MLLMs Hallucinate when Information Distribution Drifts in Synergy Heads

Anonymous authors
Paper under double-blind review

Abstract

Multimodal Large Language Models (MLLMs) often struggle with hallucinations, thus hindering their reliable practical applications. Existing attention-based mitigation methods mainly rely on indirect signals (e.g., attention weights) that fail to accurately reflect the actual information shift underlying hallucination generation. In this paper, we propose HEAL, **H**ead-l**E**vel information disent**A**nglement and ca**L**ibration for identifying and mitigating hallucinations. HEAL first employs causal noise intervention on multi-head outputs to filter out causally redundant heads. Subsequently, it disentangles information distribution within the remaining heads via the counterfactual Difference-in-Differences, categorizing heads into four types. Through analysis, we observe: *hallucinations happen when information distribution drifts away from a healthy equilibrium in synergy heads*, not strongly correlated with the quantity or strength of modality-specific heads. Motivated by this insight, HEAL injects dynamic information calibration factors into the value vectors of synergy heads and actively regulates visual-language dependencies, steering the output distribution towards factual evidence. Extensive experiments demonstrate that HEAL effectively reduces hallucinations across multiple MLLMs, offering a simple and interpretable pathway to enhance model trustworthiness.

1 Introduction

Though Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities across a wide range of complex tasks (Bai et al., 2025a; Liu et al., 2024b;c), they are frequently plagued by hallucinations (Kalai et al., 2025; Chen et al., 2026; Bai et al., 2024), which means generating plausible-sounding responses that are factually inconsistent with the visual context. This phenomenon remains a significant bottleneck for the reliable applications of MLLMs in high-precision fields like medical imaging.

To mitigate hallucinations, extensive efforts have been made from both macro- and micro-level perspectives (Bai et al., 2024). Macro-level strategies, such as retraining or fine-tuning (Zhang et al., 2024a; Yu et al., 2024; Chen et al., 2025), architectural scaling (Jain et al., 2024; Chen et al., 2024b), and reinforcement learning (Zhao et al., 2023), typically treat the model as a black box and attempt to correct hallucinations from the outside. In contrast, micro-level methods aim to alleviate hallucinations through decoding optimization (Leng et al., 2024; Park et al., 2025; Zhu et al., 2025) or attention-based enhancement (He et al., 2025; Yu et al., 2026; Li et al., 2023b). However, decoding optimization intervenes only at the output level, offering limited interpretability and insight into the internal mechanisms. Likewise, most attention-based approaches depend on indirect signals, such as attention weights or FFN activations, which may be unable to capture the actual causal information distribution and confounding synergy effects involved in hallucination generation. Moreover, uni-modal attention enhancement often leads to a "see-saw" dilemma (Yu et al., 2026): strengthening visual attention may cause truncated or less creative responses, whereas over-reliance on language can exacerbate visual hallucinations.

Motivated by these limitations, we propose HEAL, head-level information disentanglement and calibration, to identify and mitigate hallucinations during autoregressive generation. It disentangles the information distribution within attention heads, categorizing them into redundant, visual, language and synergy heads. By leveraging causal noise intervention and counterfactual Difference-in-Differences, HEAL provides a fine-grained, head-level interpretable lens into the internal mechanics

