contrastive signals, which may introduce unreliable visual information or provide insufficient visual guidance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

More recently, some studies manipulate internal attention to enhance attention toward visual tokens and improve visually grounded generation (Wang et al., 2026; Ouyang et al., 2026; Tang et al., 2026). However, existing methods often rely on fixed scaling or uniformly enhance visual tokens, lacking adaptation to different visual contents and query contexts (Liu et al., 2024c; Yin et al., 2025; Ren et al., 2026; Zhang et al., 2026). As shown in Figure 1(a), a static coefficient may provide insufficient visual grounding, causing the model to over-rely on language priors and hallucinate a non-existent 'fork". Figure 1(b) further shows that uniformly amplifying all visual tokens introduces visual noise and disperses attention, leading the model to miss existing objects such as the 'bicycle".



(a) Static Attention Manipulation during Decoding

(b) Indiscriminate Enhancement across All Visual Tokens

Figure 1: Limitations of existing VLM attention-enhancement methods. (a) Static attention manipulation may weaken visual grounding. (b) Indiscriminate enhancement of visual tokens may divert attention from salient visual features.

To address these limitations, we propose Adaptive Modulation via Enhanced Navigation Decoding (AMEND), a training-free framework that dynamically calibrates visual attention during decoding to mitigate hallucinations in LVLMs. Achieving dynamic visual attention calibration is challenging due to three aspects: determining the attention enhancement coefficient, selecting relevant image tokens, and avoiding irrelevant visual amplification. Specifically, AMEND comprises three components: (1) Uncertainty-aware Adaptive Scaling dynamically adjusts the enhancement coefficient based on the entropy of the output logit distribution. (2) Expert-Layer Attention-Driven Token Extraction identifies relevant visual information by locating expert layers that are most receptive to visual features and extracting query-relevant tokens based on their attention scores. (3) Contrastive Attention Calibration avoids irrelevant visual amplification by dynamically enhancing key image tokens while suppressing irrelevant image tokens and system prompts.

In summary, the contributions of this paper are as follows:

- We propose AMEND, a lightweight, training-free decoding strategy that dynamically modulates visual attention to reduce object and attribute hallucinations in LVLMs.
- Our AMEND contains uncertainty-aware scaling mechanism and expert-layer token selection strategy, enabling context-sensitive enhancement of visual information.
- Extensive experiments on the POPE and MME-Hallucination benchmarks demonstrate that AMEND consistently outperforms competitive decoding methods, achieving up to a 5.95-point F1 improvement on POPE and a 65.00-point gain on MME-Hallucination over regular decoding. Ablation studies further validate the effectiveness of each proposed module.

2 Related Work

2.1 Hallucination in Large Vision-Language Models

By aligning powerful LLMs with visual encoders, recent advancements in LVLMs have demonstrated remarkable capabilities across a diverse range of multimodal tasks, including visual question answering, image captioning, and complex visual reasoning (Yin et al., 2024; Liu et al., 2023b; 2024b; Bai et al., 2025; Chen et al., 2023).Despite their impressive success, current LVLMs inherently suffer from hallucination (Ji et al., 2023; Liu et al., 2024a; Bai et al., 2024). This phenomenon occurs when models generate responses that are inconsistent with visual inputs, including object, attribute, and relationship hallucinations. A key cause is insufficient cross-modal alignment, which allows the strong language prior of LLMs to dominate visual evidence during generation. (Li et al., 2023a; Huang et al., 2025; Zhibo et al., 2023).

2.2 Approaches to Hallucination Mitigation

To address the hallucination problem in VLMs, various hallucination mitigation methods have been proposed. In particular, decoding strategies used in the inference process stand out due to their lightweight and training-free nature (Leng et al., 2024; Wang et al., 2024b; Huang et al., 2024; Zhang et al., 2025). For example, Visual Contrastive Decoding (VCD) (Leng et al., 2024) compares the original output with the interference output suppress language priors and mitigate hallucinations. However, the injected noise may not accurately reflect the actual hallucination patterns. Alternatively, VAF (Yin et al., 2025) alleviates hallucinations by enhancing image attention and suppressing attention to system prompts. However, such approaches cannot dynamically adjust the coefficient of attention enhancement based on the model's reasoning, nor can they ensure that the model focuses on important visual information, which can cause the model's attention to be distracted by irrelevant visual noise. To address this, our AMEND mitigates hallucinations by dynamically adjusting the attention weights based on the entropy of the model's logit outputs and filtering the image tokens most relevant to the input query through expert layers for enhancement.

3 Preliminary

3.1 Cross-Modal Attention in Vision-Language Models

LVLMs achieve cross-modal alignment between visual and linguistic modalities through cross-modal attention mechanisms (Alayrac et al., 2022; Team, 2023). Given an an image-text pair, the visual encoder and text embedder extract their respective feature representations. Let $\mathbf{X}_v \in \mathbb{R}^{N_v \times d}$ denote the sequence of visual patch features and $\mathbf{X}_l \in \mathbb{R}^{N_l \times d}$ denote the linguistic token features, where N_v and N_l represent the lengths of the image and text tokens, and d is the hidden dimension. To achieve cross-modal alignment, the attention mechanism projects these features into Queries $\mathbf{Q}$, Keys $\mathbf{K}$, and Values $\mathbf{V}$ as follows:

$$\mathbf{Q} = \mathbf{X}_l \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}_v \mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}_v \mathbf{W}_V, \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices. The attention operation is defined as follows:

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

The attention matrix determines how much attention each language token allocates to each visual token. The accurate and correct computation of this matrix is crucial for generating responses based on image input.

3.2 Inaccurate Visual Attention and Hallucination

The response process of LVLMs is token-by-token. At each decoding step t, the next token y_t is predicted based on the textual history $y_{<t}$ and the attention-weighted visual context. However, when the attention mechanism fails to assign sufficient weight $\alpha_{t,i}$ to the crucial visual patches containing the relevant object which is denoted as $\mathcal{I}_{obj}$, the model exhibits visual misalignment, formalized as:

$$P(y_t \mid y_{<t}, \mathcal{I}) = f_\theta\left(y_{<t}, \sum_{i=1}^{N_v} \alpha_{t,i}\mathbf{V}_i\right) \xrightarrow{\sum_{i \in \mathcal{I}_{obj}} \alpha_{t,i} \to 0} P_{LLM}(y_t \mid y_{<t}), \tag{3}$$

where $\mathbf{V}_i$ is the i-th visual patch feature and f_θ represents the multimodal decoder. This equation illustrates that without reliable visual anchors, VLMs rely on the LLM backbone's parameterized memory and language priors. Consequently, the model relies on statistical correlations in the training corpus rather than visual evidence, resulting in object hallucinations. Therefore, improving visual attention accuracy is essential for maintaining faithful visual grounding.

3.3 Confidence and Visual Grounding

To mitigate hallucinations, it is essential to dynamically identify visual misalignment and adjust visual focus. The predictive uncertainty during the autoregressive decoding step serves as a reliable indicator for visual misalignment (Wang et al., 2024a; Chen et al., 2026b; Shang et al., 2025).

