Mitigating Object Hallucination in LVLMs via Attention Imbalance Rectification

Han Sun[1], Qin Li[1], Peixin Wang[1], Min Zhang[1*]
[1]Shanghai Key laboratory of Trustworthy Computing, East China Normal University
hsun@stu.ecnu.edu.cn {qli, pxwang, zhangmin}@sei.ecnu.edu.cn

A. Attention Imbalance in Object Hallucination

We provide a detailed explanation of how we identify specialized attention heads (e.g., hallucination-sensitive and hallucination-insensitive heads) in Section A.1, and describe the computational details for quantifying attention-pattern similarity in Section A.2.

A.1. Identifying Hallucination-Sensitive and -Insensitive Attention Heads

Building on erasure-based attribution [1, 8, 17, 18], a technique that evaluates the importance of model components by removing them and observing prediction changes, We propose a *hallucination-sensitive effect size* to identify attention heads that are either closely linked to hallucinations or largely unrelated. The goal is to quantify how each head differentially influences hallucinated versus non-hallucinated tokens. Formally, let

$$\mathcal{H} = \{t \mid y_t \in \mathcal{V}_{\text{hall}}\}, \quad \mathcal{N} = \{t \mid y_t \in \mathcal{V}_{\text{non}}\},$$

where $\mathcal{V}_{\text{hall}}$ and $\mathcal{V}_{\text{non}}$ denote the sets of hallucinated and non-hallucinated tokens. For a given attention head h, we first compute its sensitivity difference:

$$\mathcal{S}_h = \frac{1}{|\mathcal{H}|} \sum_{t \in \mathcal{H}} \Delta\mathbb{P}_h(y_t) - \frac{1}{|\mathcal{N}|} \sum_{t \in \mathcal{N}} \Delta\mathbb{P}_h(y_t),$$

where

$$\Delta\mathbb{P}_h(y_t) = \mathbb{P}_M(y_t \mid v, x, y_{<t}) - \mathbb{P}_{M \setminus h}(y_t \mid v, x, y_{<t})$$

measures the change in the model's probability of generating token y_t after removing head h. To further characterize the strength and consistency of this effect, we define a normalized effect size:

$$\mathcal{E}_h = \frac{\mathcal{S}_h}{\sqrt{\sigma_{\mathcal{H}}^2 + \sigma_{\mathcal{N}}^2}},$$

where $\sigma_{\mathcal{H}}^2$ and $\sigma_{\mathcal{N}}^2$ are the variances of $\Delta\mathbb{P}_h(y_t)$ across hallucinated and non-hallucinated tokens, respectively. A larger $\mathcal{E}_h$ indicates that head h exerts a stronger and more consistent influence on hallucination generation. A smaller $\mathcal{E}_h$ indicates that the head has only a weak influence on hallucinations, and removing it causes little to no change in the generation probability of hallucinated tokens.

A.2. Quantifying the Similarity between Attention Patterns

To examine whether hallucination-sensitive heads inherit the attention behaviors of their base language models, we perform a systematic similarity analysis between the attention patterns of Large Vision-Language Models (LVLMs) and its underlying Large Language Models (LLMs). Since LLMs cannot directly process image inputs, we introduce a placeholder token `<image>` while keeping the textual context identical to ensure consistency. We then focus on the attention patterns of the output tokens.

*Corresponding author

Extracting Layer–Head Attention Maps. For each image–caption pair, we provide both models with the same textual prompt: the LVLM processes the prompt together with the corresponding image, whereas the LLM receives only the textual input through its embedding layer. With `output_attentions=True`, both models return full attention matrices

$$A^{(\ell,h)} \in \mathbb{R}^{T\times T},$$

where T denotes the number of output tokens. Since our comparison concerns output's attention, we extract only the output-related submatrix:

$$\tilde{A}^{(\ell,h)} = A^{(\ell,h)}[-T:, -T:].$$

Similarity Measurement. For each layer–head pair, we flatten both attention maps into vectors and compute their cosine similarity:

$$\mathrm{Sim}^{(\ell,h)} = \mathrm{cosine}\Big(\mathrm{vec}\Big(\tilde{A}^{(\ell,h)}_{\mathrm{LVLM}}\Big), \mathrm{vec}\Big(\tilde{A}^{(\ell,h)}_{\mathrm{LLM}}\Big)\Big).$$

This metric directly quantifies how closely the LVLM head reproduces the token-level attention distribution of the LLM.

Hallucination-Sensitive vs. Hallucination-Insensitive Heads. We evaluate two type of heads: (1) *hallucination-sensitive heads*, which correlate strongly with hallucinated tokens, and (2) *hallucination-insensitive heads*, which rarely contribute to hallucinations. For each sample, we compute similarity scores for all selected heads and average them across both heads and samples:

$$\bar{S}_{\mathrm{hal}} = \mathbb{E}_{\mathrm{samples}}\mathbb{E}_{(\ell,h)\in\mathcal{H}_{\mathrm{hal}}}\mathrm{Sim}^{(\ell,h)}, \qquad \bar{S}_{\mathrm{non}} = \mathbb{E}_{\mathrm{samples}}\mathbb{E}_{(\ell,h)\in\mathcal{H}_{\mathrm{non}}}\mathrm{Sim}^{(\ell,h)}.$$

B. Theoretical Proofs

In this section, we formally establish the mathematical formulation linking the query-key parameter matrix W_{QK}, the attention matrix A, and the attention distribution.

B.1. Query–Key Parameter Matrix W_{QK} and Attention Distribution

B.1.1. Token Propagation Probability

We adopt the *token propagation probability* [3, 15] as a quantitative measure of attention localization and uniformity. To connect it with the attention computation, we begin by approximating the softmax function using a piecewise linear formulation. For a T-dimensional input $\omega \in \mathbb{R}^T$, the softmax is defined as

$$S(\omega)_i = \frac{\exp(\omega_i)}{\sum_{j\in[T]}\exp(\omega_j)}, \quad i \in [T].$$

By performing a first-order Taylor expansion around the origin, we obtain

$$\gamma^i := \nabla_i S(\mathbf{0}) = \frac{1}{T}\mathbf{e}^i - \frac{1}{T^2}\mathbf{1}, \quad \gamma_0^i := S(\mathbf{0})_i = \frac{1}{T}.$$