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

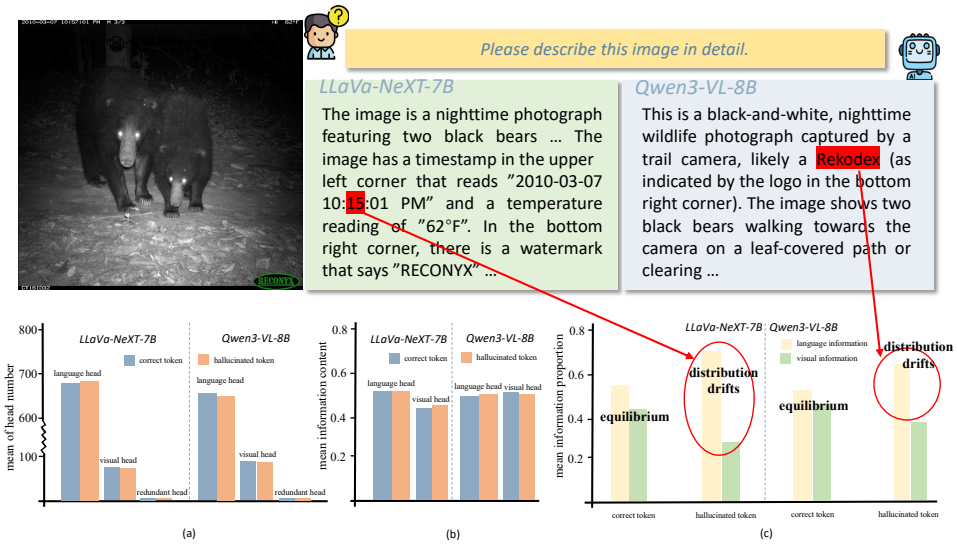


Figure 1: (a) Mean number of heads of different types when generating hallucinated versus correct tokens; (b) Mean information content of visual and language heads in hallucinated and correct token generation; (c) Mean information proportion within synergy heads for hallucinated and correct tokens.

of MLLMs. As discussed in Figure 1 and Section 4.5, through extensive analysis on models like Qwen3-VL (Bai et al., 2025a) and LLaVA-NeXT (Liu et al., 2024c), we observe critical insights into the pathology of hallucinations:

MLLMs hallucinate when information distribution drifts away from a healthy equilibrium in synergy heads, not strongly correlated with the quantity or strength of modality-specific heads.

Additionally, from Figure 2(a) and (b), we further see that head roles are highly dynamic throughout the generation. As the model generates different types of tokens, attention heads undergo a task-driven phase transition. When producing language-centric tokens, synergy heads tend to revert toward language heads; when generating visually grounded tokens, some language heads shift toward synergy heads, and a small subset of synergy heads may further move toward visual heads. This dynamic behavior highlights the adaptive nature of Transformer attention (Vaswani et al., 2017). However, when focusing specifically on visually grounded token generation, we find that hallucinations often begin with an internal drift of information distribution within synergy heads.

Based on these insights, we propose a dynamic information calibration strategy during inference. We introduce an equilibrium factor (α), akin to the temperature hyperparameter, to regulate the model's dependence on visual versus language information. By dynamically monitoring and calibrating the information distribution within synergy heads, HEAL effectively steers the output back toward factual evidence without sacrificing too much linguistic coherence. Extensive experiments across multiple MLLM benchmarks demonstrate that HEAL consistently reduces hallucinations and provides a simple, interpretable pathway toward more trustworthy multimodal systems.

Our contributions are summarized as follows:

- We propose HEAL, utilizing causal noise intervention and counterfactual Difference-in-Differences to categorize attention heads into four functional types, providing a new interpretable perspective on MLLM internals.
- We identify visual-language information disequilibrium within synergy heads as an intervenable factor that can influence hallucination behavior, rather than attributing hallucinations solely to macro-level modality heads, and characterize the "task-driven role transition" of attention heads.
- We introduce a theoretically grounded dynamic calibration strategy that uses the equilibrium factor α to regulate the visual-language information distribution, predictably steering the representation geometry toward the equilibrium direction and improving the factuality of MLLM generation.

