PAR: Training-Free Positional Perturbation and Attention Recycling for Faithful OCR

Yao Yao[1,2], Manwen Liao[4], Weitian Zhang[3], Zuchao Li[5,*], Hai Zhao[1,2,*]
[1]AGI Institute, School of Computer Science, Shanghai Jiao Tong University, Shanghai, China
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[4]School of Computing and Data Science, The University of Hong Kong, Hong Kong, China
[5] School of Artificial Intelligence, Wuhan University, Wuhan, China

yaoyao27@sjtu.edu.cn, manwen@connect.hku.hk, weitianzhang@sjtu.edu.cn,
zcli-charlie@whu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

In high-precision scenarios, vision language models suffer from Linguistic Priors Hallucination. When processing familiar text, models tend to over-rely on internal parametric knowledge, effectively "reciting" the content rather than "reading" the image. In this paper, we first systematically investigate this phenomenon by constructing the GlitchText Probing Dataset. We discover that the model's reliance on visual grounding diminishes significantly as the generation length increases. To mitigate this, we propose PAR (Positional Perturbation and Attention Recycling), a training-free, inference-time intervention framework. PAR consists of two parts: (1) Positional Perturbation (PP) injects structured phase noise into the rotary positional embeddings; (2) Foveal Attention Recycling (FAR) detects over-confident linguistic priors and dynamically redistributes attention mass back to important visual regions. Extensive experiments across state-of-the-art models, demonstrate that PAR significantly reduces hallucination rates (reducing CER by 12%), particularly in long-context scenarios, while maintaining robust generalization on standard benchmarks. Our code is publicly available at https://github.com/Zoeyyao27/PAR-for-Faithful-OCR.

1 Introduction

By deeply integrating visual encoders with Large Language Model decoders, modern Vision-Language Models (VLMs), such as DeepSeek-OCR (Wei et al., 2025) and GPT-4o (Hurst et al., 2024), have demonstrated remarkable performance in Visual Question Answering (VQA), and general scene understanding. However, as these models are increasingly deployed in high-precision tasks like Optical Character Recognition (OCR) and document analysis, a pernicious failure mode has emerged: Linguistic Priors Hallucination. Unlike traditional object hallucination, where models detect non-existent objects in an image (Shu et al., 2025; Favero et al., 2024; Li et al., 2025a), Linguistic Priors Hallucination in OCR manifests as an over-correction of textual content driven by the model's internal language distribution. When a conflict arises between visual grounding and linguistic priors, VLMs tend to disregard the visual evidence, favoring the text sequence with the highest probability according to their pre-trained knowledge, as illustrated in Figure 1. This phenomenon is particularly pronounced when processing high-familiarity text. In the pursuit of semantic fluency, models sacrifice visual fidelity, which is a trade-off that poses a severe threat to applications requiring "What You See Is What You Get" accuracy, such as archival digitization, legal forensics, and academic collation. In these scenarios, the semantic perplexity of the output is secondary to its faithfulness to the source image.

To systematically investigate and quantify this issue, we construct the GlitchText Probing Dataset. This dataset consists of high-familiarity texts (e.g., classic Chinese poetry and famous English quotes), into which we inject visual anomalies via character glyph confusion (Visual Confusion) for Chinese and keyboard proximity interference (Typos) for English. Experiments on this benchmark confirm that existing VLMs suffer from a severe dependency on priors. Furthermore, our findings

* Corresponding author. This research was supported by the Shanghai Jiao Tong University 2030 Initiative and The Major Program of Chinese National Foundation of Social Sciences under Grant 'The Challenge and Governance of Smart Media on News Authenticity' [No. 23&ZD213].

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23258–23273

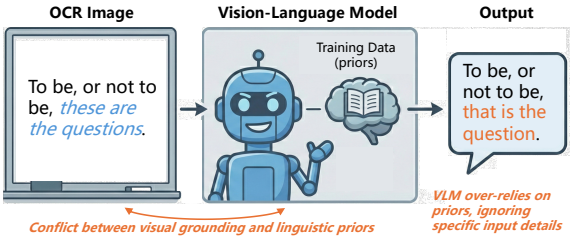

Figure 1: Linguistic Priors Hallucination: When visual input conflicts with training data priors, the model tends to ignore specific visual details and output the text it has memorized, leading to unfaithful OCR results.

align with observations in object hallucination research (Favero et al., 2024): as the generation length increases, the model's reliance on textual tokens progressively intensifies while its attention to visual tokens diminishes. This attention drift leads to a significant degradation in OCR accuracy for long sequences, as the model effectively stops "reading" and starts "reciting".

To mitigate this issue, we draw inspiration from biological vision mechanisms, specifically the foveal vision of the human eye, where visual performance degrades rapidly away from this central line of sight (Tuten and Harmening, 2021) and human visual attention operates akin to a Gaussian distribution (He et al., 2015), prioritizing the central processing of details similar to the foveal vision. Motivated by these biological insights, we propose Training-Free **P**ositional **P**erturbation and **A**ttention **R**ecycling (PAR) for faithful OCR. PAR operates on two components to mimic the human visual verification process. First, it introduces **Positional Perturbation** during the Transformer decoding process to disrupt the inertia of long-range linguistic dependencies, preventing the model from falling into the reciting mode. Second, it employs a **Foveal Attention Recycling** mechanism which detects when the model becomes overly confident in its language priors and forces it to look back at the visual features, redistributing attention using a Gaussian kernel to mimic the biological visual span. Extensive experiments demonstrate that PAR significantly reduces the hallucination rate and restores sensitivity to visual details without the need for expensive fine-tuning. Our method achieves substantial improvements on both our proposed GlitchText Probing Dataset and the document parsing benchmark OmniDocBench (Ouyang et al., 2024), proving its efficacy in balancing linguistic coherence with strict visual grounding.

2 The GlitchText Probing Dataset and Empirical Analysis

2.1 Dataset Construction

To quantitatively evaluate the conflict between visual grounding and linguistic priors, we introduce the GlitchText Probing Dataset. Unlike standard OCR benchmarks, GlitchText is specifically designed for Linguistic Priors Hallucination. The construction pipeline follows a three-stage process: familiarity-driven selection, adversarial glitches injection, and standardized visual synthesis.

Familiarity-Driven Selection The core premise of our study is that linguistic priors are most potent when the model encounters text it has memorized during training. To maximize this prior activation, we create a high-familiarity corpus. For Chinese, we utilize the *Chinese Poetry Dataset*[1], which comprises 73,281 ancient poems. For English, we select the *English Quotes Dataset* [2], which contains 2,510 quotes retrieved from goodreads quotes [3].

To ensure the selected texts strictly trigger strong priors, we employ DeepSeek-OCR (Wei et al., 2025) to calculate the Perplexity (PPL) of candidate texts. We rank the texts by PPL in ascending order (lower PPL indicates higher familiarity) and retain the top candidates. Specifically, we select the top-500 Chinese texts and the top-200 English texts. This ensures that the base model has a strong intrinsic tendency to predict the future sequence.

Adversarial Anomaly Injection To synthesize a conflict between visual grounding and linguistic priors, for each instance, we systematically inject top-20 "glitches" into high-familiarity texts. **(1) Chinese: Visual Glitch.** Utilizing the logographic nature of Chinese, we replace original characters with visually confounding counterparts (e.g., "兵" to "乒") based on visual similarity scores. **(2) English: Morphological Glitch.** For English, we introduce grammatical inconsistencies by altering tenses or pluralization (e.g., *go* → *went*) for targeted verbs and nouns, prioritizing changes that maintain high string similarity. Detailed generation pipelines, including specific toolkits and filtering algorithms, are provided in Appendix A.