Then, the piecewise linear approximation of $S(\omega)$ can be expressed as

$$S(\omega)_i \approx \tilde{S}(\omega)_i = \max\{0, \min\{1, \langle\gamma^i, \omega\rangle + \gamma_0^i\}\}.$$

For notational simplicity, we write the vector form as $\tilde{S}(\omega) = \Gamma^\top\omega + \tilde{\gamma}_0$, where $\Gamma := [\tilde{\gamma}^1, \tilde{\gamma}^2, \ldots, \tilde{\gamma}^T]$ and $\tilde{\gamma}_0 := [\tilde{\gamma}_0^1, \tilde{\gamma}_0^2, \ldots, \tilde{\gamma}_0^T]^\top$. Formally, for the i-th token, the *signal propagation probability* is defined as

$$\rho_i = \mathbb{P}\left\{\langle\gamma^i, \omega\rangle + \gamma_0^i \in [0,1]\right\}, \tag{1}$$

where $\omega = \mathbf{X}^\top W_{\mathrm{QK}}\mathbf{X}_T/\sqrt{d}$, and the randomness originates solely from the input tokens $\mathbf{X}$. When only a few ρ_i take significantly large values, the attention distribution is considered *localized softmax*; in contrast, when ρ_i values are similar across tokens, it corresponds to a *uniform softmax*.

B.1.2. Localized vs. Uniform Attention Distribution

To analytically characterize ρ_i, we follow a synthetic random walk model of token generation:

Assumption 1 (Gaussian Random Walk). The tokens $\mathbf{x}_{t(t\geq 1)}$ are assumed to follow a Gaussian random walk, analogous to the random walk described by Pearson [14], where the first token $\mathbf{x}_1$ is drawn from a Gaussian distribution with mean 0 and covariance matrix Σ:

$$\mathbf{x}_1 \sim \mathcal{N}(0, \Sigma),$$

and for each subsequent token, the value of $\mathbf{x}_{t+1}$ is drawn from a Gaussian distribution centered around the previous token $\mathbf{x}_t$, with the same covariance structure Σ:

$$\mathbf{x}_{t+1} \sim \mathcal{N}(\mathbf{x}_t, \Sigma).$$

This process models a sequence of tokens that evolve through random steps, each governed by a Gaussian distribution with mean equal to the previous token and fixed covariance Σ.

Under this assumption, we approximate the expression $\langle \gamma^i, \omega \rangle + \gamma_0^i$ as a Gaussian random variable with mean μ^i and variance v^i:

$$\langle \gamma^i, \omega \rangle + \gamma_0^i \sim \mathcal{N}(\mu^i, v^i).$$

This approximation simplifies the distribution of the term to a Gaussian, enabling easier analysis.

To estimate the probability ρ_i associated with this Gaussian random variable, we utilize the cumulative distribution function (CDF) of the normal distribution. Specifically, ρ_i can be approximated as:

$$\rho_i \approx \frac{1}{2}\left\{\operatorname{erf}\left(\frac{1-\mu^i}{\sqrt{2v^i}}\right) + \operatorname{erf}\left(\frac{\mu^i}{\sqrt{2v^i}}\right)\right\}. \tag{2}$$

Here, the error function $\operatorname{erf}(z)$ is related to the CDF of the standard normal distribution, and the quantities $\frac{1-\mu^i}{\sqrt{2v^i}}$ and $\frac{\mu^i}{\sqrt{2v^i}}$ standardize the values of μ^i and v^i to yield probabilities within the range [0,1]. This approximation thus leverages the properties of the Gaussian distribution to compute ρ_i efficiently.

Lemma 1 (Moment Formulas for Gaussian Quadratic Forms). Let $\mathbf{W} \in \mathbb{R}^{d\times d}$ be a symmetric matrix, and let $\mathbf{a}, \boldsymbol{\mu} \in \mathbb{R}^d$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ be a covariance matrix. For $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the following moment formulae hold:

$$\begin{aligned}
\mathbb{E}[\mathbf{x}^\top \mathbf{W}\mathbf{x}] &= \operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{W}\boldsymbol{\mu}, \\
\mathbb{E}[\mathbf{x}\mathbf{x}^\top] &= \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top, \\
\mathbb{E}[\mathbf{a}^\top \mathbf{W}\mathbf{x}\mathbf{x}^\top \mathbf{W}\mathbf{x}] &= 2\,\mathbf{a}^\top \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\mu} + \mathbf{a}^\top \mathbf{W}\boldsymbol{\mu}\big\{\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{W}\boldsymbol{\mu}\big\}, \\
\mathbb{E}[\mathbf{x}^\top \mathbf{W}\mathbf{x}\mathbf{x}^\top \mathbf{W}\mathbf{x}] &= 2\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}) + \{\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma})\}^2 + 4\,\boldsymbol{\mu}^\top \mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\mu} \\
&\quad + 2\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma})\,\boldsymbol{\mu}^\top \mathbf{W}\boldsymbol{\mu} + \boldsymbol{\mu}^\top \mathbf{W}\boldsymbol{\mu}\,\boldsymbol{\mu}^\top \mathbf{W}\boldsymbol{\mu}.
\end{aligned}$$

For $i \leq j$, suppose that $\mathbf{x}_i, \mathbf{x}_j$ follow Assumption 1. Then, the following formulae hold:

$$\begin{aligned}
\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}\mathbf{x}_i] &= (i-1)\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}), \\
\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}\mathbf{x}_i\mathbf{x}_i^\top \mathbf{W}\mathbf{x}_i] &= (i^2-2i+2)\big\{2\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}) + \{\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma})\}^2\big\}, \\
\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}\mathbf{x}_i\mathbf{x}_j^\top \mathbf{W}\mathbf{x}_j] &= (i^2+ij-3i-j+4)\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}) + (i^2-2i+2)\{\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma})\}^2, \\
\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}\mathbf{x}_j\mathbf{x}_j^\top \mathbf{W}\mathbf{x}_j] &= (ij-i-j+2)\big\{2\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}) + \{\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma})\}^2\big\}.
\end{aligned}$$

The formulas in Lemma 1 are standard and can also be found in prior studies [3, 4].