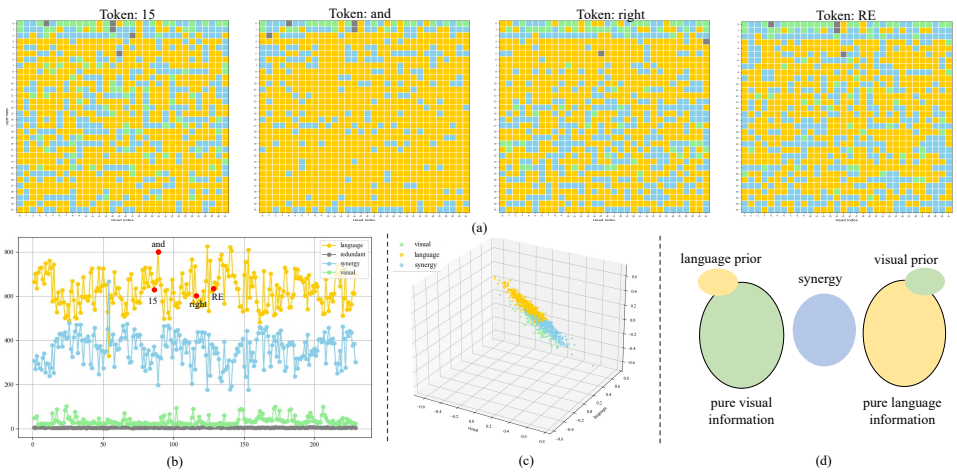


Figure 2: (a) Head distribution in the LLaVA-NeXT-7B model when generating different tokens, where green, blue, yellow and gray denote visual, synergy, language and redundant heads, respectively; (b) The evolution of the number of heads in the autoregressive generation. (c) 3D scatter plot of attention heads of different types, where the axes correspond to visual, language and synergy contributions. (d) Illustration of the internal information structure within a single attention head. The visualization in (a) and (b) is based on the same VQA example as in Figure 1. *From the figure, we observe that the number of visual heads remains nearly constant throughout the generation.*

2 Related Works

Hallucination Mitigation in MLLMs. Recent studies mitigate hallucinations in MLLMs mainly from four perspectives (Bai et al., 2024). Data-centric methods reduce spurious correlations by constructing negative, counterfactual, or cleaner supervision data (Liu et al., 2024a; Zhou et al., 2024). Training- and model-level methods strengthen visual grounding through improved alignment objectives, auxiliary supervision, preference learning, and stronger multimodal architectures (Sun et al., 2024a; Zhao et al., 2023; Zhang et al., 2024a; Yu et al., 2024; Chen et al., 2025). Inference-time methods intervene in decoding by contrastive or guided strategies to suppress language priors and encourage visual evidence usage (Leng et al., 2024; Huang et al., 2024; Park et al., 2025). In addition, detection-based and post-hoc correction methods serve as a practical complement by locating hallucinated content and revising unreliable outputs (Zhou et al., 2024).

Attention-based Mitigation Methods. Most attention-based methods mitigate hallucinations by explicitly increasing visual attention during decoding. The main intuition is that MLLMs tend to rely less on image prompts as generation proceeds, causing language priors to dominate and leading to visually ungrounded outputs (Bai et al., 2024). Representative approaches differ in how they select and strengthen visual attention. AGLA (An et al., 2025) improves prompt-relevant local attention by combining global and local attention to emphasize regions most relevant to the query. PAI (Liu et al., 2024d), EAH (Zhang et al., 2024b), and VHR (He et al., 2025) further intervene on attention heads to make decoding more image-centric and alleviate visual attention sinks (Kang et al., 2025; Xiao et al., 2023). Owl (Yu et al., 2026) introduces a dual-path contrastive decoding strategy in which one path reinforces visually grounded attention while the other amplifies hallucinated ones. CausalMM (Zhou et al., 2025) uses structural causal modeling to treat modality priors as a confounder between attention and output in MLLMs. FarSight (Tang et al., 2025b) proposes a versatile plug-and-play decoding strategy that reduces attention interference from outlier tokens merely by optimizing the causal mask. As discussed above, these approaches tend to rely on attention weights as an indirect proxy for token contribution and cannot reflect the actual information structure in attention heads, while uni-modal attention enhancement inherently struggles to balance vision and language.

3 Method

In this section, we present the proposed HEAL and introduce how it (1) identifies and categorizes attention heads, and (2) dynamically calibrates information drift within synergy heads.