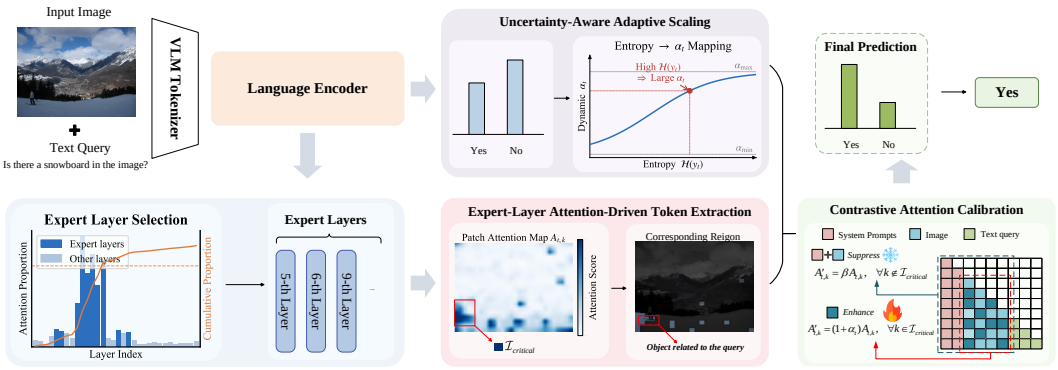


Figure 3: The overall framework of AMEND. Our training-free method mitigates hallucinations by dynamically anchoring attention to critical visual evidence. Specifically, it evaluates output uncertainty to adaptively scale the required visual reliance, and extracts query-relevant image tokens from visually sensitive expert layers. Finally, a contrastive attention calibration mechanism amplifies these critical visual tokens while suppressing irrelevant image tokens and text prompts.

Specifically, the uncertainty of generated token y_t is quantified using the Shannon entropy of its output probability:

$$\mathcal{H}(y_t) = -\sum_{v \in \mathcal{V}} P(v \mid y_{<t}, \mathbf{c}_t) \log P(v \mid y_{<t}, \mathbf{c}_t), \tag{4}$$

where $\mathcal{V}$ is the vocabulary, v is a candidate token, and $\mathbf{c}_t$ is the attention-weighted visual context at decoding step t.

To empirically validate this, we analyze the token prediction entropy on the POPE-Adversarial (Li et al., 2023a) benchmark using LLaVa-v1.5-7B (Liu et al., 2024b). As shown in Figure 2, hallucinated outputs exhibit significantly higher mean entropy than correct predictions, indicating increased uncertainty when the VLM lacks reliable visual evidence. Furthermore, False Positives yield the highest entropy among all categories. A large $\mathcal{H}(y_t)$ indicates the need for dynamically enhancing visual attention at the current generation step. Empirical observations indicate that specific transformer layers exhibit a significantly higher propensity for cross-modal alignment. The image patches that receive the highest attention in these specific layers represent the salient objects requiring extra focus, while the remaining tokens can be largely regarded as less relevant visual tokens (Chen et al., 2025). This intrinsic mechanism lays the theoretical foundation for our targeted attention enhancement strategy.

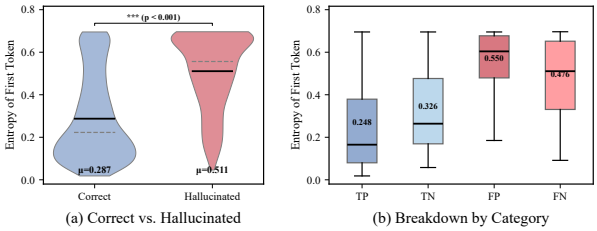


Figure 2: Distribution of token prediction entropy for correct and hallucinated responses on POPE-Adversarial. (a) Hallucinated outputs exhibit significantly higher entropy than correct ones (Welch's t-test, $p < 0.001$), suggesting uncertainty under insufficient visual evidence. (b) FP shows the highest mean entropy among the four categories, indicating that object hallucination is associated with higher uncertainty.

4 Method

4.1 Uncertainty-Aware Adaptive Scaling

To strike a dynamic balance in visual attention, we propose an Uncertainty-Aware Adaptive Scaling mechanism. This mechanism prevents excessive visual attention which disrupts linguistic balance when the model is confident, while ensuring sufficient enhancement to mitigate hallucinations when the model is uncertain. As established in our preliminary analysis, the predictive uncertainty serves

as an indicator of visual misalignment according to Equation 4. Specifically, at each decoding step t, we first calculate the normalized uncertainty score σ_t as follows:

$$\sigma_t = \text{Sigmoid}\left(\frac{\mathcal{H}(y_t) - \mu}{s}\right), \tag{5}$$

where μ and s represent the empirical center and scale factor. Given that the theoretical maximum entropy for a typical LLM vocabulary space $\mathcal{V}$ (e.g., $|\mathcal{V}| \approx 32000$) is approximately 10.4, we empirically set the center $\mu = 3.0$ and scale $s = 1.5$. Subsequently, we linearly map σ_t to a strictly bounded dynamic enhancement coefficient α_t to ensure the visual intervention remains within a safe boundary:

$$\alpha_{\text{t}} = \alpha_{\text{min}} + (\alpha_{\text{max}} - \alpha_{\text{min}}) \cdot \sigma_t. \tag{6}$$

Through this design, enhancement coefficient dynamically fluctuates within an appropriate range. During confident text generation, α_t maintains the original cross-modal attention weights. Conversely, during highly uncertain steps that are prone to produce hallucinations, α_t will increase, causing the model to pay more attention to the input visual information.

4.2 Expert-Layer Attention-Driven Token Extraction

The image itself contains redundant background information, and not all information is useful to answer the input question. To achieve query-relevant visual grounding and prevent unimportant tokens from distracting the model or exacerbating hallucinations, we propose an Expert-Layer Attention-Driven Token Extraction strategy. This component explicitly filters the subset of visual tokens most aligned with the current language context.

In the architecture of VLMs, not all transformer layers contribute equally to cross-modal alignment (He et al., 2025; Fan et al., 2025; Liu et al., 2025). Specific intermediate layers, which we denote as expert layers $\mathcal{L}_{expert}$, exhibit a pronounced sensitivity to semantic visual features. We leverage the inherent attention weights from these expert layers to evaluate the semantic importance of each visual token at the current decoding step t. Let $A_{t,i}^{(l,h)}$ denote the cross-attention weight assigned to the i-th visual token by the h-th attention head in the l-th layer. We directly define the total visual attention A_l for layer l by averaging these weights across all H attention heads and then summing them over all N_v visual tokens.

To evaluate the relative importance of each layer, we normalize the total attention A_l across all L layers, and we sort these normalized scores in descending order. We define the expert layers $\mathcal{L}_{\text{expert}}$ by selecting the minimum k layers whose cumulative normalized attention reaches a predefined threshold γ as follows:

$$\sum_{m=1}^{k} \frac{A_{(m)}}{\sum_{n=1}^{L} A_n} \geq \gamma, \tag{7}$$

where $A_{(m)}$ denotes the m-th largest total visual attention score among all L layers.

Once the k expert layers are determined, we leverage them to evaluate the semantic importance of individual visual tokens. The final attention score s_i for each image token i is calculated by averaging its head-averaged attention scores across the k expert layers:

$$s_i = \frac{1}{k} \sum_{l \in \mathcal{L}_{\text{expert}}} \left(\frac{1}{H} \sum_{h=1}^{H} A_{t,i}^{(l,h)}\right). \tag{8}$$

Instead of a rigid top-K selection, we employ a dynamic thresholding mechanism to separate critical visual tokens from redundant ones. The dynamic threshold τ is defined as follows:

$$\tau = \mu_s + \lambda \cdot \sigma_s, \tag{9}$$

where μ_s and σ_s are the mean and standard deviation of the final attention scores $\{s_i\}_{i=1}^{N_v}$ respectively, and λ is a control hyperparameter. Based on this threshold, we filter the visual tokens to form the critical visual subset $\mathcal{I}_{\text{critical}}$, which encapsulates the most informative visual anchors $\mathcal{I}_{\text{obj}}$ for the current generation step:

$$\mathcal{I}_{\text{critical}} = \{i \mid s_i \geq \tau, i \in \{1, \ldots, N_v\}\}. \tag{10}$$

By dynamically locating the expert layers and filtering the most attended visual tokens with them, we strive to ensure that subsequent attention enhancement accurately targets core visual information.