Lemma 2. Suppose that $\mathbf{W}_{QK}$ is symmetric (for the asymmetric case, the probability is calculated using the symmetrized matrix $\frac{1}{2}(W_{\mathrm{QK}} + W_{\mathrm{QK}}^\top)$) and independent from $\mathbf{X}$, and let $\mathbf{W} := \mathbf{W}_{QK}\Sigma$. Under Assumption 1, for $i \in [T]$, the mean μ^i and variance v^i of $\langle \gamma^i, \omega \rangle + \gamma_0^i$ with the input $\omega := \mathbf{X}^\top \mathbf{W}_{QK}\mathbf{x}_T/\sqrt{d}$ are given as follows:

$$\mu^i = \left(\frac{i}{T} - \frac{1}{2}\right)\frac{\operatorname{tr}(\mathbf{W})}{\sqrt{d}} + o(1), \tag{3}$$

$$v^i = \left(\frac{2i^2}{T^2} + \frac{7}{12}\right)\frac{\operatorname{tr}(\mathbf{W}^2)}{d} + o(1). \tag{4}$$

Proofs. We use Lemma 1 to derive the μ^i:

$$\begin{aligned}\mu^i &= \frac{1}{\sqrt{d}T}\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{QK}\mathbf{x}_T] - \frac{1}{\sqrt{d}T^2}\sum_{j\in[T]}\mathbb{E}[\mathbf{x}_j^\top \mathbf{W}_{QK}\mathbf{x}_T] + o(1)\\ &= \frac{i-1}{\sqrt{d}T}\mathrm{tr}(\mathbf{W}) - \frac{\sum_{j\in[T]}(j-1)}{\sqrt{d}T^2}\mathrm{tr}(\mathbf{W}) + o(1)\\ &= \left(\frac{2i-T+1}{2\sqrt{d}T}\right)\mathrm{tr}(\mathbf{W}) + o(1).\end{aligned}$$

To derive the variance, we first evaluate the expectation $\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{QK}\mathbf{x}_T\mathbf{x}_j^\top \mathbf{W}_{QK}\mathbf{x}_T]$ (for $i \le j \le T$):

$$\begin{aligned}\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{QK}\mathbf{x}_T\mathbf{x}_j^\top \mathbf{W}_{QK}\mathbf{x}_T] &= \mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{QK}(\mathbf{x}_T\mathbf{x}_T^\top)\mathbf{W}_{QK}\mathbf{x}_j]\\ &= \mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{QK}((T-j)\Sigma + \mathbf{x}_j\mathbf{x}_j^\top)\mathbf{W}_{QK}\mathbf{x}_j]\\ &= (T-j)\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{QK}\Sigma\mathbf{W}_{QK}\mathbf{x}_j] + \mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{QK}\mathbf{x}_j\mathbf{x}_j^\top \mathbf{W}_{QK}\mathbf{x}_j]\\ &= (T-j)(i-1)\mathrm{tr}(\mathbf{W}^2) + (ij-i-j+2)\{2\mathrm{tr}(\mathbf{W}^2) + \mathrm{tr}(\mathbf{W})^2\}\\ &= \left[(i-1)(T+j-2)+2\right]\mathrm{tr}(\mathbf{W}^2) + \left[(i-1)(j-1)+1\right]\mathrm{tr}(\mathbf{W})^2.\end{aligned}$$

Then, the expectation of the squared term is expanded:

$$\begin{aligned}\mathbb{E}\Big[\langle\gamma^i, \mathbf{X}^\top \mathbf{W}_{QK}\mathbf{x}_T\rangle^2\Big] &= \mathbb{E}\left[\left(\frac{1}{T}\mathbf{x}_i^\top \mathbf{W}_{QK}\mathbf{x}_T - \frac{1}{T^2}\sum_{j\in[T]}\mathbf{x}_j^\top \mathbf{W}_{QK}\mathbf{x}_T\right)^2\right]\\ &= \frac{1}{T^2}\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{QK}\mathbf{x}_T\mathbf{x}_i^\top \mathbf{W}_{QK}\mathbf{x}_T] - \frac{2}{T^3}\sum_{j\in[T]}\mathbb{E}[\mathbf{x}_i^\top \mathbf{W}_{QK}\mathbf{x}_T\mathbf{x}_j^\top \mathbf{W}_{QK}\mathbf{x}_T]\\ &\quad + \frac{1}{T^4}\sum_{j,j'\in[T]}\mathbb{E}[\mathbf{x}_j^\top \mathbf{W}_{QK}\mathbf{x}_T\mathbf{x}_{j'}^\top \mathbf{W}_{QK}\mathbf{x}_T].\end{aligned}$$

By simplifying the equations, we obtain:

$$\mathbb{E}\Big[\langle\gamma^i, \mathbf{X}^\top \mathbf{W}_{QK}\mathbf{x}_T\rangle^2\Big] = \left(\frac{7}{12} + \frac{2i^2}{T^2}\right)\mathrm{tr}(\mathbf{W}^2) + \left(\frac{1}{4} - \frac{i}{T} + \frac{i^2}{T^2}\right)\mathrm{tr}(\mathbf{W})^2 + o(1).$$

We derived the v^i:

$$\begin{aligned}v^i &= \mathbb{V}[\langle\gamma^i, \omega\rangle]\\ &= \frac{1}{d}\left[\mathbb{E}\Big[\langle\gamma^i, \mathbf{X}^\top \mathbf{W}_{QK}\mathbf{x}_T\rangle^2\Big] - (\mu^i)^2\right]\\ &= \frac{1}{d}\left(\frac{7}{12} + \frac{2i^2}{T^2}\right)\mathrm{tr}(\mathbf{W}^2) + o(1).\end{aligned}$$

By extending μ^i and v^i smoothly over $\theta = i/T \in [0,1]$, we define the *token propagation probability* $\rho(\theta)$ as

$$\rho(\theta) = \Phi\big((\theta - \tfrac{1}{2})\,\xi; \theta\big) - \Phi\Big((\theta - \tfrac{1}{2})\,\xi - \tfrac{1}{\eta}, \theta\Big), \tag{5}$$

where

$$\xi = \frac{\mathrm{tr}(\mathbf{W})}{\sqrt{\mathrm{tr}(\mathbf{W}^2)}}, \quad \eta = \frac{\sqrt{\mathrm{tr}(\mathbf{W}^2)}}{\sqrt{d}},$$
$$\Phi(z;\theta) = \frac{1}{2}\left[\mathrm{erf}\left(\frac{z}{\sqrt{(4\theta^2 + \frac{7}{6})}}\right)\right].$$

