Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory

Quanjiang Li [*1] **Zhiming Liu** [*2] **Wei Luo** [3] **Tingjin Luo** [1] **Chenping Hou** [1]

Abstract

Multimodal large language models (MLLMs) frequently suffer from object hallucinations, yet the visual perceptual mechanism underlying this failure remains poorly understood. In this work, we reveal that hallucinations are strongly associated with a human-like attention distraction phenomenon, where humans under divided focus experience degraded visual clarity and produce inaccurate descriptions, while in models the same mechanism manifests as spatial inconsistency in multi-head attention and temporal fading of attention to image tokens during decoding. We further provide theoretical insights that attention dispersion increases model complexity and degrades classification generalization. Motivated by these findings, we propose an Attention-Focused Approach for Improved Image Perception (AFIP), which corrects attention distraction via cross-head attention enrichment and reinforces visual grounding through dynamic historical attention enhancement. Extensive experiments on multiple benchmarks and models validate the effectiveness of AFIP without additional training.

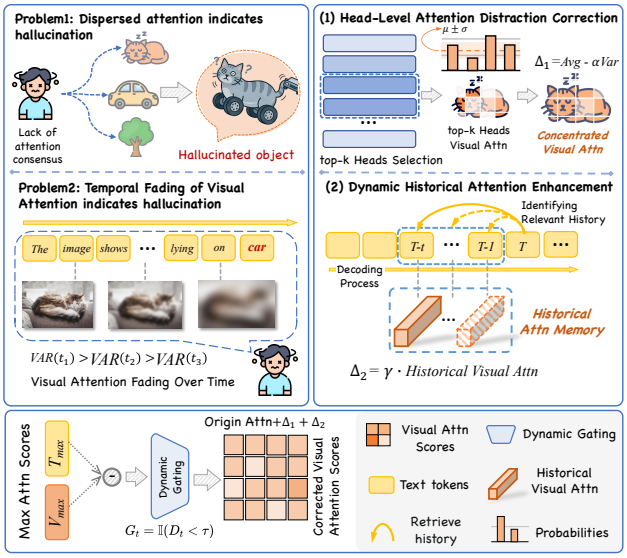


Figure 1. Motivation illustration and overview of AFIP.

1. Introduction

MLLMs have recently made substantial progress and have been applied to a wide range of tasks, including medical analysis (Lin et al., 2025; Li et al., 2025b), creative content generation (Wang et al., 2024a; Huang et al., 2023) and autonomous driving (Sima et al., 2024; Guo et al., 2024). Such progress is enabled by joint modeling of visual and textual information, which facilitates integrated multimodal understanding and reasoning across diverse real-world scenarios. Despite this impressive versatility (Li et al., 2026; 2025a), modern MLLMs still suffer from a critical limitation, as they may generate outputs that are insufficiently grounded in the actual visual input. In particular, models can produce confident yet erroneous descriptions, such as attributing nonexistent objects or incorrectly specifying visual attributes (e.g., color, number, or spatial relations). This phenomenon is commonly referred to as hallucination and poses a substantial obstacle to the reliable deployment of MLLMs in risk-sensitive and high-precision applications.

To address the hallucination issue in MLLMs, recent researches have explored a diverse set of mitigation strategies that can be summarized into four paradigms. The first line of work exploits external knowledge sources through the incorporation of auxiliary databases or retrieval modules into the generation process, thereby supplying explicit factual or visual cues (Qu et al., 2024). The second direction prioritizes model fine-tuning and aims to improve generation reliability via parameter updating driven by task-specific or consistency-oriented supervision (Yu et al., 2024b). The third category concentrates on attention adjustment mech-

[*]Equal contribution [1]National University of Defense Technology, China [2]Harbin Institute of Technology (Shenzhen), China [3]Xi'an Jiaotong University, China. Correspondence to: Chenping Hou <hcpnudt@hotmail.com>, Tingjin Luo <tingjinluo@hotmail.com>.

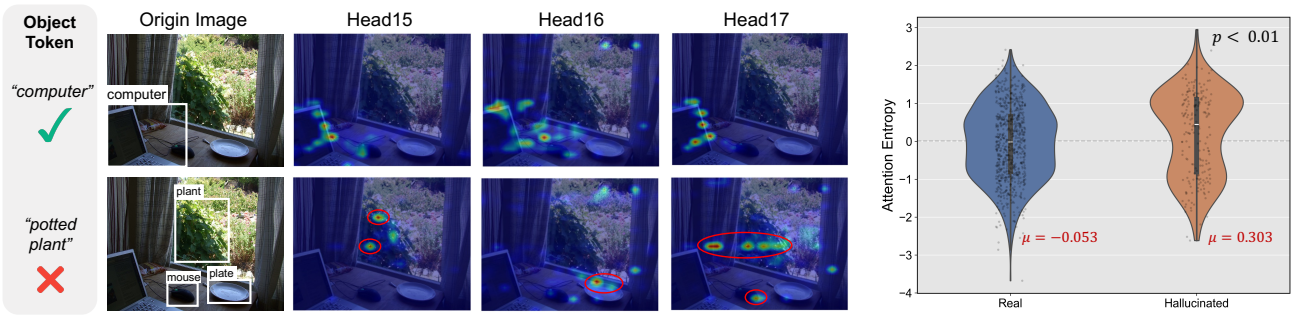
Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

anisms, which modifies attention distributions across different tokens during inference to reduce spurious correlations(Huang et al., 2024). The fourth stream builds upon contrastive learning principles, which compares alternative responses or representations so that outputs with stronger visual support are preferred (Leng et al., 2024). The fifth paradigm focuses on hidden-state intervention, which modifies the internal representations of the model during inference so that the generated outputs better align with true visual perception. (Zou et al., 2025)

Despite the proposal of a wide range of methods aimed at mitigating hallucinations, they still exhibit the following major limitations. (i) Existing methods often entail a reliance on supplementary resources or extended processing time. In particular, approaches based on external knowledge incorporation or model fine-tuning require auxiliary components, such as large-scale instruction-oriented datasets, to correct hallucinated outputs in a post-hoc manner (Yang et al., 2024; Liu et al., 2023). Although attention adjustment methods avoid dependence on external data, they often involve complex inference-time procedures, such as retrospection-based decoding or multi-step attention reallocation, which inevitably increase memory consumption (Tu et al., 2026). (ii) The visual perceptual mechanism underlying hallucinations remains insufficiently explored. Compared with text, images carry denser information and exhibit richer structural patterns, which makes it difficult for MLLMs to faithfully encode, maintain, and exploit visual evidence (Liu et al., 2024b). Therefore, hallucinations should not be explained solely by language-side biases, since distortions introduced during visual interpretation can also play a crucial role. Although prior studies mainly associate hallucinations with language-related factors, such as text inertia (Liu et al., 2024b), the influence of visual perception has not been sufficiently examined. In practice, failures in visual information processing can directly weaken the grounding of generated tokens and trigger hallucinated descriptions. (iii) Existing explanations of hallucination causes remain limited from both spatiotemporal and theoretical perspectives. The generation process of MLLMs involves spatial grounding over image regions as well as temporal reasoning during decoding. Errors arising from incorrect spatial attention may be progressively amplified in subsequent decoding steps. However, few studies have jointly examined hallucinations from the perspective of spatial attention dynamics and temporal degradation of visual grounding. Moreover, the analogy between model hallucination and human perceptual phenomena has rarely been investigated, despite its potential to provide intuitive insights into model behavior. From a theoretical standpoint, hallucinations are still insufficiently characterized. Since the final decoding stage can be viewed as a token classification process, the factors that impair token discrimination deserve further investigation.

To address these challenges, we begin by establishing a strong correspondence between hallucination phenomena in MLLMs and human physiological behavior under dispersed attention, supported by extensive statistical validation. As illustrated in Fig. 1, when humans describe visual content under divided focus, frequent shifts in attention render the cognitive representation of the target image increasingly indistinct and less salient, leading to confabulated descriptions. A comparable pattern emerges in MLLMs, manifested as inconsistent multi-head attention distributions across spatial positions and a progressive decline of visual focus during autoregressive reasoning. These spatiotemporal misalignments impair the integration of visual information and compromise the fidelity of image perception. Besides, theoretical analyses of the multi-head attention architecture, together with the view that attention heads act as independent information sources whose fusion governs generalization, suggest that cross-head attention inconsistency increases model complexity and weakens generalization capability. Motivated by these insights, we propose the Attention Focused Approach for Improved Image Perception (AFIP), a training-free framework that imposes no additional inference overhead. AFIP calibrates high-confidence focused regions by aggregating multi-head information, while penalizing and eliminating irrelevant areas with a variance-induced regularization term. Furthermore, reliable visual focus derived from prior decoding steps is adaptively fused into the current state to mitigate temporal decay in perception. A dynamic gating mechanism, guided by the difference between peak visual and textual attention, prevents excessive adjustment and ensures response stability. Together, these designs enable robust visual grounding and substantially reduce hallucination. The primary contributions of this paper are summarized as follows:

- For the first time, we reveal a strong analogy between the underlying causes of machine hallucinations in MLLMs and human visual blur resulting from attention distraction. This physiological phenomenon is further decomposed into two factors intrinsic to the MLLMs, i.e., spatial inconsistency of multi-head attention and temporal attenuation of visual attention.
- Through analyzing the architecture of the multi-head attention mechanism and the information aggregation capabilities it enables, we theoretically establish that discrepancies across attention heads increase model complexity and impair token-level discrimination, thereby reducing the overall generalization performance.
- We propose AFIP, a plug-and-play module that incurs no additional cost to mitigate object hallucinations. Extensive experiments demonstrate that AFIP surpasses mainstream baselines and attains SOTA status.



(a) Visualization of Spatial Divergence **(b) Distribution Shift in Attention Entropy**

Figure 2. Correspondence between spatial attention inconsistency and object hallucination. **(a)** Attention maps for the correct token computer" and the hallucinated token potted plant". Attention linked to the hallucinated token is broadly dispersed across the image, whereas attention for the correct token remains sharply concentrated on the target object. Additional examples are provided in Appendix D.6. **(b)** Distribution shift in attention entropy. Tokens are categorized as *Real* or *Hallucinated*. Real tokens predominantly exhibit low entropy, while hallucinated tokens correspond to substantially higher entropy values. The result $p < 0.01$ indicates a statistically significant difference in mean entropy between the two groups.

2. Related work

2.1. Vision Language models

In recent years, vision-language modeling has undergone a significant paradigm shift. Early vision-language models primarily relied on BERT-style language decoders (Devlin et al., 2019; Koroteev, 2021) to integrate visual representations with text, enabling unified representation learning but offering limited generative flexibility. (Li et al., 2020; Sun et al., 2019). The emergence of large language models (LLMs) (Brown et al., 2020; Achiam et al., 2023; Bai et al., 2023a) significantly enhanced the expressive power of language backbones and spurred the development of MLLMs (Alayrac et al., 2022; Wang et al., 2024b; Driess et al., 2023). Leveraging end-to-end training and joint decoding over visual and textual tokens, MLLMs exhibit substantially improved adaptability. Contemporary MLLMs generally follow a unified processing pipeline in which images are first encoded by powerful vision backbones such as CLIP (Radford et al., 2021) or EVA (Fang et al., 2023), aligned through lightweight projection modules, and decoded by pretrained LLMs. Nevertheless, the specific design of projection modules differs substantially across models. LLaVA-1.5 (Liu et al., 2024a) adopts a linear projection for modality alignment, while MiniGPT-4 (Zhu et al., 2023b) utilizes a querying transformer to compress visual representations. More recent MLLMs, including Qwen-VL (Bai et al., 2023b), InternVL (Chen et al., 2024a), and LLaVA-NeXT (Li et al., 2024), further improve multimodal reasoning through enhanced alignment and high-resolution visual instruction tuning.

2.2. Object Hallucination in MLLMs

Hallucination remains a persistent and consequential challenge in modern MLLMs. Such inaccuracies are particularly concerning in high-stakes domains, including medical imaging (Kraljevic et al., 2021; Zheng et al., 2025; He et al., 2023) and autonomous driving (Cui et al., 2024), where unreliable outputs may lead to severe downstream consequences. In the vision-language setting, hallucination often manifests as object hallucination, where models generate plausible descriptions that mention objects missing from the image (Li et al., 2023; Rohrbach et al., 2018). A wide range of mitigation strategies has emerged in recent years. Early studies mainly relied on representation-level techniques, including contrastive learning, ROI-level feature fusion (Biten et al., 2021), and data augmentation for reducing spurious cross-modal correlations (Kim et al., 2023). Later efforts introduced visual instruction tuning (Jiang et al., 2024; Yu et al., 2024a; Zhu et al., 2023a) and external expert modules for auxiliary verification and logical constraints during generation (Chen et al., 2024b; Wu et al., 2024; Zhao et al., 2024). Meanwhile, substantial progress has also been made in hallucination evaluation and detection, with fine-grained benchmarks and diagnostic datasets proposed to assess object-level faithfulness and relational consistency (Li et al., 2023; Wang et al., 2023). Building on these advances, an increasing number of training-free methods have been proposed to mitigate hallucination in real-time deployments. Representative directions included decoding-time interventions for improving reasoning fidelity (Huang et al., 2024), attention modulation (Tu et al., 2026) for restoring proper token-level correlations, and hidden-layer adjustments for enhancing visual encoding (Zou et al., 2025). Although these approaches have shown promising results, the visual mechanisms underlying hallucination remain insufficiently understood. In particular, it is still unclear how MLLMs perceive visual content and where visual evidence becomes distorted during generation. This gap motivates us to investigate hallucination from the perspective of visual information processing and develop lightweight interventions that im-

prove visual grounding at the root level.

3. Uncover The Core Reasons Behind Hallucinations

In this section, we begin by delineating the fundamental operational mechanisms of MLLMs and introduce two pivotal metrics, i.e., Attention Entropy and Visual Attention Ratio, which are employed to probe internal signals and characterize model behavior. Utilizing these metrics, we conduct a comprehensive set of empirical analyses and identify two principal factors that drive hallucinations, including spatial inconsistency across multi-head attention and temporal attenuation of visual attention. Our case studies are based on a randomly sampled subset of the COCO validation split (Lin et al., 2014), primarily focusing on LLaVA-1.5-7B (Liu et al., 2024a), with other models evaluated as well.

3.1. Preliminary

Notation. We consider a multimodal large language model $\mathcal{F}_\theta$ parameterized by θ, whose architecture is composed of a text embedding component, a vision encoder, a cross-modal alignment module, and a Transformer-based decoder. Given a sequence of input image tokens $\{v_1, \ldots, v_n\}$, together with prompt tokens $\{t_1, \ldots, t_m\}$, followed by the $t-1$ previously generated tokens $\{y_1, \ldots, y_{t-1}\}$, the model produces the current output token y_t in an autoregressive manner. Define L as the total number of Transformer layers, each comprising H attention heads. At layer $l \in [L]$, the attention weights in head $h \in [H]$ are represented as $\boldsymbol{A}_t^{(l,h)} \in \mathbb{R}^{a_t \times a_t}$, where $a_t = n + m + t - 1$. Besides, let s and e denote the start and end indices of the image tokens.

Attention Entropy. At layer l, we obtain the head-averaged attention distribution over the n visual patches for the currently generated token by $\boldsymbol{A}_t^l = \frac{1}{H}\sum_{h=1}^{H} \boldsymbol{A}_t^{(l,h)}$. Then, we define the attention entropy as below:

$$\mathcal{D}_t^l \;=\; -\sum_{i=1}^{n} \boldsymbol{A}_t^l(a_t, i)\log(\boldsymbol{A}_t^l(a_t, i) + \epsilon), \qquad (1)$$

where $\boldsymbol{A}_t^l(a_t, i)$ denotes the head-averaged attention weight from the current decoding position a_t to the i-th visual token, and $\epsilon > 0$ is a small constant for numerical stability. The entropy $\mathcal{D}_t^l$ captures the dispersion of attention across image patches. A smaller $\mathcal{D}_t^l$ corresponds to a concentrated attention distribution over a limited subset of visual patches, indicating a focused and selective visual association for the current token. Conversely, a larger entropy value reflects a spatially dispersed attention pattern, suggesting diffuse and uncertain visual grounding.

Visual Attention Ratio. Regarding the t-th output token y_t at layer l and head h, we quantify its reliance on visual information by measuring the total attention mass assigned to the visual prefix:

$$\mathrm{VAR}_t^{(l,h)} \;\triangleq\; \sum_{i=1}^{n} \boldsymbol{A}_t^{(l,h)}(a_t, i). \qquad (2)$$

$\mathrm{VAR}_t^{(l,h)}$ represents the proportion of attention devoted to visual tokens during the generation of y_t. A larger value indicates stronger dependence on image information, whereas a smaller value suggests that the token is generated predominantly from textual context.

3.2. Finding 1: Spatial Inconsistency in Multi-head Attention indicates Hallucination

Inspired by humans' stable visual fixation on salient regions, visually grounded tokens in MLLMs are expected to elicit concentrated attention with strong cross-head spatial consistency. In contrast, hallucinated tokens tend to be associated with diffuse visual attention and spatial misalignment across heads. To empirically examine this distinction, we extract the attention entropy $\mathcal{D}_t^l$ at each decoding step. Following the CHAIR criteria (Rohrbach et al., 2018), tokens are categorized into *Real* and *Hallucinated* groups, and their entropy distributions are compared through statistical significance testing. As illustrated in Fig. 3.2, grounded tokens, such as *computer*, exhibit sharply localized attention and strong spatial alignment across heads, whereas hallucinated tokens display dispersed attention patterns and pronounced head-wise divergence. Quantitatively, hallucinated tokens consistently yield higher H_t^l than real tokens. Moreover, the entropy distribution of real tokens is unimodal, whereas hallucinated tokens exhibit a bimodal, long-tailed distribution, indicating diffuse and unstable visual perception.

Building on prior observations that hallucinated tokens are often accompanied by heightened cross-head disagreement, we further characterize the role of head-wise inconsistency by examining its statistical association with the confidence of token prediction. During decoding, we normalize $\boldsymbol{A}_t^{(l,h)}(a_t, i)$ to a probability vector $P_t^{(l,h)}$ with $\sum_{i=1}^{n} P_t^{(l,h)}(a_t, i) = 1$. Then, we compute the head-averaged probability distribution as $P_t^l = \frac{1}{H}\sum_{h=1}^{H} P_t^{(l,h)}$. From an information-theoretic perspective, head-wise inconsistency is quantified using the Kullback–Leibler divergence between each individual distribution and P_t^l:

$$D_{kl} = \frac{1}{H}\sum_{h=1}^{H} P_t^{(l,h)}\big(\log P_t^{(l,h)} - \log P_t^l\big), \qquad (3)$$

where D_{kl}^l provides a principled measure of the extent to which individual attention heads deviate from the collective attention pattern, with larger values indicating weaker cross-head agreement. As for quantifying token-level uncertainty,

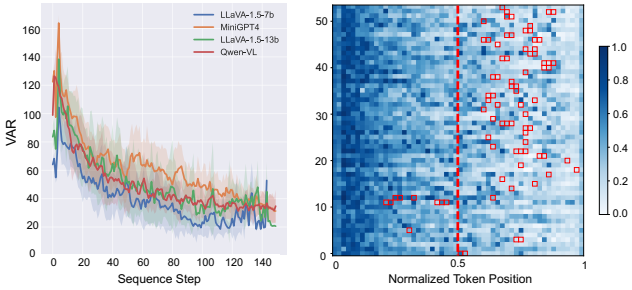


Figure 3. Temporal Fading of Visual Attention. **(Left)** VAR decreases as generation proceeds across multiple MLLMs, indicating progressive weakening of visual grounding in long-form responses. The shaded area represents the standard deviation band, evaluated on the COCO dataset. **(Right)** Heatmap shows VAR over normalized token position, with red boxes denoting hallucinated tokens.