4.3 CONTRASTIVE ATTENTION CALIBRATION

Having derived the uncertainty-aware scaling factor α_t and identified the critical visual tokens $\mathcal{I}_{critical}$, the final stage of AMEND involves recalibrating the attention distribution. To effectively bridge the cross-modal gap and mitigate over-reliance on language priors, we introduce a contrastive modulation mechanism. This mechanism selectively amplifies the attention for the selected important image tokens while suppressing attention for other image tokens and system prompts.

Formally, let $A_{t,k}$ denote the original attention weight assigned to the k-th context token at the t-th decoding step. For the tokens within the critical visual subset $\mathcal{I}_{critical}$, we significantly amplify their attention weights using the dynamically calculated enhancement coefficient α_t. Simultaneously, we introduce a suppression coefficient, denoted as β, to scale down the attention weights for all unselected visual tokens and system prompts as follows:

$$A'_{t,k} = \begin{cases} (1+\alpha_t)A_{t,k}, & \text{if } k \in \mathcal{I}_{critical}, \\ \beta A_{t,k}, & \text{if } k \notin \mathcal{I}_{critical}. \end{cases} \tag{11}$$

By dynamically enhancing the attention weights of key image tokens and suppressing attention to system prompts, our method enables the model to focus on core visual evidence when encountering high prediction uncertainty, while balancing attention between text and images during certain predictions. This dual-action regulation effectively corrects inaccurate attention allocation, thereby effectively mitigating object hallucinations.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Baselines. To comprehensively evaluate the effectiveness of hallucination mitigation, we compare our method with regular decoding and five representative state-of-the-art decoding strategies, including **VCD** (Leng et al., 2024), **ICD** (Wang et al., 2024b), **ASCD** (Wang et al., 2026), **PAI** (Liu et al., 2024c), **Self-Aug** (Im et al., 2026), and **VAF** (Yin et al., 2025).

Models. We conduct experiments on three representative LVLMs to evaluate our method across different architectures and parameter scales. We select **LLaVA-v1.5-7B** and **LLaVA-v1.5-13B** (Liu et al., 2024b), both of which utilize a CLIP-ViT-L/14-336px visual encoder and a Vicuna-v1.5 language backbone. Additionally, we evaluate the methods on **Qwen2.5-VL-7B** (Bai et al., 2025), which features native dynamic resolution and advanced modality interaction mechanisms.

Benchmarks and Evaluation Metrics. We follow previous work and conduct experiments on discriminative and comprehensive benchmarks to evaluate hallucination, focusing on both object-level hallucinations and fine-grained perception capabilities. Specifically, **POPE** (Li et al., 2023b) is employed to assess object existence hallucinations. We conduct evaluations across three distinct subsets: COCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson & Manning, 2019), each including Random, Popular, and Adversarial settings. Following the standard evaluation protocol of POPE, we report Accuracy, Precision, Recall, and F1-score. **MME-Hallucination** (Fu et al., 2023) is utilized to comprehensively measure the model's object-level and attribute-level hallucinations through four sub-tasks, including existence, count, position and color.

Implementation Details. Following empirical observations (Yin et al., 2025), we set the enhancement bounds to $\alpha_{\text{min}} = 0.0$ and $\alpha_{\text{max}} = 0.25$, and set β to 0.9. For the hyperparameter settings in our method, although the optimal hyperparameters may vary slightly depending on the data distribution, we set $\gamma = 0.6$, $\lambda = 0.1$. The density of visual information and the required reasoning depth vary across datasets and tasks. Therefore, tuning the layer selection and token retention ratio per task-type allows the model to optimally allocate attention. The optimal layers for dynamic attention enhancement depend heavily on model architecture. Specifically, we modulate layers 8 to 13 for LLaVA-v1.5, whereas Qwen2.5-VL requires enhancing a broader span across layers 7 to 19. Details are provided in Appendix A.1. All experiments are conducted on a single NVIDIA A100 GPU.

Table 1: Average POPE results across COCO, A-OKVQA, and GQA under Random, Popular, and Adversarial settings. Best results are bolded and second-best results are underlined.

Setup	Method	LLaVA-v1.5-7B			LLaVA-v1.5-13B			Qwen2.5-VL-7B		
		Acc.↑	Prec.↑	F1↑	Acc.↑	Prec.↑	F1↑	Acc.↑	Prec.↑	F1↑
Random	Regular	84.46	89.76	83.35	84.77	89.65	83.76	88.49	<u>96.83</u>	87.31
	VCD	87.07	87.35	87.06	86.99	89.21	86.64	89.27	95.89	88.39
	VAF	<u>88.91</u>	87.72	<u>89.13</u>	**89.54**	89.75	**89.55**	90.18	96.40	89.44
	ASCD	86.47	82.44	87.36	80.84	73.63	83.58	86.22	**97.85**	84.26
	PAI	87.11	83.42	87.90	86.93	82.22	87.96	<u>90.32</u>	96.65	<u>89.59</u>
	ICD	84.84	87.91	84.24	85.46	89.97	84.94	89.50	96.39	88.63
	Self-Aug	87.74	**92.81**	86.94	87.59	**93.11**	86.73	89.87	95.84	89.11
	AMEND	**89.57**	<u>91.69</u>	**89.30**	<u>89.53</u>	<u>92.51</u>	<u>89.16</u>	**91.46**	95.83	**91.00**
Popular	Regular	80.94	<u>83.34</u>	80.36	82.39	85.22	81.70	85.79	90.99	84.87
	VCD	82.66	80.49	83.44	84.92	85.50	84.84	86.75	90.74	86.11
	VAF	82.44	78.57	<u>83.96</u>	85.21	82.84	<u>85.88</u>	87.65	91.27	87.12
	ASCD	79.05	73.58	81.88	75.74	68.51	80.13	84.31	**93.39**	82.49
	PAI	80.59	75.23	82.95	80.19	74.01	82.88	<u>87.89</u>	91.71	<u>87.33</u>
	ICD	80.95	81.31	81.07	82.13	83.48	81.87	87.30	<u>91.83</u>	86.60
	Self-Aug	<u>83.31</u>	**84.64**	83.11	<u>85.68</u>	**89.20**	85.02	86.75	89.57	86.32
	AMEND	**84.39**	83.05	**84.87**	**86.36**	<u>86.84</u>	**86.36**	**88.82**	90.60	**88.49**
Adversarial	Regular	76.85	<u>76.84</u>	77.10	78.41	78.77	78.44	82.91	85.83	82.37
	VCD	77.75	73.73	79.69	80.12	77.99	80.96	83.63	85.05	83.43
	VAF	76.87	71.61	<u>79.92</u>	79.33	74.96	<u>81.36</u>	<u>84.70</u>	**86.01**	<u>84.60</u>
	ASCD	71.70	65.85	76.90	68.69	62.31	75.72	81.85	88.37	80.34
	PAI	72.93	66.91	77.65	73.02	66.77	78.03	84.64	<u>85.99</u>	84.52
	ICD	76.39	74.66	77.48	78.00	77.06	78.67	83.92	85.77	83.67
	Self-Aug	<u>79.02</u>	**78.24**	79.64	<u>81.08</u>	**81.34**	81.17	83.57	84.44	83.58
	AMEND	**80.10**	76.77	**81.46**	**81.14**	<u>78.85</u>	**82.12**	**84.96**	84.33	**85.28**