Substituting the expression for $\Phi(z;\theta)$ into the definition of $\rho(\theta)$ gives

$$\rho(\theta) = \frac{1}{2}\left[\mathrm{erf}\left(\frac{(\theta-\frac{1}{2})\xi}{\sqrt{4\theta^2+\frac{7}{6}}}\right) - \mathrm{erf}\left(\frac{(\theta-\frac{1}{2})\xi - \frac{1}{\eta}}{\sqrt{4\theta^2+\frac{7}{6}}}\right)\right]. \tag{6}$$

Next, by substituting ξ and η, we obtain the compact closed-form expression:

$$\rho(\theta) = \tfrac{1}{2}\left[\operatorname{erf}\left(\frac{(\theta - \frac{1}{2})\operatorname{tr}(\mathbf{W})}{\sqrt{\operatorname{tr}(\mathbf{W}^2)}\sqrt{4\theta^2 + \frac{7}{6}}}\right) - \operatorname{erf}\left(\frac{(\theta - \frac{1}{2})\operatorname{tr}(\mathbf{W}) - \sqrt{d}}{\sqrt{\operatorname{tr}(\mathbf{W}^2)}\sqrt{4\theta^2 + \frac{7}{6}}}\right)\right]. \tag{7}$$

Results. From the above derivation, the behavior of $\rho(\theta)$ is governed by the trace terms $\operatorname{tr}(\mathbf{W})$ and $\operatorname{tr}(\mathbf{W}^2)$, which correspond to the first- and second-order moments of the eigenvalue spectrum of W_{QK}. Specifically:

- **Localized regime.** $\rho(\theta)$ exhibits a peaked (localized) pattern when $|\operatorname{tr}(\mathbf{W})| \gtrsim \sqrt{d}$ so that the peak position $\theta^\star = \frac{1}{2} + \frac{\sqrt{d}}{2\operatorname{tr}(\mathbf{W})}$ lies within $(0, 1)$, and when $\operatorname{tr}(\mathbf{W}^2)$ is close to zero. In this case, the attention distribution shows clear concentration.
- **Uniform regime.** When $\operatorname{tr}(\mathbf{W})$ is close to zero and $\operatorname{tr}(\mathbf{W}^2)$ is finite, the function $\rho(\theta)$ becomes smooth without a directional bias, indicating an approximately uniform attention distribution.

B.2. Attention Matrix A and Attention Distribution

Given the query–key interaction $QK^\top$, the attention matrix is defined as

$$A = \operatorname{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right),$$

where each row $A_i = (A_{i1}, A_{i2}, \ldots, A_{iT})$ represents the attention distribution of token i over all contextual tokens. The statistical property of A_i reflects how the model allocates its focus across the input sequence.

Variance as a measure of concentration. To quantify the concentration of attention, we compute the row-wise variance of A:

$$\sigma_A^2 = \frac{1}{T}\sum_{j=1}^{T}(A_{ij} - \tfrac{1}{T})^2.$$

A small σ_A^2 indicates a nearly uniform attention distribution, where each token receives comparable importance. Conversely, a large σ_A^2 implies that the attention weights are sharply peaked, signaling localized or selective attention. This variance measure serves as a simple yet effective descriptor of the attention distribution's shape.

Connection to entropy. The variance of A_i is inversely related to its entropy $H(A_i) = -\sum_j A_{ij} \log A_{ij}$:

$$H(A_i) \uparrow \Leftrightarrow \sigma_A^2 \downarrow, \qquad H(A_i) \downarrow \Leftrightarrow \sigma_A^2 \uparrow .$$

This negative correlation reflects a fundamental principle from information theory: for a discrete distribution over a fixed support, the entropy reaches its maximum when the distribution is uniform and decreases as the distribution becomes more concentrated [9, 11]. Accordingly, the variance serves as a complementary descriptor of attention entropy, providing an analytic proxy for quantifying how "localized" or "uniform" the attention distribution is.

C. Experimental Details

C.1. Detailed Descriptions and Experimental Setup of Baselines

We provide detailed descriptions and experimental settings of the baseline methods used for comparison in our experiments.

FarSight [16] FarSight is a decoding-based hallucination mitigation method designed to enhance information propagation during generation. It constructs causal masks to reinforce effective interactions among multimodal tokens and to suppress attention drift toward outlier tokens. By dynamically adjusting the causal mask through an attention-register structure and a positional-aware masking mechanism, FarSight effectively reduces both initial hallucinations and snowball hallucinations across image and video tasks.

VCD [12] Visual Contrastive Decoding (VCD) mitigates object hallucination by introducing contrastive visual signals during decoding. Specifically, VCD constructs a positive image (the original input) and a negative image (a visually perturbed version), and contrasts the logits produced under these two conditions. Tokens whose logits do not sufficiently depend on the true visual input are down-weighted, reducing visually ungrounded generations. VCD is lightweight and model-agnostic, requiring no training or model modification.

DoLA [7] Decoding by Contrasting Layers (DoLa) improves factuality in LLMs by contrasting the logits of deep layers (which are more factual) with shallow layers (which may contain stylistic or biased information). The final logits are obtained by subtracting the shallow-layer distribution from the deep-layer distribution. Although originally proposed for LLMs, DoLa can be applied to LVLMs by performing layer-wise logit contrast during multimodal decoding, thus reducing hallucinated or unsupported content.

HALC [6] Hallucination Reduction via Adaptive Focal-Contrast Decoding (HALC) is a decoding method designed specifically for object hallucination in LVLMs. It applies a focal-style modulation to down-weight uncertain or weakly grounded visual tokens, and introduces a contrastive correction that penalizes over-confident hallucinations. HALC adaptively adjusts the correction strength based on token-level confidence, enabling fine-grained control over hallucination mitigation without harming general performance.

OPERA [10] Over-trust Penalty and a Retrospection-Allocation strategy (OPERA) reduces hallucination via two key mechanisms: *Over-Trust Penalty* and *Retrospection-Allocation*. The first penalizes the model when its textual prediction becomes overly confident in the absence of sufficient visual grounding. The second reallocates attention across vision tokens by retrospectively checking whether generated tokens are visually supported. Together, these mechanisms enforce alignment between the model's confidence and its visual evidence.

AD-HH [20] AD-HH analyzes hallucination sources in LVLMs via causal mediation analysis and identifies specific hallucination heads within Multi-Head Attention as key contributors. It introduces a training-free decoding intervention and a lightweight fine-tuning strategy that reduce the over-reliance of these heads on text tokens, achieving substantial improvements in hallucination mitigation across captioning benchmarks.