3.1 Causal Noise Intervention on Multi-head Outputs

In MLLMs, the image is first encoded by a visual encoder into visual embeddings, which are then projected into the language space via a projector. The resulting visual tokens are concatenated with text tokens and fed into the language backbone for autoregressive generation. Generally, the language backbone consists of multiple transformer decoder layers and each layer l performs a multi-head attention operation among tokens. Under the KV-cache setting, the attention head i at generation step t is formulated as:

$$\mathbf{o}^{(t,l,i)} = \mathrm{Softmax}\left(\frac{\mathbf{q}^{(t,l,i)}(\mathbf{K}^{(t,l,i)})^\top}{\sqrt{d}}\right)\begin{bmatrix}\mathbf{V}_{vis}^{(t,l,i)}\\ \mathbf{V}_{lang}^{(t,l,i)}\end{bmatrix}, \tag{1}$$

where $\mathbf{q}^{(t,l,i)}$ is the query at the current step, d is the dimension of the query, and $\mathbf{K}^{(t,l,i)}$, $\mathbf{V}_{vis}^{(t,l,i)}$, $\mathbf{V}_{lang}^{(t,l,i)}$ denote the cached key, visual value, language value matrices. The outputs of all heads are concatenated and linearly projected:

$$\mathbf{y}^{(t,l)} = \mathbf{x}^{(t,l)} + \mathbf{W}_O^{(l)}\left[\mathbf{o}^{(t,l,1)};\cdots;\mathbf{o}^{(t,l,H)}\right], \tag{2}$$

where $\mathbf{x}^{(t,l)}$ is the residual input and $\mathbf{W}_O^{(l)}$ is the output projection matrix.

Our goal is to identify insignificant attention heads that contribute less to the outputs. Instead of relying on the corresponding projection weights in $\mathbf{W}_O^{(l)}$, we directly intervene on the head outputs and measure the resulting differences in the layer representation. This is necessary because a head with a large activation may still contribute substantially even if its corresponding projection weights are small, while a head with large weights but near-zero activations may have a limited effect. In addition, a head's contribution cannot be faithfully captured by its own weights alone, since the final output may depend on the cooperative interaction among multiple heads.

To estimate the actual effect of the i-th head in layer l, we replace its output with distribution-matched Gaussian noise in Equation 2:

$$\tilde{\mathbf{y}}_{(-i)}^{(t,l)} = \mathbf{x}^{(t,l)} + \mathbf{W}_O^{(l)}\left[\mathbf{o}^{(t,l,1)};\cdots;\tilde{\mathbf{o}}^{(t,l,i)};\cdots;\mathbf{o}^{(t,l,H)}\right], \tag{3}$$

$$\tilde{\mathbf{o}}^{(t,l,i)} \sim \mathcal{N}(\boldsymbol{\mu}_{t,l,i}, \boldsymbol{\Sigma}_{t,l,i}), \tag{4}$$

where $\boldsymbol{\mu}_{t,l,i}$ and $\boldsymbol{\Sigma}_{t,l,i}$ are the corresponding mean and covariance in $\mathbf{o}^{(t,l,i)}$. Before quantifying the significance of the head, we need define the vector similarity metric as follows:

$$\mathrm{Sim}(\mathbf{a},\mathbf{b}) = \frac{1}{2}\left(1+\cos(\mathbf{a},\mathbf{b})\right) = \frac{1}{2}\left(1+\frac{\mathbf{a}^\top\mathbf{b}}{\|\mathbf{a}\|_2\,\|\mathbf{b}\|_2}\right). \tag{5}$$

Thus, the information contribution of the head is

$$I^{(t,l,i)} = 1 - \mathrm{Sim}\left(\mathbf{y}^{(t,l)}, \tilde{\mathbf{y}}_{(-i)}^{(t,l)}\right). \tag{6}$$

After calculating the scores of all heads in layer l, we classify a head as **causally redundant** if $I(t,l,i) < \mu(I^{(t,l,:)}) - 3\sigma(I^{(t,l,:)})$, where μ and σ denote the mean and standard deviation operators. A head may exhibit rich internal information patterns while having negligible influence on the final output; such heads are not informative for downstream intervention. Therefore, before performing subsequent finer-grained decomposition, we first exclude causally redundant heads.