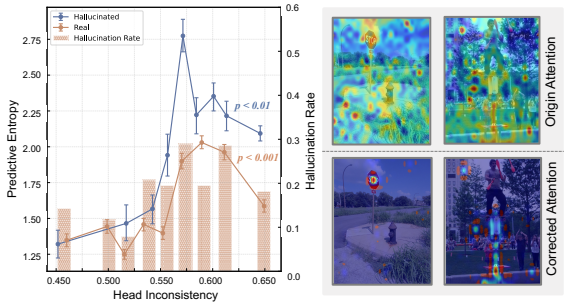


Figure 4. **(Left)** Predictive entopy under varying head inconsistency. Increasing head inconsistency is associated with elevated uncertainty in token prediction, with hallucinated tokens consistently displaying higher entropy than real tokens. The bars denote the hallucination rate, which also increases as head inconsistency grows. **(Right)** Case study for attention correction. Attention maps before (top) and after (bottom) correction during the generation of sign and platform reveal that the correction enhances visual focus on the relevant regions. More cases are shown in Appendix D.7.

we compute the following entropy of the output distribution:

$$E_{tok} = -\sum_{e} p(e) \log p(e), \tag{4}$$

where $p(e)$ denotes the predicted probability of each candidate word e. Then, we sort tokens tokens according to the layer-averaged cross-head divergence D_{kl}^{l}, where $D_{kl} = \frac{1}{L}\sum_{l=1}^{L} D_{kl}^{l}$. The ranked tokens are subsequently divided into quantile-based groups. For each group, we compute the mean token-level predictive entropy E_{tok} separately for real and hallucinated tokens, with bootstrap confidence intervals estimated from 2,000 resampling trials. As illustrated in Fig. 4 (Left), increasing values of D_{kl} tends to coincide with higher E_{tok} for both hallucinated and real tokens. Moreover, correlation tests are conducted to indicate a positive correlation between cross-head divergence and predictive entropy, with small p-values supporting its statistical reliability. These results indicate that stronger head-wise inconsistency is consistently accompanied by lower token prediction confidence and reduced reasoning accuracy.

3.3. Finding 2: Temporal Fading of Visual Attention indicates Hallucination

Human's long-horizon reasoning relies on sustained attention to maintain coherent progress. However, in long-sequence visual question answering tasks performed by MLLMs, attention directed to visual regions typically diminishes as decoding advances (Tu et al., 2026). To quantify this effect, we prompt the model to generate long-form responses while recording layer-wise attention distributions at each decoding step for VAR computation. These values are then averaged across layers to produce a VAR profile that tracks visual attention throughout the sequence of yielded tokens. As illustrated in Fig. 3 (Left), the resulting mean VAR exhibits a consistent downward trajectory during generation across all model backbones, which reflects a progressively reduced reliance on visual evidence as decoding proceeds.

Building on the preceding exploration, we further investigate the interplay between the attenuation of visual attention and hallucination during long-form generation. As decoding unfolds, the model gradually shifts focus away from visual inputs and increasingly relies on linguistic priors, increasing the likelihood of generating hallucinated tokens. To quantitatively reveal this inherent association, we track the occurrence of real and hallucinated tokens during decoding, linking each token to its corresponding VAR value. Outputs are normalized by total length to account for varying response lengths, and tokens are positioned according to their relative location along the normalized trajectory. The resulting patterns are visualized in Fig. 3 (Right), with color intensity representing VAR magnitude, which shows that hallucinated tokens predominantly cluster in low-VAR regions and emerge in the later stages of generation. In particular, tokens falling within the lowest 30% of VAR values exhibit a markedly higher incidence of hallucination compared to the remainder. Taken together, these observations uncover a clear statistical association between visual blur and hallucination, suggesting that sustained reduction of visual grounding during extended responses coincides with increased hallucination risk.

4. Object Hallucination Mitigation

4.1. Head-Level Attention Distraction Correction

As established in the Section 3.2, the emergence of hallucinations primarily arises from attention distraction, characterized by diffuse attention allocation to irrelevant regions. Consequently, accurate grounding of target tokens requires a highly concentrated attention distribution over salient positions, with multiple heads jointly focusing on genuine visual objects. Therefore, the attention weights $\{A_t^{(l,h)}(a_t, i)\}_{i=1}^{n}$

assigned to image tokens should exhibit both high confidence and cross-head consistency. To this end, we aggregate strong attention signals from multiple heads to identify reliable perceptual cues. Specifically, the attention score matrix $\boldsymbol{B}_t^{(l,h)}$ before softmax operation is extracted by

$$\boldsymbol{B}_t^{(l,h)} = \left(\frac{\boldsymbol{Q}_{l,h}\boldsymbol{K}_{l,h}^{\top}}{\sqrt{d_t}} \right)_t, \tag{5}$$

where $\boldsymbol{Q}_{l,h}$ and $\boldsymbol{K}_{l,h} \in \mathbb{R}^{a_t \times d_t}$ represent the query and key matrices of dimension d_t, respectively. To mitigate visual divergence caused by human-like attention shift, the cross-head mean attention is used as the criterion for identifying faithful image tokens, and the corresponding variance is adopted as a regularization term to penalize insufficient concentration. Moreover, attention heads that manifest pronounced responses to image tokens can offer valuable cues for calibrating attention distributions, where such response intensity is measured by the magnitude of $|\boldsymbol{B}_t^{(l,h)}(a_t, i)|$. Then, the top-$k$ heads in each layer are selected based on the maximum absolute response each head exhibits across all image tokens, forming $\mathcal{H}_k^{(l)} \subseteq \{1, 2, \ldots, H\}$ with $|\mathcal{H}_k^{(l)}| = k$. For image token i, the head-wise mean and variance formulations of the attention scores are given as follows:

$$\begin{cases} Avg^{(l)} = \frac{1}{H} \sum_{h \in \mathcal{H}_k^{(l)}} \left| \boldsymbol{B}_t^{(l,h)}(a_t, i) \right| \\ Var^{(l)} = \frac{1}{H} \sum_{h \in \mathcal{H}_k^{(l)}} \left(\left| \boldsymbol{B}_t^{(l,h)}(a_t, i) \right| - Avg^{(l)} \right)^2. \end{cases} \tag{6}$$

Formally, we adjust the visual attention scores by enriching beneficial information across multiple heads and suppressing inconsistent attention responses:

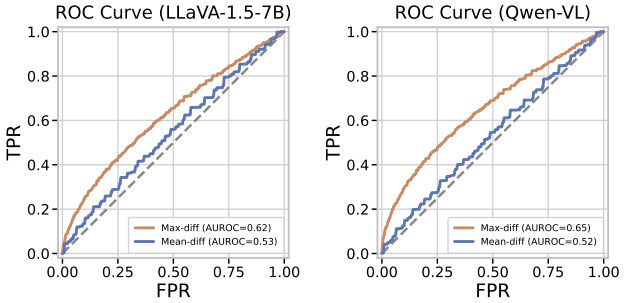
$$\hat{\boldsymbol{B}}_t^{(l,h)}(a_t, i) = \boldsymbol{B}_t^{(l,h)}(a_t, i) + Avg^{(l)} - \alpha Var^{(l)}, \tag{7}$$

where the incremental term $Avg^{(l)} - \alpha Var^{(l)}$ is abbreviated as $\Delta_1^{(l)}$. As illustrated in Fig. 4 (Right), regulating visual attention by Eq. (7) leads to a concentrated head-level attention distribution, which facilitates the model in aligning visual content with textual semantics and identifying the correct generated words.

4.2. Dynamic Historical Attention Enhancement

The progressive degradation of visual perception accompanied by attention distraction, as discussed in the Section 3.3, underscores the importance of reintegrating historical visual attention into the decoding process to correct visual blur and prevent the loss of image content. Accordingly, historical visual attention scores are first stored within a fixed-length history window as below:

$$\mathcal{C}_t = \{\boldsymbol{C}_{t-i}^{(l,h)}\}_{i=1}^{\mathcal{W}_t}, \quad \boldsymbol{C}_{t-i}^{(l,h)} = \{\boldsymbol{B}_{t-i}^{(l,h)}(a_t, j)\}_{j=s}^{e}, \tag{8}$$



(a) LLaVA-1.5-7b *(b)* Qwen-VL

Figure 5. Hallucination occurrence with different attention-based measures. The ROC (Receiver Operating Characteristic) curves of two models are presented for hallucination detection.

where $\mathcal{W}_t$ denotes the size of the history window, fixed at 8 to constitute a sufficiently extended temporal context. Subsequently, the historical text token exerting the greatest influence on the decoding process is identified by

$$r_t^{(l,h)} = \arg\max_{j \in \mathcal{W}_t} \boldsymbol{B}_t^{(l,h)}(a_t, a_t - 1 - j). \tag{9}$$

Subsequently, the attention over the visual content conditioned on the most relevant historical text is fused into the current attention score to strengthen the comprehension of the visual information:

$$\tilde{\boldsymbol{B}}_t^{(l,h)}(a_t, j) = \hat{\boldsymbol{B}}_t^{(l,h)}(a_t, j) + \gamma \boldsymbol{C}_{t-1-r_t^{(l,h)}}^{(l,h)}(s + j), \tag{10}$$

where the incremental term $\gamma \boldsymbol{C}_{t-1-r_t^{(l,h)}}^{(l,h)}(s+j)$ is denoted as $\Delta_2^{(l,h)}$. As shown in Appendix D.1, the incorporation of $\Delta_2^{(l,h)}$ effectively strengthens visual attention and mitigates the temporal perceptual decay, which ensures that visual representations remain accurately interpreted over time.

Although the application of Eq. (10) substantially strengthens the model's sensitivity to image tokens, it is crucial to mitigate abnormal generation behaviors caused by excessive visual retracing, while maintaining awareness to necessary textual context. To this end, we propose a dynamic gating mechanism that controls the activation conditions of attention adjustment. Specifically, we compute the maximum attention score over the visual contents as $V_t^{(l,h)} = \max_{j \in (s,e)} \boldsymbol{B}_t^{(l,h)}(a_t, j)$. Similarly, the maximum attention score over the textual regions can be obtained by

$$T_t^{(l,h)} = \max\left(\max_{j \in (0,s)} \boldsymbol{B}_t^{(l,h)}(a_t, j), \max_{j \in [e, a_t)} \boldsymbol{B}_t^{(l,h)}(a_t, j) \right). \tag{11}$$

Based on $V_t^{(l,h)}$ and $T_t^{(l,h)}$, we define the text–visual dominance gap as $D_t^{(l,h)} = T_t^{(l,h)} - V_t^{(l,h)}$. A larger $D_t^{(l,h)}$ indicates stronger reliance on textual context and may correspond to a decoding regime where subsequent token generation is tightly constrained by the existing texts, as manifested

by structural markers or end-of-sequence tokens. As shown in Fig. 5, we investigate the relationship between hallucination occurrence and two attention-based measures, namely the disparity between peak attention over visual and textual regions and the corresponding gap in their average attention. The results indicate that the peak-level measure possesses greater discriminative capability for identifying hallucinations. Consequently, selecting D_t as the gating criterion is well motivated, as it provides an accurate indicator of hallucinations induced by textual inertia. Then, we introduce a gating function $G_t^{(l,h)} = \mathbb{I}(D_t^{(l,h)} < \tau)$ to selectively regulate attention adjustment. Accordingly, in regimes where textual attention outweighs visual perception, reinforcing visual attention helps alleviate hallucination risks arising from textual inertia. However, when textual attention becomes overly dominant, i.e., $D_t^{(l,h)} \geq \tau$, the decoding process should proceed freely to ensure fluent generation while avoiding repetitive outputs. After incorporating the dynamic gating mechanism, the attention modulation scheme for each image token can be expressed as

$$\tilde{\boldsymbol{B}}_t^{(l,h)}(a_t, j) = \boldsymbol{B}_t^{(l,h)}(a_t, j) + G_t^{(l,h)}(\Delta_1^{(l)} + \Delta_2^{(l,h)}). \quad (12)$$

It is noteworthy that we only perform the correction operation in Eq. (12) at the textual compilation layers, i.e., $l \in [l_e, L]$. Prior work has shown that earlier layers primarily function as modality fusion stages, which facilitates cross-modal alignment rather than supports high-level semantic reasoning (Yin et al., 2025; Chen et al., 2025). Guided by this insight, our approach operates attention adjustment only on the intermediate and deeper layers. This design preserves the intrinsic modality alignment established in the shallow layers while calibrating the attention patterns in the reasoning-dominant layers. We further conduct a layer-wise ablation study in Appendix D.3, which confirms that AFIP attains its optimal performance when applied to layers with $l \in [l_e, L]$. In summary, the procedure of our proposed AFIP is presented in Algorithm 1.

Algorithm 1 Attention-Focused Approach for Improved Image Perception (AFIP)

Require: MLLM $\mathcal{F}_\theta$, input image tokens $\{v_1, \ldots, v_n\}$, prompt tokens $\{t_1, \ldots, t_m\}$.
Output: Model response y_t at the decoding step t.
Store historical visual attention scores in the window $\mathcal{C}_t$.
foreach l *in* $[l_e, L]$ **do**
  Identify the the top-k attention heads $\mathcal{H}_k^{(l)}$;
  Compute the mean and variance value of the attention scores by Eq. (6);
  Compute the term $\Delta_1^{(l)}$ by $Avg^{(l)} - \alpha Var^{(l)}$;
  Determine the most reliant text by Eq. (9);
  Compute the term $\Delta_2^{(l,h)}$ by $\gamma \boldsymbol{C}^{(l,h)}_{t-1-r_t^{(l,h)}}(s+j)$;
  Obtain the maximum attention score over the visual and textual region as $V_t^{(l,h)}$ and $T_t^{(l,h)}$, respectively;
  if $T_t^{(l,h)} - V_t^{(l,h)} < \tau$ **then**
    Conduct attention modulation for each layer and attention head by Eq. (12)
  end
end
MLLM $\mathcal{F}_\theta$ decoding, obtain the current token y_t.

5. The Theoretical Discovery of Hallucination

In this section, we offer a theoretical explanation for the emergence of hallucinations. By examining the structural properties of the attention layers and interpreting multi-head attention as diverse information from multiple sources, we demonstrate that inconsistencies in attention distributions across different heads can impair the model's generalization capability, thereby increasing its vulnerability to hallucination behaviors. The following conclusion encapsulates our central theoretical insights:

Theorem 5.1. *Let $\mathcal{F}$ denote the function class induced by a Transformer architecture with self-attention as its core mechanism. The input sequence is $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t]^\top \in \mathbb{R}^{t\times d}$, augmented with a classification token $\boldsymbol{x}_{cls} \in \mathbb{R}^d$ to produce a scalar output. For the h-th head, the value and query–key projections are represented by $\boldsymbol{W}_V^{(h)} \in \mathbb{R}^{d\times d_v}$ and $\boldsymbol{W}_{QK}^{(h)} \in \mathbb{R}^{d\times d}$. The scalar output is given by*

$$g(\boldsymbol{X}) = \boldsymbol{w}^\top \boldsymbol{W}_C^\top \sigma\left(\boldsymbol{W}_O^\top \operatorname{concat}\left(z^{(1)}(\boldsymbol{X}), \ldots, z^{(H)}(\boldsymbol{X})\right)\right), \quad (13)$$

where the updated representation of the h-th head is $z^{(h)}(\boldsymbol{X}) = \boldsymbol{W}_V^{(h)^\top} \boldsymbol{X}^\top \alpha^{(h)}(\boldsymbol{X})$, with attention weights defined by $\alpha^{(h)}(\boldsymbol{X}) = \operatorname{softmax}\left(\boldsymbol{X}\boldsymbol{W}_{QK}^{(h)^\top} \boldsymbol{x}_{cls}\right)$. Assume the parameter norms satisfy $\|\boldsymbol{w}\|_2 \leq B_w$, $\|\boldsymbol{W}_C\|_F \leq B_{WC}$, $\|\boldsymbol{W}_O\|_F \leq B_O$, $\|\boldsymbol{W}_V^{(h)}\|_F \leq B_{WV}^{(h)}$, and $\|\boldsymbol{W}_{QK}^{(h)}\|_F \leq B_{WQK}^{(h)}$, which aligns with standard regularization practices to ensure stable training. Suppose the activation function σ is L_σ-Lipschitz and all input tokens are bound as $\|\boldsymbol{x}_t\|_2 \leq R$. Then, the upper bound of the Rademacher complexity $\hat{\mathfrak{R}}_D(\mathcal{F})$ can be expressed as

$$\hat{\mathfrak{R}}_D(\mathcal{F}) \leq \frac{4B_w B_{WC} B_O L_\sigma R^3}{\sqrt{n}} \sqrt{H\bar{a}^2 + \sum_{h=1}^{H}(a_h - \bar{a})^2}, \quad (14)$$

where $a_h = B_{WV}^{(h)} B_{WQK}^{(h)}$ and $\bar{a} = \frac{1}{H}\sum_{h=1}^{H} a_h$. The upper bound attains its minimum if and only if $a_1 = \cdots = a_H = \bar{a}$, corresponding to uniform contributions from all attention heads. Any deviation from this regime introduces additional dispersion, enlarges the complexity and weakens generalization guarantees. This finding formalizes the principle that maintaining homogeneity across multi-head attention supports optimal generalization performance.

Table 1. CHAIR hallucination evaluation on five models with *max new token* set to 512. † denotes using the greedy decoding strategy. $\Delta\%$ denotes the relative performance improvement with respect to the second-best method.

Methods	LLaVA-1.5-7B			LLaVA-1.5-13B			Shikra-7B			MiniGPT-4-7B			Qwen-VL			Avg.	
	$C_S \downarrow$	$C_I \downarrow$	Precision ↑	$C_S \downarrow$	$C_I \downarrow$	Precision ↑	$C_S \downarrow$	$C_I \downarrow$	Precision ↑	$C_S \downarrow$	$C_I \downarrow$	Precision ↑	$C_S \downarrow$	$C_I \downarrow$	Precision ↑	$C_S \downarrow$	$C_I \downarrow$
Decoding Strategy																	
Greedy	55.4	14.4	74.0	49.7	14.7	71.0	62.0	17.5	70.9	39.4	11.0	80.5	28.2	8.9	84.9	46.9	13.3
Beam	56.2	15.1	74.0	50.4	15.2	76.3	59.2	16.2	74.2	39.2	12.2	81.7	30.0	10.7	83.8	47.0	13.9
OPERA	48.2	13.1	76.9	41.3	14.1	77.2	41.9	13.8	72.1	30.9	11.2	77.5	29.6	9.5	79.4	38.4	12.3
Contrastive Decoding																	
VCD†	54.6	17.3	70.1	57.2	16.1	71.3	56.4	15.5	75.2	41.4	12.6	68.2	45.4	18.1	70.0	51.0	15.9
Hidden states-intervention																	
MemVR	50.4	14.3	74.2	53.2	13.5	75.6	48.8	16.5	70.9	33.6	9.7	83.5	38.2	13.6	74.9	44.8	13.5
Attention-intervention																	
HGAI†	24.6	6.4	86.8	27.1	7.9	85.1	32.8	10.4	81.0	28.0	7.5	85.1	20.9	5.2	89.1	26.7	7.5
Ours†	**16.8**	**4.4**	91.3	**18.4**	**6.4**	89.2	**18.0**	**6.6**	89.8	**17.8**	**7.3**	89.5	**16.6**	**5.0**	91.1	**17.5**	**5.9**
$\Delta\%$	↓7.8%	↓2.0%	↑4.5%	↓8.7%	↓1.5%	↑4.1%	↓14.8%	↓3.8%	↑8.8%	↓10.2%	↓0.2%	↑4.4%	↓4.3%	↓0.2%	↑2.0%	↓9.2%	↓1.6%

Theorem 5.2. *From a generalization analysis perspective, the representation captured at the h-th attention head is $z^{(h)}(\boldsymbol{X}) \in \mathbb{R}^{d_v}$, and the target token to be predicted is denoted as $\boldsymbol{Y}$. After concatenating the representations from all attention heads, the model produces a scalar output through the subsequent network $\psi : \mathbb{R}^{Hd_v} \to \mathbb{R}$, yielding $f(\boldsymbol{X}) = \psi(z(\boldsymbol{X}))$. The training dataset $D = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i) : i \in [n]\}$ is drawn from a distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ denotes the corpus of C words. Besides, the expected risk and empirical risk w.r.t. the training dataset D can be denoted as $R(f) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}[l(f(\boldsymbol{X}), \boldsymbol{Y})]$ and $\widehat{R}_D(f) = \frac{1}{N}\sum_{i=1}^{N} l(f(\boldsymbol{X}_i), \boldsymbol{Y}_i)$, respectively. With probability at least $1-\delta$, we have the following generalization error bound:*

$$R(f) - \widehat{R}_D(f) \le \frac{\widetilde{\mathcal{K}}_1}{n^{1/2}H^{-1/2}} + \frac{\widetilde{\mathcal{K}}_2}{n^{3/4}H^{-1/4}} + \widetilde{\mathcal{K}}_3\sqrt{\frac{\left(-\sum_{h=1}^{H} I(z^{(h)}(\boldsymbol{X}); \boldsymbol{Y}) + \widetilde{\mathcal{K}}_4\right)}{nH}},$$

where $\widetilde{\mathcal{K}}_1 = \mathcal{O}(H)$, $\widetilde{\mathcal{K}}_2 = \mathcal{O}(\sqrt{C\log C})$, $\widetilde{\mathcal{K}}_3 = \mathcal{O}(\sqrt{C}H)$, $\widetilde{\mathcal{K}}_4$ is constant of order $\widetilde{\mathcal{O}}(1)$ as $n, H \to \infty$. Moreover, the generalization error bound becomes tighter as the visual evidence captured by each attention head aligns more closely with the target token. This conclusion suggests that strong generalization emerges when multiple heads maintain highly consistent semantic focus.