5.2 Results and Discussions

Results on POPE. As shown in Table 1, we evaluate AMEND against various decoding baselines on the POPE benchmark under three settings: Random, Popular, and Adversarial. AMEND consistently achieves the highest F1-score across all VLMs and evaluation settings, demonstrating its effectiveness in mitigating object hallucinations. Compared with regular decoding, AMEND obtains substantial improvements across all models. For example, on LLaVA-v1.5-7B, AMEND improves the F1-score by 5.95%, 4.51%, and 4.36% under the three settings, respectively. Moreover, AMEND maintains strong performance under the challenging Adversarial setting. It achieves F1-scores of 81.46%, 82.12%, and 85.28% on LLaVA-v1.5-7B, LLaVA-v1.5-13B, and Qwen2.5-VL-7B, respectively, outperforming the competitive Self-Aug baseline by 1.82%, 0.95%, and 1.70%. These results demonstrate that AMEND effectively mitigates hallucinations by dynamically enhancing informative visual tokens and improving the alignment between generated responses and visual content. Detailed POPE results on each dataset are provided in Appendix A.3.

Results on MME-Hallucination. Table 2 presents the results on the MME-Hallucination benchmark. AMEND achieves strong overall performance across all three evaluated VLMs, obtaining the highest total score on LLaVA-v1.5-7B and Qwen2.5-VL-7B and the second-highest score on LLaVA-v1.5-13B. Specifically, compared with regular decoding, AMEND improves the total score by 11.66, 65.00, and 62.09 points across the three evaluated VLMs, respectively. Beyond overall performance, AMEND shows consistent improvements across object-level and attribute-level hallu-

Table 2: Results on the MME-Hallucination benchmark. We report MME scores on two hallucination dimensions: object-level (Existence and Count) and attribute-level (Position and Color), together with the Total Score. The experiments compare AMEND with various decoding strategies across three VLMs. The best performances within each setting are bolded, and the second-best performances are underlined.

Model	Method	Object-level		Attribute-level		Total Score
		Existence	Count	Position	Color	
LLaVA-v1.5-7B	Regular	185.00	158.33	141.67	**175.00**	660.00
	VCD	190.00	153.33	125.00	165.00	633.33
	VAF	190.00	153.33	145.00	155.00	643.33
	ASCD	50.00	140.00	**168.33**	110.00	468.33
	PAI	**195.00**	153.33	133.33	155.00	636.66
	ICD	185.00	133.33	116.67	131.67	566.67
	Self-Aug	180.00	146.67	123.33	155.00	605.00
	AMEND	190.00	**163.33**	148.33	170.00	**671.66**
LLaVA-v1.5-13B	Regular	175.00	131.67	110.00	145.00	561.67
	VCD	180.00	140.00	**141.67**	160.00	621.67
	VAF	188.33	**150.00**	123.33	160.00	621.66
	ASCD	**200.00**	110.00	150.00	150.00	618.34
	PAI	190.00	133.33	120.00	155.00	598.33
	ICD	175.00	118.33	126.67	135.00	555.00
	Self-Aug	195.00	146.67	128.33	165.00	**635.00**
	AMEND	190.00	145.00	121.67	**170.00**	626.67
Qwen2.5-VL-7B	Regular	174.58	138.33	143.33	185.00	641.24
	VCD	190.00	143.33	160.00	185.00	678.33
	VAF	190.00	148.33	153.33	190.00	681.66
	ASCD	175.00	98.33	106.67	173.33	553.33
	PAI	195.00	**155.00**	140.00	190.00	680.00
	ICD	190.00	143.33	153.33	190.00	676.66
	Self-Aug	195.00	140.00	**175.00**	185.00	695.00
	AMEND	**200.00**	145.00	163.33	**195.00**	**703.33**

Table 3: Component ablation on POPE-Adversarial.

Dynamic Scaling	Token Selection	MSCOCO		A-OKVQA		GQA	
		Acc.	F1	Acc.	F1	Acc.	F1
		82.34	83.07	73.60	77.74	74.67	78.69
✓		83.10	83.29	75.43	78.73	76.70	79.80
	✓	82.40	82.91	77.17	79.63	79.17	81.10
✓	✓	**83.23**	**83.31**	**77.53**	**79.78**	**79.53**	**81.30**

Table 4: Inference efficiency on POPE.

Method	F1	T (s)	Lat. (s)	Mem. (GB)
Regular	78.40	635.01	0.17	14.04
VCD	81.45	2506.59	0.66	14.87
ASCD	79.13	10944.07	3.65	15.12
VAF	83.07	443.73	0.13	13.95
AMEND	83.31	980.72	0.31	15.38
Self-Aug	81.50	7124.21	2.37	14.52

cination dimensions. For example, on Qwen2.5-VL-7B, AMEND achieves the best scores on object existence (200.00) and color (195.00), together with a competitive position score of 163.33, resulting in the highest total score of 703.33. These results demonstrate that dynamically identifying expert layers and enhancing relevant visual tokens improves visual grounding while reducing interference from irrelevant visual context, thereby effectively mitigating hallucinations.

Robustness Analysis. To thoroughly evaluate AMEND, we conduct experiments across varying model sizes and architectures on both the POPE and MME benchmarks. Regarding model scaling, AMEND remains effective as model capacity increases, improving LLaVA-v1.5-13B by 65.00 points over regular decoding on MME. Furthermore, AMEND generalizes well to different architectures such as Qwen2.5-VL-7B, achieving state-of-the-art performance among the compared methods on POPE and a total MME score of 703.33. These consistent improvements demonstrate the robustness and generalizability of AMEND across different model sizes and architectures.

5.3 Ablation Study

We conduct ablation studies to demonstrate the contributions of uncertainty-aware adaptive scaling and expert-layer visual token extraction, which respectively determine the attention enhancement coefficient and the specific image tokens to be enhanced. To rigorously verify that our method effectively counteracts language priors rather than relying on statistical shortcuts, we conduct ablation experiments using LLaVA-v1.5-7B on three datasets (MSCOCO, A-OKVQA, GQA) under the demanding adversarial setting of the POPE benchmark.

As shown in Table 3, all proposed components contribute significantly to mitigating hallucinations, with their combination achieving the optimal performance. Specifically, integrating only the Uncertainty-Aware Adaptive Scaling (Dynamic Scaling) consistently improves accuracy across all datasets by counteracting language priors during high-uncertainty generation. Meanwhile, applying Expert-Layer Visual Token Extraction (Token Selection) alone yields substantial gains, such as a 4.50% accuracy increase on GQA dataset, proving that selectively enhancing question-relevant visual cues is crucial to avoid the noise introduced by indiscriminate attention. Ultimately, combining both modules to perform Contrastive Attention Calibration maximizes this synergistic effect, achieving the best overall results and validating the robustness of our training-free framework under adversarial conditions. Ablation studies on other hyperparameters are provided in Appendix A.4.