3.2 Counterfactual Difference-in-Differences in Attention Head

Inspired by Partial Information Decomposition Theory (Williams & Beer, 2010), for a multimodal head, its internal information can be viewed as a combination of pure visual information, pure language information, prior information and multimodal synergy, as illustrated in Figure 2(d). To practically and easily approximate these components, we construct four counterfactual inputs by independently masking visual and language tokens with Gaussian noise preserving the first- and second-order statistics:

$$\mathbf{H}_{11} = H_{t,l,i}(\mathbf{V},\mathbf{T}),\quad \mathbf{H}_{01} = H_{t,l,i}(\bar{\mathbf{V}},\mathbf{T}),\quad \mathbf{H}_{10} = H_{t,l,i}(\mathbf{V},\bar{\mathbf{T}}),\quad \mathbf{H}_{00} = H_{t,l,i}(\bar{\mathbf{V}},\bar{\mathbf{T}}), \tag{7}$$

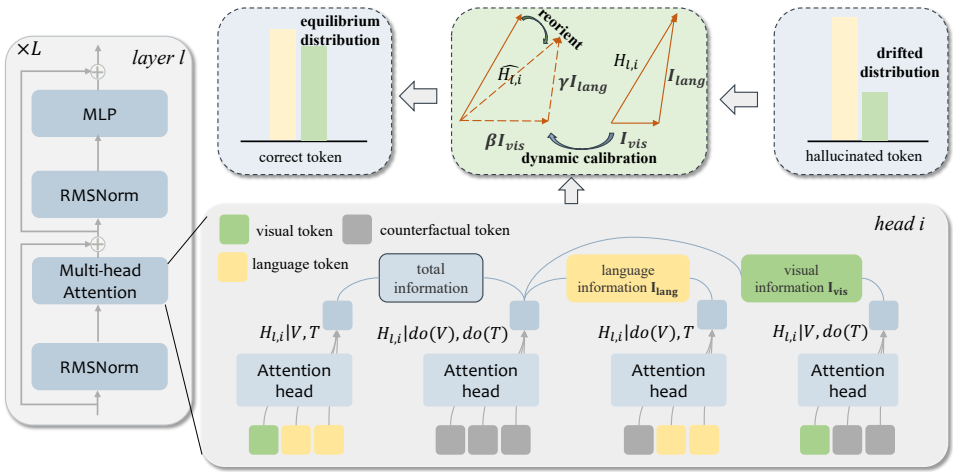


Figure 3: Illustration of the proposed HEAL. It employs the counterfactual difference-in-differences mechanism to disentangle the information distribution within attention heads and dynamically calibrates the information drift in synergy heads.

where $\bar{\mathbf{V}}$ and $\bar{\mathbf{T}}$ denote masked visual and language tokens, respectively. We define the total information content of this head as the difference between the full and fully counterfactual settings:

$$I_{\text{total}} = \underbrace{1 - \text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{00})}_{\text{first-order difference}}, \tag{8}$$

Based on the two single-modal counterfactual cases, the visual and language information content can be calculated as

$$I_{\text{lang}} = \underbrace{(1 - \text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{00})) - (1 - \text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{01}))}_{\text{second-order difference}} = \underbrace{\text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{01}) - \text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{00})}_{\text{pure language information+language prior}}, \tag{9}$$

$$I_{\text{vis}} = \underbrace{(1 - \text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{00})) - (1 - \text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{10}))}_{\text{second-order difference}} = \underbrace{\text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{10}) - \text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{00})}_{\text{pure visual information+visual prior}}. \tag{10}$$

Then the synergy effect is approximately measured by:

$$I_{\text{syn}} = I_{\text{total}} - I_{\text{vis}} - I_{\text{lang}} = \underbrace{1 + \text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{00}) - \text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{01}) - \text{Sim}(\mathbf{H}_{11}, \mathbf{H}_{10})}_{\text{synergy effect}-\text{evidence supported prior}} \tag{11}$$