6. Experiment

6.1. Experimental Setup

Model Backbones. We assess the efficacy of our approach on four representative MLLMs, i.e., LLaVA-1.5 (Liu et al., 2024a), Shikra (Chen et al., 2023), MiniGPT-4 (Zhu et al., 2023b), and Qwen-VL (Bai et al., 2023b). To probe the influence of model scale, we evaluate both the 7B and 13B variants of LLaVA-1.5. Furthermore, by encompassing models with diverse alignment mechanisms, such as Q-Former-based designs and encoder-based approaches, we can further demonstrate the robustness of our method.

Table 2. POPE hallucination evaluation on five models with *max new token* set to 512. † denotes using the greedy decoding strategy.

Methods	Random		Popular		Adversarial		Average	
	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑
LLaVA-1.5-7B								
Greedy	89.20	89.12	85.77	84.33	79.63	81.26	84.87	84.90
Beam	87.40	87.13	86.27	86.55	81.70	81.86	85.12	85.18
VCD†	87.73	87.16	85.38	85.06	80.88	81.33	84.66	84.52
MemVR	89.37	89.56	86.33	86.88	79.83	81.78	85.18	86.07
HGAI†	89.70	89.42	86.97	86.98	80.50	81.63	85.72	86.01
Ours†	**90.77**	**90.90**	**87.70**	**88.24**	**82.87**	**84.33**	**87.11**	**87.82**
LLaVA-1.5-13B								
Greedy	89.93	90.16	85.97	86.80	81.23	83.29	85.71	86.75
Beam	89.13	89.08	84.80	85.37	82.40	83.80	85.44	86.08
VCD†	87.39	86.55	85.74	85.06	81.92	81.78	85.02	84.46
MemVR	88.20	88.53	87.10	85.33	81.23	82.40	85.51	85.42
HGAI†	89.93	90.10	84.87	85.82	80.67	82.57	85.16	86.16
Ours†	**91.20**	**91.29**	**87.97**	**88.46**	**83.57**	**84.87**	**87.80**	**86.90**
Shikra-7B								
Greedy	84.50	84.76	83.53	83.96	79.67	80.87	82.57	83.20
Beam	84.90	82.75	83.46	84.26	79.93	80.94	82.76	82.65
VCD†	84.50	84.76	83.53	83.96	79.67	80.87	82.79	83.20
MemVR	84.76	85.04	83.86	84.30	79.76	81.01	82.79	83.45
HGAI†	87.06	86.36	84.13	83.77	82.43	82.32	84.54	84.15
Ours†	**87.47**	**87.08**	**86.17**	**85.93**	**87.07**	**86.36**	**86.90**	**86.46**
MiniGPT-4-7B								
Greedy	81.73	79.36	75.50	72.14	74.20	**73.27**	77.14	74.92
Beam	80.63	77.36	**76.40**	74.86	72.00	68.84	76.34	73.69
VCD†	**83.33**	**83.30**	72.80	**75.35**	69.87	73.43	75.33	**77.36**
MemVR	83.00	81.25	75.40	74.97	72.13	72.61	76.84	76.28
HGAI†	80.27	80.87	71.23	74.35	69.20	73.09	73.57	76.10
Ours†	82.17	79.73	76.30	74.74	**74.23**	73.24	**77.57**	75.90

Baselines. Our method is compared with baselines covering different categories of hallucination mitigation strategies. Greedy decoding and beam search serve as standard decoding methods (Sutskever et al., 2014). Additional baselines include OPERA (Huang et al., 2024), a beam-search-based method, HGAI (Jiang et al., 2025), an attention intervention approach, VCD (Leng et al., 2024), a contrastive decoding strategy, and MemVR (Zou et al., 2025), which performs hidden-state intervention. Beam search and OPERA employ a beam size of 5, while VCD, MemVR and HGAI are applied at inference with default hyperparameters.

Benchmark Following (Huang et al., 2024; Jiang et al., 2025), we evaluate our method on complementary benchmarks with distinct task types. We adopt POPE (Li et al., 2023), a standardized protocol for short-horizon binary question answering. For long-form generation, we employ CHAIR (Rohrbach et al., 2018), which measures hallucination severity based on the proportion of generated objects absent from ground-truth annotations. Additional details are provided in Appendix C.

6.2. Comparative Experiment

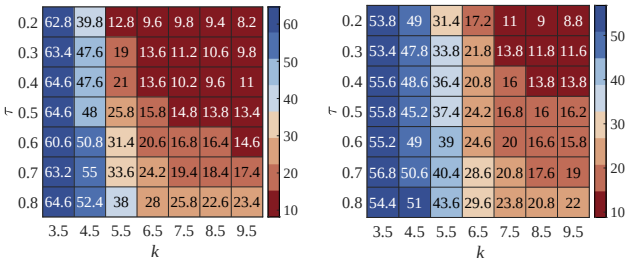
We conduct hallucination evaluations on the CHAIR and POPE benchmarks, and present the corresponding results in Tables 1 and 2. Besides, to evaluate the impact of generated token length, experiments are performed by varying the *max new token* among $\{64, 128, 256\}$, with outcomes reported in Table 3. From these comparative results, we can derive the following observations: (i) Our method substantially reduces hallucination compared with original model responses. As shown in Table 1, relative to the vanilla decoding strategies Greedy and Beam, our approach achieves over 50% improvements on both C_I and C_S metrics across LLaVA-1.5 and Qwen-VL. (ii) Under diverse evaluation criteria, AFIP demonstrates consistently superior performance. It ranks first across all metrics and models on the CHAIR benchmark, and attains leading results on POPE under the Random, Popular, and Adversarial settings for all models except MiniGPT-4. (iii) As reported in Table 3, AFIP remains robust across varying generation lengths. Besides, it achieves more pronounced gains in long-sequence tasks. This finding shows that continuously reinforcing visual perception during long-form generation is essential for achieving stronger resistance to hallucination. (iv) Compared with different baselines, our approach exhibits clear advantages, reflecting its comprehensive consideration of hallucination causes across both spatial and temporal dimensions. Unlike VCD and HGAI, which focus on contrastive decoding and attention control, our method reinjects historical visual information to counteract temporal degradation and maintain sustained visual grounding. Compared with OPERA and MemVR, AFIP corrects attention inconsistencies across multiple heads to enhance visual reliability.

6.3. Parameter Sensitivity Analysis

In our proposed AFIP, there are four hyperparameters in total, i.e., the coefficient α for the variance regularization term, the coefficient γ for the historical attention enhancement term, the Top-k head selection and the threshold τ for the dynamic gating mechanism. As evidenced by the parameter analysis results in Appendix D.5, AFIP is largely insensitive to α and γ. Consequently, we focus on investigating the joint effect of k and τ on the effectiveness of hallucination mitigation. Specifically, k is evaluated over the discrete set $\{3.5, 4.5, 5.5, 6.5, 7.5, 8.5, 9.5\}$, while τ is varied across $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. The results under different parameter combinations are summarized in the heatmap, as shown in Fig. 6. Analysis of these results indicates that optimal performance is generally obtained when k falls within $[5.5, 7.5]$ and τ lies in the range $[0.4, 0.6]$. Careful tuning of k and τ is crucial, as the confidence levels of the top-ranked attention heads and the corresponding threshold estimates can vary substantially across different models.

Table 3. CHAIR hallucination evaluation on LLaVA-1.5-7B and Qwen-VL with varying *max new token* in $\{64, 128, 256\}$. † denotes using the greedy decoding strategy.

Method	*max new token 64*			*max new token 128*			*max new token 256*			Avg.	
	$C_S \downarrow$	$C_I \downarrow$	Precision↑	$C_S \downarrow$	$C_I \downarrow$	Precision↑	$C_S \downarrow$	$C_I \downarrow$	Precision↑	$C_S \downarrow$	$C_I \downarrow$
LLaVA-1.5-7B											
Greedy	22.4	6.4	88.9	49.4	13.1	76.4	50.0	13.1	76.2	40.6	10.9
Beam	19.2	5.9	90.4	51.6	13.8	75.0	52.4	14.5	73.8	41.1	11.4
VCD†	29.2	9.7	84.5	58.8	17.5	70.1	54.6	17.3	71.3	47.5	14.8
MemVR	24.0	7.3	87.4	42.6	14.2	74.7	43.4	14.3	77.2	36.7	11.9
HGAI†	18.6	5.3	90.5	25.2	6.9	70.1	26.6	7.4	86.2	23.5	6.5
Ours†	**13.0**	**4.0**	92.6	**21.6**	**4.9**	89.3	**15.0**	**3.6**	92.0	**16.5**	**4.2**
Qwen-VL											
Greedy	19.6	10.6	86.0	21.8	11.7	84.1	28.2	14.7	79.4	23.2	12.3
Beam	20.2	7.2	88.9	23.2	7.5	85.5	27.2	8.5	82.5	23.5	7.7
VCD†	19.2	9.8	83.6	39.0	15.8	76.5	41.4	15.6	73.8	33.2	13.7
MemVR	25.0	7.7	87.1	27.4	13.2	75.5	32.3	11.6	75.7	28.2	10.8
HGAI†	20.8	10.7	85.8	16.4	7.4	90.2	15.4	6.7	90.6	17.5	8.3
Ours†	**14.4**	**6.6**	90.0	**15.6**	**7.2**	89.7	**12.2**	**6.0**	91.7	**14.1**	**6.6**



(a) LLaVA-1.5-7b *(b)* LLaVA-1.5-13b

Figure 6. Parameter sensitivity analysis of k and τ.

7. Conclusion

This paper proposes an Attention-Focused Approach for Improved Image Perception named AFIP to mitigate hallucinations in MLLMs. Extensive statistical analyses reveal that model hallucinations mainly arise from spatial inconsistencies in multi-head visual attention and temporal degradation in image perception, a pattern analogous to visual blur induced by attention distraction in human perception. We further theoretically demonstrate that attention distraction undermines the model's discriminative capacity during vocabulary selection. Moreover, AFIP is introduced as a plug-and-play module that incurs no additional inference latency, while extensive experiments demonstrate its effectiveness in correcting head-level attention deviations and alleviating the temporal decay of visual attention.

Acknowledgments

This work was partially supported by the NSF for Distinguished Young Scholars under Grant No. 62425607, the Key NSF of China under Grant No. 62136005, Hunan Provincial Natural Science Outstanding Youth Fund under Grant No. 2026JJ20066 and the National Natural Science Foundation of China Grant No. 62376281.

Impact Statement

This paper aims to advance the field of machine learning. As our work focuses on mitigating hallucination in Multimodal large language models (MLLMs), it may give rise to a range of societal implications. Overall, the primary impact of this research is to promote the safer and more reliable deployment of MLLMs across diverse application domains. We do not anticipate that the proposed method introduces additional ethical concerns or poses risks related to privacy.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *NIPS*, 2022.

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

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023b.

Biten, A. F., Gomez, L., and Karatzas, D. Let there be a clock on the beach: Reducing object hallucination in image captioning, 2021. URL https://arxiv.org/abs/2110.01705.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *NIPS*, 2020.

Chen, H., Lin, J., Chen, X., Fan, Y., Dong, J., Jin, X., Su, H., Fu, J., and Shen, X. Multimodal language models see better when they look shallower. In *EMNLP*, 2025.

Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pp. 24185–24198, 2024a.

Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., and Zhou, J. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024b.

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., et al. A survey on multimodal large language models for autonomous driving. In *WACV*, 2024.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., et al. Palm-e: An embodied multimodal language model. 2023.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. Inductive biases and variable creation in self-attention mechanisms. In *ICML*, 2022.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023.

Guo, Z., Yagudin, Z., Lykov, A., Konenkov, M., and Tsetserukou, D. Vlm-auto: Vlm-based autonomous driving assistant with human-like behavior and understanding for complex road scenes. In *2nd International Conference on Foundation and Large Language Models,FLLM*, 2024.

He, Y., Yang, G., Ge, R., Chen, Y., Coatrieux, J.-L., Wang, B., and Li, S. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *CVPR*, 2023.

Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., and Yu, N. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, 2024.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, N. J. B., Chaudhary, V., Som, S., Song, X., and Wei, F. Language is not all you need: Aligning perception with language models. In *NIPS*, 2023.

Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., and Zhang, S. Hallucination augmented contrastive learning for multimodal large language model. In *CVPR*, 2024.

Jiang, Z., Chen, J., Zhu, B., Luo, T., Shen, Y., and Yang, X. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *CVPR*, 2025.

Kawaguchi, K., Deng, Z., Luh, K., and Huang, J. Robustness implies generalization via data-dependent generalization bounds. In *ICML*, 2022.

Kim, J. M., Koepke, A., Schmid, C., and Akata, Z. Exposing and mitigating spurious correlations for cross-modal retrieval. In *CVPR*, 2023.

Koroteev, M. V. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*, 2021.

Kraljevic, Z., Shek, A., Bean, D., Bendayan, R., Teo, J., and Dobson, R. Medgpt: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*, 2021.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, 2024.

Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., and Li, C. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *ACL*, 2020.

Li, Q., Luo, T., Jiang, M., Jiang, Z., Hou, C., and Li, F. Semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18430–18438, 2025a.

Li, Q., Luo, T., and Liao, J. Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20706–20715, 2025b.

Li, Q., Xu, T., Luo, T., Zhong, Y., Li, Y., Zhou, Y., and Hou, C. Theory-driven label-specific representation for incomplete multi-view multi-label learning. *Advances in Neural Information Processing Systems*, 38:121996–122037, 2026.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Lin, T., Zhang, W., Li, S., Yuan, Y., Yu, B., Li, H., He, W., Jiang, H., Li, M., Song, X., Tang, S., Xiao, J., Lin, H., Zhuang, Y., and Ooi, B. C. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *CoRR*, 2025.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *CVPR*, 2024a.

Liu, S., Zheng, K., and Chen, W. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *ECCV*, 2024b.

Lust-Piquard, F. and Pisier, G. Non commutative khintchine and paley inequalities. *Arkiv för matematik*, 1991.

Qu, X., Chen, Q., Wei, W., Sun, J., Liu, D., and Dong, J. Alleviating hallucination in large vision-language models with active retrieval augmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. *EMNLP*, 2018.

Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., and Li, H. Drivelm: Driving with graph visual question answering. In *ECCV*, 2024.

Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *NIPS*, 2014.

Tu, C., Ye, P., Zhou, D., Bai, L., Yu, G., Chen, T., and Ouyang, W. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *ICCV*, 2026.

Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., Pei, Y., Nourbakhsh, A., and Liu, X. Docllm: A layout-aware generative language model for multimodal document understanding. In *ACL*, 2024a.

Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Wen, W., Gong, T., Dong, Y., Yu, S., and Zhang, W. Towards the generalization of multi-view learning: An information-theoretical analysis. *arXiv preprint arXiv:2501.16768*, 2025.

Wu, J., Liu, Q., Wang, D., Zhang, J., Wu, S., Wang, L., and Tan, T. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*, 2024.

Yang, D., Rao, J., Chen, K., Guo, X., Zhang, Y., Yang, J., and Zhang, Y. Im-rag: Multi-round retrieval-augmented generation through learning inner monologues. In *SIGIR*, 2024.

Yin, H., Si, G., and Wang, Z. Lifting the veil on visual information flow in mllms: Unlocking pathways to faster inference. In *CVPR*, 2025.

Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., and Zhuang, Y. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *CVPR*, 2024a.

Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*, 2024b.

Zhao, L., Deng, Y., Zhang, W., and Gu, Q. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv e-prints*, 2024.

Zheng, Y., Gan, W., Chen, Z., Qi, Z., Liang, Q., and Yu, P. S. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*, 2025.

Zhu, B., Niu, Y., Lee, S., Hur, M., and Zhang, H. Debiased fine-tuning for vision-language models by prompt regularization. In *AAAI*, 2023a.

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

Zou, X., Wang, Y., Yan, Y., Lyu, Y., Zheng, K., Huang, S., Chen, J., Jiang, P., Liu, J., Tang, C., et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *ICML*, 2025.

A. Proof of The Theorem 5.1

We begin the proof of Theorem 5.1 by introducing the following lemmas:

Lemma A.1 ((Edelman et al., 2022)). *For vectors $\theta_1, \theta_2 \in \mathbb{R}^p$, we have*

$$\|\mathrm{softmax}(\theta_1) - \mathrm{softmax}(\theta_2)\|_1 \leq 2\|\theta_1 - \theta_2\|_\infty . \tag{15}$$

Lemma A.2 ((Lust-Piquard & Pisier, 1991)). *Consider vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with $\|\cdot\|$ being the associated p-th norm. Let $\epsilon_1, \ldots, \epsilon_n$ be a sequence of independent Rademacher variables. $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space with $\|\cdot\|$ being the associated p-th norm. Let $\epsilon_1, \ldots, \epsilon_n$ be a sequence of independent Rademacher variables. Then, for any $p \geq 1$, we have*

$$\min(\sqrt{p-1}, 1)\left[\sum_{i=1}^n \|\boldsymbol{v}_i\|^2\right]^{\frac{1}{2}} \leq \left[\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum_{i=1}^n \epsilon_i \boldsymbol{v}_i\right\|^p\right]^{\frac{1}{p}} \leq \max(\sqrt{p-1}, 1)\left[\sum_{i=1}^n \|\boldsymbol{v}_i\|^2\right]^{\frac{1}{2}}, \tag{16}$$

and

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum_{i=1}^n \epsilon_i \boldsymbol{v}_i\right\| \geq 2^{-\frac{1}{2}}\left[\sum_{i=1}^n \|\boldsymbol{v}_i\|^2\right]. \tag{17}$$

We adopt the definitions and notation for self-attention and Transformers introduced in (Edelman et al., 2022). Accordingly, we disregard other architectural components of the Transformer, such as normalization and feed-forward networks, and concentrate solely on the role of the attention layer in shaping generalization behavior. The input sequence is denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t]^\top \in \mathbb{R}^{t\times d}$, which contains t d-dimensional tokens. Additionally, for notational convenience, the t-th token of the i-th sample is denoted by $\boldsymbol{x}_t^i$. Suppose there are H attention heads, For the h-th head, the query, key, and value are denoted as $\boldsymbol{W}_Q^{(h)} \in \mathbb{R}^{d\times d_v}$, $\boldsymbol{W}_K^{(h)} \in \mathbb{R}^{d\times d_v}$, $\boldsymbol{W}_V^{(h)} \in \mathbb{R}^{d\times d_v}$, respectively. Moreover, we combine the query and key and write them as $\boldsymbol{W}_{QK}^{(h)} \in \mathbb{R}^{d\times d}$. To enable a principled generalization analysis, we further adopt the standard classification-token (CLS) construction. Specifically, we prepend a special token $\boldsymbol{x}_{cls} \in \mathbb{R}^d$ to the input sequence and define the model output solely based on the representation associated with this token. Concretely, the CLS token serves as a global query that attends to all input tokens through the self-attention mechanism, thereby aggregating sequence-level information into a single vector. As a consequence, the norm of the aggregated representation can be controlled independently of the sequence length. From a functional perspective, introducing the CLS token does not reduce the expressive power of the model for sequence-level tasks. Any attention-based pooling or global aggregation over tokens can be equivalently implemented by an appropriately parameterized CLS query. Instead, the CLS construction internalizes the pooling operation into the attention layer itself, yielding a single vector representation that naturally supports scalar-valued prediction. This scalar-output formulation is crucial for complexity analysis, as it allows the Transformer to be viewed as a real-valued function class, enabling the application of Rademacher complexity and related generalization bounds. Such a setup is standard in practice and is employed in BERT (Devlin et al., 2019). Then, we can obtain the attention weights for the h-th head as

$$\alpha^{(h)}(\boldsymbol{X}) = \mathrm{softmax}\left(\boldsymbol{X}\boldsymbol{W}_{QK}^{(h)^\top}\boldsymbol{x}_{cls}\right) \in \mathbb{R}^t, \tag{18}$$

where the $\mathrm{softmax}$ function is applied row-wise, which ensures that $\alpha_t^{(h)}(\boldsymbol{X}) \geq 0$ and $\sum_{t=1}^T \alpha_t^{(h)}(\boldsymbol{X}) = 1$. Hence, the output corresponding to the h-th head can be expressed as

$$z^{(h)}(\boldsymbol{X}) = \boldsymbol{W}_V^{(h)^\top}\boldsymbol{X}^\top\alpha^{(h)}(\boldsymbol{X}) \in \mathbb{R}^{d_v}. \tag{19}$$

By concatenating the outputs of all heads, we can obtain

$$z(\boldsymbol{X}) = \mathrm{concat}\left(z^{(1)}(\boldsymbol{X}), \ldots, z^{(H)}(\boldsymbol{X})\right) \in \mathbb{R}^{Hd_v}. \tag{20}$$

Subsequently, applying the projection matrix of the multi-head output layer yields the following output:

$$u(\boldsymbol{X}) = \boldsymbol{W}_O^\top z(\boldsymbol{X}) \in \mathbb{R}^d, \tag{21}$$

where $\boldsymbol{W}_O \in \mathbb{R}^{Hd_v\times d}$. Finally, through the combined effect of the activation function and the classification head, the following scalar output is obtained:

$$g(\boldsymbol{X}) = \boldsymbol{w}^\top\boldsymbol{W}_C^\top\sigma(u(\boldsymbol{X})), \tag{22}$$

where $\boldsymbol{w} \in \mathbb{R}^d$ and $\boldsymbol{W}_C \in \mathbb{R}^{d\times d}$. Besides, σ denotes an element-wise activation function.