Standardized Image Synthesis To isolate the impact of textual content, all texts are rendered

[1] https://github.com/javayhu/poetry
[2] https://huggingface.co/datasets/Abirate/english_quotes
[3] https://www.goodreads.com/quotes

	Chinese	English
#Instance	500	200
Max ppl	2.40	5.69
#Glitches	20	20
Average Length	180.76	146.31

Table 1: Detailed statistics of the GlitchText Probing Dataset. The symbol "#" denotes the count. Note that the average length is measured in characters for the Chinese subset and words for the English subset.

on a pure white background with black font using standardized layouts and high-resolution settings.

Table 1 summarizes the key statistics of the constructed benchmark, including the number of instances, the strict Perplexity thresholds used for familiarity filtering, and the density of injected glitches. It is worth noting that to maintain consistency in difficulty, we inject a fixed number of glitches per sample.

2.2 Experimental Setup

Based on the GlitchText Probing Dataset, we conducted a comprehensive evaluation of six representative VLMs, including open-source models (DeepSeek-OCR (Wei et al., 2025), Qwen3-VL-2B/8B (Bai et al., 2025), InternVL3.5-8B(Wang et al., 2025)) and closed-source models (GPT-4o (Hurst et al., 2024), Claude-3.5-Sonnet).

To rigorously disentangle the contribution of visual perception from linguistic prediction, we constructed a `Shuffle` variant of our dataset. In this variant, the character (Chinese) or word order (English) is randomized. This setup allows us to probe the model's pure visual grounding capability without linguistic priors.

Metrics We employ three key metrics to diagnose the failure modes:

- **Character Error Rate (CER)** $\downarrow$ A standard measure of edit distance between the prediction and the ground truth (visual content) (Morris et al., 2004).
- **Identification Rate (Ident)** $\uparrow$ The proportion of glitches that are correctly transcribed as anomalies. This reflects visual faithfulness.
- **Correction Rate (Cor)** $\downarrow$ The proportion of glitches that are corrected back to the original characters. This reflects the intensity of linguistic hallucination.

Note that Ident and Cor do not necessarily sum to 100%, as models may also omit characters or produce unrelated errors.

2.3 Empirical Analysis

As presented in Table 2, regardless of the language and model type, shuffling the text consistently leads to a sharp decline in Correction Rate. While the Cor is significantly higher in Chinese than in English, likely due to the logographic nature of Chinese where visual confusion maps directly to valid semantic tokens, the relative trend remains universal. This universally confirms that **Linguistic Priors Hallucination is a bottleneck for advanced OCR models that are capable of seeing the details but choose to ignore them**. Once the semantic flow is disrupted, the model's inclination to fix the text vanishes, allowing the visual evidence to prevail. On Shuffle, the DeepSeek-OCR's CER drops while the Iden doubles ($44.81\% \rightarrow 82.39\%$). A similar pattern in Qwen3-VL suggests that for models with strong visual encoders, linguistic priors act as a distractor in high-fidelity OCR. Conversely, generalist models like GPT-4o show increased CER on Shuffle ($0.2283 \rightarrow 0.5587$), indicating they rely on linguistic context to compensate for weaker fine-grained visual grounding.

Attention Decay Analysis To investigate the internal mechanism driving this hallucination, we analyze the dynamic dependency of the model on visual modality versus linguistic modality during the generation process. We select two architectures, DeepSeek-OCR and Qwen3-VL-8B, and extract their attention weights from the last transformer layer during inference on the GlitchText dataset.

For each generated token t, we calculate the Image-to-Text Focus Ratio (R_{focus}), defined as:

$$R_{focus}^{(t)} = \frac{\sum_{i \in \mathcal{V}} A_{t,i}}{\sum_{j \in \mathcal{L}} A_{t,j}} \tag{1}$$

where $\mathcal{V}$ and $\mathcal{L}$ represent the set of image tokens and historical text tokens, respectively, and $A_{t,k}$ denotes the attention weight assigned by token t to token k. This metric serves as a proxy for the model's instantaneous reliance on visual evidence relative to linguistic context. As illustrated in Figure 2, we observe a distinct attention decay pattern across both models. This decay suggests that the accumulation of text history creates an inertia that progressively suppresses the visual signal. In long-context OCR tasks, the model effectively shifts

Models	Chinese						English					
	Ori			Shuffle			Ori			Shuffle		
	CER↓	Iden↑	Cor↓	CER↓	Iden↑	Cor↓	CER↓	Iden↑	Cor↓	CER↓	Iden↑	Cor↓
Deepseek-OCR (Wei et al., 2025)	0.1000	44.81	49.57	0.0546	82.39	11.52	0.0016	99.31	0.69	0.0019	69.88	0.07
Qwen3-VL-2B (Bai et al., 2025)	0.0883	79.46	14.91	0.0422	90.71	3.95	0.0023	98.28	1.37	0.0035	69.25	0.21
Qwen3-VL-8B (Bai et al., 2025)	0.0729	58.68	37.87	0.0293	92.80	3.90	0.0022	98.35	1.51	0.0033	69.53	0.00
InternVL3.5-8B (Wang et al., 2025)	0.2837	63.63	34.03	1.3008	64.25	2.41	0.0047	97.18	2.20	0.0056	69.25	0.35
GPT-4o (Hurst et al., 2024)	0.2283	24.83	53.53	0.5587	30.99	6.16	0.0320	93.88	0.62	0.1890	51.59	0.55
Claude-Sonnet-4.5 [4]	0.6365	15.61	78.15	0.7297	60.51	16.58	0.0002	99.93	0.07	0.1164	57.89	0.00

Table 2: Quantitative results on the GlitchText Benchmark. Iden and Cor are reported in percentages (%). `Ori` represents the original high-familiarity texts, whereas `Shuffle` represents texts with randomized word order.

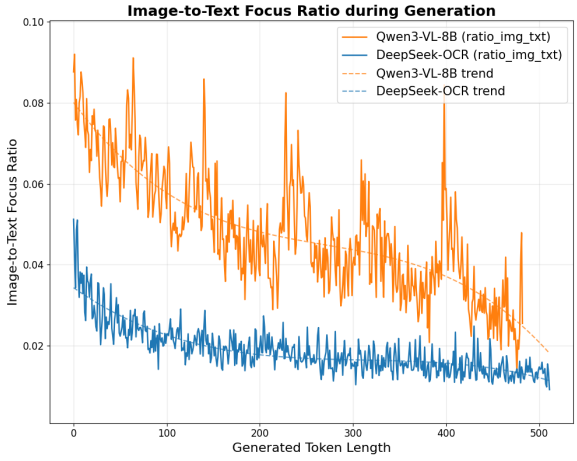

Figure 2: Visualization of the Image-to-Text Focus Ratio during generation. Both DeepSeek-OCR and Qwen3-VL-8B exhibit a significant Attention Decay: as the generated sequence lengthens, the model's attention drifts away from image tokens and becomes increasingly dominated by textual history.

from a "reading mode" to a "reciting mode", indicating that hallucinations and correction behaviors can be exacerbated in the later parts of texts.

3 PAR: Positional Perturbation and Attention Recycling

To tackle the Linguistic Priors Hallucination identified in Section 2.3, we propose **P**ositional Perturbation and **A**ttention **R**ecycling (PAR), a training-free, inference-time intervention mechanism, as illustrated in Figure 3. PAR consists of two components: Positional Perturbation (PP), and Foveal Attention Recycling (FAR).

3.1 Diagnose Layer-wise Attention Dynamics

Before applying interventions, it is crucial to diagnose where the model loses visual grounding. We utilize an attention visualization probe to map the layer-wise distribution of attention weights. For each layer l and head h, we calculate the Image-to-Text Focus Ratio R_{focus} as defined in Equation 1. The resulting heatmaps are visualized in Figure 4 (with additional architectures provided in Appendix B). From these visualizations, we identify distinct attention signatures across different models. Some models exhibit a "Decay" pattern (e.g., DeepSeek-OCR, InternVL). Shallow layers maintain high R_{focus}, whereas deep layers, which are responsible for semantic integration, show a sharp drop in visual attention. Another type of models exhibits a "Late-Fusion" pattern (e.g., Qwen3-VL). Shallow layers are text-dominated, while deeper layers re-attend to visual features to integrate multimodal information.