Note that I_{syn} is not the true synergy measure and may be positive or negative, since it reflects the synergy effect after discounting the overlap between prior-induced and evidence-supported information. However, our subsequent categorization primarily rely on the visual and language information content. Based on these scores, we classify attention heads into different types. First, heads with negligible total information content are defined as **information redundant heads**:

$$I_{\text{total}} < \mu_{\text{total}} - 3\sigma_{\text{total}}, \tag{12}$$

where μ_{total} and σ_{total} denote the mean and standard deviation of the total information scores across all heads. For the non-redundant heads, we further distinguish modality-specific heads. If $I_{\text{vis}} > 0$ and $I_{\text{lang}} \leq 0$, the head is regarded as a visual head. Conversely, if $I_{\text{vis}} \leq 0$ and $I_{\text{lang}} > 0$, the head is regarded as a language head. When both I_{vis} and I_{lang} are positive, we compute the modality ratio:

$$\alpha_{\text{vis}} = \frac{I_{\text{vis}}}{I_{\text{vis}} + I_{\text{lang}}}, \qquad \alpha_{\text{lang}} = \frac{I_{\text{lang}}}{I_{\text{vis}} + I_{\text{lang}}}. \tag{13}$$

Since the distribution of modality ratios is typically skewed(He et al., 2025; Wang et al., 2025a), we first apply a Logit transformation to reduce skewness and then employ the robust median absolute deviation (MAD) to determine the thresholds:

$$\text{MAD}(x) = \text{median}\big(|x - \text{median}(x)|\big). \tag{14}$$

A head is classified as a visual head if $\mathrm{Logit}(\alpha_{\mathrm{vis}}) > \mathrm{median}(\mathrm{Logit}(\alpha_{\mathrm{vis}})) + \lambda \cdot \mathrm{MAD}(\mathrm{Logit}(\alpha_{\mathrm{vis}}))$, and analogously for a language head. The remaining heads are categorized as synergy heads. Synergy heads can be further divided into visual-preferred and language-preferred ones according to whether $\alpha_{\mathrm{vis}} > \alpha_{\mathrm{lang}}$ or $\alpha_{\mathrm{vis}} < \alpha_{\mathrm{lang}}$. Figure 2(c) visualizes the resulting head taxonomy in a 3D space, where the three axes correspond to visual, language, and synergistic information, respectively.

3.3 Dynamic Information Calibration

Figure 3 shows the procedure of the proposed dynamic information calibration strategy. From the above analysis, hallucinated tokens are often triggered by a significant information drift in synergy heads. Therefore, we introduce an equilibrium factor $\alpha \in (0, 1)$ to characterize the desired visual-language equilibrium when generating correct tokens. Intuitively, α controls the model's reliance on visual information, while $1 - \alpha$ controls its reliance on language information. To reorient the information distribution in the head toward the target equilibrium α, we need two calibration factors so that

$$\frac{\beta I_{\mathrm{vis}}}{\beta I_{\mathrm{vis}} + \gamma I_{\mathrm{lang}}} = \alpha. \tag{15}$$

The typical choice is

$$\beta = \frac{\alpha}{\alpha_{vis}}, \qquad \gamma = \frac{1 - \alpha}{\alpha_{lang}}. \tag{16}$$

After estimating the information distribution at each generation step, we dynamically calibrate the value vectors corresponding to visual and language tokens by β and γ, respectively. Importantly, this operation is applied *after* the KV cache update and *before* the attention kernel, so that the internal implementation of FlashAttention (Dao, 2024) or PagedAttention (Kwon et al., 2023) is not affected. The calibrated attention head can be reformulated as:

$$\mathbf{o}^{(t,l,i)} = \mathrm{Softmax}\left(\frac{\mathbf{q}^{(t,l,i)}(\mathbf{K}^{(t,l,i)})^\top}{\sqrt{d}}\right)\begin{bmatrix}\beta\mathbf{V}_{vis}^{(t,l,i)} \\ \gamma\mathbf{V}_{lang}^{(t,l,i)}\end{bmatrix}. \tag{17}$$

Theorem 1 (Equivalence between value space and information distribution calibration)**.** Under a local additive approximation, in an attention head, calibrating the value space is equivalent to calibrating its modality-wise information distribution toward a target proportion.