Next, we impose appropriate bounds on the input magnitude and the norms of the model parameters:

Assumption A.3. There exists a constant $R > 0$, such that, for all input tokens, the embeddings satisfy

$$\|\boldsymbol{x}_t\|_2 \le R, \quad \|\boldsymbol{x}_{cls}\|_2 \le R. \tag{23}$$

Assumption A.4. There exist positive constants B_w, B_{WC}, and B_O, as well as head-specific constants $B_{WV}^{(h)}$ and $B_{WQK}^{(h)}$, such that the corresponding norms of the model parameters are bounded as follows:

$$\begin{cases} \|\boldsymbol{w}\|_2 \le B_w, \quad \|\boldsymbol{W}_C\|_F \le B_{WC}, \quad \|\boldsymbol{W}_O\|_F \le B_O, \\ \|\boldsymbol{W}_V^{(h)}\|_F \le B_{WV}^{(h)}, \quad \|\boldsymbol{W}_{QK}^{(h)}\|_F \le B_{WQK}^{(h)}, \quad h = 1, \dots, H. \end{cases} \tag{24}$$

Assumption A.5. The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is L_σ-Lipschitz and $\sigma(0) = 0$.

Let $\mathcal{F}$ denote the real-valued function class induced by a Transformer architecture with self-attention as its core mechanism. Then, we can get the Rademacher complexity of the scalar one layer Transformer as below:

$$\hat{\mathfrak{R}}_D(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f(\boldsymbol{X}_i)\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \boldsymbol{w}^\top W_C^\top \sigma\left(\boldsymbol{W}_O^\top \operatorname{concat}\left(z^{(1)}(\boldsymbol{X}_i), \dots, z^{(H)}(\boldsymbol{X}_i)\right)\right)\right]. \tag{25}$$

According to the Assumption A.4 and the Cauchy–Schwarz inequality, we can upper bound $\hat{\mathfrak{R}}_D(\mathcal{F})$:

$$\begin{aligned} \hat{\mathfrak{R}}_D(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\|\boldsymbol{w}\|_2\le B_{\boldsymbol{w}},\sigma} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \left\langle \boldsymbol{w}, \boldsymbol{W}_C^\top \sigma\left(\boldsymbol{W}_O^\top \operatorname{concat}\left(z^{(1)}(\boldsymbol{X}_i), \dots, z^{(H)}(\boldsymbol{X}_i)\right)\right)\right\rangle\right] \\ &\le B_{\boldsymbol{w}} \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\|\boldsymbol{W}_C\|_F\le B_{WC},\sigma} \frac{1}{n}\left\|\sum_{i=1}^{n} \epsilon_i \boldsymbol{W}_C^\top \sigma\left(\boldsymbol{W}_O^\top \operatorname{concat}\left(z^{(1)}(\boldsymbol{X}_i), \dots, z^{(H)}(\boldsymbol{X}_i)\right)\right)\right\|_2\right] \\ &\le \frac{B_{\boldsymbol{w}} B_{W_C}}{n} \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\sigma}\left\|\sum_{i=1}^{n} \epsilon_i \sigma\left(\boldsymbol{W}_O^\top \operatorname{concat}\left(z^{(1)}(\boldsymbol{X}_i), \dots, z^{(H)}(\boldsymbol{X}_i)\right)\right)\right\|_2\right]. \end{aligned} \tag{26}$$

According to the Assumption A.5, we can further get

$$\begin{aligned} \hat{\mathfrak{R}}_D(\mathcal{F}) &\le \frac{2B_{\boldsymbol{w}} B_{W_C} L_\sigma}{n} \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\|\boldsymbol{W}_O\|_F\le B_O}\left\|\sum_{i=1}^{n} \epsilon_i \boldsymbol{W}_O^\top \operatorname{concat}\left(z^{(1)}(\boldsymbol{X}_i), \dots, z^{(H)}(\boldsymbol{X}_i)\right)\right\|_2\right] \\ &\le \frac{2B_{\boldsymbol{w}} B_{W_C} B_O L_\sigma}{n} \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup\left\|\sum_{i=1}^{n} \epsilon_i \operatorname{concat}\left(z^{(1)}(\boldsymbol{X}_i), \dots, z^{(H)}(\boldsymbol{X}_i)\right)\right\|_2\right]. \end{aligned} \tag{27}$$

Then, we define S_h as the aggregate sum over all samples at the h-th head, with respect to the function z^h, i,e., $S_h = \sum_{i=1}^{n} \varepsilon_i z^{(h)}(\boldsymbol{X}_i) \in \mathbb{R}^{d_v}$. Since $z(\boldsymbol{X}_i) = \operatorname{concat}\left(z^{(1)}(\boldsymbol{X}_i), \dots, z^{(H)}(\boldsymbol{X}_i)\right)$, we have

$$\left\|\sum_{i=1}^{n} \varepsilon_i z(\boldsymbol{X}_i)\right\|_2^2 = \sum_{h=1}^{H} \|S_h\|_2^2. \tag{28}$$

Based on Eq. (27), we arrive at the following inequality:

$$\sup\left\|\sum_{i=1}^{n} \varepsilon_i z(X_i)\right\|_2 \le \left(\sum_{h=1}^{H}\left(\sup\|S_h\|_2^2\right)\right)^{1/2}. \tag{29}$$

Consequently, we further obtain the following bound on $\hat{\mathfrak{R}}_D(\mathcal{F})$:

$$\hat{\mathfrak{R}}_D(\mathcal{F}) \le \frac{2B_w B_{WC} B_O L_\sigma}{n} \mathbb{E}_\varepsilon\left(\sum_{h=1}^{H}\left(\sup\|S_h\|_2^2\right)\right)^{1/2}. \tag{30}$$

Regarding the head-specific term S_h, we obtain the following estimate:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup\|S_h\|_2\right] &= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup\left\|\sum_{i=1}^{n}\varepsilon_i z^{(h)}\left(\boldsymbol{X}_i\right)\right\|_2\right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup\left\|\sum_{i=1}^{n}\varepsilon_i \boldsymbol{W}_V^{(h)^\top}\boldsymbol{X}_i^\top\alpha^{(h)}(\boldsymbol{X}_i)\right\|_2\right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup\left\|\sum_{i=1}^{n}\varepsilon_i \boldsymbol{W}_V^{(h)^\top}\boldsymbol{X}_i^\top\operatorname{softmax}\left(\boldsymbol{X}_i\boldsymbol{W}_{QK}^{(h)^\top}\boldsymbol{x}_{cls}\right)\right\|_2\right] \\
&\leq B_{WV}^{(h)}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\|\boldsymbol{W}_{QK}^{(h)}\|_F\leq B_{WQK}^{(h)}}\left\|\sum_{i=1}^{n}\varepsilon_i\boldsymbol{X}_i^\top\operatorname{softmax}\left(\boldsymbol{X}_i\boldsymbol{W}_{QK}^{(h)^\top}\boldsymbol{x}_{cls}\right)\right\|_2\right]
\end{aligned} \tag{31}$$

Using the inequality $\|\boldsymbol{Q}\boldsymbol{x}\| \leq \|\boldsymbol{Q}\|_{2,\infty}\|\boldsymbol{x}\|_1$, we can get

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup\|S_h\|_2\right] &\leq B_{WV}^{(h)}\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{\boldsymbol{W}_{QK}^{(h)}\leq B_{WQK}^{(h)}}\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{X}_i^\top\right\|_{2,\infty}\left\|\operatorname{softmax}\left(\boldsymbol{X}_i\boldsymbol{W}_{QK}^{(h)^\top}\boldsymbol{x}_{cls}\right)\right\|_1\right] \\
&\leq B_{WV}^{(h)}\sup_{\boldsymbol{W}_{QK}^{(h)}\leq B_{WQK}^{(h)}}\left\|\operatorname{softmax}\left(\boldsymbol{X}_i\boldsymbol{W}_{QK}^{(h)^\top}\boldsymbol{x}_{cls}\right)\right\|_1\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{X}_i^\top\right\|_{2,\infty}\right].
\end{aligned} \tag{32}$$

By applying the Lemma A.2, the following result is derived:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup\|S_h\|_2\right] &\leq 2B_{WV}^{(h)}\sup_{\boldsymbol{W}_{QK}^{(h)}\leq B_{WQK}^{(h)}}\left\|\boldsymbol{X}_i\boldsymbol{W}_{QK}^{(h)^\top}\boldsymbol{x}_{cls}\right\|_\infty\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{X}_i^\top\right\|_{2,\infty}\right] \\
&= 2B_{WV}^{(h)}\sup_{\boldsymbol{W}_{QK}^{(h)}\leq B_{WQK}^{(h)}}\max_t\left\|\boldsymbol{x}_t^{i^\top}W_{QK}^{(h)^\top}\boldsymbol{x}_{cls}\right\|_2\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{X}_i^\top\right\|_{2,\infty}\right] \\
&\leq 2B_{WV}^{(h)}\sup_{\boldsymbol{W}_{QK}^{(h)}\leq B_{WQK}^{(h)}}\max_t\left\|W_{QK}^{(h)}\boldsymbol{x}_t^i\right\|_2\|\boldsymbol{x}_{cls}\|_2\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{X}_i^\top\right\|_{2,\infty}\right] \\
&\leq 2B_{WV}^{(h)}B_{WQK}^{(h)}\max_t\left\|\boldsymbol{x}_t^i\right\|_2\|\boldsymbol{x}_{cls}\|_2\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{X}_i^\top\right\|_{2,\infty}\right].
\end{aligned} \tag{33}$$

According to the Assumption A.5, the result of Eq. (33) can be further scaled to

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup\|S_h\|_2\right] &\leq 2B_{WV}^{(h)}B_{WQK}^{(h)}R^2\max_t\mathbb{E}_{\boldsymbol{\epsilon}}\left[\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{X}_i^\top\right\|_{2,\infty}\right] \\
&= 2B_{WV}^{(h)}B_{WQK}^{(h)}R^2\max_t\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{x}_t^i\right\|_2.
\end{aligned} \tag{34}$$

By leveraging the Jensen's Inequality and Lemma A.2, we obtain the following bound on S_h for a single head:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup\|S_h\|_2\right] &\leq 2B_{WV}^{(h)}B_{WQK}^{(h)}R^2\max_t\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left\|\sum_{i=1}^{n}\epsilon_i\boldsymbol{x}_t^i\right\|_2^2\right]^{\frac{1}{2}} \\
&\leq 2B_{WV}^{(h)}B_{WQK}^{(h)}R^2\max_t\left[\sum_{i=1}^{n}\left\|\boldsymbol{x}_t^i\right\|_2^2\right]^{\frac{1}{2}} \\
&\leq 2B_{WV}^{(h)}B_{WQK}^{(h)}R^3\sqrt{n}.
\end{aligned} \tag{35}$$

Since the square root function is concave, we have

$$\mathbb{E}_\varepsilon \left(\sum_{h=1}^{H} \left(\sup \|S_h\|_2^2 \right) \right)^{1/2} \leq \left(\mathbb{E}_\varepsilon \sum_{h=1}^{H} \left(\sup \|S_h\|_2^2 \right) \right)^{1/2} = \left(\sum_{h=1}^{H} \mathbb{E}_\varepsilon \left(\sup \|S_h\|_2^2 \right) \right)^{1/2}. \tag{36}$$

Based on Eqs. (30) and (36), we have the following bound on $\hat{\mathfrak{R}}_D(\mathcal{F})$:

$$\begin{aligned} \hat{\mathfrak{R}}_D(\mathcal{F}) &\leq \frac{2B_w B_{WC} B_O L_\sigma}{n} \left(\sum_{h=1}^{H} \mathbb{E}_\varepsilon \left(\sup \|S_h\|_2^2 \right) \right)^{1/2} \\ &\leq \frac{2B_w B_{WC} B_O L_\sigma}{n} \left(\sum_{h=1}^{H} \left(2B_{WV}^{(h)} B_{WQK}^{(h)} R^3 \sqrt{n} \right)^2 \right)^{1/2} \\ &\leq \frac{4B_w B_{WC} B_O L_\sigma R^3}{\sqrt{n}} \left(\sum_{h=1}^{H} \left(B_{WV}^{(h)} B_{WQK}^{(h)} \right)^2 \right)^{1/2}. \end{aligned} \tag{37}$$

Since gradient explosion does not occur during normal training, the norms of the model parameters are bounded and admit finite upper bounds, while not simultaneously collapsing to zero. Therefore, we denote $a_h = B_{WV}^{(h)} B_{WQK}^{(h)}$, and the mean value of a_h necessarily lies between the minimum value and the maximum value. Then, we can obtain the following equality:

$$\sum_{h=1}^{H} a_h^2 = \sum_{h=1}^{H} (a_h - \bar{a})^2 + H\bar{a}^2, \quad \bar{a} = \frac{1}{H} \sum_{h=1}^{H} a_h. \tag{38}$$

Thus, the following result holds:

$$\left(\sum_{h=1}^{H} a_h^2 \right)^{1/2} \geq \sqrt{H}\bar{a}, \tag{39}$$

where the condition for the equality to hold is that $a_1 = \cdots = a_H = \bar{a}$. Therefore, we can get the final bound on $\hat{\mathfrak{R}}_D(\mathcal{F})$:

$$\hat{\mathfrak{R}}_D(\mathcal{F}) \leq \frac{4B_w B_{WC} B_O L_\sigma R^3}{\sqrt{n}} \sqrt{H\bar{a}^2 + \sum_{h=1}^{H} (a_h - \bar{a})^2}. \tag{40}$$

Eq. (40) shows that the upper bound of $\hat{\mathfrak{R}}_D(\mathcal{F})$ attains its minimum if and only if $a_1 = \cdots = a_H = \bar{a}$, which corresponds to a high degree of alignment in the parameter scales of the query, key and value projection matrices across different heads. Consequently, increased heterogeneity in the attention distributions among heads leads to a looser complexity upper bound, which in turn is associated with weaker generalization performance and reduced accuracy in predicting the next token.

B. Proof of The Theorem 5.2

We begin the proof of Theorem 5.2 by introducing the following lemmas:

Lemma B.1. *(Kawaguchi et al., 2022) The vector $X = (X_1, \ldots, X_k)$ is defined to follow the multinomial distribution with parameters $p = (p_1, \ldots, p_k)$. Let $\bar{a}_1, \ldots, \bar{a}_k \geq 0$ be fixed such that $\sum_{i=1}^{k} \bar{a}_i p_i \neq 0$. Then, for any $\epsilon > 0$, the following inequality holds:*

$$\mathbb{P}\left(\sum_{i=1}^{k} \bar{a}_i \left(p_i - \frac{X_i}{m}\right) > \epsilon\right) \leq \exp\left(-\frac{m\epsilon^2}{\beta}\right),$$

where $\beta = 2\sum_{i=1}^{k} \bar{a}_i^2 p_i$.

Lemma B.2. *For any $y \in \mathcal{Y}$, if the loss function $l(\cdot, y)$ is L_l-Lipschitz, the following inequality exists:*

$$|l(u, y) - l(v, y)| \leq L_l \|u - v\|_2, \quad \forall u, v \in \mathbb{R}. \tag{41}$$

Lemma B.3. *If the function ψ is L_l-Lipschitz, with respect to the Euclidean norm $\|\cdot\|_2$, the following inequality exists:*

$$|\psi(a) - \psi(b)| \leq L_\psi \|a - b\|_2, \quad \forall a, b \in \mathbb{R}^{d_v}. \tag{42}$$

We consider a Transformer model with multi-head self-attention and a single prediction target. Let $\boldsymbol{X}$ denote the input sequence and $\boldsymbol{Y}$ the target token to be predicted. For each head $h \in \{1, \ldots, H\}$, we define a feature extraction mapping $\phi^{(h)} : \mathcal{X} \to \mathbb{R}^{d_v}$. The output representation of the h-th head is defined as $z^{(h)}(\boldsymbol{X}) = \phi^{(h)}(\boldsymbol{X}) \in \mathbb{R}^{d_v}$. $z^{(h)}(\boldsymbol{X})$ summarizes information extracted from $\boldsymbol{X}$ through the h-th attention mechanism. Since the Transformer fuses the multi-head representations by concatenation, the resulting concatenated representation is denoted by $z(\boldsymbol{X}) = \text{concat}\left(z^{(1)}(\boldsymbol{X}), \ldots, z^{(H)}(\boldsymbol{X})\right) \in \mathbb{R}^{Hd_v}$. By defining the subsequent networks (a composition of linear layers and activation functions) as $\psi : \mathbb{R}^{Hd_v} \to \mathbb{R}$, the model's scalar output is then given by $f(\boldsymbol{X}) = \psi(z(\boldsymbol{X})) \in \mathbb{R}$. Next, from the perspective of model training, we define the training dataset $D = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i) : i \in [n]\}$ drawn from the distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ denotes the training corpus of the model, consisting of C words. The goal of learning is to find a hypothesis $f \in \mathcal{F}$ with good generalization performance by minimizing the loss l on the dataset D. Then, the expected risk and empirical risk w.r.t. the training dataset D can be denoted as $R(f) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}[l(f(\boldsymbol{X}), \boldsymbol{Y})]$ and $\widehat{R}_D(f) = \frac{1}{N}\sum_{i=1}^{N} l(f(\boldsymbol{X}_i), \boldsymbol{Y}_i)$, respectively. The generalization error can be written as

$$\text{Gen}(f) = R(f) - \hat{R}_D(f) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}[l(f(\boldsymbol{X}), \boldsymbol{Y})] - \frac{1}{N}\sum_{i=1}^{N} l(f(\boldsymbol{X}_i), \boldsymbol{Y}_i). \tag{43}$$

Eq. (43) can be expanded as

$$\begin{aligned}\text{Gen}(f) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}\left[l\left(\psi\left(\text{concat}\left(z^{(1)}(\boldsymbol{X}), \ldots, z^{(H)}(\boldsymbol{X})\right)\right), \boldsymbol{Y}\right)\right] \\ - \frac{1}{n}\sum^{n} l\left(\psi\left(\text{concat}\left(z^{(1)}((\boldsymbol{X}_i)), \ldots, z^{(H)}(\boldsymbol{X}_i)\right)\right), \boldsymbol{Y}_i\right).\end{aligned} \tag{44}$$

According to Lemma B.2 and in view of the Lipschitz continuity of $l(\cdot; \boldsymbol{Y})$, it follows that

$$l(\psi(z(\boldsymbol{X})), \boldsymbol{Y}) \leq l(\psi(0), \boldsymbol{Y}) + L_l|\psi(z(\boldsymbol{X})) - \psi(0)|. \tag{45}$$

Since the activation function in ψ is Lipschitz continuous, it follows that ψ is also Lipschitz continuous. Moreover, applying Lemma B.3 yields that

$$|\psi(z(\boldsymbol{X})) - \psi(0)| \leq L_\psi \|z(\boldsymbol{X})\|_2. \tag{46}$$

Furthermore, noting that $z(\boldsymbol{X}) = \text{concat}\left(z^{(1)}(\boldsymbol{X}), \ldots, z^{(H)}(\boldsymbol{X})\right)$, we obtain the following norm relation induced by concatenation:

$$\|z(\boldsymbol{X})\|_2 = \left(\sum_{h=1}^{H} \left\|z^{(h)}(\boldsymbol{X})\right\|_2^2\right)^{1/2} \leq \sum_{h=1}^{H} \left\|z^{(h)}(\boldsymbol{X})\right\|_2. \tag{47}$$

Combining Eqs. (45), (46) and (47), we can obtain the following result:

$$l(\psi(z(\boldsymbol{X})), \boldsymbol{Y}) \leq l(\psi(0), \boldsymbol{Y}) + L_l L_\psi \sum_{h=1}^{H} \left\|z^{(h)}(\boldsymbol{X})\right\|_2. \tag{48}$$