Adaptive Layer Selection Strategy Motivated by these observations, we adopt an Adaptive Layer Selection strategy, dynamically locates the optimal intervention windows, maximizing efficacy while minimizing degradation: (1) We apply Positional Perturbation primarily in the shallow-to-middle layers (e.g., Layers 0-5). Our experiments in the section 4.1 indicate that disrupting positional dependency early prevents the formation of rigid linguistic patterns. (2) We apply Foveal Attention Recycling in the deep, text-dominated layers. Our experiments show that since these layers correspond to the phase where the drift to text generation occurs, recycling is more effective.

3.2 Positional Perturbation

Preliminaries Standard LLMs typically employ Rotary Positional Embeddings (RoPE) (Su et al., 2024) to encode sequence order. For a query or key vector $\mathbf{x} \in \mathbb{R}^d$ at position m, RoPE operation can be computationally realized via the following efficient formulation:

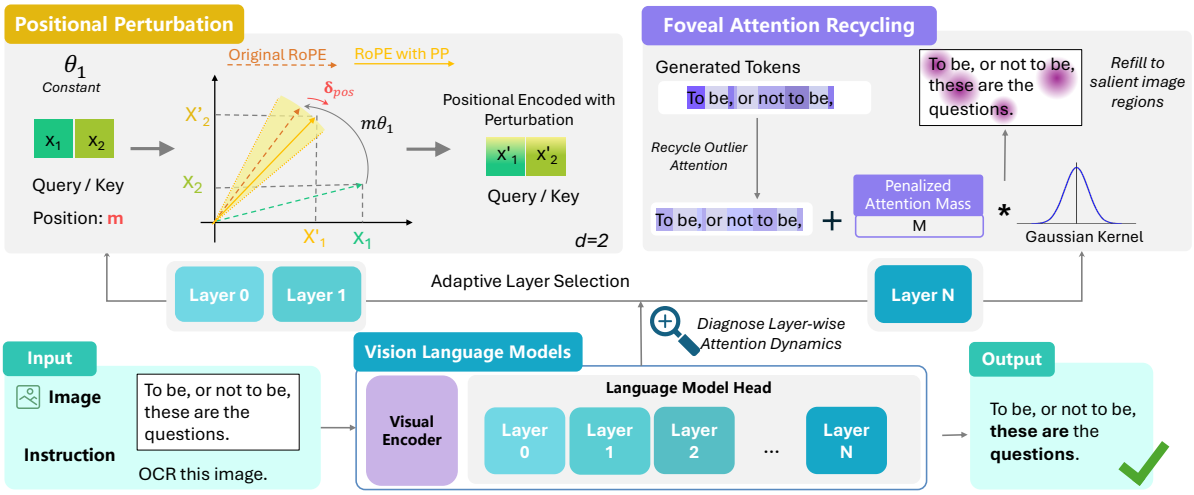

Figure 3: Overview of the PAR Framework. The mechanism consists of two key components applied dynamically during inference. Positional Perturbation (PP) introduces a phase shift δ_{pos} into the Rotary Positional Embeddings (RoPE), blurring precise position indices to break "reciting" inertia. Foveal Attention Recycling (FAR) detects outlier attention on text tokens and redistributing this mass to salient image regions via a Gaussian kernel.

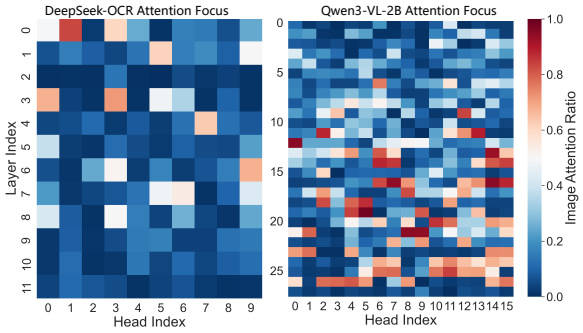

Figure 4: Layer-wise attention heatmaps illustrating the Image-to-Text Focus Ratio (R_{focus}) for DeepSeek-OCR and Qwen3-VL-2B.

$$\begin{aligned}\mathbf{R}^{d}_{\Theta,m}\mathbf{x} =& \mathbf{x} \otimes \mathbf{Cos} + \bar{\mathbf{x}} \otimes \mathbf{Sin} \\ =& \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \otimes \begin{pmatrix} \cos m\theta_1 \\ \cos m\theta_1 \\ \vdots \\ \cos m\theta_{d/2} \end{pmatrix} \\ &+ \begin{pmatrix} -x_2 \\ x_1 \\ \vdots \\ x_{d-1} \end{pmatrix} \otimes \begin{pmatrix} \sin m\theta_1 \\ \sin m\theta_1 \\ \vdots \\ \sin m\theta_{d/2} \end{pmatrix}\end{aligned} \tag{2}$$

where $\otimes$ denotes the element-wise product, $\bar{\mathbf{x}} = (-x_2, x_1, \ldots, x_{d-1})^\top$ is the conjugate pairs of $\mathbf{x}$ and $\Theta = \{\theta_i = 10000^{-2(i-1)/d}, i \in [1, d/2]\}$ represents the pre-defined frequency basis.

Perturbing the Rotary Basis To disrupt this positional inertia without corrupting the semantic content, we introduce Positional Perturbation (PP). Our core intuition is to modulate the basis functions (cos and sin) in Equation 2 while leaving the input feature $\mathbf{x}$ untouched.

Specifically, we introduce a structured noise δ_{pos} to the position index m, effectively blurring the rotation angle from $m\theta_i$ to $m\theta_i - \delta_{pos}$. Applying the first-order Taylor expansion, we can have:

$$\begin{aligned}\cos(m\theta_i - \delta_{pos}) &\approx \cos(m\theta_i) + \delta_{pos} \sin\theta_i \\ \sin(m\theta_i - \delta_{pos}) &\approx \sin(m\theta_i) - \delta_{pos} \cos\theta_i\end{aligned} \tag{3}$$

Therefore, we define the Perturbed Basis Vectors $\mathbf{Cos}'$ and $\mathbf{Sin}'$ as:

$$\begin{aligned}\mathbf{Cos}' &= \mathbf{Cos} + \delta_{pos} \otimes \mathbf{Sin} \\ \mathbf{Sin}' &= \mathbf{Sin} - \delta_{pos} \otimes \mathbf{Cos}\end{aligned} \tag{4}$$

where $\mathbf{Cos}$ and $\mathbf{Sin}$ are the original cosine and sine vectors from Eq. 2, and δ_{pos} is the broadcasted noise vector. Consequently, the **PP-enhanced RoPE** is formally defined as:

$$\mathrm{PP}(\mathbf{x}, m) = \mathbf{x} \otimes \mathbf{Cos}' + \bar{\mathbf{x}} \otimes \mathbf{Sin}' \tag{5}$$

By injecting noise strictly into the rotation basis, PP creates a positional blur effect that forces the attention mechanism to rely on visual grounding rather than precise token indices, preserving the norm and semantic integrity of $\mathbf{x}$.

Structured Harmonic Noise Generation Prior work (Melo et al., 2025) suggests that incorporating harmonic signals into positional embeddings offers superior stability compared to random noise.

Consequently, to maintain local coherence, for the generated token t, the perturbation $\delta_{\text{pos}}(t)$ is not modeled as random white noise but rather as a structured harmonic signal, defined as:

$$\delta_{pos}(t) = \alpha \cdot \mathcal{S}(t) \cdot \sin(\omega \cdot t + \phi)$$

where ω is a fixed frequency, and $\phi \sim \mathcal{N}(0, 1)$ is a random phase noise applied per sequence to ensure diversity. $\mathcal{S}(t)$ is a Peak-Hold-Decay Scheduler designed to protect the initial system prompt while targeting the hallucination-prone generation phase:

$$\mathcal{S}(t) = \tanh\left(\frac{t}{\tau_{rise}}\right) \cdot \left[\beta + (1-\beta)e^{-\frac{\max(0, t-t_p)}{\tau_{decay}}}\right]$$

Here, τ_{rise} allows a smooth warm-up to peak t_p, and τ_{decay} stabilizes the noise amplitude β for long sequences.