5.4 Inference Efficiency Analysis

To evaluate the computational efficiency of AMEND during inference, we conduct experiments on the COCO-Adversarial subset of the POPE benchmark using the LLaVA-v1.5-7B model. We compare AMEND with other decoding-based hallucination mitigation methods in terms of F1 score, total inference time, latency per sample, and GPU memory consumption. The total inference time measures the overall cost of processing the evaluation set, while latency and GPU memory reflect the per-sample inference overhead and resource requirements, respectively.

As shown in Table 4, AMEND achieves the highest F1 score of 83.31 while maintaining moderate inference overhead. Compared with regular decoding, AMEND introduces additional computation due to dynamic attention modulation, increasing the latency from 0.17s to 0.31s per sample. Nevertheless, AMEND remains substantially more efficient than VCD, ASCD, and Self-Aug in both total inference time and latency. Although VAF achieves lower inference latency and memory consumption, its hallucination mitigation performance is inferior to AMEND. AMEND requires slightly higher GPU memory, reaching 15.38 GB, but remains practical for inference. Overall, these results demonstrate that AMEND achieves a favorable balance between hallucination mitigation performance and computational efficiency.

6 Conclusion

In this paper, we propose AMEND, a training-free inference strategy that dynamically adjusts visual attention during decoding to mitigate hallucinations in VLMs. By enhancing informative visual tokens while suppressing irrelevant visual context, AMEND improves the alignment between generated responses and visual content. Extensive experiments on multiple VLMs and benchmarks demonstrate that AMEND consistently outperforms existing decoding-based hallucination mitigation methods. Furthermore, our efficiency analysis shows that AMEND introduces acceptable inference overhead while maintaining strong mitigation capability. Its training-free and plug-and-play design enables integration into existing VLMs without additional training costs. This work provides a practical direction for improving the reliability and faithfulness of multimodal generation systems.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with language polishing and editing of the manuscript, including improving clarity and readability, refining wording, and correcting grammatical issues. All AI-assisted text was carefully reviewed and revised by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To support the reproducibility of our work, we provide anonymous access to the implementation code through the link given at the end of the abstract. Detailed descriptions of the datasets used in our experiments are provided in Appendix A.2. Additional implementation details and hyperparameter settings are reported in Appendix A.1, while further ablation studies on the hyperparameters are presented in Appendix A.4. These materials provide the necessary information to reproduce our experimental results.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Zongsheng Cao, Yangfan He, Anran Liu, Jun Xie, Zhepeng Wang, and Feng Chen. Cofi-dec: Hallucination-resistant decoding via coarse-to-fine generative feedback in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10709–10718, 2025.

Guoqing Chen, Fu Zhang, Bingqian Liu, Chenglong Lu, and Jingwei Cheng. Vcgd: Visual clue guided decoding with caption model for mitigating hallucination in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20041–20049, 2026a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Xinlong Chen, Yuanxing Zhang, Qiang Liu, Junfei Wu, Fuzheng Zhang, and Tieniu Tan. Mixture of decoding: An attention-inspired adaptive decoding strategy to mitigate hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8525–8542, 2025.

Zhiyuan Chen, Yuecong Min, Jie Zhang, Bei Yan, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. A survey of multimodal hallucination evaluation and detection. *International Journal of Computer Vision*, 134(3):131, 2026b.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 958–979, 2024.

Ziyun Dai, Xiaoqiang Li, Shaohua Zhang, Yuanchen Wu, and Jide Li. See different, think better: Visual variations mitigating hallucinations in lvlms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3310–3319, 2025.

Yingqi Fan, Anhao Zhao, Jinlan Fu, Junlong Tong, Hui Su, Yijie Pan, Wei Zhang, and Xiaoyu Shen. Visipruner: Decoding discontinuous cross-modal dynamics for efficient multimodal llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18896–18913, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 12709–12720, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Zoe Wanying He, Sean Trott, and Meenakshi Khosla. Seeing through words, speaking through pixels: Deep representational alignment between vision and language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35645–35660, 2025.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Eun Woo Im, Muhammad Kashif Ali, and Vivek Gupta. Self-aug: Query and entropy adaptive decoding for large vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 136345–136368, 2026.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends in Computer Graphics and Vision*, 16(1-2):1–214, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd international conference on computer, artificial intelligence and control engineering*, pp. 405–409, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024b.

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

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *arXiv preprint arXiv:2510.17771*, 2025.

Shuyi Ouyang, Hongyi Wang, Gongfan Fang, Xinyin Ma, Lanfen Lin, and Xinchao Wang. Taming the phantom: Token-asymmetric filtering for hallucination mitigation in large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8206–8214, 2026.

Hairui Ren, Zixuan Wang, Yibo Yang, He Zhao, Fan Tang, Dandan Guo, and Yi Chang. Imitating the truth: Attention-aware truth-guided enhancement for hallucination mitigation in large vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 43994–44020, 2026.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Wenzhen Yuan, Ting Liu, and Yuzhuo Fu. Mpicd: Multi-path information contrastive decoding for mitigating hallucinations in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4251–4260, 2025.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Yuying Shang, Xinyi Zeng, Yutao Zhu, Xiao Yang, Zhengwei Fang, Jingyuan Zhang, Jiawei Chen, Zinan Liu, and Yu Tian. From pixels to tokens: Revisiting object hallucinations in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10496–10505, 2025.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Lexiang Tang, Xianwei Zhuang, Bang Yang, Zhiyuan Hu, Hongxiang Li, Lu Ma, Jinghan Ru, and Yuexian Zou. Not all tokens and heads are equally important: Dual-level attention intervention for hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9439–9447, 2026.

Qwen Team. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Jiaqi Wang, Yifei Gao, and Jitao Sang. Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding. *arXiv preprint arXiv:2411.15839*, 2024a.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, 2024b.

Yujun Wang, Jinhe Bi, Soren Pirk, Yunpu Ma, et al. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10306–10314, 2026.

Yuanchen Wu, Lu Zhang, Hang Yao, Junlong Du, Ke Yan, Shouhong Ding, Yunsheng Wu, and Xiaoqiang Li. Antidote: A unified framework for mitigating lvlm hallucinations in counterfactual presupposition and object perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14646–14656, 2025.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10610–10620, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14625–14634, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Liu Yu, Zhonghao Chen, Ping Kuang, Zhikun Feng, Fan Zhou, Lan Wang, and Gillian Dobbie. Causally-grounded dual-path attention intervention for object hallucination mitigation in lvlms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36021–36029, 2026.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12944–12953, 2024.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2502.06130*, 2025.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, et al. Hallucination begins where saliency drops. *arXiv preprint arXiv:2601.20279*, 2026.

Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. Prompt highlighter: Interactive control for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13215–13224, 2024.

Qiyan Zhao, Xiaofeng Zhang, Yiheng Li, Yun Xing, Xiaosong Yuan, Feilong Tang, Sinan Fan, Xuhang Chen, Da-Han Wang, and Xu-Yao Zhang. Mca-llava: Manhattan causal attention for reducing hallucination in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3981–3990, 2025.

Ren Zhibo, Wang Huizhen, Zhu Muhua, Wang Yichao, Xiao Tong, and Zhu Jingbo. Overcoming language priors with counterfactual inference for visual question answering. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pp. 600–610, 2023.

A Appendix

A.1 Additional Details on Experimental Setup

The attention adjustment module in our implementation was developed based on the official source code of ClearSight (Yin et al., 2025), which introduces the Visual Amplification Fusion (VAF) method. Following the attention suppression setting in ClearSight, we set the attention suppression coefficient β to 0.9, such that the attention weights of the suppressed tokens are scaled by β during decoding. For the LLaVA models, we also adopt the same attention adjustment layer configuration as ClearSight.