Then, the head-wise surrogate loss is denoted as follows:

$$\hat{l}\left(\psi \circ \phi^{(h)}(\boldsymbol{X}), \boldsymbol{Y}\right) = \frac{1}{H} l(\psi(0), \boldsymbol{Y}) + L_l L_\psi \left\| z^{(h)}(\boldsymbol{X}) \right\|_2, \quad \phi^{(h)}(\boldsymbol{X}) = z^{(h)}(\boldsymbol{X}). \tag{49}$$

Besides, the mean surrogate loss across multiple heads is given by

$$\hat{l}_{\text{avg}}((\boldsymbol{X}, \boldsymbol{Y}); \psi, \phi) = \frac{1}{H} \sum_{h=1}^{H} \hat{l}\left(\psi \circ \phi^{(h)}(\boldsymbol{X}), \boldsymbol{Y}\right). \tag{50}$$

Based on Eqs. (49) and (50), we can get the following equality:

$$H\hat{l}_{\text{avg}}((\boldsymbol{X}, \boldsymbol{Y}); \psi, \phi) = l(\psi(0), \boldsymbol{Y}) + L_l L_\psi \sum_{h=1}^{H} \left\| z^{(h)}(\boldsymbol{X}) \right\|_2. \tag{51}$$

Consequently, we can obtain the following relation:

$$l(\psi(z(\boldsymbol{X})), \boldsymbol{Y}) \leq H\hat{l}_{\text{avg}}((\boldsymbol{X}, \boldsymbol{Y}); \psi, \phi). \tag{52}$$

From Eq. (52), we can get the following bound of the expected risk $R(f)$:

$$R(f) = \mathbb{E}_{(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{X} \times \mathcal{Y}}[l(\psi(z(\boldsymbol{X})), \boldsymbol{Y})] \leq H \cdot \mathbb{E}_{(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{X} \times \mathcal{Y}} \left[\hat{l}_{\text{avg}}((\boldsymbol{X}, \boldsymbol{Y}); \psi, \phi)\right]. \tag{53}$$

Substituting the upper bound in Eq. (53) back into (43), we can obtain:

$$\text{Gen}(f) \leq H\mathbb{E}_{(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{X} \times \mathcal{Y}} \left[\hat{l}_{\text{avg}}((\boldsymbol{X}, \boldsymbol{Y}); \psi, \phi)\right] - \hat{R}_D(f). \tag{54}$$

Next, we add and subtract the same term $H \cdot \frac{1}{n} \sum_{i=1}^{n} \hat{l}_{\text{avg}}(X_i, Y_i)$ on the right-hand side of Eq. (54), and obtain the following expression via algebraic manipulation:

$$\begin{aligned} &H\mathbb{E}_{(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{X} \times \mathcal{Y}} \left[\hat{l}_{\text{avg}}((\boldsymbol{X}, \boldsymbol{Y}); \psi, \phi)\right] - \hat{R}_D(f) \\ &= H\left(\mathbb{E}_{(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{X} \times \mathcal{Y}} \left[\hat{l}_{\text{avg}}((\boldsymbol{X}, \boldsymbol{Y}); \psi, \phi)\right] - \frac{1}{n} \sum_{i=1}^{n} \hat{l}_{\text{avg}}\ (\boldsymbol{X}_i, \boldsymbol{Y}_i)\right) + \underbrace{\left(H\frac{1}{n} \sum_{i=1}^{n} \hat{l}_{\text{avg}}\ (\boldsymbol{X}_i, \boldsymbol{Y}_i) - \hat{R}_D(f)\right)}_{\Delta_D(f)}. \end{aligned} \tag{55}$$

Based on the inequality (52), we can obtain the result $H\hat{l}_{\text{avg}}\ (\boldsymbol{X}_i, \boldsymbol{Y}_i) \geq l(f(\boldsymbol{X}_i), \boldsymbol{Y}_i)$ for each sample point. Therefore, the conclusion $\hat{R}_D(f) \geq 0$ holds. Then, the upper bound of $\text{Gen}(f)$ can be reformulated as

$$\begin{aligned} \text{Gen}(f) &\leq H\left(\mathbb{E}_{(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{X} \times \mathcal{Y}} \left[\hat{l}_{\text{avg}}((\boldsymbol{X}, \boldsymbol{Y}); \psi, \phi)\right] - \frac{1}{n} \sum_{i=1}^{n} \hat{l}_{\text{avg}}\ (\boldsymbol{X}_i, \boldsymbol{Y}_i)\right) \\ &= H\left(\mathbb{E}_{(\boldsymbol{X}, \boldsymbol{Y}) \sim \mathcal{X} \times \mathcal{Y}} \left[\hat{l}_{\text{avg}}((\boldsymbol{X}_i, \boldsymbol{Y}_i); \psi, \phi)\right] - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{H} \sum_{h=1}^{H} \hat{l}\left(\psi \circ \phi^{(h)}(\boldsymbol{X}_i), \boldsymbol{Y}_i\right)\right). \end{aligned} \tag{56}$$

The representation extracted via the multi-head attention mechanism can be expressed as $\boldsymbol{S} = \phi(\boldsymbol{X}) = (\boldsymbol{S}^{(1)}, \ldots, \boldsymbol{S}^{(H)}) = (\phi^{(1)}(\boldsymbol{X}), \ldots, \phi^{(H)}(\boldsymbol{X})) = (z^{(1)}(\boldsymbol{X}), \ldots, z^{(H)}(\boldsymbol{X}))$. Let the row input data $\boldsymbol{X}$ be generated with a hidden category-specific function θ^c by $\boldsymbol{X} = \theta^c(\boldsymbol{y}^c, \boldsymbol{V})$, where $\boldsymbol{y}^c \in \mathcal{Y}$ denotes the c-th word in training corpus, and $\boldsymbol{V} = \{\boldsymbol{V}^{(h)} = (\boldsymbol{V}_1^{(h)}, \ldots, \boldsymbol{V}_d^{(h)})\}_{h=1}^{H} \in \mathbb{R}^{H \times d}$ are nuisance variables. The conditional random variables of $\boldsymbol{X}$ and $\boldsymbol{S}$ given the category $\boldsymbol{Y} = \boldsymbol{y}^c$ are denoted as $\boldsymbol{X}_{\boldsymbol{y}^c}$ and $\boldsymbol{S}_{\boldsymbol{y}^c}$, respectively. For any $\boldsymbol{y}^c \in \mathcal{Y}$, we define the sensitivity $c_\phi^{\boldsymbol{y}^c}$ of the mapping $\phi = \{\phi^{(h)}\}_{h=1}^{H}$ with respect to the nuisance variable $\boldsymbol{V}_i^{(h)}$ is defined as

$$\begin{aligned} c_\phi^{\boldsymbol{y}^c} = \sup_{h \in [H]} \sup_{\boldsymbol{v}_1^{(h)}, \ldots, \hat{\boldsymbol{v}}_i^{(h)}, \ldots, \boldsymbol{v}_d^{(h)}} &|\log(p_{\boldsymbol{s}}(\phi^{(h)} \circ \theta_{\boldsymbol{y}_c}(\boldsymbol{v}_1^{(h)}, \ldots, \boldsymbol{v}_i^{(h)}, \ldots, \boldsymbol{v}_d^{(h)}))) \\ &- \log(p_{\boldsymbol{s}}(\phi^{(h)} \circ \theta_{\boldsymbol{y}_c}\left(\boldsymbol{v}_1^{(h)}, \ldots, \hat{\boldsymbol{v}}_i^{(h)}, \ldots, \boldsymbol{v}_d^{(h)}\right)))|, \end{aligned} \tag{57}$$

where $\theta_{\boldsymbol{y}^c}(\boldsymbol{v}^{(h)}) = \theta(\boldsymbol{y}^c, \boldsymbol{v}^{(h)})$ and $p_{\boldsymbol{s}}(\boldsymbol{s}) = \mathbb{P}(\boldsymbol{S} = \boldsymbol{s})$. Following Eq. (57), the global sensitivity of ϕ is defined as the supremum of its class-wise sensitivities, i.e., $c_\phi = \sup_{c\in[C]} c_\phi^{\boldsymbol{y}^c}$. Let $\mathcal{S}_c^x = \{\theta_{\boldsymbol{y}^c}(\boldsymbol{v}), \phi \circ \theta_{\boldsymbol{y}^c}(\boldsymbol{v}) : \boldsymbol{v} \in \mathcal{V}, \boldsymbol{y}^c \in \mathcal{Y}\}$ denote the complete set of row data and its corresponding head-wise representations. For any $\gamma > 0$, we construct the following category-specific typical representation subset (Wen et al., 2025):

$$\mathcal{S}_{c,\gamma}^x = \left\{ \boldsymbol{X}, \boldsymbol{S} \in \mathcal{S}_c^x : -\log p_{\boldsymbol{s}|\boldsymbol{y}^c}(\boldsymbol{s}) - H(\boldsymbol{S}_{\boldsymbol{y}^c}) \le c_\phi \sqrt{\frac{d\log(\sqrt{nh}/\gamma)}{2}} \right\}. \tag{58}$$

Define the function $h'_{\boldsymbol{y}^c}(\boldsymbol{v}) = -\log p_{\boldsymbol{s}|\boldsymbol{y}^c}(h''_{\boldsymbol{y}^c}(\boldsymbol{v}))$, where $h''_{\boldsymbol{y}^c}(\boldsymbol{v}) = \phi \circ (\theta_{\boldsymbol{y}^c}(\boldsymbol{v}))$. Let $p_{\boldsymbol{v}}(\boldsymbol{v}) = P(\boldsymbol{V} = \boldsymbol{v})$ and $h_{\boldsymbol{y}^c}^{-1}(\boldsymbol{s}) = \{\boldsymbol{v} \in \mathcal{V} : h_{\boldsymbol{y}^c}(\boldsymbol{v}) = \boldsymbol{s}\}$, we can get

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{V}}\left[h'_{\boldsymbol{y}^c}(\boldsymbol{V})\right] &= -\sum_{\boldsymbol{v}\in\mathcal{V}} p_{\boldsymbol{v}}(\boldsymbol{v}) \log p_{\boldsymbol{s}|\boldsymbol{y}^c}\left(h''_{\boldsymbol{y}^c}(\boldsymbol{v})\right) \\
&= -\sum_{\boldsymbol{s}\in\mathcal{S}_c^x} \sum_{\boldsymbol{v}\in h_{\boldsymbol{y}^c}^{-1}(\boldsymbol{s})} p_{\boldsymbol{v}}(\boldsymbol{v}) \log p_{\boldsymbol{s}|\boldsymbol{y}^c}\left(h''_{\boldsymbol{y}^c}(\boldsymbol{v})\right) \\
&= -\sum_{\boldsymbol{s}\in\mathcal{S}_c^x} \left(\sum_{v\in h_{\boldsymbol{y}^c}^{-1}(\boldsymbol{s})} p_{\boldsymbol{v}}(\boldsymbol{v}) \right) \log p_{\boldsymbol{s}|\boldsymbol{y}^c}(\boldsymbol{s}) \\
&= -\sum_{\boldsymbol{s}\in\mathcal{R}_c^x} p_{\boldsymbol{s}|\boldsymbol{y}^c}(\boldsymbol{s}) \log p_{\boldsymbol{s}|\boldsymbol{y}^c}(\boldsymbol{s}) \\
&= H\left(S_{\boldsymbol{y}^c}\right).
\end{aligned}$$

Using McDiarmid's inequality for the function $h'_{\boldsymbol{y}^c}(\boldsymbol{V})$, we derive

$$\mathbb{P}\left(-\log p_{\boldsymbol{s}|\boldsymbol{y}^c}(\boldsymbol{s}) - H(\boldsymbol{S}_{\boldsymbol{y}^c}) \ge \epsilon\right) \le \exp\left(-\frac{2\epsilon^2}{d(c_\phi)^2}\right).$$

Taking $\delta = \exp\left(-\frac{2\epsilon^2}{d(c_\phi)^2}\right)$, we have $\epsilon = c_\phi\sqrt{\frac{d\log(1/\delta)}{2}}$. By setting $\delta = \gamma/\sqrt{nH}$, we can obtain $\mathbb{P}\left(\boldsymbol{X}, \boldsymbol{S} \notin \mathcal{S}_{c,\gamma}^x\right) \le \delta = \frac{\gamma}{\sqrt{nH}}$. Since $\exp(-H(\boldsymbol{S}_{\boldsymbol{y}^c}) - \epsilon) \le p_{\boldsymbol{s}|\boldsymbol{y}^c}(\boldsymbol{s})$, we can further derive the following transformation:

$$\left|\mathcal{S}_{c,\gamma}^x\right| \exp(-H(\boldsymbol{S}_{\boldsymbol{Y}^c}) - \epsilon) = \sum_{\boldsymbol{s}\in\mathcal{S}_{c,\gamma}^x} \exp(-H(\boldsymbol{S}_{\boldsymbol{y}^c}) - \epsilon) \le \sum_{r\in\mathcal{S}_{c,\gamma}^x} p_{\boldsymbol{s}}(\boldsymbol{s}) = 1. \tag{59}$$

From Eq. (59), we have the following result:

$$\left|\mathcal{S}_{c,\gamma}^x\right| \le \exp\left(H(\boldsymbol{S}_{\boldsymbol{y}^c}) + c_\phi\sqrt{\frac{d\log(\sqrt{nH}/\gamma)}{2}}\right). \tag{60}$$

We further define $T^{\boldsymbol{y}^c} = |\mathcal{S}_{c,\gamma}^x|$, where the typical subset $\mathcal{S}_{c,\gamma}^x$ is given by $\mathcal{S}_{c,\gamma}^x = \{(a_{c,1}^x, a_{c,1}^s), \ldots, (a_{c,T}^x, a_{c,T}^s)\}$. To facilitate the subsequent derivations, we denote $\boldsymbol{x}_{\boldsymbol{i}}^{(\boldsymbol{h})}$=$\boldsymbol{X}_i$. Based on this notation, we introduce the following category-specific sets:

$$\begin{aligned}
\mathcal{U}^{\boldsymbol{y}^c} &= \left\{ i \in [N], h \in [H] : \boldsymbol{x}_i^{(h)}, \phi^{(h)}(\boldsymbol{x}_i^{(h)}) \notin \mathcal{S}_{c,\gamma}^x, \boldsymbol{Y}_i = \boldsymbol{y}^c \right\}, \\
\mathcal{U}_k^{\boldsymbol{y}^c} &= \left\{ i \in [N], h \in [H] : \boldsymbol{x}_i^{(h)} = a_k^{c,x}, \phi^{(h)}(\boldsymbol{x}_i^{(h)}) = a_k^{c,s}, \boldsymbol{Y}_i = \boldsymbol{y}^c \right\}.
\end{aligned} \tag{61}$$

Then, we conduct an analysis of the cumulative loss for individual heads:

$$\begin{aligned}
&\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\hat{l}\left(\psi\circ\phi^{(h)}(\boldsymbol{X}_i),\boldsymbol{Y}_i\right)\\
&=\frac{1}{nH}\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\left(\sum_{i,h\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)+\sum_{k=1}^{T^{\boldsymbol{y}^c}}\sum_{i,v\in\mathcal{U}_k^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)\right)\\
&=\frac{1}{nH}\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\sum_{i,h\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)+\frac{1}{nH}\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\sum_{k=1}^{T^{\boldsymbol{y}^c}}\sum_{i,v\in\mathcal{U}_k^{\boldsymbol{y}^c}}\hat{l}(\psi(a_k^{c,s}),\boldsymbol{y}^c)\\
&=\frac{1}{nH}\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\sum_{i,h\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)+\frac{\left|\mathcal{U}_k^{y^c}\right|}{nH}\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\sum_{k=1}^{T^{\boldsymbol{y}^c}}\hat{l}(\psi(a_k^{c,s}),\boldsymbol{y}^c).
\end{aligned}\tag{62}$$

Similarly, the population risk $\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}\left[\hat{l}_{\text{avg}}((\boldsymbol{X},\boldsymbol{Y});\psi,\phi)\right]$ can be decomposed into the following components:

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}\left[\hat{l}_{\text{avg}}((\boldsymbol{X},\boldsymbol{Y});\psi,\phi)\right]\\
&=\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\mathbb{P}(\boldsymbol{Y}=y^c)\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y}^c)\sim\mathcal{X}\times\mathcal{Y}}[\hat{l}_{\text{avg}}((\boldsymbol{X},\boldsymbol{Y});\psi,\phi)|\boldsymbol{Y}=y^c]\\
&=\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\mathbb{P}(\boldsymbol{Y}=y^c,(\boldsymbol{X},\boldsymbol{S})\notin\mathcal{S}_{c,\gamma}^x)\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}[\hat{l}_{\text{avg}}((\boldsymbol{X},\boldsymbol{Y});\psi,\phi)|\boldsymbol{Y}=y^c,(\boldsymbol{X},\boldsymbol{S})\notin\mathcal{S}_{c,\gamma}^x]\\
&+\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\mathbb{P}(\boldsymbol{Y}=y^c,(\boldsymbol{X},\boldsymbol{S})\in\mathcal{S}_{c,\gamma}^x)\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}[\hat{l}_{\text{avg}}((\boldsymbol{X},\boldsymbol{Y});\psi,\phi)|\boldsymbol{Y}=y^c,(\boldsymbol{X},\boldsymbol{S})\in\mathcal{S}_{c,\gamma}^x]\\
&=\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\mathbb{P}(\boldsymbol{Y}=y^c,(\boldsymbol{X}^{(j)},\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}_{c,\gamma}^x)\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y},\boldsymbol{X}^j\sim\boldsymbol{X}}[\hat{l}(\psi\circ\phi^j(\boldsymbol{X}^j),y^c)|\boldsymbol{Y}=y^c,(\boldsymbol{X}^{(j)},\\
&\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}_{c,\gamma}^x]+\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\sum_{k=1}^{T^{\boldsymbol{y}^c}}\mathbb{P}(\boldsymbol{Y}=y^c,\boldsymbol{X}^{(j)}=a_k^{c,x},\phi^j(\boldsymbol{X}^{(j)})=a_k^{c,s})\hat{l}(\psi a_k^{c,s},y^c)).
\end{aligned}\tag{63}$$

Putting Eqs. (62) and (63) back into Eq. (56), we can obtain

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}\left[\hat{l}_{\text{avg}}((\boldsymbol{X}_i,\boldsymbol{Y}_i);\psi,\phi)\right]-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{H}\sum_{h=1}^{H}\hat{l}\left(\psi\circ\phi^{(h)}(\boldsymbol{X}_i),\boldsymbol{Y}_i\right)\\
&\leq\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\mathbb{P}(\boldsymbol{Y}=y^c,(\boldsymbol{X}^{(j)},\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}_{c,\gamma}^x)\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y}^c)\sim\mathcal{X}\times\mathcal{Y},\boldsymbol{X}^j\sim\boldsymbol{X}}[\hat{l}(\psi\circ\phi^j(\boldsymbol{X}^j),y^c)|\boldsymbol{Y}=y^c,(\boldsymbol{X}^{(j)},\\
&\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}_{c,\gamma}^x]-\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\mathbb{P}(\boldsymbol{Y}=y^c,(\boldsymbol{X}^{(j)},\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}_{c,\gamma}^x)\frac{1}{|\mathcal{U}^{\boldsymbol{y}^c}|}\sum_{i,h\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)\\
&+\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\mathbb{P}(\boldsymbol{Y}=y^c,(\boldsymbol{X}^{(j)},\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}_{c,\gamma}^x)\frac{1}{|\mathcal{U}^{\boldsymbol{y}^c}|}\sum_{i,h\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)\\
&-\frac{1}{nH}\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\sum_{i,h\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)\\
&+\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\sum_{k=1}^{T^{\boldsymbol{y}^c}}\mathbb{P}(\boldsymbol{Y}=y^c,\boldsymbol{X}^{(j)}=a_k^{c,x},\phi^j(\boldsymbol{X}^{(j)})=a_k^{c,s})\hat{l}(\psi(a_k^{c,s}),y^c))\\
&-\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\sum_{k=1}^{T^{\boldsymbol{y}^c}}\frac{\left|\mathcal{U}_k^{y^c}\right|}{nH}\hat{l}(\psi(a_k^{c,s}),\boldsymbol{y}^c)
\end{aligned}\tag{64}$$