Experimental Setup For a fair and consistent comparison, we adopt the official configurations and recommended hyperparameters for all baseline methods. For **FarSight**, we set the decay rate σ to $\log_\alpha(\text{seq})$, where $\alpha = 1024$ denotes the typical maximum token limit. **DoLA** is implemented with an adaptive plausibility threshold of 0.1 and an early-exit schedule applied to layers $[0, 2, 4, \ldots, 32]$. **OPERA** uses a self-attention scaling factor of 50, an attending-retrospection threshold of 15, a beam size of 5, and a penalty weight of 1.0. For **VCD**, we follow the standard configuration with an amplification factor of 1, an adaptive plausibility threshold of 0.1, and 500 diffusion noise steps. **HALC** is implemented with the official DINO-based detector; we set the JSD buffer size to $m = 6$, beam size to 1, number of sampled fields-of-view to $n = 4$, exponential growth ratio of contextual fields to 0.6, bounding-box threshold to 0.4, and adaptive plausibility threshold to 0.1. For **AD-HH**, we reweight the top 20 attention heads and use 0.5 as the reweighting threshold.

C.2. Hyperparameter Sensitivity Analysis

Effect of the visual reallocation factors λ. Figure 1 (a) illustrates the effects of the visual reallocation factor λ on hallucination mitigation. For the factor λ, a moderate increase enhances the model's attention to visual information during decoding, thereby effectively reducing hallucinations. The model achieves its best performance at $\lambda = 3.5$. However, once λ exceeds the threshold of 12.5, the performance begins to deteriorate noticeably, and when $\lambda > 17.5$, the model collapses entirely, producing repetitive outputs.

Effect of the textual reallocation factors γ. Figure 1 (b) illustrates the effects of the textual reallocation factor γ on hallucination mitigation. As γ decreases, we observe a gradual reduction in object hallucinations. Since reallocation is applied only to hallucination-sensitive attention heads rather than all attention heads, setting γ to a lower value does not lead to abnormal model behavior. On the contrary, by suppressing excessive attention to the text, the modality-wise attention imbalance (MAI) across the average attention heads is reduced, achieving a better balance.

Effect of the attention reallocation threshold τ_{text}. Figure 1 (c) shows how the attention reallocation threshold τ_{text} influences the model's behavior. As the threshold decreases from 1.0, both C_S and C_I scores show an overall downward

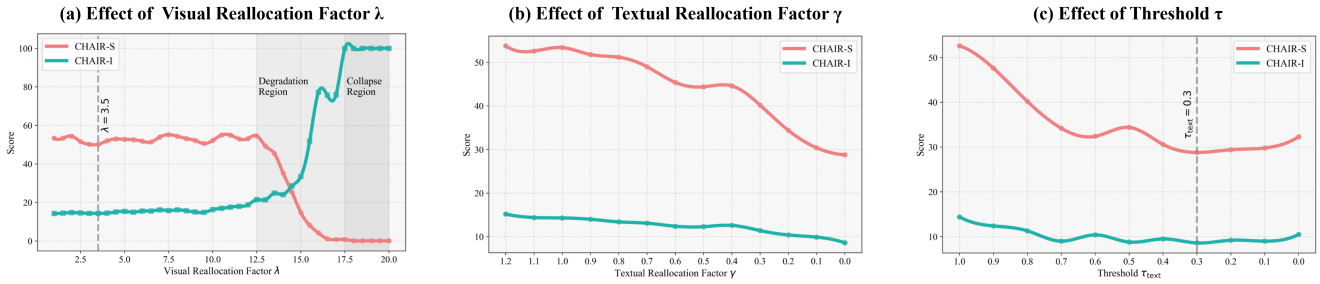

Figure 1. Sensitivity of the attention reallocation mechanism to hyperparameters. (a) shows the impact of the visual reallocation factor λ. (b) illustrates the effect of the textual reallocation factor γ. (c) demonstrates the influence of the attention reallocation threshold $\tau_{\texttt{text}}$.

trend, indicating that a lower threshold helps trigger attention reallocation earlier and suppress hallucinations caused by excessive focus on textual modalities. The model achieves optimal performance when $\tau_{\texttt{text}}$ achieves 0.3. However, further decreasing the threshold leads to overly frequent reallocation, introducing noise perturbations and causing slight performance fluctuations.

Effect of the regularization coefficient β. Table 1 reports the effect of different regularization coefficients β. As β increases from 0 to 0.3, all metrics improve notably, indicating that moderate mean-shift regularization helps stabilize the attention distribution and suppress hallucinations. When β is raised to 0.6, the model shows a slight performance drop but remains close to the optimal setting. In contrast, further increasing β to 0.9 leads to excessive homogenization, weakening the expressive capacity of attention and resulting in clear degradation on MM-Vet. Overall, a moderate level of regularization (e.g., $\beta = 0.3$–0.6) proves most effective.

Table 1. Effect of the regularization coefficient β.

	$C_S \downarrow$	$C_I \downarrow$	$F_1 \uparrow$	**MM-Vet Overall** $\uparrow$
$\beta = 0$	32.1	9.9	0.85	30.5
$\beta = 0.3$	**28.8**	**8.6**	**0.86**	**32.0 (+4.9%)**
$\beta = 0.6$	30.4	9.0	0.86	31.9 (+4.6%)
$\beta = 0.9$	32.0	9.2	0.85	27.4 (-10.2%)

C.3. Additional Experimental Results

Comparison of generation time. Figure 2 presents the average decoding time per sample for all compared methods. Greedy decoding provides the fastest speed at 2.8 seconds. AIR achieves a similarly efficient latency of **3.4** seconds, which is close to AD-HH at 3.2 seconds. Other lightweight baselines, including FarSight (4.1 seconds), DoLA (4.2 seconds), and VCD (5.3 seconds), also remain below the 6-second range. In contrast, HALC and OPERA introduce substantial computational cost, requiring 28.6 seconds and 25.0 seconds per sample respectively. These results indicate that AIR offers competitive efficiency while maintaining strong performance in hallucination mitigation.