Theorem 2 (Monotonic directional effect of equilibrium factor)**.** With non-collinear visual and language components, the cosine alignment between the calibrated attention head output and visual components is strictly increasing regarding α.

Theorem 1 shows that applying the factors β and γ to calibrate the value vectors corresponding to visual and language tokens is equivalent to calibrating their information distribution. Theorem 2 further proves that, by using the equilibrium factor to calibrate the information distribution, the geometric alignment between the attention head output and visual information changes monotonically and increasingly with respect to α. We provide the proofs in Appendix B, and the theoretical analysis of HEAL in Appendix C.

3.4 Periodic, Parallelized and Batched Implementation

Periodic Type Update. The type distribution of heads and their internal information are inherently dynamic across generation steps. However, our empirical observations find a temporal locality: the macroscopic type distribution of heads shifts minimally within short generation windows (e.g., 10 to 15 steps). Consequently, we update the global head types periodically at a step interval S, while computing the information calibration factor β and γ at every decoding step.

Parallelized Causal Intervention. During the head type update, we must compute the causal noise intervention for all H attention heads. To avoid sequential evaluation, we accelerate this process via parallelized tensor operations:

$$\widetilde{\mathbf{Y}}^{(t,l)} = \mathbf{1}_H(\mathbf{x}^{(t,l)})^\top + \begin{bmatrix}\tilde{\mathbf{o}}^{(t,l,1)} & \mathbf{o}^{(t,l,2)} & \cdots & \mathbf{o}^{(t,l,H)} \\ \mathbf{o}^{(t,l,1)} & \tilde{\mathbf{o}}^{(t,l,2)} & \cdots & \mathbf{o}^{(t,l,H)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{o}^{(t,l,1)} & \mathbf{o}^{(t,l,2)} & \cdots & \tilde{\mathbf{o}}^{(t,l,H)}\end{bmatrix}(\mathbf{W}_O^{(l)})^\top. \tag{18}$$

Batched Difference-in-Differences Analysis. Similarly, the DiD calculation requires evaluating four distinct counterfactual states to disentangle the information distribution. Instead of computing these sequentially, we form the batched tensors $\widehat{\mathbf{q}}^{t,l,i} \in \mathbb{R}^{4\times B\times H\times 1\times d_h}$, $\widehat{\mathbf{K}}^{t,l,i}$, $\widehat{\mathbf{V}}^{t,l,i} \in \mathbb{R}^{4\times B\times H\times L\times d_h}$ and execute a unified attention operation. This allows the hardware accelerator to process the factual and counterfactual attention matrices simultaneously.

4 Experiments

4.1 Experimental Setting

Baselines. To evaluate the generalizability and effectiveness of our method, we conduct experiments on several representative MLLMs, including LLaVA series (LLaVA-1.5-7B (Liu et al., 2024b) and LLaVA-NeXT-7B (Liu et al., 2024c)), Qwen series (Qwen2-VL-7B (Wang et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025b) and Qwen3-VL-8B(Bai et al., 2025a)), and InternVL series (InternVL-7B (Chen et al., 2024b) and InternVL3.5-8B (Wang et al., 2025b)).

Evaluation Benchmarks. We perform comprehensive evaluations across two primary categories of benchmarks to assess both general multimodal capabilities and specific hallucination tendencies:

(1) Comprehensive Benchmarks: we use LLaVA-Bench (Liu et al., 2024b), MME (Fu et al., 2023), and BLINK-Twice (JIANG et al., 2026) to measure the impact of our method on the models' core reasoning and perception abilities.

(2) Hallucination Benchmarks: to specifically quantify hallucination reduction, we employ POPE (Li et al., 2023c) for object existence, CHAIR (Rohrbach et al., 2018) for fine-grained image captioning, and MMHal-Bench (Sun et al., 2024b) for complex actions and spatial relationships.