Let $M_{x,y} = \sup_{(\boldsymbol{X},\boldsymbol{Y})\in\mathcal{X}\times\mathcal{Y}} \sum_{h=1}^{H} \hat{l}\left(f\left(\boldsymbol{X}^{(h)}\right),\boldsymbol{Y}\right)$ represent the maximum attainable loss and $M_{x,y}^{s} = \sup_{i\in[n]} \sum_{h=1}^{H} \hat{l}\left(f(\boldsymbol{X}_i^{(h)}),\boldsymbol{Y}_i\right)$ denote the maximum instance-level loss. The expression in Eq. (64) can be partitioned into three distinct terms. The upper bound corresponding to the first term is derived as follows:

$$
\begin{aligned}
\mathcal{P}_1 &= \sum_{\boldsymbol{y}^c\in\mathcal{Y}} \mathbb{P}(\boldsymbol{Y}=y^c,(\boldsymbol{X}^{(j)},\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}_{c,\gamma}^{x})(\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y}^c)\sim\mathcal{X}\times\mathcal{Y},\boldsymbol{X}^j\sim\boldsymbol{X}}[\hat{l}(\psi\circ\phi^j(\boldsymbol{X}^j),y^c)|\boldsymbol{Y}=y^c,\\
&(\boldsymbol{X}^{(j)},\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}_{c,\gamma}^{x}]-\frac{1}{|\mathcal{U}^{\boldsymbol{y}^c}|}\sum_{i,h\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c))\\
&\le \sum_{\boldsymbol{y}^c\in\mathcal{Y}} \mathbb{P}(\boldsymbol{Y}=y^c,(\boldsymbol{X},\boldsymbol{S})\notin\mathcal{S}_{c,\gamma}^{x})\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y}^c)\sim\mathcal{X}\times\mathcal{Y},\boldsymbol{X}^j\sim\boldsymbol{X}}[\sum_{h=1}^{H}\hat{l}(f(\boldsymbol{X}^{(h)},\boldsymbol{Y}|\boldsymbol{Y}=y^c,(\boldsymbol{X},\boldsymbol{S})\notin\mathcal{S}_{c,\gamma}^{x}]\\
&\le \sum_{\boldsymbol{y}^c\in\mathcal{Y}} \mathbb{P}(\boldsymbol{Y}=y^c)\frac{\gamma}{\sqrt{nH}}\mathcal{M}_{x,y}=\frac{\gamma}{\sqrt{nH}}\mathcal{M}_{x,y}.
\end{aligned}
\tag{65}
$$

Define $p_k^c = \mathbb{P}(\boldsymbol{Y}=y^c,\boldsymbol{X}^{(j)}=a_k^{c,x},\phi^j(\boldsymbol{X}^{(j)})=a_k^{c,s})$ for $k\in[T^{\boldsymbol{y}^c}]$ and $p_{T^{\boldsymbol{y}^c}+1}^c = \mathbb{P}(\boldsymbol{Y}=y^c,(\boldsymbol{X}^{(j)},\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}_{c,\gamma}^{x})$. Moreover, the associated losses are denoted as $b_k^c=\hat{l}(\psi(a_k^{c,s}),y^c)$, where $k\in[T^{\boldsymbol{y}^c}+1]$. Then, we can get the following term:

$$
\mathcal{P}_{3,k}^c = \sum_{t=1}^{T^{\boldsymbol{y}^c}}(p_t^c-\frac{|\mathcal{U}_t^{\boldsymbol{y}^c}|}{nH})b_t^c-(p_k^c-\frac{|\mathcal{U}_k^{\boldsymbol{y}^c}|}{nH})b_k^c. \tag{66}
$$

By applying Lemma A.2, we have the following result for any $\epsilon>0$ and $k\in\left[T^{\boldsymbol{y}^c}\right]$:

$$
\mathbb{P}\left(\mathcal{P}_{3,k}^c\ge\epsilon\right)\le\exp\left(-\frac{nH\epsilon^2}{2\left(\sum_{t=1}^{T^{\boldsymbol{y}^c}}p_t^c\left(b_t^c\right)^2-p_k^c\left(b_k^c\right)^2\right)}\right). \tag{67}
$$

Besides, we can get the following inequality regarding $k=T^{\boldsymbol{y}^c}+1$:

$$
\mathbb{P}\left(p_{T^{\boldsymbol{y}^c}+1}^c-\frac{|\mathcal{U}^{\boldsymbol{y}^c}|}{nH}\ge\epsilon\right)\le\exp\left(-\frac{nH\epsilon^2}{2p_{T^{\boldsymbol{y}^c}+1}^c}\right). \tag{68}
$$

Upon assigning δ to the right-hand sides of Eqs. (67) and (68), respectively, the following variants emerge:

$$
\begin{cases}
\mathbb{P}\left(\mathcal{P}_{3,k}^c\ge\sqrt{\sum_{t=1}^{T^{\boldsymbol{y}^c}}p_t^c\left(b_t^c\right)^2-p_k^c\left(b_k^c\right)^2}\sqrt{\frac{2\log(1/\delta)}{nH}}\right)\le\delta,\ \forall k\in[T^{\boldsymbol{y}^c}]\\
\mathbb{P}\left(p_{T^{\boldsymbol{y}^c}+1}^c-\frac{|\mathcal{U}^{\boldsymbol{y}^c}|}{nH}\ge\sqrt{\frac{2p_{T^{\boldsymbol{y}^c}+1}^c\log(1/\delta)}{nH}}\right)\le\delta.
\end{cases}
\tag{69}
$$

By taking union bounds over all $\boldsymbol{y}^c\in\mathcal{Y}$ ($|\mathcal{Y}|=C$) and $k\in[T^{\boldsymbol{y}^c}]$, we can get

$$
\begin{cases}
\mathbb{P}\left(\mathcal{P}_{3,k}^c\ge\sqrt{\sum_{t=1}^{T^{\boldsymbol{y}^c}}p_t^c\left(b_t^c\right)^2-p_k^c\left(b_k^c\right)^2}\sqrt{\frac{2\log(CT^{\boldsymbol{y}^c}/\delta)}{nH}}\right)\le\delta,\ \forall k\in[T^{\boldsymbol{y}^c}]\\
\mathbb{P}\left(p_{T^{\boldsymbol{y}^c}+1}^c-\frac{|\mathcal{U}^{\boldsymbol{y}^c}|}{nH}\ge\sqrt{\frac{2p_{T^{\boldsymbol{y}^c}+1}^c\log(C/\delta)}{nH}}\right)\le\delta.
\end{cases}
\tag{70}
$$

Then, for any $\delta>0$, with probability at least $1-\delta$, the second term in Eq. (64) admits the following scaling, as a

consequence of Jensen's inequality:

$$\begin{aligned}
\mathcal{P}_2 &= \frac{1}{|\mathcal{U}^{\boldsymbol{y}^c}|}\sum_{y^c\in\mathcal{Y}}\left(\mathbb{P}\left(\boldsymbol{Y}=y^c,(\boldsymbol{X}^{(j)},\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}^x_{c,\gamma}\right)-\frac{|\mathcal{U}^{\boldsymbol{y}^c}|}{nH}\right)\sum_{i,h\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)\\
&\le\sum_{y^c\in\mathcal{Y}}\sqrt{\mathbb{P}(\boldsymbol{Y}=y^c,(\boldsymbol{X}^{(j)},\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}^x_{c,\gamma})}\frac{\sum_{i,j\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)}{|\mathcal{U}^{\boldsymbol{y}^c}|}\sqrt{\frac{2\log(C/\delta)}{nH}}\\
&=\sum_{y^c\in\mathcal{Y}}\sqrt{\mathbb{P}(\boldsymbol{Y}=y^c)}\sqrt{\mathbb{P}(\boldsymbol{X}^{(j)},\phi^j(\boldsymbol{X}^{(j)}))\notin\mathcal{S}^x_{c,\gamma})}\frac{\sum_{i,j\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)}{|\mathcal{U}^{\boldsymbol{y}^c}|}\sqrt{\frac{2\log(C/\delta)}{nH}}\\
&\le\frac{\sqrt{\gamma}}{(nH)^{1/4}}\frac{\sum_{i,j\in\mathcal{U}^{\boldsymbol{y}^c}}\hat{l}(\psi\circ\phi^{(h)}(\boldsymbol{x}_i^{(h)}),\boldsymbol{y}^c)}{|\mathcal{U}^{\boldsymbol{y}^c}|}\sqrt{|\mathcal{Y}|}\sqrt{\sum_{y^c\in\mathcal{Y}}\mathbb{P}(\boldsymbol{Y}=y^c)}\sqrt{\frac{2\log(C/\delta)}{nH}}\\
&\le M^s_{x,y}\frac{\sqrt{\gamma}}{(nH)^{1/4}}\sqrt{\frac{2C\log(C/\delta)}{nH}}.
\end{aligned}\tag{71}$$

Similarly, with probability at least $1-\delta$, we have the following comparison for $\mathcal{P}^c_{3,k}$:

$$\begin{aligned}
\mathcal{P}^c_{3,k}&\le\sqrt{\sum_{t=1}^{T^{\boldsymbol{y}^c}}p_t^y\left(b_t^y\right)^2-p_k^c\left(b_k^y\right)^2}\sqrt{\frac{2\log(CT^{\boldsymbol{y}^c}/\delta)}{nH}}\\
&\le M_{x,y}\sqrt{\sum_{t=1}^{T^{\boldsymbol{y}^c}}p_t^y-p_k^c}\sqrt{\frac{2\log(CT^{\boldsymbol{y}^c}/\delta)}{nH}}\\
&=M_{x,y}\sqrt{\mathbb{P}\left(\boldsymbol{Y}^c=\boldsymbol{y}^c\bigcap(\boldsymbol{X},\boldsymbol{S})\in\mathcal{S}^x_{c,\gamma}\bigcap\left(\boldsymbol{X}^{(j)},\phi^{(j)}\left(\boldsymbol{X}^{(j)}\right)\right)\neq(a_k^{c,x},a_k^{c,s})\right)}\sqrt{\frac{2\log(CT^{\boldsymbol{y}^c}/\delta)}{nH}}\\
&\le M_{x,y}\sqrt{\mathbb{P}(\boldsymbol{Y}^c=\boldsymbol{y}^c)}\sqrt{\frac{2\log(CT^{\boldsymbol{y}^c}/\delta)}{nH}}.
\end{aligned}\tag{72}$$

Next, we can derive the scaling result for the third term:

$$\begin{aligned}
\mathcal{P}_3&=\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\sum_{k=1}^{T^{\boldsymbol{y}^c}}\left(\mathbb{P}(\boldsymbol{Y}=y^c,\boldsymbol{X}^{(j)}=a_k^{c,x},\phi^j(\boldsymbol{X}^{(j)})=a_k^{c,s})-\frac{|\mathcal{U}_k^{\boldsymbol{y}^c}|}{NC}\right)\hat{l}(\psi(a_k^{c,s}),\boldsymbol{y}^c)\\
&=\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\frac{1}{T^{\boldsymbol{y}^c}-1}\sum_{k=1}^{T^{\boldsymbol{y}^c}}\mathcal{P}^c_{3,k}\\
&\le\sum_{\boldsymbol{y}^c\in\mathcal{Y}}\frac{T^{\boldsymbol{y}^c}}{T^{\boldsymbol{y}^c}-1}M_{x,y}\sqrt{\mathbb{P}(\boldsymbol{Y}^c=\boldsymbol{y}^c)}\sqrt{\frac{2\log(CT^{\boldsymbol{y}^c}/\delta)}{nH}}\\
&\le 2\sum_{\boldsymbol{y}^c\in\mathcal{Y}}M_{x,y}\sqrt{\mathbb{P}(\boldsymbol{Y}^c=\boldsymbol{y}^c)}\sqrt{\frac{2\log(CT^{\boldsymbol{y}^c}/\delta)}{nH}}.
\end{aligned}\tag{73}$$

Based on Eq. (60), the property of $T^{\boldsymbol{y}^c}$ can be expressed as

$$T^{\boldsymbol{y}^c}=|\mathcal{S}^x_{c,\gamma}|\le\exp\left(H(\boldsymbol{S}_{\boldsymbol{Y}^c})+c_\phi\sqrt{\frac{d\log(\sqrt{nH}/\gamma)}{2}}\right).\tag{74}$$

Combining Eqs. (73) and (74), we can get the following result by applying Jensen's inequality:

$$\begin{aligned}
\mathcal{P}_3 &\leq 2\mathcal{M}_{x,y} \sum_{\boldsymbol{y}^c \in \mathcal{Y}} \sqrt{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}^c)} \sqrt{\frac{2\left(H(\boldsymbol{S}_{\boldsymbol{Y}^c}) + c_\phi \sqrt{\frac{d\log(\sqrt{nH}/\gamma)}{2}}\right) + 2\log(C/\delta)}{nH}} \\
&\leq 2\mathcal{M}_{x,y}\sqrt{|\mathcal{Y}|} \sqrt{\sum_{y^c \in \mathcal{Y}} \mathbb{P}(\boldsymbol{Y} = y^c)} \sqrt{\frac{2\left(H(\boldsymbol{S}_{\boldsymbol{Y}^c}) + c_\phi \sqrt{\frac{d\log(\sqrt{nH}/\gamma)}{2}}\right) + 2\log(C/\delta)}{nH}} \\
&= 2\sqrt{2C}\mathcal{M}_{x,y} \sqrt{\frac{H(\boldsymbol{S}_{\boldsymbol{Y}^c}) + c_\phi \sqrt{\frac{d\log(\sqrt{nH}/\gamma)}{2}} + \log(C/\delta)}{nH}} \\
&\leq 2\sqrt{2C}\mathcal{M}_{x,y} \sqrt{\frac{H(\boldsymbol{S}|\boldsymbol{Y}^c) + c_\phi \sqrt{\frac{d\log(\sqrt{nH}/\gamma)}{2}} + \log(C/\delta)}{nH}}. \\
&\leq 2\sqrt{2C}\mathcal{M}_{x,y} \sqrt{\frac{H(\boldsymbol{S}) - I(\boldsymbol{S};\boldsymbol{Y}) + c_\phi \sqrt{\frac{d\log(\sqrt{nH}/\gamma)}{2}} + \log(C/\delta)}{nH}}.
\end{aligned} \tag{75}$$

Therefore, with probability at least $1-\delta$, the generalization error bound of our model is

$$\begin{aligned}
&\mathrm{Gen}(f) = R(f) - \hat{R}_D(f) = \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}[l(f(\boldsymbol{X}),\boldsymbol{Y})] - \frac{1}{N}\sum_{i=1}^{N} l(f(\boldsymbol{X}_i),\boldsymbol{Y}_i) \\
&\leq= H\left(\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{X}\times\mathcal{Y}}\left[\hat{l}_{\mathrm{avg}}((\boldsymbol{X}_i,\boldsymbol{Y}_i);\psi,\phi)\right] - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H}\sum_{h=1}^{H}\hat{l}\left(\psi\circ\phi^{(h)}(\boldsymbol{X}_i),\boldsymbol{Y}_i\right)\right) \\
&\leqq \frac{H\gamma}{\sqrt{nH}}\mathcal{M}_{x,y} + M^s_{x,y}\frac{H\sqrt{\gamma}}{(nH)^{1/4}}\sqrt{\frac{2C\log(C/\delta)}{nH}} \\
&+ 2\sqrt{2C}H\mathcal{M}_{x,y}\sqrt{\frac{H(\boldsymbol{S}) - I(\boldsymbol{S};\boldsymbol{Y}) + c_\phi\sqrt{\frac{d\log(\sqrt{nH}/\gamma)}{2}} + \log(C/\delta)}{nH}} \\
&= \frac{\widetilde{\mathcal{K}}_1}{n^{1/2}H^{-1/2}} + \frac{\widetilde{\mathcal{K}}_2}{n^{3/4}H^{-1/4}} + \widetilde{\mathcal{K}}_3\sqrt{\frac{\left(-\sum_{h=1}^{H} I(\boldsymbol{S}^{(h)};\boldsymbol{Y}) + \widetilde{\mathcal{K}}_4\right)}{nH}},
\end{aligned} \tag{76}$$

where

$$\begin{aligned}
\widetilde{\mathcal{K}}_1 &= H\gamma\mathcal{M}_{x,y} \\
\widetilde{\mathcal{K}}_2 &= M^s_{x,y}\sqrt{2\gamma C\log(C/\delta)} \\
\widetilde{\mathcal{K}}_3 &= 2\sqrt{2C}H\mathcal{M}_{x,y} \\
\widetilde{\mathcal{K}}_4 &= H(\boldsymbol{S}) + c_\phi\sqrt{\frac{d\log(\sqrt{nH}/\gamma)}{2}} + \log(C/\delta).
\end{aligned} \tag{77}$$

Based on the term $-\sum_{h=1}^{H} I(\boldsymbol{S}^{(h)}$ in Eq. (76), it follows that constraining the mutual information between each representation extracted by the multi-head attention mechanism and the target $\boldsymbol{Y}$ leads to lower generalization error. Under this condition, the model attains greater precision in retrieving the correct lexical roots from the vocabulary. In other words, greater consistency in the image information attended to by different heads leads to stronger overall classification generalization and facilitates the selection of correct descriptive terms, thereby suppressing hallucination.

C. Details of benchmarks

To supplement the main text, we provide the evaluation details for the two hallucination benchmarks used in our captioning study. We follow the standard setup and evaluate on 500 images randomly sampled from the MSCOCO 2014 validation split, which provides ground-truth annotations for 80 object categories. For CHAIR (Rohrbach et al., 2018), we obtain captions by prompting the MLLM with "Please describe this image in detail" and compute both sentence-level and object-level hallucination rates as

$$C_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, C_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}$$

For POPE (Li et al., 2023), we probe object existence using binary VQA queries of the form "Is [object] in this image?" and append "Please answer yes or no." to enforce binary outputs. Following the protocol, the queried objects are constructed under three sampling strategies (random, popular, and adversarial), and we report results across all sampling options.

D. Additional Results

D.1. Functionality of Dynamic Historical Attention Enhancement

Fig. 7 illustrates the impact of dynamic historical attention enhancement on visual attention retention during long-form generation. In the left panel, we plot VAR on LLaVA-1.5-7B as a function of token position and compare the standard forward propagation with the enhanced variant. Throughout the decoding process, the enhancement mechanism consistently yields higher VAR values, indicating that the model maintains a greater proportion of attention on visual tokens.

On the right, we report VAR across multiple models (LLaVA-1.5-13B, MiniGPT-4, and Qwen-VL) under the same experimental setting. Solid bars represent the original attention mechanism, while the lighter segments denote the VAR gains introduced by the enhancement strategy. The results demonstrate that the proposed method leads to consistent improvements across diverse MLLMs and model scales.

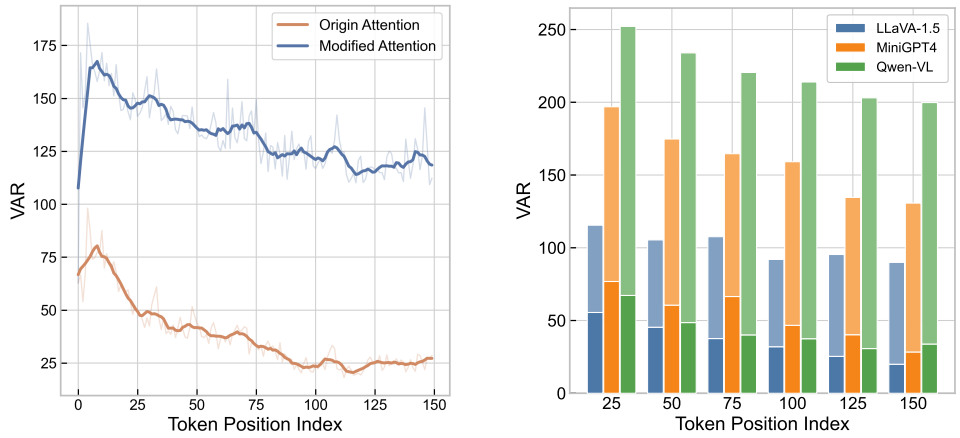


Figure 7. Visualization of visual attention retention with dynamic historical attention enhancement.

D.2. Ablation Studies

The ablation experiments are conducted to deeply investigate the effect of the three crucial modules of AFIP, i.e.,a dynamic gating mechanism driven by the discrepancy between attention allocated to visual and textual content (S_1), the integration of historical visual attention information (S_2), and a head-level attention distraction correction strategy (S_3). Based on the ablation results reported in Table 4, we observe that removing any individual component consistently leads to performance degradation, thereby validating the effectiveness and careful design of the proposed AFIP. Moreover, the attention distraction correction strategy plays a pivotal role in boosting performance, as it mitigates inconsistencies in attention distributions across multiple heads and reinforces the model's ability to focus on the salient visual content. In addition, the dynamic gating mechanism and the integration of historical visual attention further enhance performance by strengthening the model's visual perception capability while preventing excessive attention intervention.

Table 4. Ablation study on components across different MLLMs.