Evaluation of additional LVLM models. We further evaluate AIR on four additional LVLMs to assess its generalizability across different architectures. As shown in Table 2, AIR consistently lowers both the object-level (C_S) and image-level (C_I) hallucination scores compared with greedy decoding. The improvements are substantial across all evaluated models, demonstrating that AIR is not tied to a specific backbone. In particular, the largest relative reductions are observed on DeepSeek-VL2-4.5B, where hallucination rates decrease by around 40% on both metrics. Similar gains are also achieved on LLaVA-NeXT-7B, Qwen-2.5-VL-7B, and InternVL-3.5-8B, indicating that AIR effectively mitigates hallucinations across models with distinct training paradigms and visual encoders. These results highlight the robustness of AIR for improving the reliability of LVLM-generated responses.

AIR's performance across different scales of LVLMs. Across LLaVA models of varying capacities (7B, 13B, and 34B), AIR consistently reduces hallucination rates under the CHAIR metric. As shown in Table 3, AIR delivers substantial improvements over the greedy decoding baseline, achieving up to **44.4%** and **37.2%** relative reductions in C_S and C_I on the 7B model. The improvements remain robust as model size increases: AIR reduces hallucinations by **33.0%** and **28.0%** on the 13B model, and continues to provide measurable benefits on the stronger 34B LLaVA-NeXT model. These results

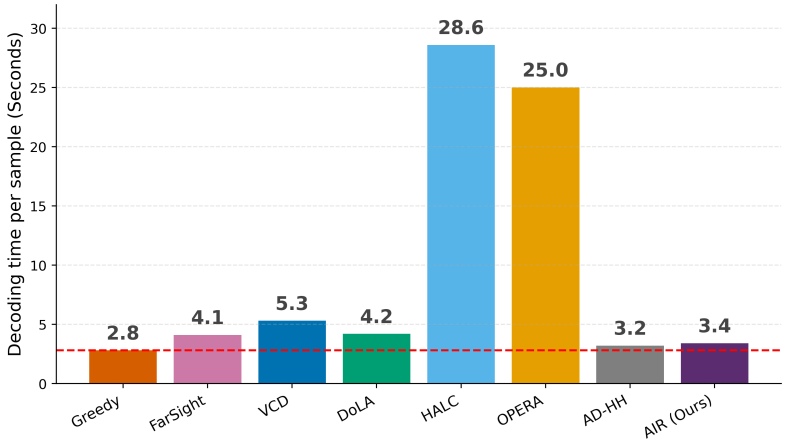

Figure 2. Comparison of inference time across baselines. Except for HALC and OPERA, most methods maintain an average decoding time of under 6 seconds per sample.

Table 2. **CHAIR** hallucination evaluation results on additional four LVLMs across baselines. Smaller C_S and C_I indicate fewer hallucinations. Results are reported with maximum new tokens set to 256. $\Delta\%$ represents the relative improvement over the baseline. Our method AIR consistently enhances hallucination mitigation across all additional LVLMs.

Max New Tokens: 256								
Methods	LLaVA-NeXT-7B [13]		Qwen-2.5-VL-7B [2]		InternVL-3.5-8B [5]		DeepSeek-VL2-4.5B [19]	
	$C_S\downarrow$	$C_I\downarrow$	$C_S\downarrow$	$C_I\downarrow$	$C_S\downarrow$	$C_I\downarrow$	$C_S\downarrow$	$C_I\downarrow$
Greedy	31.2	8.6	38.0	9.4	32.9	9.0	42.4	11.6
AIR (Ours)	**22.5**	**6.1**	**29.3**	**8.8**	**25.1**	**8.3**	**25.8**	**6.9**
$\Delta\%$	↓27.9%	↓29.1%	↓22.9%	↓6.4%	↓23.7%	↓7.8%	↓39.2%	↓40.5%

Table 3. **CHAIR** hallucination evaluation results across LVLMs of different scales. Smaller C_S and C_I indicate fewer hallucinations. All results are reported with the maximum new token limit set to 256. $\Delta\%$ denotes the relative improvement over the baseline. Our method AIR consistently improves hallucination mitigation across models of varying scales.

Max New Tokens: 256						
Methods	LLaVA-1.5-7B		LLaVA-1.5-13B		LLaVA-NeXT-34B	
	$C_S\downarrow$	$C_I\downarrow$	$C_S\downarrow$	$C_I\downarrow$	$C_S\downarrow$	$C_I\downarrow$
Greedy	51.8	13.7	46.3	12.5	25.7	6.1
AIR (Ours)	**28.8**	**8.6**	**31.0**	**9.0**	**23.2**	**6.0**
$\Delta\%$	↓44.4%	↓37.2%	↓33.0%	↓28.0%	↓9.7%	↓1.6%

indicate that AIR scales effectively with model capacity and remains beneficial even for large LVLMs that already exhibit lower hallucination levels.

Visual analysis of TAI across layers and attention heads. We visualize attention maps under hallucination-generation settings across multiple layers and attention heads to further examine Token-wise Attention Imbalance (TAI), and present three representative cases for demonstration.(Figure 3, Figure 4, and Figure 5) for illustration. Our key observations are summarized as follows.

1. **TAI spans the entire generation sequence rather than only the final output tokens.** Visualizations across multiple layers and attention heads show that TAI does not appear solely within the output token sequence; instead, it emerges over large portions of the entire sequence and is particularly dense in the visual-token region.
2. **TAI is more concentrated and pronounced in specific attention heads.** Different heads demonstrate markedly different attention patterns: certain heads (e.g., head 16) show stronger striped or slanted regularities, while others exhibit much weaker effects. This indicates that TAI is *not uniformly distributed* across heads but manifests as *head-specific biases*.
3. **The strength of TAI varies structurally across layers.** From lower layers (Layer 0) to middle layers (Layer 15) and

higher layers (Layer 31), attention patterns become increasingly regular and biased toward fixed token regions. This reveals that internal biases are progressively amplified or consolidated with depth, indicating a *layer-wise accumulation or reinforcement* of TAI.