3.3 Foveal Attention Recycling (FAR)

While PP weakens the prior, Foveal Attention Recycling actively steers the attention mechanism. Inspired by human biological foveal vision, when linguistic prediction fails, FAR dynamically redistributes attention mass.

Outlier Detection & Suppression In deep layers, attention heads often collapse onto specific text tokens (e.g., punctuation or repeated words). We identify these outliers using dynamic statistics. For the text attention distribution A_{txt} for token t, we calculate the mean μ_{txt} and standard deviation σ_{txt}. A penalty mask $P_{penalty}$ is generated:

$$P_{penalty} = \text{ReLU}(A_{txt} - (\mu_{txt} + k \cdot \sigma_{txt})) \cdot \lambda(t)$$

$$\lambda(t) = s_{base} + s_{add}\sqrt{t/L_{sat}}$$

where $\lambda(t)$ is a progressive strength factor that increases with generation length to counteract growing hallucination as discovered in Section 2.3. Specifically, s_{base} represents the base penalty strength, s_{add} denotes the additional penalty amplitude, and L_{sat} is the saturation length that controls the growth rate. This ensures that the suppression mechanism becomes more aggressive as the sequence lengthens and the prior becomes stronger.

Foveal Refill via Gaussian Diffusion To simulate the biological mechanism of foveal vision, we first identify salient image regions by thresholding the visual attention:

$$\hat{A}_{img} = \text{ReLU}(A_{img} - (\mu_{img} + 1.5\sigma_{img})) \quad (6)$$

Models		Chinese			English		
		CER↓	Iden↑	Cor↓	CER↓	Iden↑	Cor↓
Deepseek-OCR	Ori	0.1000	44.81	49.57	0.9040	38.35	**10.86**
	PAR	**0.0959**	**47.13**	**47.09**	**0.6343**	**46.05**	13.20
Qwen3-VL-8B	Ori	0.0729	58.68	37.87	4.1021	93.20	**1.51**
	PAR	**0.0617**	**69.86**	**26.90**	**2.9075**	**93.54**	1.72
InternVL3.5-8B	Ori	0.2837	63.63	34.03	3.2550	76.56	4.40
	PAR	**0.2806**	**64.38**	**33.28**	**3.2400**	**79.24**	**3.44**

Table 3: Main results on the GlitchText Probing Dataset. We compare the standard inference baseline (Ori) with PAR. PAR achieves consistent improvements in CER across different architectures without training.

To preserve the spatial structure of these regions, we apply a 1D Gaussian convolution, mimicking the biological point spread function to diffuse the salient attention:

$$A'_{img} = A_{img} + \gamma \cdot M \cdot \frac{\hat{A}_{img} * G_\sigma}{\sum(\hat{A}_{img} * G_\sigma)} \quad (7)$$

where G_σ is a Gaussian kernel with a window size proportional to σ, and γ is a refill factor that modulates the intensity of the re-injected attention. This strategy ensures that the recycled attention reinforces coherent, high-confidence visual regions rather than scattering mass onto isolated pixels.

4 Experiments

Experimental Setup We evaluated the proposed PAR mechanism on three state-of-the-art open-source VLMs with distinct architectures: DeepSeek-OCR (Wei et al., 2025), Qwen3-VL-2B/8B (Bai et al., 2025), and InternVL3.5-8B (Wang et al., 2025). All experiments were conducted on our proposed GlitchText Probing Dataset. For the Chinese subset, we utilized the original dataset configuration. For the English subset, given the relatively short length of the quotes, each sample was left-padded with 200 words of the corresponding glitch-free text. The specific hyperparameter configurations for each model are detailed in Appendix C, and the evaluation metrics follow the definitions provided in Section 2.3.

Main Results Table 3 presents the comparative results of standard inference (Ori) versus PAR.

Improvement in General Fidelity (CER) As shown in Table 3, PAR consistently reduces the CER across all models and languages without requiring any parameter updates. For instance, Qwen3-VL-8B achieves a significant CER reduction on the English subset (4.10 → 2.91) and the

Method	CER ↓	Iden ↑	Cor ↓
Baseline (Original)	0.0729	58.68	37.87
VCD (Leng et al., 2024)	0.0865	63.65	31.82
TVC-7B (Sun et al., 2025)	36.80	60.21	16.04
PAR (Ours)	**0.0617**	**69.86**	**26.90**

Table 4: Comparison with training-free (VCD) and training-based (TVC) on the GlitchText dataset using Qwen3-VL-8B.

Chinese subset (0.0729 → 0.0617). This confirms that PAR does not disrupt valid semantic processing but rather enhances the model's grounding on visual evidence.

Mitigation of Hallucination (Iden vs. Cor): A higher Iden with lower Cor indicates successful anomaly detection. First, note that Iden and Cor do not sum to unity; the remainder represents cases where the model omits the glitch or outputs unrelated mismatches. Thus, a rise in Cor does not necessarily imply worse performance if accompanied by a significant drop in omissions and CER. On the Chinese subset, PAR demonstrates a remarkable ability to suppress linguistic priors. For Qwen3-VL-8B, the Identification Rate increases by over 11 points (58.68% → 69.86%), while the Correction Rate drops significantly (37.87% → 26.90%), proving effective alleviation of the auto-correction effect. Similar trends are observed in DeepSeek-OCR and InternVL3.5, validating the efficacy of the proposed noise injection and attention recycling mechanisms. For Qwen3-VL-8B on the English subset, we observe a slight increase in Cor. However, this is outweighed by the substantial reduction in CER (4.10 → 2.91) and the increase in Iden. This indicates that PAR successfully recalls previously ignored tokens. While a minor fraction of this recovered text was auto-corrected, the majority are correctly identified, significantly enhancing overall fidelity.

4.1 Further Explorations

Comparison with Visual Conditioning Methods As shown in Table 4, PAR consistently outperforms both training-free and training-based baselines across all key metrics.

VCD (Leng et al., 2024) is a representative inference-time method that mitigates hallucinations via contrastive decoding, where predictions from the original image are contrasted against those from a visually perturbed (e.g., blurred) version to penalize language-driven generations. Compared to VCD, PAR achieves a lower CER. We attribute this to the fact that VCD relies on visual perturbations (e.g., blurring), which can destroy fine-grained glyph features critical for OCR.

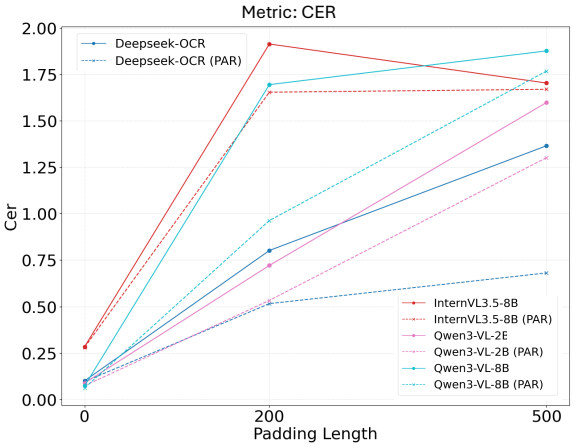

Figure 5: Impact of Context Length on CER. We evaluate model performance with increasing padding lengths. As the sequence lengthens, the baselines (solid lines) suffer from catastrophic error increases, whereas PAR (dashed lines) maintains significantly better stability.

TVC (Sun et al., 2025) is a recent training-based approach that addresses visual forgetting in long CoT reasoning by re-injecting visual tokens during decoding to maintain global visual context. Interestingly, despite this design, TVC exhibits substantially higher CER in our setting. We further analyze its failure mode through qualitative examples. In cases where the input image contains only a partial text fragment (e.g., the first half of a classical poem), TVC often completes the entire sequence based on its parametric memory.