S1	S2	S3	**LLaVA-7b**		**Shikra**		**Qwen-VL**		**LLaVA-13b**		**MiniGPT-4**	
			C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
✗	✓	✓	22.6	8.8	27.4	11.4	22.8	7.6	26.4	11.6	27.3	9.5
✓	✗	✓	19.3	7.2	22.6	9.3	19.1	6.9	22.5	10.2	21.6	8.4
✓	✓	✗	49.6	13.2	46.8	14.3	25.8	8.2	45.3	13.8	32.4	10.2
✓	✓	✓	**16.8**	**4.4**	**18.0**	**6.6**	**16.6**	**5.0**	**18.4**	**6.4**	**17.8**	**7.3**
Beam			56.2	15.1	59.2	16.2	30.0	10.7	50.4	15.2	39.2	12.2
Greedy			55.4	14.4	62.0	17.5	28.2	8.9	49.7	14.7	39.4	11.0

D.3. Layer Selection for Our Method

We apply our method to multiple candidate layer ranges $(0\text{–}32, 5\text{–}18, 18\text{–}32, 0\text{–}5 and 5\text{–}32)$ across five models under a unified experimental setting, and report the corresponding CHAIR results in Table 5. From the results, we observe that the 5–32 configuration (Ours) consistently achieves the lowest $\mathcal{C}_S$ and $\mathcal{C}_I$, and demonstrates a substantial improvement over the Greedy selection strategy. Notably, intervening in the early layers (0–5) clearly degrades performance, which is consistent with our analysis. Specifically, the early layers of MLLMs primarily perform modality aggregation and alignment, as discussed in the main text. These ablation results further validate the practicality of our layer-selection strategy, showing that it can reliably identify the layers that require attention modulation and effectively reduce hallucinations.

Table 5. CHAIR hallucination evaluation across different layer ranges on various models.

Models	**Layers**	**0–32**	**5–18**	**18–32**	**0–5**	**Greedy**	**5–32 (Ours)**
LLaVA-7B	$\mathcal{C}_S \downarrow$	28.9	23.7	36.7	53.2	55.4	16.8
	$\mathcal{C}_I \downarrow$	8.8	7.2	10.9	13.7	14.4	4.4
Shikra	$\mathcal{C}_S \downarrow$	26.1	21.5	42.8	58.6	62.0	18.0
	$\mathcal{C}_I \downarrow$	10.9	8.9	13.7	15.9	17.5	6.6
MiniGPT	$\mathcal{C}_S \downarrow$	25.3	20.5	29.4	37.5	39.4	17.8
	$\mathcal{C}_I \downarrow$	8.5	7.9	9.9	10.6	11.0	7.3
Qwen	$\mathcal{C}_S \downarrow$	21.8	19.5	23.6	26.3	28.2	16.6
	$\mathcal{C}_I \downarrow$	6.9	6.3	7.7	8.4	8.9	5.0

D.4. Case study of Temporal Fading of Visual Attention

The Fig. 8 provides a qualitative case study of *temporal fading of visual attention*. As the generation process unfolds, the proportion of visual attention assigned to successive tokens exhibits a clear overall downward trend. Notably, hallucinated tokens predominantly emerge in regions where visual attention attains local minima, revealing a strong correlation between diminished visual grounding and the onset of hallucinations.

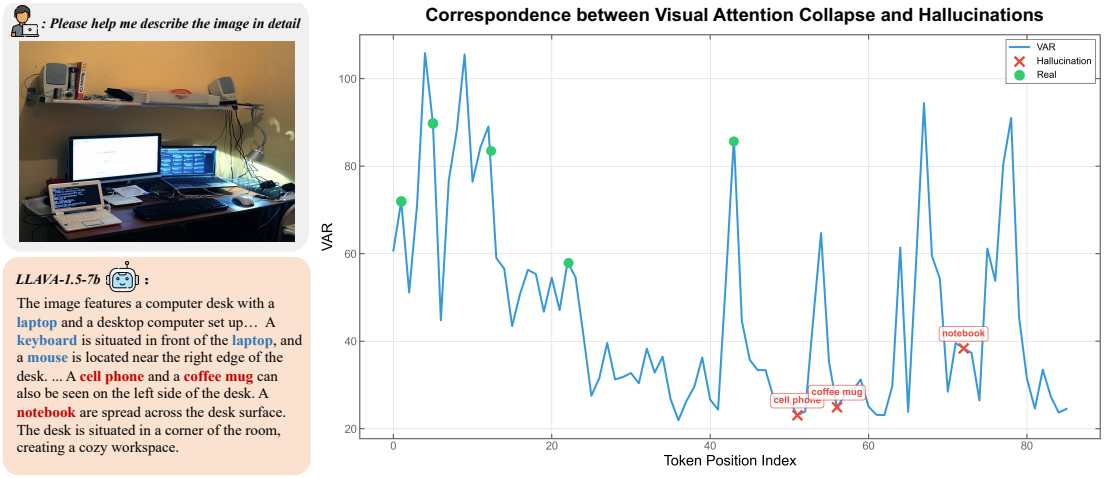
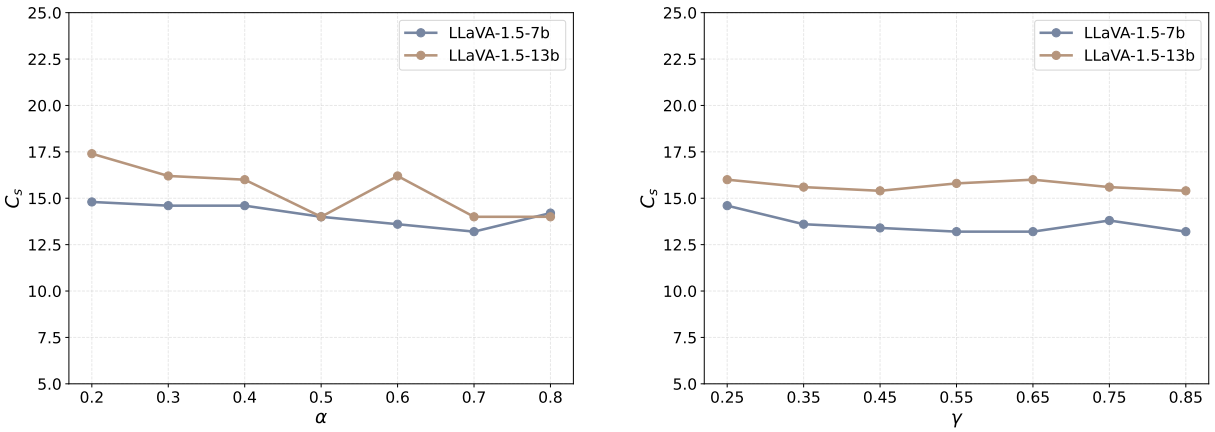


Figure 8. Case study of temporal fading of visual attention.

D.5. Parameter studies on α and γ

We present the parameter studies on α and γ here. From the result, we can observe that our AFIP is not sensitive to α and γ, and adjusting α and γ does not lead to significant improvements in performance.



(a) Performance change across different α

(b) Performance change across different γ

Figure 9. Parameter sensitivity analysis of α and γ.

D.6. Visualization of Attention Spatial Inconsistency

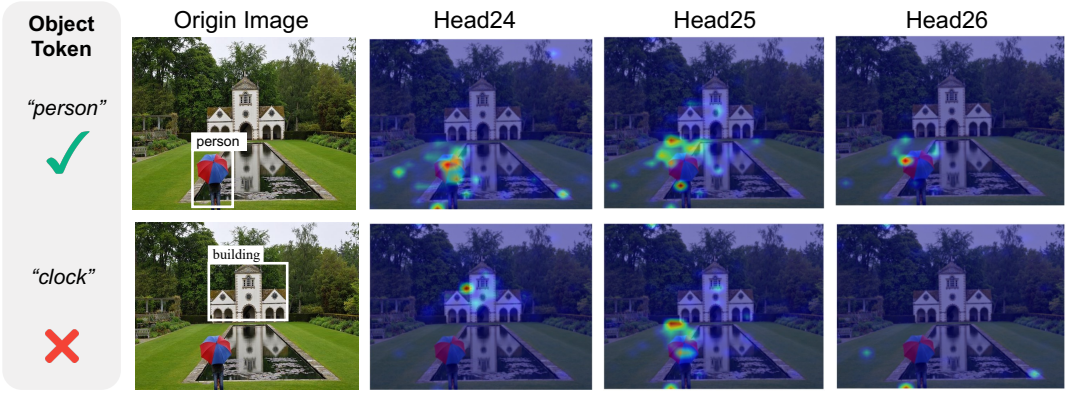
We present additional visualization examples of LLaVA-1.5-7B to demonstrate that attention spatial inconsistency contributes to the generation of hallucinated object tokens.



User: Please help me describe the image in detail.

LLaVA-1.5-7b:

The image features a **person** standing in a garden, holding a red umbrella. The person is positioned near a pond, which is surrounded by a beautiful garden setting. The umbrella is open, providing shelter from the elements. In the background, there is a building with a **clock** tower, adding to the picturesque atmosphere of the scene. The garden is filled with various **potted plants**, some of which are placed near the pond, while others are scattered throughout the area. The combination of the person, the umbrella, the pond, and the garden creates a serene and inviting environment.





User: Please help me describe the image in detail.

LLaVA-1.5-7b:

The image features a dining table with a person sitting at it, enjoying a meal. The person is eating a **pizza**, which is placed on a plate in front of them. There are also a few **carrots** on the table, adding a healthy touch to the meal. Various utensils and items are present on the table, including a fork, a knife, and a **spoon**. There are also two cups, one near the person and another further away. A bottle can be seen on the left side of the table. Another person is partially visible in the background, possibly waiting for their turn to eat or simply observing the scene.

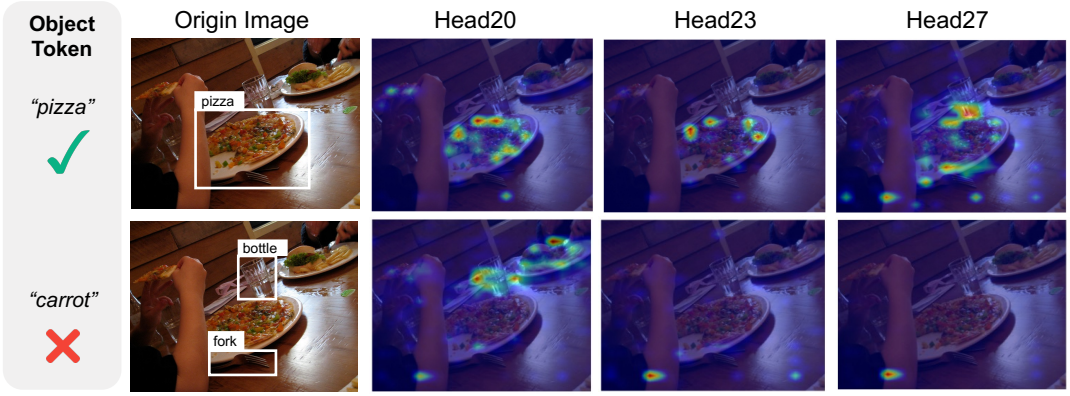


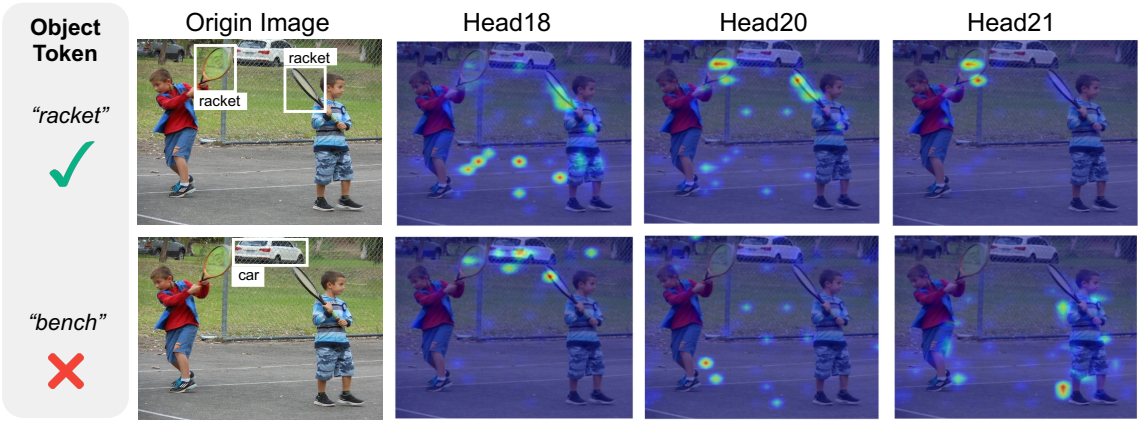
Figure 10. Visualization of attention maps over image for real and hallucinated object tokens.



User: Please help me describe the image in detail.

LLaVA-1.5-7b:

The image features two young boys playing tennis on a court. One boy is holding a tennis **racket**, while the other is holding a **tennis ball.** They are both standing on the court, ready to play. In the background, there are two cars parked, one on the left side and the other on the right side of the court. Additionally, there is a **bench** located near the right side of the court, providing a place for spectators to sit and watch the game.





User: Please help me describe the image in detail.

LLaVA-1.5-7b:

The image depicts a bedroom with a bed situated in the center of the room. The bed is covered with a red blanket, and there is a red curtain hanging over the bed. The room also features a **chair** and a dresser. In addition to the bedroom furniture, there are two **dogs** in the room. One **dog** is located near the left side of the bed, while the other dog is positioned closer to the right side. A **vase** can be seen on the dresser, and a clock is mounted on the wall. The room appears to be a cozy and comfortable space for relaxation and rest.

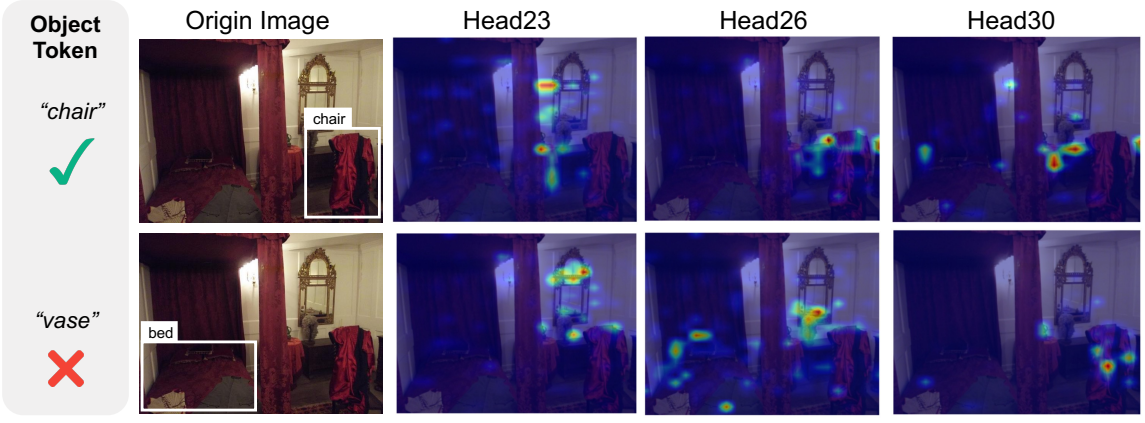


Figure 11. Visualization of attention maps over image for real and hallucinated object tokens.



User: Please help me describe the image in detail.

LLaVA-1.5-7b:

The image features a young **girl** standing in front of a row of motorcycles parked on a dirt road. There are five motorcycles in total, with some of them being parked closer to the girl and others further away. The girl appears to be posing for a picture, possibly with the motorcycles as the main subject. In the background, there is a bus parked on the left side of the scene, and a **truck** can be seen further back on the right side. The scene seems to be a gathering or event involving motorcycles and vehicles.





User: Please help me describe the image in detail.

LLaVA-1.5-7b:

The image features a person riding a horse in a park, with the horse standing next to a fence. The person on the horse appears to be watching a group of children playing baseball on the other side of the fence. There are several children in the scene, with some of them holding **baseball bats** and a **sports ball**. In addition to the horse and the children, there is a **bicycle** parked near the fence, and a **backpack** can be seen placed on the ground. The scene captures a lively atmosphere with people enjoying outdoor activities in the park.

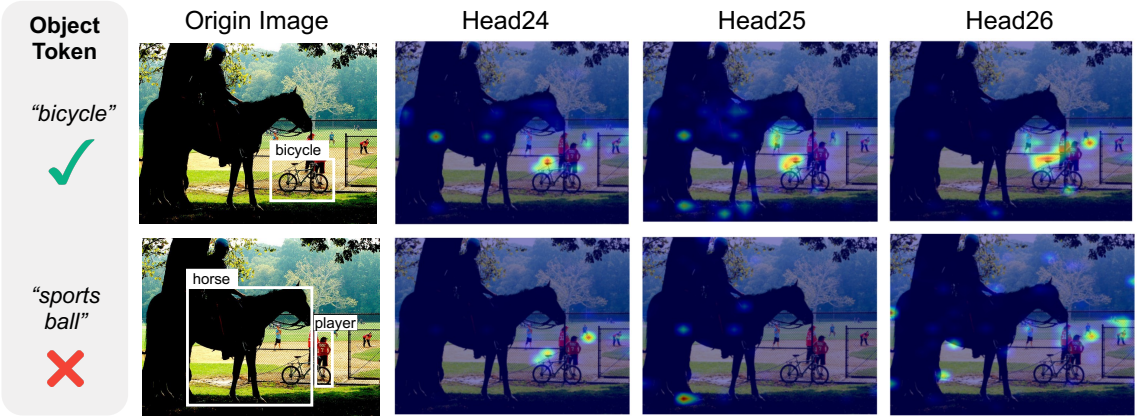


Figure 12. Visualization of attention maps over image for real and hallucinated object tokens.

D.7. Case study for Attention Distraction Correction

We present visualizations of the model's visual attention before and after applying our head-level attention distraction correction method on LLaVA-1.5-7B. All figures depict attention scores from the 20-th layer, which demonstrates that our approach effectively refines the model's focus toward semantically relevant visual regions.

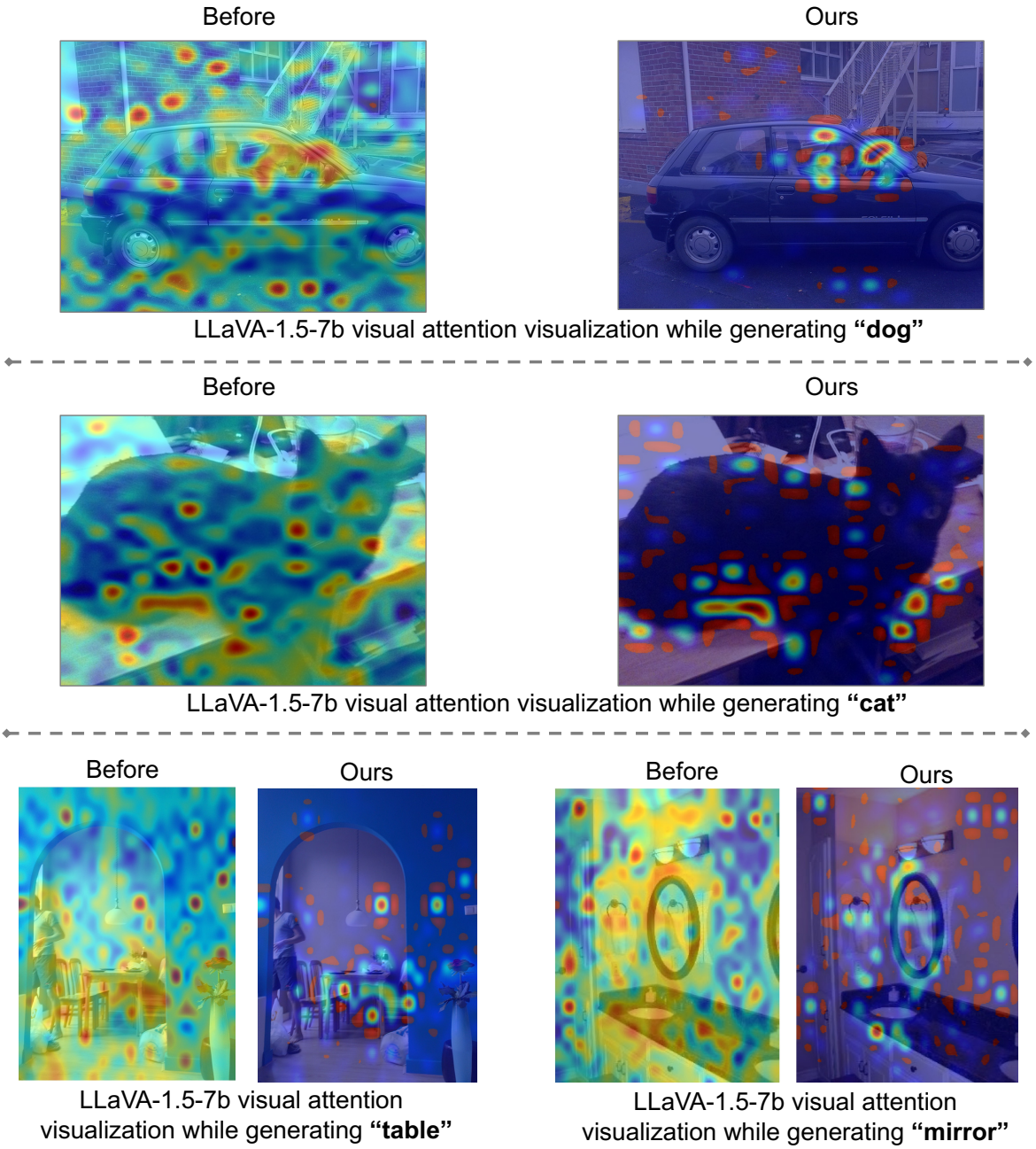


Figure 13. Qualitative comparison of visual attention maps before and after applying Head-Level Attention Distraction Correction.