4. **Averaged attention reveals global biases more clearly than individual heads.** Compared with per-head visualizations, the mean attention maps exhibit a more coherent slanted structure, showing that although individual heads behave differently, multi-head attention collectively forms a consistent sequence-level bias. This implies that TAI arises from a *multi-head synergistic effect* rather than the behavior of isolated heads.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, pages 4190–4197, 2020. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 8

[3] Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when qk-eigenspectrum concentrates. In *ICML*, 2024. 2, 3

[4] M. Brookes. The matrix reference manual. http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html, 1998. [Online; accessed 01-September-2023]. 3

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 8

[6] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *ICLR*, pages 7824–7846, 2024. 6

[7] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *ICLR*, 2024. 6

[8] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *ACL Workshop*, pages 276–286, 2019. 1

[9] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, USA, 2nd edition, 2006. 5

[10] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, pages 13418–13427, 2024. 6

[11] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957. 5

[12] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pages 13872–13882, 2024. 6

[13] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 8

[14] P. K. Rawlings. Modes of a gaussian random walk. *Journal of Statistical Physics*, 111(5):769–788, 2003. 3

[15] Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs. In *ICLR*, 2025. 2

[16] Fangzhi Tang, Chen Liu, Zhen Xu, Meng Hu, Zhenyu Peng, Zhihao Yang, Jinsong Su, Ming Lin, Yuxin Peng, Xi Cheng, Imran Razzak, and Zongyuan Ge. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *CVPR*, 2025. 5

[17] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *ACL*, pages 5797–5808, 2019. 1

[18] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *EMNLP-IJCNLP*, pages 11–20, 2019. 1

[19] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 8

[20] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *ICLR*, 2025. 6

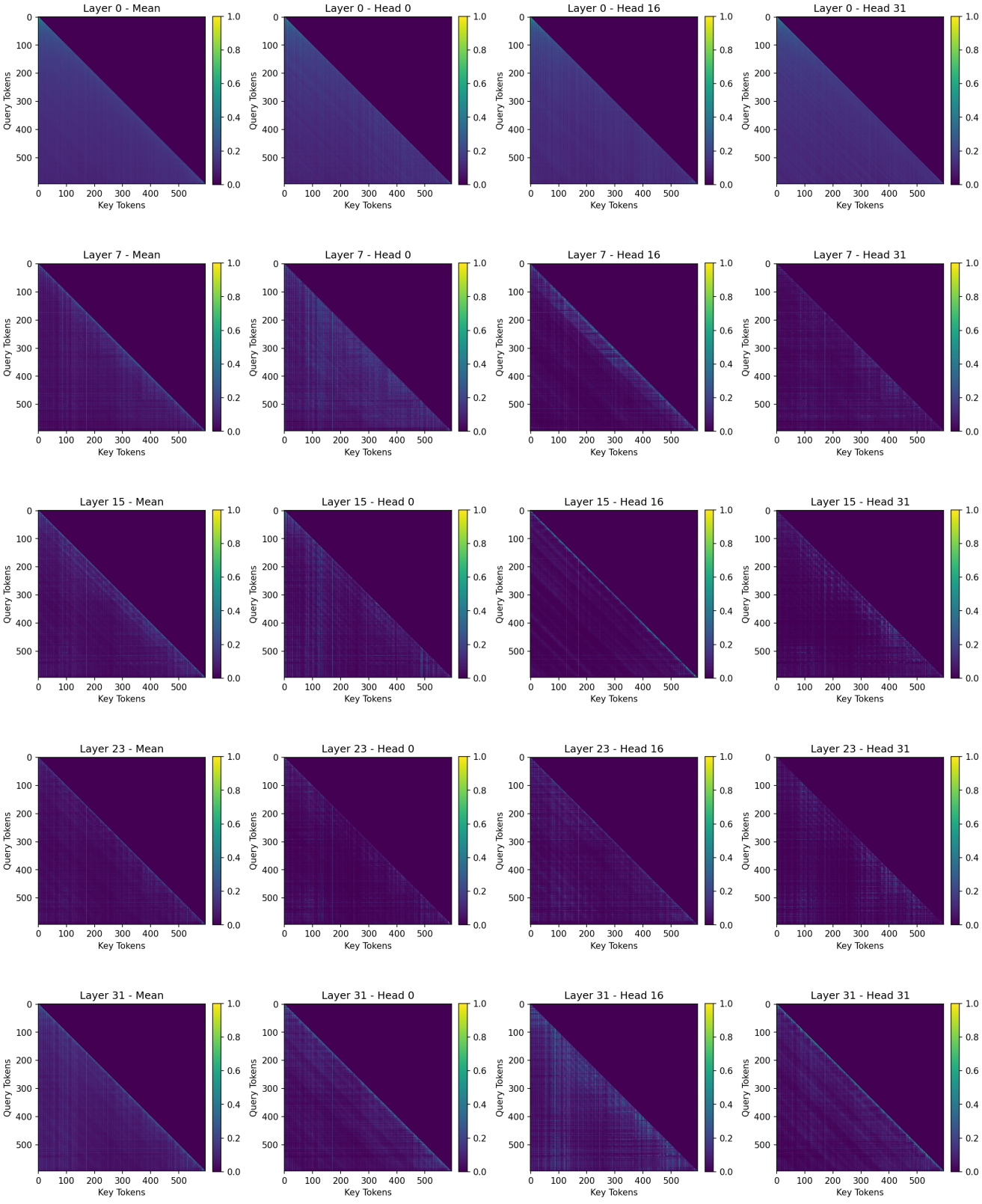

Figure 3. Attention maps across layers and heads for Case One.