This highlights a key distinction between global visual retention and fine-grained visual grounding. PAR directly addresses this issue by intervening in the decoding dynamics, effectively enforcing token-level visual alignment.

Robustness to Context Length Building upon the "Attention Decay" phenomenon analyzed in Section 2.3, we hypothesize that the hallucination effect intensifies as the generated sequence lengthens. To verify this and evaluate the robustness of our method, we extend the context by left-padding each sample with its original poems or quotes without glitches. This setup effectively delays the occurrence of visual glitches, pushing them deeper into the generation sequence (e.g., after 200 or 500 tokens) to simulate extreme prior accumulation.

Figure 5 illustrates the CER trends across dif-

ferent padding lengths. The results reveals that as the padding length increases, all baseline models exhibit a catastrophic surge in CER. This confirms that as the linguistic context accumulates, the model enters "reciting mode", becoming increasingly blind to visual evidence. In contrast, models equipped with PAR (Dashed Lines) demonstrate superior stability. PAR consistently maintains a much lower CER than the baseline. Crucially, the performance gap widens as the length increases, providing compelling evidence that **PAR effectively counteracts the cumulative inertia of linguistic priors and preserves visual fidelity even in long-context scenarios.** A comparative analysis within the same architecture family reveals a counter-intuitive scaling phenomenon. As shown in Figure 5, the smaller Qwen3-VL-2B maintains a lower CER compared to the larger Qwen3-VL-8B across extended contexts. This suggests that **a more powerful Language Model head imposes stronger linguistic priors, thereby exacerbating the hallucination effect** when visual signals conflict with high-probability semantic predictions.

Ablation Studies We conduct a series of ablation studies on the DeepSeek-OCR model using the Chinese subset. As summarized in Table 5, removing either FAR or PP leads to a degradation in all metrics. We also replace our structured harmonic noise with random noise and Gaussian-based foveal refill with a naive uniform distribution strategy. The results both show a clear performance drop. Although average refill achieves a slightly lower Correction Rate, it comes at the cost of a higher CER. This indicates that simply scattering attention mass uniformly disrupts the local coherence of visual features, impairing the fluency and fidelity of generation. Furthermore, we replace our Peak-Hold-Decay scheduler with a Constant strategy (fixed α without warm-up/decay) and a Linear strategy (linearly increasing with position). As shown in Table 5, the Linear scheduler yields the highest Identification Rate (50.56%), suggesting that intensifying noise effectively shatters linguistic inertia. However, this aggressiveness severely compromises the primary OCR capability, resulting in the worst CER (0.1002). The Constant scheduler also fails to match the baseline, confirming that a dynamic regulation, protecting the initial prompt and stabilizing long-tail generation, is essential. Finally, we verify the efficiency of our layer-wise application strategy; experiments where intervention layers are shifted result in significant performance drops, as detailed in Appendix D. For a detailed analysis of the computational efficiency, please refer to Appendix G.

Models	CER ↓	Iden↑	Cor↓
Deepseek-OCR			
PAR	**0.0959**	47.13	47.09
w/o FAR	0.0981	45.65	48.86
w/o PP	0.0971	46.68	47.32
w/ Random Perturbation	0.0986	46.94	47.53
w/ Average Refill	0.0973	47.71	46.34
w/ Constant Scheduler	0.0966	47.01	47.28
w/ Linear Scheduler	0.1002	50.56	43.46

Table 5: Ablation studies on the Chinese subset using DeepSeek-OCR. The full PAR achieves the lowest CER.

4.2 Generalization on OmniDocBench

Beyond the adversarial GlitchText dataset, we further evaluate the generalization capability of PAR on **OmniDocBench** (Ouyang et al., 2024), a comprehensive benchmark designed for diverse document parsing tasks. We adopt the official End-to-End evaluation setting and standard metrics as defined in the benchmark. As presented in Table 6, applying PAR yields consistent improvements across different architectures. These results suggest that the benefits of PAR extend to general document understanding. By mitigating the over-reliance on linguistic priors, PAR encourages the model to attend more faithfully to complex layout structures (such as table alignments), thereby enhancing performance on standard benchmarks without any fine-tuning. Detailed analysis and case studies can be found in Appendix E and F.

5 Related Work

Hallucination in VLMs typically refers to the generation of content not present in the visual input (Tonmoy et al., 2024). Extensive research has focused on Object Hallucination, where models describe non-existent objects or incorrect attributes (Li et al., 2023; Wang et al., 2023; Rawte et al., 2025; Augustin et al., 2025; Park et al., 2025). Li et al. (2023) discovered that objects that frequently appear in the visual instructions or co-occur with the image objects are obviously prone to be hallucinated by LVLMs. Metrics like POPE (Li et al., 2023) and CHAIR (Rohrbach et al., 2018) have been established to quantify this phenomenon, and

Models		Text$^{Edit}\downarrow$	Table$^{TEDS}\uparrow$	Table$^{TEDS-S}\uparrow$	Read Order$^{Edit}\downarrow$	Overall $\uparrow$
DeepSeek-OCR	Ori	0.3812	83.6621	**88.2035**	0.1080	48.5140
	PAR	**0.3802**	**83.9923**	88.0966	**0.1066**	**48.6574**
Qwen3-VL-8B	Ori	**0.0585**	66.5593	70.6068	**0.0722**	53.5698
	PAR	0.0675	**71.5517**	**76.3676**	0.0824	**54.9339**
Qwen3-VL-2B	Ori	**0.0922**	48.379	51.6288	0.1136	46.3863
	PAR	0.0941	**49.5061**	**52.6606**	**0.1132**	**46.6987**
InternVL3.5-8B	Ori	**0.1538**	65.8361	72.0031	**0.1416**	50.1520
	PAR	0.1586	**66.7146**	**73.0953**	0.1434	**50.2849**

Table 6: Performance on the OmniDocBench dataset. PAR achieve consistent improvements in Overall scores.

mitigation strategies often involve reinforcement learning (He et al.; Gunjal et al., 2024) or visual evidence prompting (Li et al., 2025b). Sun et al. (2025) further identified that vision-language models suffer from visual forgetting in long chain-of-thought reasoning, where visual evidence is gradually overshadowed by accumulated textual context and proposed Take-along Visual Conditioning (TVC), which re-injects visual tokens into the decoding process to refresh visual memory during long reasoning chains. TVC requires additional training or fine-tuning to integrate visual conditioning, whereas PAR is a purely training-free, inference-time intervention.

However, Textual Hallucination, specifically in OCR, remains underexplored. Unlike object hallucination, OCR hallucination often manifests as an "auto-correction" driven by semantic priors. Shu et al. (2025) highlighted how semantic context can mislead vision in scene text spotting, causing models to output semantically plausible but visually incorrect words. Liang and Zhang (2025) introduced ReViCo, a benchmark for real-world visual spelling correction, and proposes Background Information Enhancement to mitigate the recognition failures. Our work extends this line of works to the domain of dense document OCR, specifically focusing on the Linguistic Priors Hallucination where models ignore visual glitches in high-familiarity texts due to excessive reliance on parametric memory.

Recent inference-time interventions employ Contrastive Decoding (CD) (Wang et al., 2024; Leng et al., 2024) or attention manipulation (Kang et al.; Zhang et al., 2025) to curb hallucinations. VAR (Kang et al.) identified the Visual Attention Sink phenomenon and proposed Visual Attention Redistribution to identify image-centric heads and recycle attention back to relevant visual features. Zhang et al. (2025) proposed Enhancing Vision Attention Sinks (EVAS), demonstrating that maintaining dense vision attention sinks in shallow layers is critical for grounding. While methods like VAR and EVAS optimize visual grounding by redistributing attention or reinforcing shallow sinks, they primarily target general Object Hallucination. Direct application to OCR proves suboptimal for two reasons. First, CD approaches rely on global image distortions which obliterate the fine-grained glyph features essential for OCR; furthermore, they penalize linguistic priors at the probability level, risking the suppression of correct text where priors align with ground truth. Second, unlike EVAS which reinforces shallow patterns, our task requires breaking the autoregressive inertia. In contrast, our PAR framework targets Linguistic Priors Hallucination in OCR. PAR actively disrupts positional inertia and synergizes this with a Foveal-Vision recycling mechanism to recover neglected details without compromising linguistic fluency.