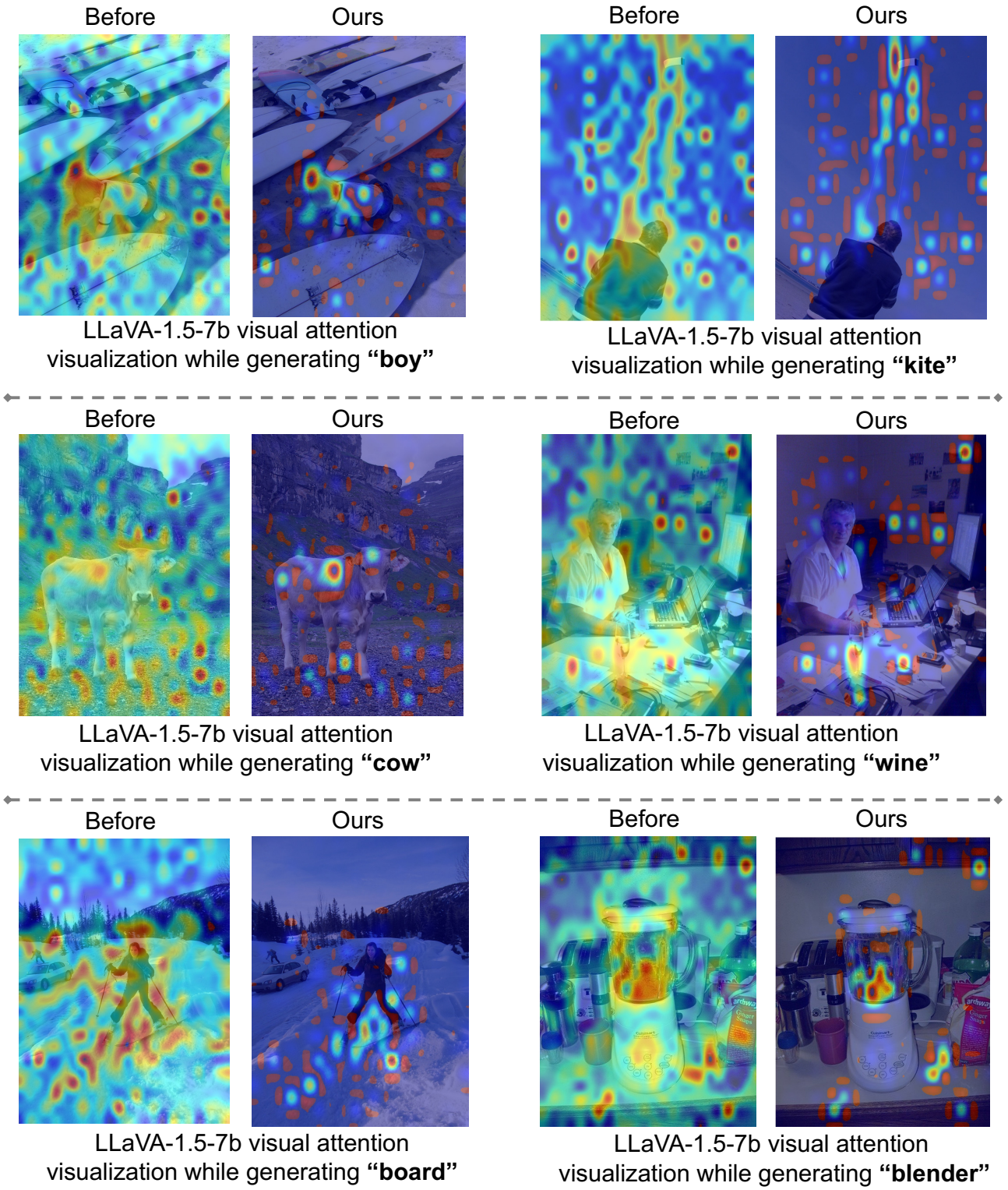


Figure 14. Qualitative comparison of visual attention maps before and after applying head-level attention distraction correction.

D.8. Qualitative Results of Hallucination Mitigation

We further present sample results from our hallucination mitigation method, as detailed in Section 4. Our prompt is "Please help me describe the image in detail" for all samples. The real and hallucinated object words are marked in blue and red, respectively.



Greedy: The image shows an elderly man sitting on a blue **bench** in a quiet park, holding a book and looking down as if reading. He wears a blue **cap**, a pair of glasses, and a light blue jacket, sitting calmly with a relaxed posture. A **wooden cane** leans against the bench close to his leg, suggesting it is used for support. The **bench** stands on a smooth paved path, surrounded by greenery. Nearby, a brown **leather bag** rests on the ground, and a small **radio** plays quietly beside the bench.

Ours: The image shows an elderly man sitting quietly on a blue wooden **bench** in a park, holding a book in his hands and looking. He is wearing a blue **cap**, glasses, and a light blue. A wooden **cane** rests against the bench near his leg. The **bench** is placed on a paved walkway. In the background, there is an open grassy area and distant trees, and above them a clear blue sky with a few light clouds can be seen, giving the scene a calm and peaceful outdoor atmosphere.



Greedy: The image feature a **train** travel down the track , passing by a **train** station . The **train** be position in the middle of the scene , with it yellow engine pull several **car** behind it . The **train** appear to be a freight **train** , a it be carry a variety of cargo . There be several **people** visible in the scene , likely wait for the train or working at the station . One **person** be stand near the train , while others be scatter around the platform . A **backpack** can be see place on the ground , possibly belong to one of the individual wait for the train .

Ours: The image shows a blue and yellow **train** traveling along the tracks, approaching a small station platform where a **man** in an orange safety vest is standing. The train is moving steadily through a rural. The overcast sky adds a calm, muted tone to the scene, suggesting a quiet moment in a suburban or semi-rural railway setting. A second set of tracks runs parallel, hinting at possible freight or additional passenger services in the area.



Greedy: The image feature a group of **woman** gather around a dining **table** , celebrate a special occasion . They be all wear hat and be in the process of cut a **cake** together . There be three **cake** visible on the **table** , with one be the main focus of the group . The **woman** be use a knife to cut the cake , and they be all smile and enjoy the moment . The **table** be set with a few **cup** and a **bowl** , indicate that they might be share the **cake** with others or enjoy some additional treat . The scene be warm and inviting , showcasing the camaraderie among the **woman** .

Ours: The image shows a group of **women** gathered around a **table**, enjoying a festive celebration with **cake** and candles. Two of the **women** are wearing hats. The third **woman** is dressed in a patterned blouse and wears a pink hat. All three appear joyful and engaged as they prepare to cut into a large white **cake** decorated with colorful flowers and a single tall candle. The warm atmosphere suggests a special occasion, possibly a birthday or anniversary, marked by elegance and camaraderie. A smaller decorative **cake** is also visible on the table, adding to the celebratory display.



Greedy: The image feature a zoo enclosure with a large **elephant** standing on a dirt field . The **elephant** be position in the middle of the enclosure , surround by a few rock and a body of water . There be two **bench** in the scene , one locate near the left side of the enclosure and the other closer to the right side . In addition to the **elephant** , there be two **people** visible in the enclosure , likely zookeepers or visitor observe the animal . They be stand near the left side of the enclosure , close to the **elephant** .

Ours: The image shows a large **elephant** standing in a zoo enclosure, surrounded by a body of water and landscaping. The **elephant** is near a tree and several large rocks, with a small waterway running through the habitat. Although it is not currently in the water, the setting suggests it may be cooling off or preparing to enter. In the background, a cable car glides overhead, indicating the location is part of a larger park or wildlife sanctuary. The enclosure is designed to mimic a natural environment, providing shade, open space, and opportunities for enrichment.

Figure 15. Qualitative results of hallucination mitigation on LLaVA-1.5-7B



Greedy: The image feature a clear glass vase fill with water and place on a table . the vase contain a beautiful arrangement of white flower , include daisy and baby 's breath . the flower be arrange in such a way that they create a visually appealing display . the glass vase allow the viewer to see the water and the stem of the flower , add to the overall aesthetic of the scene .

Ours: The image feature a beautiful glass vase fill with water , which be place on a table . the vase contain a few yellow flower , and the sunlight shine through the glass , create a warm and invite atmosphere. The vase be make of glass , and it be fill with water , which help to keep the flower fresh and hydrate .



Greedy: The image feature a large , delicious sandwich cut in half and place on a wooden cut board . the sandwich be fill with various ingredient , include lettuce , tomato , and cheese . the cut board be position on a dining table , which occupy most of the scene . In the background , there be a computer keyboard and a computer mouse , indicate that the table be likely use for work or leisure activity . the presence of these electronic device suggest that the person enjoy the sandwich might be take a break from their computer-related task to have a meal .

Ours: The image features a delicious sandwich on a wooden plate, which is placed on a wooden cutting board. The cut board be position on a dining table. The sandwich is a large stack of sandwiches, with a variety of ingredients, including tomatoes, lettuce, and cheese. The sandwich is served on a wooden platter, and it is accompanied by a side of chips.



Greedy: The image depict a cozy living room with a large flat-screen tv mount on the wall . the room be furnish with a brown leather couch and a chair , both place in front of the tv. A coffee table be situate between the couch and the chair , provide a comfortable space for place item. In addition to the main furniture , there be several book scatter around the room , some on the coffee table and others on the floor . a remote control can be see on the couch , likely for operate the tv. A potted plant be also present in the room , add a touch of greenery to the space .

Ours: The image showcases a living room with a cozy and comfortable atmosphere. The room feature a large leather couch , a television , and a coffee table . the living room be furnish with a television, and a coffee table . the room be decorate with a green carpet , a brown rug. The living room also includes a wooden entertainment unit with built-in shelves, providing both storage and display space for decorative items and media devices.



Greedy: The image depicts a bathroom interior centered around a sink and a mirror. The sink occupies the foreground and is situated slightly left of the image's center; it features two decorative silver faucets. Mounted above the sink, in the upper-left quadrant of the frame, is a mirrored medicine cabinet framed in white with visible metal hinges. To the right of the sink stands a black oscillating fan, plugged into a wall outlet, its blades exposed and motionless. In the mirror, a toothbrush rests beside the sink, though only one is visible in the actual room. The walls are uniformly painted white, and on the far right edge of the composition, a narrow, floor-to-ceiling door can be seen.

Ours: The image features a bathroom scene with a sink and a mirror. The sink is located in the foreground of the image, and it is positioned in the middle of the scene. The mirror is situated in the upper left part of the image, and the sink is situated in the middle of the image. The sink is white and is positioned in the middle of the image. The bathroom is equipped with a sink, a mirror, and a fan.

Figure 16. Qualitative results of hallucination mitigation on LLaVA-1.5-13B



Greedy: Asmall wooden **outhouse** sits alone in a grassy field, with its **door** partly open to reveal a simple toilet seat inside. Wild plants and a few flowers grow around it, and trees and distant hills line the background under a cloudy sky. A **red bicycle** leans against the wall, and a **dog** rests in the grass nearby.

Ours: The image shows a small wooden **outhouse** standing alone in the middle of a field. The outhouse looks simple and slightly weathered, with rough wooden panels and a large **wooden door** that is swung open to the side. The ground around it is uneven and covered with dry-looking vegetation, giving the scene a rural and isolated feeling. In the background, there are several trees and thick green bushes, and farther away there are low hills or mountains.



Greedy: A young **girl** sit on a **chair** , hold a brown **teddy bear** in her lap . she appear to be enjoy her time with the stuffed animal , possibly even lick her finger . the girl be wear a white sweater , which add to the cozy atmosphere of the scene . in the background , there be a din **table** with a few item on it , include a **cup** and a **bowl** . another chair can be see place near the table . the overall setting suggest a comfortable and playful environment for the young **girl** and her **teddy bear**.

Ours: The image show a young **girl** sit in front of a Christmas tree , hold a stuffed animal toy in her hand, which is a brown **teddy bear**. The girl be wear a white sweater and have brown hair and brown eye . the background be a room with a Christmas tree in the corner and a pink wall in the background. The girl appears to be engaged in a moment of quiet contemplation or play, gently touching her face as she looks at the **teddy bear**. The warm lighting and festive surroundings suggest a cozy holiday atmosphere, likely during a Christmas celebration.



Greedy: A**baby** sit on a **bed** , reach out to touch a **laptop computer**. The baby be position in the center of the scene , with the **laptop** place on the bed in front of them . the baby appear to be curious and interested in the **laptop** , possibly try to interact with it. In the background , there be a **couch** visible , and a **teddy bear** be place on the right side of the bed . The scene capture a moment of curiosity and exploration for the baby a they engage with the **laptop** .

Ours: The image show a small **baby** sit on a **bed** , hold a laptop computer on the bed . The baby be wear yellow pajama and be sit on top of a **laptop computer** on the **bed** . the bed be cover in white sheet and there be a white **teddy bear** on the bed next to the laptop computer . The baby appears to be curiously exploring the **laptop**, reaching out toward the keyboard with one hand while sitting upright on the bed. The soft lighting and cozy bedroom setting create a warm and intimate atmosphere, capturing a tender moment of early childhood discovery.



Greedy: A small white **dog** stands on a red brick sidewalk beside a green metal post and a red **fire hydrant**. Its thin, fluffy fur is slightly lifted by the breeze, making the dog look alert and energetic. Bright sunlight illuminates the area, casting clear, sharp shadows of the dog and nearby posts onto the ground. Behind it is a red brick **wall** with a visible **metal pipe**, reinforcing the urban street setting. The surroundings look empty and peaceful, suggesting a quiet afternoon, while a yellow **skateboard** rests against the wall and a **paper coffee cup** sits near the hydrant.

Ours: The image shows a small white **dog** standing on a red brick sidewalk near a green metal post and a red **fire hydrant**. The dog's long, wispy fur is being blown by the wind, giving it a slightly messy but lively look. Sunlight falls directly onto the scene, creating sharp shadows from the posts and the dog on the ground. The background features a red brick building **wall**, adding an urban touch to the setting. The street appears quiet and clean, with no people or cars visible nearby, giving the impression of a calm, sunny afternoon.

Figure 17. Qualitative results of hallucination mitigation on MiniGPT-4



Greedy: The image shows a woman is shown relaxing on a red **bed** set up outdoors in front of a large public building. She lies across it in a playful pose, lifting one leg and raising an arm while smiling broadly toward the camera. The **bed** is covered with a red **blanket** and matching **pillow** and is supported by a wheeled base, giving it the appearance of a mobile bed. In the background, brick walls mark the building entrance, with a few people nearby adding to the everyday public atmosphere. A **small bedside lamb** sits near the pillow, and a **cat** is curled up at the foot of the bed.

Ours: The image shows a woman lying on a red **bed** placed outside in front of a large building. She is stretched out playfully on the **bed**, with one leg raised in the air and one arm lifted, smiling widely. The **bed** has a red **blanket** and a red **pillow**, and it appears to be mounted on a wheeled frame, making it look like a movable bed. Behind her, the building has glass **doors** and brick walls, and a few people can be seen walking or standing near the entrance, giving the scene a casual public setting.



Greedy: The image shows a cutting board with a **sandwich** and chips placed in front of a **computer desk**. The sandwich is made with various ingredients, and it is cut in half, making it easier to eat. The **sandwich** is made with a variety of ingredients, including meat, lettuce, and tomato. The **sandwich** is placed on a plate, which is sitting on the desk next to the computer. The **sandwich** is made with various ingredients, including meat, lettuce, and tomato. The **sandwich** is sitting on a plate, which is placed on the desk next to the **computer**. The **sandwich** is cut in half, making it easier to eat

Ours: The image shows a **sandwich** placed on a table in an outdoor setting, sitting inside a red plastic **basket** lined with white paper. The **sandwich** is cut in half and filled with several ingredients, including sliced meat, lettuce, and other fillings that are visible from the side. It is positioned on a green tabletop, with a folded **napkin** and a small wrapped item placed next to it. In the background, there is another **basket** with food on a nearby table, along with chairs and parts of the sidewalk, suggesting a casual outdoor café or street dining area. The overall scene looks informal and ready to eat, with the sandwich clearly prepared for convenience.



Greedy: The image feature a large **clock** mount on the side of a tall building . the clock be prominently display , take up a significant portion of the building 's facade . the building itself appear to be a hotel , a indicate by the presence of a **clock** on the exterior . in the background , there be several **car** park or drive along the street , add to the urban atmosphere of the scene . the combination of the **clock** and the **car** create a sense of a bustling city environment .

Ours: The image feature a very large , gold-colored **clock** , which appear to be an oversized , mount to the side of a tall building . the **clock** be quite prominent against the sky , make it an eye-catching feature in the city skyline. The **clock** is mounted vertically on the building's facade, with an ornate design and a long pendulum-like extension below the face, giving it a distinctive and decorative appearance.



Greedy: The image depict a **man** and a **woman** stand next to each other in a hallway . The **man** be look at his **cell phone** , while the **woman** appear to be look at him . They seem to be engage in a conversation or share information on the **phone** . There be a few other **people** in the scene , with one **person** stand further back and another **person** closer to the right side of the image . The hallway appear to be a public space , possibly a subway or a **train** station , a there be a **clock** visible on the wall .

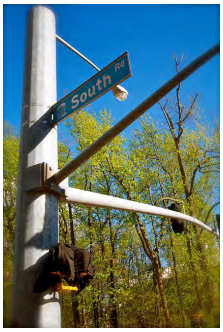
Ours: The image be a black and white photo of a **man** and **woman** walk down a busy subway platform . The **man** be check his **cell phone** while the **woman** walk ahead of him . The black and white photo capture the **man** and **woman** a they walk down the subway . In the image , there be multiple **people** in the subway station , indicate a busy and bustling environment . the **people** be engage in various activity while they navigate through the subway station .

Figure 18. Qualitative results of hallucination mitigation on Shikra



Greedy: A bunch of fruits and vegetables are stacked on top of each other in wooden crates and on display at a sunny marketplace. There are oranges, pineapples, lemons, tangerines, apples, grapes, mangoes, and other fruits and vegetables neatly arranged under an open awning.

Ours: A bunch of fruits and vegetables are stacked on top of each other in boxes and on display at a market place. there are apples, oranges, pineapples, limes, tangerines, and other fruits and vegetables at the market place.



Greedy: A street corner with a traffic light and a street sign . The traffic light be position on the left side of the scene , while the street sign be locate above the traffic light . the sign be blue and read south. There be a backpack hang on the pole of the traffic light , possibly belong to a pedestrian or a commuter . the backpack be situate near the center of the scene , and it presence add a sense of everyday life to the urban setting .

Ours: A street sign reading "South Rd" is mounted on a metal pole, accompanied by a streetlight and surrounded by trees in the background. A pair of shoes is hanging from the pole, adding an unusual touch to the scene. At the corner of South Road, a streetlight stands near the sign, with lush green foliage visible behind it.



Greedy: A woman sit on a couch in a living room , hold a controller in her hand . She appear to be surprise or excite about something , possibly relate to the video game she be play . The living room be furnish with a chair and a dining table , and there be a clock on the wall . A remote control can be see on the couch , and a bottle be place nearby . The woman be the main focus of the scene , and her expression conveys her engagement with the video game .

Ours: A woman is sitting on a couch in a cozy living room, holding a remote in her hand, her face showing a look of surprise or excitement. The room is warmly lit by a table lamp, and decorative elements like a patterned wall mirror and a soft blanket add to the homely atmosphere. She appears fully engaged in a video game, reacting with animated expression as she plays.



Greedy: A person stand in a room with a large screen tv mount on the wall . the person be hold a remote , likely play a video game on the tv . the room appear to be a living area with a couch and several chair scatter around . There be two din table in the room , one near the center and another towards the right side . A handbag can be see place on the floor , close to the person . the room also feature a few bottle , with one on the left side and another on the right side of the room .

Ours: A woman play a video game on a projector screen in a graffitied room with graffiti cover wall and chair and couch in front of a large screen . The woman appears focused on the game, standing near a retro-style orange couch that adds a pop of color to the space, while wooden tables and chairs are arranged around the area for casual seating.

Figure 19. Qualitative results of hallucination mitigation on Qwen-VL