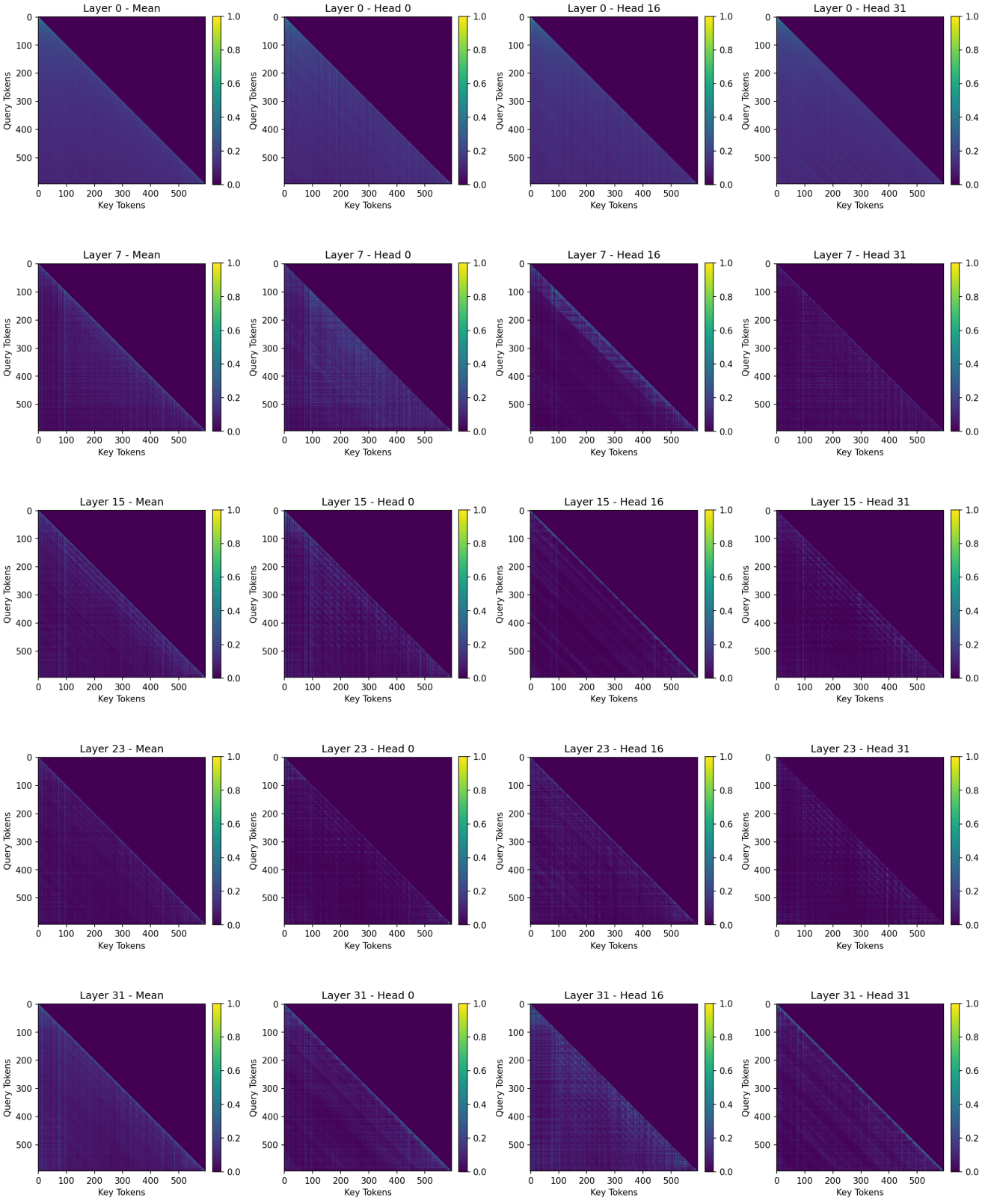

Figure 4. Attention maps across layers and heads for Case Two.

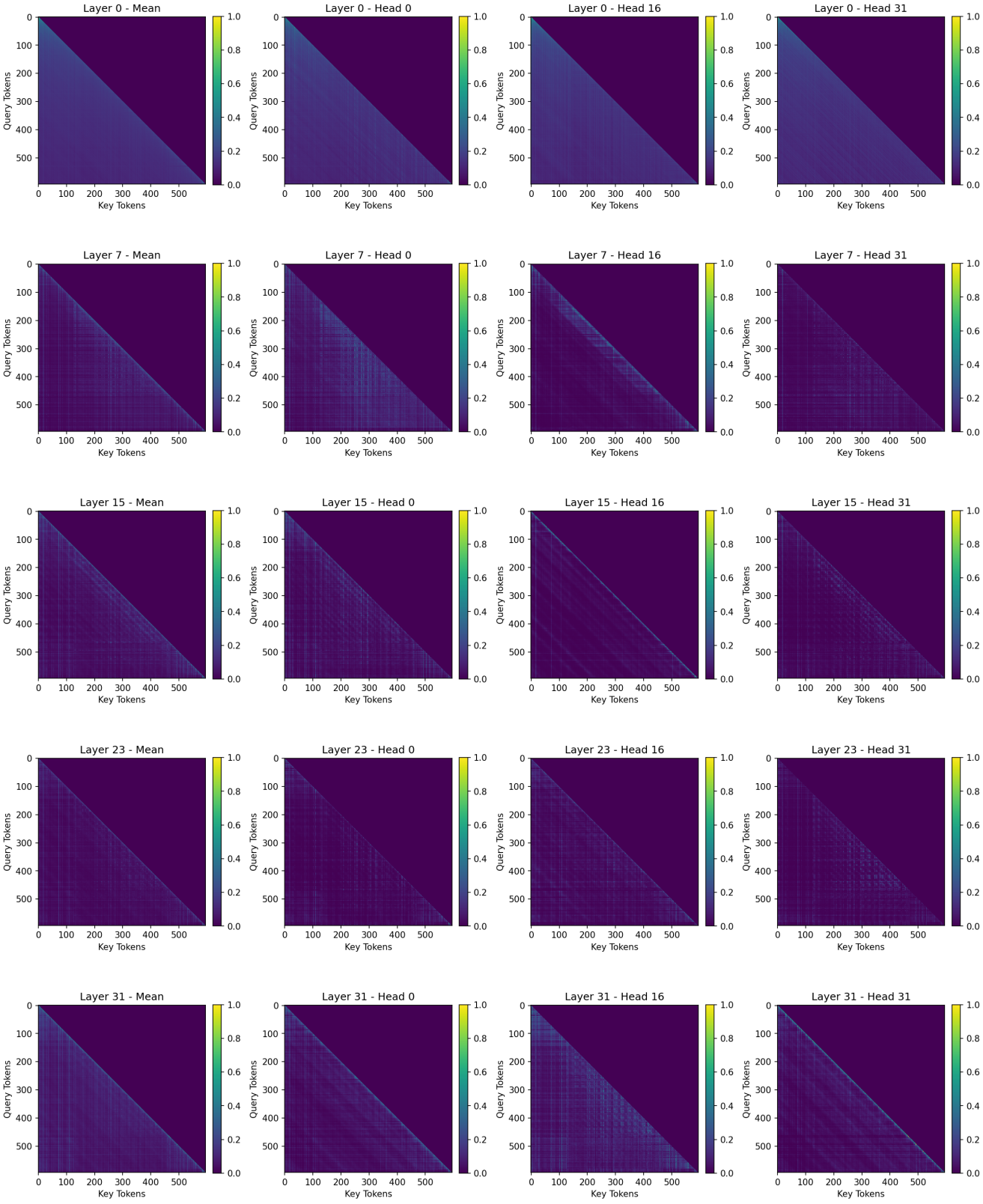

Figure 5. Attention maps across layers and heads for Case Three.