6 Conclusion

In this work, we identify and quantify the bottleneck of Linguistic Priors Hallucination in VLMs. Through experiments on our proposed GlitchText probing datset, we find that longer linguistic contexts and more powerful language models often tend to ignore visual details and output the text sequence it has memorize. To address this, we introduce PAR, a novel training-free mechanism which consists of Positional Perturbation and Foveal Attention Recycling to enforce faithful visual grounding. PAR effectively balances semantic fluency with visual fidelity. Experiment results confirm that PAR not only rectifies hallucinations in adversarial settings but also enhances robustness in general document parsing tasks without extra fine-tuning.

Limitations

Different VLM architectures exhibit distinct attention dynamics. While PAR effectively mitigates this issue by diagnosing specific models, it implies that applying PAR to a completely new architecture might require a preliminary diagnostic run to identify the optimal intervention layers for Positional Perturbation and Attention Recycling. Furthermore, our current implementation operates at the layer level. Future research could refine the granularity of PAR from the layer level to the individual attention head level. By specifically targeting hallucination-prone heads for recycling while leaving semantic-rich heads untouched for Positional Perturbation, we could achieve a more precise balance between visual fidelity and linguistic fluency.

Ethics Statement

The GlitchText probing dataset constructed in this study relies entirely on synthetic generation using public dictionaries and standard font libraries, ensuring the absence of Personally Identifiable Information and strict compliance with copyright regulations. By actively mitigating linguistic hallucinations, this research enhances the faithfulness of automated document digitization, thereby reducing the societal risks associated with AI-generated misinformation in high-precision applications.

References

Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. 2025. Dash: Detection and assessment of systematic hallucinations of vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22748–22759.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Paul E. Black. 2021. Ratcliff/Obershelp pattern recognition. Dictionary of Algorithms and Data Structures. Accessed: 2025-12-25.

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*, pages 14303–14312.

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

Lehan He, Zeren Chen, Zhelun Shi, Tianyu Yu, Lu Sheng, and Jing Shao. Systematic reward gap optimization for mitigating vlm hallucinations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yingchen He, Jennifer M Scholz, Rachel Gage, Christopher S Kallie, Tingting Liu, and Gordon E Legge. 2015. Comparing the visual spans for faces and letters. *Journal of vision*, 15(8):7–7.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

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

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. 2025a. Visual evidence prompting mitigates hallucinations in large vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4048–4080.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. 2025b. Visual evidence prompting mitigates hallucinations in large vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4048–4080. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Junhong Liang and Bojun Zhang. 2025. Vision language models are not (yet) spelling correctors. *arXiv preprint arXiv:2509.17418*.

Rui Melo, Rui Abreu, and Corina S Pasareanu. 2025. Microsaccade-inspired probing: Positional encoding perturbations reveal llm misbehaviours. *arXiv preprint arXiv:2510.01288*.

Andrew Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2024. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. *Preprint*, arXiv:2412.07626.

Eunkyu Park, Minyeong Kim, and Gunhee Kim. 2025. Halloc: Token-level localization of hallucinations for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29893–29903.

Vipula Rawte, Aryan Mishra, Amit Sheth, and Amitava Das. 2025. Defining and quantifying visual hallucinations in vision-language models. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 501–510, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Yan Shu, Hangui Lin, Yexin Liu, Yan Zhang, Gangyan Zeng, Yan Li, Yu Zhou, Ser-Nam Lim, Harry Yang, and Nicu Sebe. 2025. When semantics mislead vision: Mitigating large multimodal models hallucinations in scene text spotting and understanding. *arXiv preprint arXiv:2506.05551*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. 2025. Mitigating visual forgetting via take-along visual conditioning for multi-modal long CoT reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5158–5171, Vienna, Austria. Association for Computational Linguistics.

S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *CoRR*, abs/2401.01313.

William S Tuten and Wolf M Harmening. 2021. Foveal vision. *Current Biology*, 31(11):R701–R703.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023. Evaluation and analysis of hallucination in large vision-language models. *CoRR*, abs/2308.15126.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

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

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. 2025. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3512–3534.

A Dataset

A.1 Dataset Construction

Adversarial Anomaly Injection To synthesize a conflict between visual grounding and linguistic priors, we systematically inject "glitches", including subtle visual or semantic deviations, into the high-familiarity texts. **(1) Chinese: Visual Glitch.** Given the logographic nature of Chinese, errors frequently manifest as visually similar characters. We leverage the *Confused_Chinese* dictionary [5] to map original characters to their visually confounding counterparts (e.g., replacing "兵" with "乒"). For each sample, we iterate through the text and prioritize replacements based on visual similarity scores provided by the dictionary, selecting the top-20 hardest-to-distinguish characters to maximize visual discrimination difficulty. **(2) English: Morphological Glitch.** For English, we employ `SpaCy` [6] for Part-of-Speech (POS) tagging and `lemminflect` [7] for morphological transformation. We specifically target Verbs and Nouns to introduce grammatical inconsistencies, such as altering tenses (e.g., *go* → *went*) or pluralization (e.g., *apple* → *apples*). To refine the selection, we utilize Gestalt Pattern Matching (Black, 2021) to calculate string similarity between the original and transformed words, retaining the top-20 anomalies to replace the original tokens.

Standardized Image Synthesis To isolate the impact of textual content and rule out background noise, the text is rendered on a pure white background with black font. We use SimHei for Chinese and DejaVuSans for English to ensure glyph clarity. The layout is standardized with character-level line breaking for Chinese and word-level line breaking for English, generating high-resolution PNG images ready for probing.

Rationale for Synthetic Design A potential concern regarding GlitchText is the artificial nature of the injected anomalies compared to natural OCR noise (e.g., blur, occlusion, or low resolution). We explicitly clarify that the primary objective of GlitchText is not to simulate general in-the-wild degradation, but to serve as a specialized diagnostic probe for Conflict Maximization. While relying on linguistic priors to infer missing content is often employed as an error correction mechanism in standard noisy OCR, this dependency acts as a double-edged sword. In high-precision scenarios requiring strict fidelity, such reliance often backfires, leading to unfaithful "auto-corrections" where visual evidence is overridden by semantic expectations. By isolating "visual clarity" from "linguistic familiarity," GlitchText creates an ideal testbed to diagnose this specific **Linguistic Priors Hallucination**, forcing the model to demonstrate whether it chooses to "read" the actual visual glyphs or "recite" the memorized text when a conflict arises.

[5] https://github.com/Macielyoung/Confused_Chinese

[6] https://pypi.org/project/spacy/

[7] https://pypi.org/project/lemminflect/0.1.0/

Figure 6: Examples of the GlitchText Probing Dataset

A.2 Detailed Examples

Figure 6 visualizes representative examples from both the Chinese and English subsets. Each visual sample is paired with its ground-truth annotation (formatted in JSON), which details the specific `error_positions`, both original and anomaly characters, and their similarity scores.

B Attention Heatmaps

In this appendix, we provide supplementary attention visualizations for other representative architectures. Figure 7 and Figure 8 illustrate the layer-wise **Image-to-Text Focus Ratio** heatmaps for InternVL3.5-8B and Qwen3-VL-2B, respectively. These visualizations further validate the classification of models into "Decay" and "Late-Fusion" patterns discussed in Section 3.1.