To determine the layers for image attention enhancement to Qwen2.5-VL-7B, we conducted a layer-wise attention analysis on 50 randomly sampled images from MSCOCO dataset. For each image, we performed a forward pass with attention output enabled and computed the attention proportion that each layer allocates to image tokens. The analysis focused on the attention weights from the last token position to all image token positions, averaged across attention heads and samples.

Figure 4 presents the layer-wise image attention distribution for Qwen2.5-VL-7B. Unlike LLaVA models where image attention concentrates in middle layers (Liu et al., 2024b; Yin et al., 2025), Qwen2.5-VL exhibits a more dispersed attention pattern. We observe multiple peak layers throughout the network, including Layer 7 with 8.1% attention, Layer 14 with 5.2%, Layer 16 with 5.3%, and Layer 19 with 7.4%. The horizontal dashed line indicates the mean attention proportion of 3.6% across all layers. Based on this distribution, we select layers 7 through 19 as the enhancement range for Qwen2.5-VL-7B. This continuous span covers 54.5% of the total image attention while maintaining computational efficiency compared to sparse layer selection. The chosen range encompasses major peak layers where vision-language interaction is most intensive, ensuring that out method can effectively modulate the attention patterns in the regions most critical for visual understanding.

A.2 Dataset Details

POPE POPE dataset (Li et al., 2023b) is a structured visual question-answering dataset designed specifically to evaluate object hallucinations in Large Vision-Language Models (LVLMs), which contains an image-question-answer pair in each sample. By asking VLMs simple Yes-or-No questions about whether specific probing objects are present in an image, the evaluation is transformed into a binary classification task. POPE dataset constructs these polling questions using both ground-truth objects and nonexistent objects, sampling the latter through three difficulty settings. Random sampling selects absent objects completely at random; Popular sampling selects the most frequently appearing objects in the entire dataset that do not exist in the current image; and Adversarial sampling identifies absent objects that most frequently co-occur with the actual ground-truth objects present in the image. Furthermore, this dataset is highly scalable and has been adapted across three visual datasets: MSCOCO ()lin2014microsoft, A-OKVQA (Schwenk et al., 2022), and GQA (Hudson & Manning, 2019).

MME-Hallucination MME dataset (Fu et al., 2023) is used for comprehensive evaluation suite designed to quantitatively measure VLMs across two primary aspects: perception and cognition abilities, containing 14 subtasks in total. To strictly avoid data leakage from public datasets, all

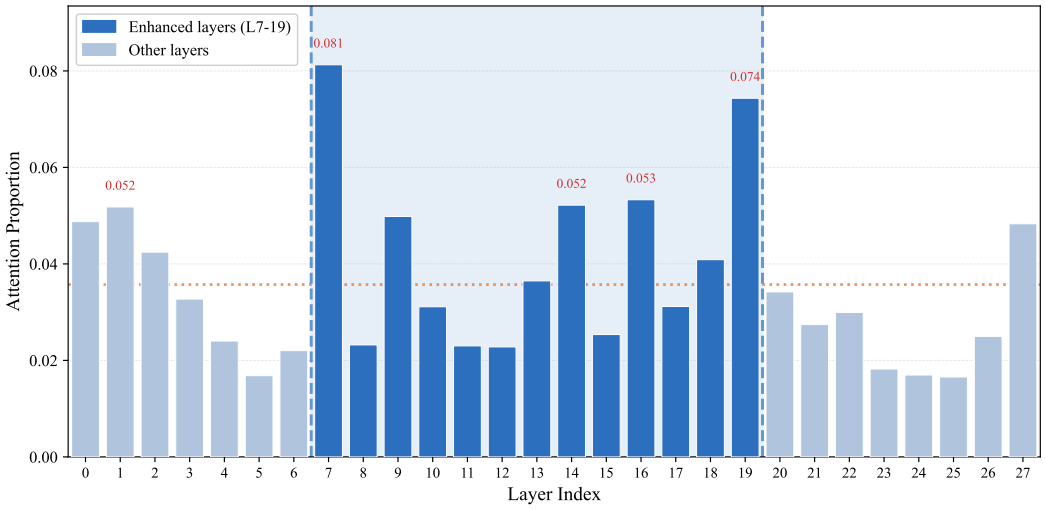


Figure 4: Layer-wise image attention distribution for Qwen2.5-VL-7B averaged over 50 samples from the MSCOCO dataset. Dark blue bars indicate the selected enhancement layers (L7–L19), while light blue bars represent other layers. Red numbers mark the top-5 peak layers with the highest attention proportions. The orange dashed line indicates the mean attention proportion across all layers.

instruction-answer pairs are manually constructed, utilizing concise instructions to avoid affecting the model's output. Within the perception aspect, MME specifically evaluate hallucination problem through four subtasks, which including existence, count, position and color.

A.3 Additional Results on POPE

The results shown in Tables 5, 6, and 7 provide detailed POPE evaluations across three datasets. These fine-grained results further validate the effectiveness of our proposed AMEND method.

Consistent Superiority Across Datasets Across different datasets, models, and evaluation settings, AMEND consistently achieves state-of-the-art or competitive performance, obtaining the best or second-best F1-score in 25 out of 27 cases. Notably, on GQA with LLaVA-v1.5-7B (Liu et al., 2024b), AMEND improves Accuracy by 4.94 points and F1-score by 5.51 points over regular decoding under the Random setting. It also outperforms VAF (Yin et al., 2025), PAI (Liu et al., 2024c), and Self-Aug (Im et al., 2026) by 1.27, 3.54, and 1.50 points in Accuracy, and by 0.63, 2.22, and 1.85 points in F1-score, respectively. Similarly, on A-OKVQA with Qwen2.5-VL-7B (Bai et al., 2025), AMEND achieves the best Accuracy of 92.27 and F1-score of 92.05 under the Random setting, outperforming Self-Aug by 1.40 and 1.60 points, respectively.

Model Scaling and Architecture Comparison The results further demonstrate that AMEND remains effective across different model scales and architectures. On both LLaVA-v1.5-7B and LLaVA-v1.5-13B, AMEND maintains strong performance across the three datasets and evaluation settings. When applied to Qwen2.5-VL-7B, AMEND achieves the best F1-score in seven of the nine settings and the second-best in the remaining two. These results indicate that AMEND generalizes well across different model architectures and scales. By identifying question-relevant visual tokens and dynamically adjusting their attention, AMEND effectively mitigates hallucinations across diverse evaluation scenarios.

A.4 Additional Ablation Studies

Effect of Token Selection Coefficients. We conduct ablation studies to investigate the impact of two critical hyperparameters introduced in the Expert-Layer Attention-Driven Token Extraction module

of our proposed AMEND framework: γ and λ. Expert layer selection threshold γ dictates the cumulative normalized attention threshold used to determine the expert layers for image token selection. A smaller γ restricts the selection to a few highly visually-sensitive layers, thereby ensuring the attention weights are derived from the most concentrated visual semantics. Conversely, a larger γ relaxes this constraint, incorporating more layers into the calculation, which may introduce noisy or redundant cross-modal attention patterns. Token filtering parameter λ is a control hyperparameter that modulates the standard deviation σ_s to establish the dynamic threshold τ. This threshold determines which visual tokens are included in the critical subset $\mathcal{I}_{\text{critical}}$. A higher λ enforces a stricter filtering criterion, retaining only the most heavily attended visual anchors ($\mathcal{I}_{\text{obj}}$) and effectively pruning background noise.