Models	Instructions	Tasks	PP Layers	FAR Layers	α	s_{base}	t_p	τ_{decay}	β	L_{sat}	s_{add}	γ
Deepseek-OCR	<image> Free OCR.	OCR	0	0,11	0.1	0.1	300	500	0.5	200	0.8	2.5
		Omnidoc	0	9,10,11	0.01	0.05	300	500	0.3	200	0.3	1.5
Qwen3-VL-2B	Read all the text in the image.	OCR	0,1,2,3	0,1,2,3	0.05	0.1	300	500	0.05	300	0.7	1.5
		Omnidoc	0,1,2	3	0.05	0.1	300	500	0.05	300	0.5	1
Qwen3-VL-8B	Read all the text in the image.	OCR	0,1,2,3	0,1,2,3	0.05	0.1	300	500	0.05	300	0.7	1.5
		Omnidoc	0,1,2,3	0,1,2,3	0.1	0.1	300	1000	0.5	300	0.8	2.5
InternVL3.5-8B	Read all the text in the image.	OCR	27	31	0.02	0.1	300	500	0.5	200	0.7	1.5
		OCR_padding	3,4	33	0.02	0.1	300	500	0.05	300	0.7	1.5
		Omnidoc	3,4	33	0.1	0.1	300	500	0.5	300	0.7	2.5

Table 7: Detailed hyperparameter settings for PAR across various models and tasks. PP Layers and FAR Layers denote the indices of transformer layers where Positional Perturbation and Attention Recycling are applied, respectively. Fixed parameters are omitted for brevity ($\omega = 1, k = 3, \tau_{rise} = 100, \sigma = 20, G_\sigma = 80$).

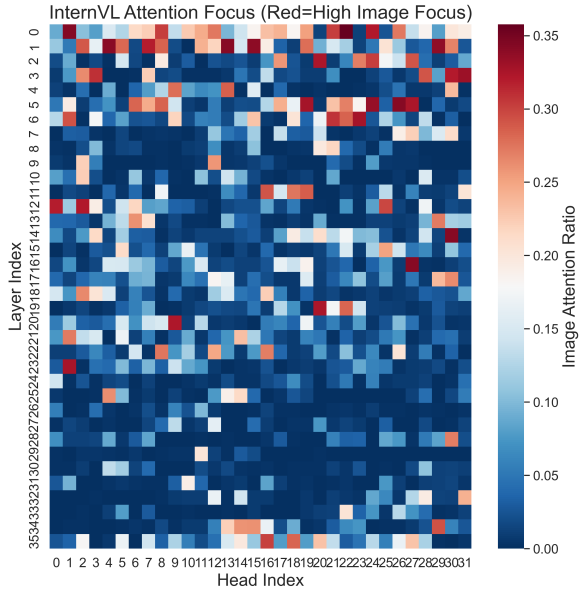

Figure 7: Layer-wise attention heatmap for **InternVL3.5-8B**, exhibiting a Type A (Decay) pattern where visual attention diminishes in deeper layers.

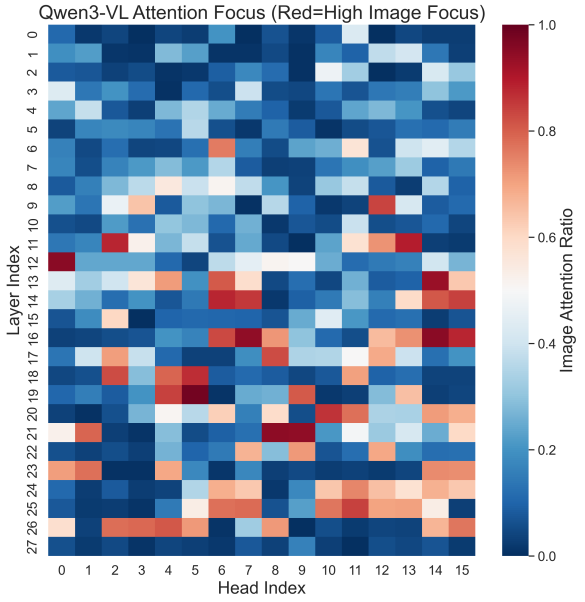

Figure 8: Layer-wise attention heatmap for **Qwen3-VL-2B**, exhibiting a Type B (Late-Fusion) pattern with varying visual focus across layers.

B.1 Visualization of Attention Redistribution

To explore how PAR alters the internal attention dynamics, we visualize the *differential* layer-wise attention maps between the PAR-enhanced model and the baseline Qwen3-VL-2B. We define the Attention Focus Shift as $\Delta R = R_{focus}^{\text{PAR}} - R_{focus}^{\text{Baseline}}$, where R_{focus} is the Image-to-Text Focus Ratio.

As illustrated in Figure 9, the heatmap does not show a uniform increase in visual attention. Instead, it facilitates a functional specialization of attention heads: reinforcing visual attention where grounding is critical, while suppressing it in heads responsible for language modeling. This validates that our method achieves a refined balance between "reading" (visual) and "reciting" (linguistic), rather than simply applying a global penalty to text tokens.

C Hyperparameters

Table 7 presents the detailed hyperparameter configurations used for each model in our experiments. To ensure a fair comparison and reduce the search space, we universally set the following parameters across all architectures and tasks: the perturbation frequency $\omega = 1$, the outlier detection threshold multiplier $k = 3$, $\tau_{rise} = 100$, the Gaussian deviation $\sigma = 20$, and the kernel window size $G_\sigma = 80$. For closed-source models (e.g., GPT-4o, Claude), we use "Read all the text in the image." as instructions.

The remaining parameters, including the target layers for Positional Perturbation (PP) and Foveal

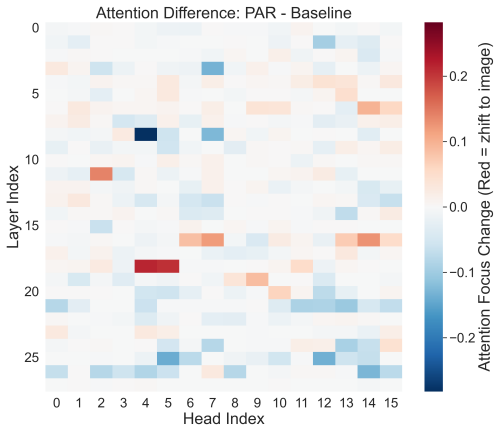

Figure 9: **Attention Focus Difference Heatmap (PAR - Baseline)** on Qwen3-VL-2B. The heatmap confirms that PAR does not uniformly boost visual weights but rather optimizes attention allocation, enabling specific heads to specialize in either visual grounding or linguistic modeling.

Attention Recycling (FAR), as well as the noise amplitude scheduler (α, β, τ), were tuned to adapt to the specific attention dynamics of each model (as diagnosed in Section 3.1).

D Layer Ablation

To validate the necessity of our layer selection strategy, we conducted a counter-factual experiment. We inverted the target layers for the DeepSeek-OCR and InternVL models (applying PP to deep layers and FAR to shallow layers) and shifted the intervention window to the deep layers for the Qwen3-VL models. As presented in Table 8, these sub-optimal layer configurations lead to a consistent degradation in performance. Specifically, the Identification Rate (Iden) drops and the Correction Rate (Cor) increases, indicating that the model fails to break the linguistic inertia when PP is applied too late, or fails to recover visual details when FAR is applied too early. This confirms that applying PP to disrupt early-stage inertia and FAR to rectify late-stage attention drift is crucial for the framework's success.

E Performance Analysis across Generation Lengths on OmniDocBench

To evaluate the stability of our method during long-context generation, we analyzed the distribution of errors relative to the generation position. Document parsing tasks often suffer from error accumulation, where the model's performance degrades as the sequence length increases due to attention drift.

Models	PP Layers	FAR Layers	CER ↓	Iden ↑	Cor ↓
Deepseek-OCR	0	0,11	0.0959	47.13	47.09
	0,11	0	0.1011	46.15	48.18
Qwen3-VL-2B	0,1,2,3	0,1,2,3	0.0737	81.86	12.64
	24,25,26,27	24,25,26,27	0.0838	78.83	15.33
Qwen3-VL-8B	0,1,2,3	0,1,2,3	0.0617	69.86	26.90
	24,25,26,27	24,25,26,27	0.0615	63.09	33.52
InternVL3.5-8B	27	31	0.2806	64.38	33.28
	31	27	0.2814	63.88	33.75

Table 8: Ablation study on Adaptive Layer Selection. We compare the proposed layer configuration (first row per model) against counter-intuitive configurations (second row), such as swapping the application layers for PP and SAR or shifting the intervention window to deep layers. The results demonstrate that deviating from the adaptively selected layers leads to performance degradation.

Figure 10 visualizes the Average Table Edit Distance (ED) of the DeepSeek-OCR model across different predicted position ranges (from 0 to over 2000 tokens) on the OmniDocBench dataset. We observe that the PAR-enhanced model (solid line) consistently achieves a lower Edit Distance compared to the Baseline (dashed line) across all position intervals. This indicates that our inference-time intervention is robust and does not destabilize the generation at any stage.

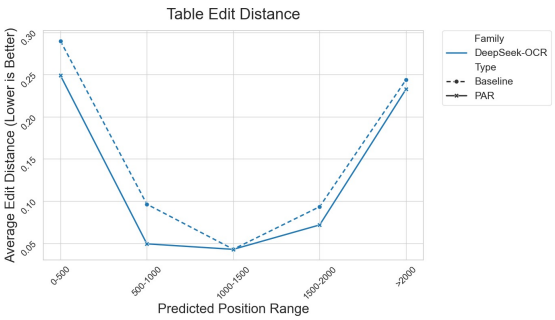

Figure 10: Average Table Edit Distance (ED) across different predicted position ranges on OmniDocBench. The solid line (PAR) consistently stays below the dashed line (Baseline), demonstrating that PAR reduces recognition errors throughout the generation process.

F Case Studies

To qualitatively evaluate the efficacy of PAR in real-world document parsing scenarios, we visualize the inference results of DeepSeek-OCR on the OmniDocBench dataset. As illustrated in Figure 11 and Figure 12, the baseline model often succumbs to two primary failure modes: (1) Linguistic Priors Hallucination, where the model "auto-corrects" or fabricates content based on semantic fluency

rather than visual evidence, and (2) Visual Omission, where fine-grained details (e.g., footnotes, complex layouts) are ignored due to attention decay. In contrast, by incorporating PAR, the model effectively suppresses these hallucinations and recovers neglected visual details, demonstrating superior faithfulness to the original document.

Context Length	Baseline (ms)	PAR (ms)
1,024	0.0092	0.2809
2,048	0.0064	0.2773
4,096	0.0065	0.2788
8,192	0.0080	0.2829

Table 9: Inference latency per decoding step (in milliseconds) across different context lengths. Results are averaged over 100 runs. Notably, since PAR is only activated in a few deep layers (e.g., 1-2 layers out of 32), the cumulative overhead for the entire model is negligible.

G Computational Overhead Analysis

Although PAR is a training-free framework that avoids the prohibitive cost of parameter updates, the inference-time intervention, specifically the Foveal Attention Recycling mechanism which introduces additional computational steps. In this section, we analyze the theoretical complexity and empirical latency of our method.

Theoretical Complexity In the standard autoregressive decoding phase, the attention mechanism scales linearly $O(L)$ with respect to the sequence length L due to the KV cache. Our PAR framework introduces two additional operations:

- **Positional Perturbation (PP):** This operation applies element-wise transformations to the Rotary Positional Embeddings. Since it operates solely on the current query token's position index, its complexity is $O(d)$ where d is head dimension, which is computationally negligible.
- **Foveal Attention Recycling (FAR):** This component involves calculating statistics (mean/std) over the attention logits and applying a 1D Gaussian convolution. The statistical computation iterates over the sequence length L, resulting in $O(L)$. The Gaussian convolution is applied to the attention logits of size L. Given a fixed kernel window size K (where $K \ll L$), the complexity of the convolution is $O(K \cdot L)$, which simplifies to $O(L)$.

Therefore, PAR maintains the same **linear complexity class** $O(L)$ as the standard decoding step, merely adding a constant factor to the inference time.

Empirical Latency Evaluation To quantify the actual computational cost, we benchmarked the inference latency of a single decoding step using an NVIDIA A100 GPU. We compared the standard Attention mechanism against our PAR-enhanced Attention across varying context lengths.

As detailed in Table 9, the PAR mechanism introduces a measurable latency increase at the operator level (rising to $\sim$0.28ms) due to the requisite statistical aggregation and Gaussian convolution operations. However, this operator-level overhead does not translate to a bottleneck in the overall model inference. **Crucially, PAR is designed as a sparse intervention strategy**, typically activated only in 2-4 specific deep layers (determined by our Adaptive Layer Selection) out of the 35 layers in standard VLMs (e.g., Qwen3-VL-8B). Consequently, the cumulative overhead for a full forward pass is strictly bounded.

To address potential concerns regarding latency accumulation over long generation sequences (e.g., $>$1000 tokens), we further conducted an End-to-End Wall-Clock Time evaluation on the OmniDocBench dataset. When generating 100 long-document samples, the total inference time for the baseline was 4h 02m 36s, while the PAR-enhanced model took 4h 21m 24s. This result confirms that the cumulative overhead remains strictly bounded and practical for real-world document processing pipelines.

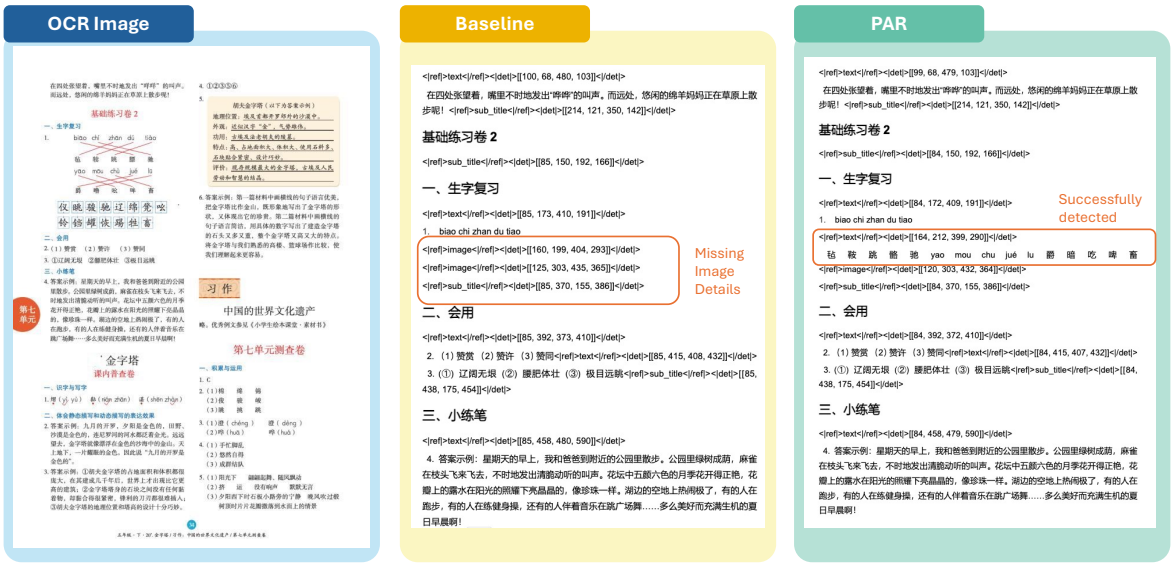

Figure 11: Qualitative comparison on OmniDocBench (Case 1). By applying PAR (right column), the model successfully rectifies these errors, faithfully recognizing the specific visual glyphs and correctly transcribing the text as it appears in the image.

Figure 12: Qualitative comparison on OmniDocBench (Case 2). This example highlights the model's ability to handle complex layouts. The baseline model fails to ground the visual content effectively, missing entire blocks of text and image details. With PAR, the model exhibits stronger visual grounding, successfully detecting and transcribing the fine-grained footnotes and layout structures that were previously ignored.