To evaluate the impact of γ and λ, we conducted ablation experiments utilizing the LLaVA-v1.5-7B model. The evaluations were performed on the COCO-Adversarial dataset split from the POPE benchmark. We report the F1-score as the primary evaluation metric to assess the model. As shown in Figure 5, setting γ to 0.6 and λ to 0.1 yields the best results and the highest F1-score. First, regarding the expert layer selection threshold γ, the results indicate that moderate values yield the most robust performance. A low threshold restricts the selection to too few layers, potentially capturing unstable or overly narrow semantic alignments, while a high threshold incorporates too many intermediate layers, which results in overly broad criteria for selecting experts. Second, concerning the token filtering parameter λ, as it increases from 0.0 towards 1.0, there is a general declining trend in the F1-score across almost all settings of γ. This phenomenon indicate that excessively high values of λ enforce overly strict filtering, aggressively discarding contextually useful visual tokens, whereas lower λ values allow the model to retain a sufficiently broad and informative subset of critical visual features, yielding the highest F1-scores.

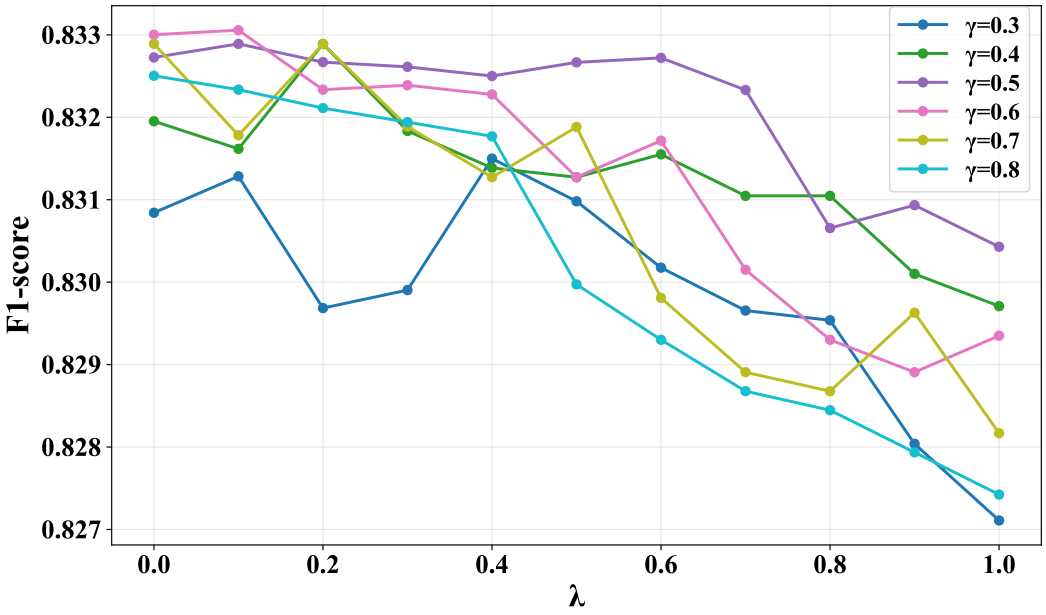


Figure 5: Ablation study of cumulative attention threshold γ and token filtering parameter λ for expert layer selection and visual token identification.

A.5 Case Study

To demonstrate AMEND's effectiveness, Figure 6 compares it with the VAF baseline, which enhances all image tokens and suppresses system prompts using fixed coefficient, and the highlights in the figure indicate the top 20% of image tokens with the highest attention. In contrast, the highlights for AMEND directly represent the specific image tokens enhanced using dynamic enhancement coefficients. In the first example, VAF focuses not only on the relevant "bird" but also on the irrelevant "cow" in the background, which leads to confusion between the colors of the cow and the bird,

ultimately leading to object hallucination. The second example exhibits a similar issue, where VAF attends to objects other than the "bottle", again causing hallucination. The examples above demonstrate that our method can dynamically enhance the image tokens most relevant to the problem and without introducing background noise, thus effectively mitigating hallucinations.

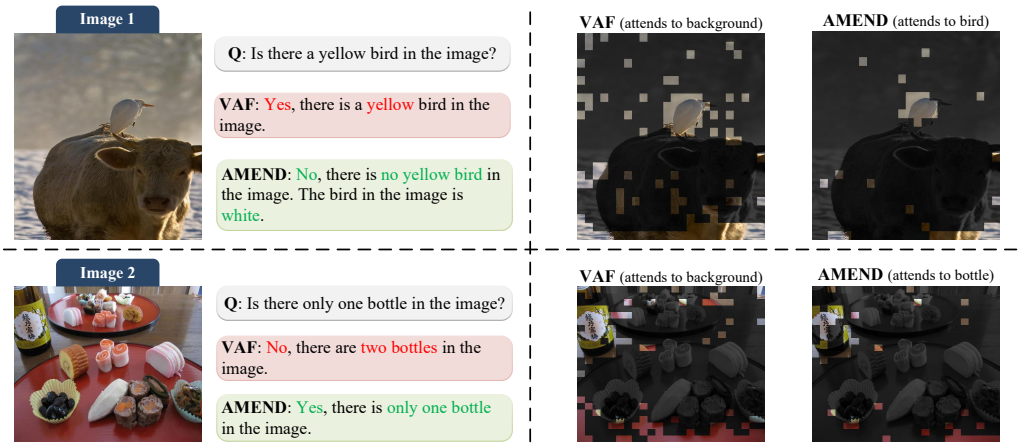


Figure 6: Case study comparing AMEND with the VAF baseline for hallucination mitigation in LVLMs. VAF enhances all image tokens, and its highlights visualize the tokens with the top 20% of global visual attention. The highlights for AMEND indicate the tokens selected for attention enhancement. Hallucinated content is marked in red, while correct content generated by AMEND is marked in green.

Table 5: POPE results on MSCOCO dataset. The best results are highlighted in bold, and the second-best results are underlined.

	Setup	Method	LLaVA-v1.5-7B			LLaVA-v1.5-13B			Qwen2.5-VL-7B		
			Acc.↑	Prec.↑	F1↑	Acc.↑	Prec.↑	F1↑	Acc.↑	Prec.↑	F1↑
MSCOCO	*Random*	Regular	83.73	92.31	81.90	84.00	91.94	82.33	86.54	98.59	84.63
		VCD	88.17	91.94	87.61	87.63	93.19	86.78	87.67	98.34	86.14
		VAF	**89.83**	92.53	**89.50**	**90.27**	94.15	89.82	88.77	98.99	87.46
		ASCD	88.47	88.31	88.49	85.23	79.61	86.51	83.97	**99.42**	81.00
		PAI	89.27	89.58	89.22	90.00	88.61	**90.18**	88.77	98.82	87.48
		ICD	85.33	92.26	84.02	85.97	93.83	84.58	88.13	98.56	86.71
		Self-Aug	86.57	**95.75**	85.07	86.90	95.71	85.50	88.07	98.64	86.61
		AMEND	89.50	94.65	88.86	89.37	**96.10**	88.53	**90.00**	98.47	**89.04**
	Popular	Regular	82.57	89.68	80.85	82.77	89.35	81.19	85.92	96.63	84.09
		VCD	85.67	87.15	85.37	85.83	89.49	85.15	86.77	96.05	85.28
		VAF	**87.33**	87.84	**87.25**	**88.03**	89.76	**87.77**	87.90	96.64	86.65
		ASCD	85.57	83.53	85.99	81.77	75.25	83.85	83.57	**98.18**	80.63
		PAI	86.07	84.24	86.43	86.20	82.59	86.92	87.83	96.55	86.85
		ICD	83.86	88.95	82.73	83.40	89.08	82.10	87.13	96.42	85.70
		Self-Aug	84.90	**91.19**	83.65	85.50	**93.26**	84.07	86.83	95.32	85.47
		AMEND	87.16	89.91	86.71	87.93	92.96	87.18	**88.73**	95.53	**87.82**
	Adversarial	Regular	79.70	83.78	78.40	80.63	84.84	79.39	84.72	94.12	82.90
		VCD	80.93	79.29	81.45	82.17	82.80	81.99	85.04	92.01	83.68
		VAF	82.34	79.75	83.07	84.20	83.14	**84.45**	86.93	94.25	85.76
		ASCD	78.87	74.16	80.74	75.60	68.53	79.51	83.10	**96.88**	80.19
		PAI	79.23	74.57	81.04	81.03	75.56	82.87	86.87	94.31	85.66
		ICD	79.86	81.64	79.29	81.80	84.92	80.95	86.10	93.92	84.74
		Self-Aug	82.53	**86.64**	81.50	83.17	**88.24**	81.97	85.93	93.33	84.62
		AMEND	**83.23**	82.95	**83.31**	**84.77**	86.80	84.33	**87.43**	92.63	**86.62**

Table 6: POPE results on A-OKVQA dataset. The best results are highlighted in bold, and the second-best results are underlined.

	Setup	Method	LLaVA-v1.5-7B			LLaVA-v1.5-13B			Qwen2.5-VL-7B		
			Acc.↑	Prec.↑	F1↑	Acc.↑	Prec.↑	F1↑	Acc.↑	Prec.↑	F1↑
A-OKVQA	*Random*	Regular	84.93	89.16	84.07	85.77	89.54	85.05	89.90	95.83	89.20
		VCD	86.27	85.10	86.49	86.63	87.92	86.40	90.38	94.99	89.87
		VAF	88.50	85.85	88.91	88.93	87.68	89.11	90.60	95.04	90.11
		ASCD	85.53	79.94	86.77	78.97	71.02	82.31	87.80	**97.01**	86.47
		PAI	85.93	80.49	87.09	85.85	79.78	87.18	90.97	95.28	90.51
		ICD	85.00	86.00	84.78	84.87	88.27	85.16	89.87	95.23	89.23
		Self-Aug	88.47	**91.75**	87.99	87.93	**92.40**	87.26	90.87	94.74	90.45
		AMEND	**89.53**	90.23	**89.44**	**89.43**	91.16	**89.21**	**92.27**	94.71	**92.05**
	Popular	Regular	80.87	81.71	80.61	83.20	84.73	82.82	86.97	90.04	86.45
		VCD	82.17	78.84	83.14	84.63	84.43	84.68	88.15	90.46	87.80
		VAF	81.70	76.20	83.44	84.27	80.41	85.20	88.47	90.69	88.14
		ASCD	78.37	71.33	81.43	74.73	66.91	79.48	86.37	93.67	85.12
		PAI	80.37	73.54	82.85	78.83	71.51	81.91	88.67	90.67	88.38
		ICD	80.40	78.79	80.93	81.87	81.91	81.85	88.30	89.11	88.18
		Self-Aug	83.30	**82.75**	83.44	85.80	**87.82**	85.41	**90.87**	**94.74**	**90.45**
		AMEND	**84.57**	81.95	**85.17**	**85.97**	85.01	**86.16**	90.33	90.44	89.98
	Adversarial	Regular	74.77	72.58	75.93	76.87	75.41	77.51	81.47	80.81	81.66
		VCD	75.33	70.21	78.11	78.47	75.00	79.86	82.33	80.66	82.80
		VAF	73.60	67.20	77.74	76.03	70.16	79.08	82.57	80.67	83.09
		ASCD	68.20	61.87	74.89	65.60	59.48	73.99	81.23	**83.39**	80.61
		PAI	69.63	63.05	75.75	69.30	62.60	75.74	82.77	80.66	83.34
		ICD	74.50	71.01	76.45	75.40	72.57	76.85	81.83	80.59	82.20
		Self-Aug	76.30	**72.95**	77.91	**79.57**	**77.39**	**80.35**	81.90	79.30	82.67
		AMEND	**77.53**	72.52	**79.78**	78.27	73.93	80.07	**83.37**	79.70	**84.33**

Table 7: POPE results on GQA dataset. The best results are highlighted in bold, and the second-best results are underlined.

	Setup	Method	LLaVA-v1.5-7B			LLaVA-v1.5-13B			Qwen2.5-VL-7B		
			Acc.↑	Prec.↑	F1↑	Acc.↑	Prec.↑	F1↑	Acc.↑	Prec.↑	F1↑
GQA	*Random*	Regular	84.73	87.81	84.09	84.53	87.48	83.90	89.02	96.07	88.11
		VCD	86.77	85.02	87.09	86.70	86.53	86.73	89.77	94.35	89.21
		VAF	88.40	84.78	88.97	89.43	87.41	89.71	91.17	95.17	90.76
		ASCD	85.40	79.08	86.83	78.33	70.26	81.93	86.90	**97.11**	85.31
		PAI	86.13	80.18	87.38	84.93	78.26	86.52	91.23	95.85	90.77
		ICD	84.20	85.48	83.91	85.53	87.80	85.09	90.50	95.37	89.96
		Self-Aug	88.17	**90.92**	87.75	87.93	**91.23**	87.43	90.67	94.14	90.28
		AMEND	**89.67**	90.20	**89.60**	**89.80**	90.28	**89.74**	**92.10**	94.32	**91.90**
	Popular	Regular	79.37	78.62	79.63	81.20	81.58	81.09	84.47	86.31	84.06
		VCD	80.13	75.48	81.80	84.30	82.58	84.70	85.34	85.72	85.26
		VAF	78.30	71.67	81.18	83.33	78.34	84.68	86.57	86.49	86.58
		ASCD	73.20	65.88	78.22	70.73	63.37	77.05	83.00	**88.31**	81.73
		PAI	75.33	67.92	79.56	75.53	67.93	79.81	87.17	87.90	87.04
		ICD	78.60	76.19	79.54	81.13	79.45	81.66	86.63	87.51	86.48
		Self-Aug	**81.73**	**79.97**	82.25	**85.73**	**86.51**	85.58	85.13	84.27	85.32
		AMEND	81.43	77.30	**82.74**	85.17	82.54	**85.74**	**87.40**	85.82	**87.67**
	Adversarial	Regular	76.07	74.17	76.97	77.73	76.07	78.42	82.55	82.56	82.55
		VCD	77.00	71.68	79.51	79.73	76.17	81.02	83.53	82.47	83.80
		VAF	74.67	67.89	78.69	77.77	71.57	80.56	**84.60**	83.10	**84.94**
		ASCD	68.03	61.53	75.07	64.87	58.91	73.66	81.23	**84.83**	80.21
		PAI	69.93	63.10	76.15	68.74	62.14	75.47	84.27	82.99	84.57
		ICD	74.80	71.33	76.70	76.80	73.70	78.22	83.83	82.81	84.08
		Self-Aug	78.23	**75.13**	79.50	**80.50**	**78.40**	81.20	82.87	80.70	83.45
		AMEND	**79.53**	74.83	**81.30**	80.37	75.81	**81.96**	84.07	80.67	84.90