Hyperparameters. We use one model from a specific family to characterize the operating range of α, and directly transfer the resulting family-level setting to the remaining models within the same or familiar family without further tuning. For LLaVA and InternVL series, we set α to 0.5 for simple benchmarks such as POPE and CHAIR, and to 0.6 for other challenging and comprehensive benchmarks. For Qwen families, α is set to 0.4 for simpler benchmarks and 0.5 for others. The update interval S is consistently set to 10 generation steps across all experiments. Detailed guidelines and empirical patterns for determining these hyperparameter values are provided in Appendix D.4.

Table 1: Comparison of HEAL with other SOTA methods on POPE, CHAIR and MME datasets. The best performances are **bolded** and baseline model is LLaVA-1.5-7B.

Method	POPE		CHAIR				MME				
	F1↑	Acc↑	C_S↓	C_I↓	Recall↑	Length	Exist.↑	Count↑	Pos.↑	Color↑	Total↑
Beam Search	85.4	84.0	51.0	15.2	75.2	102.2	175.67	124.67	114.00	151.00	565.34
DoLa (Chuang et al., 2023)	80.2	83.1	57.0	15.2	78.2	97.5	180.10	127.40	119.30	154.60	594.10
VCD (Leng et al., 2024)	85.3	85.0	51.0	14.9	77.2	101.9	184.66	137.33	128.67	153.00	603.66
OPERA (Huang et al., 2024)	84.2	85.2	47.0	14.6	78.5	95.3	180.67	133.33	111.67	123.33	549.00
DOPRA (Wei & Zhang, 2024)	84.6	84.3	46.3	13.8	78.2	96.1	185.67	138.33	120.67	133.00	577.67
HALC (Chen et al., 2024a)	83.9	84.0	50.2	12.4	78.4	97.2	190.00	143.30	128.30	160.00	621.60
EAH (Zhang et al., 2024b)	85.7	86.0	**36.4**	**9.9**	74.9	97.7	190.00	108.33	145.00	160.66	603.99
SID (Huo et al., 2025)	85.6	85.8	44.2	12.2	73.0	99.4	183.90	132.20	127.80	155.90	599.80
VISTA (Li et al., 2025)	86.3	86.2	-	-	-	-	-	-	-	-	-
TAME (Tang et al., 2025a)	85.4	85.7	41.3	12.2	74.4	98.8	193.00	137.33	139.00	164.67	634.00
VAR (Kang et al., 2025)	86.0	86.5	52.4	14.5	79.1	103.0	190.00	148.33	138.33	155.00	631.33
CausalLLM (Zhou et al., 2025)	86.0	86.5	-	-	-	-	195.00	156.00	135.00	170.00	656.00
AGLA (An et al., 2025)	84.6	85.5	43.0	14.1	78.9	98.8	195.00	153.89	129.44	161.67	640.00
FarSight(Tang et al., 2025b)	-	-	41.6	13.2	75.5	100.6	-	-	-	-	-
MemVR (Zou et al., 2024)	87.1	87.4	46.6	13.0	**80.8**	99.6	190.00	155.00	133.33	170.60	648.30
ONLY (Wan et al., 2025)	85.5	85.1	49.8	14.3	75.9	99.7	191.67	145.55	136.66	161.66	635.55
LocoRE (Zhang et al., 2026)	86.9	87.3	38.4	11.2	75.4	98.2	190.00	158.33	133.33	**175.00**	656.66
VHR (He et al., 2025)	85.5	-	38.6	12.3	-	81.3	-	-	-	-	-
HEAL	**87.7**$_{\pm 0.3}$	**88.3**$_{\pm 0.2}$	36.7$_{\pm 0.4}$	10.7$_{\pm 0.0}$	79.1$_{\pm 0.1}$	99.8$_{\pm 0.7}$	**195.45**$_{\pm 0.23}$	**158.53**$_{\pm 0.35}$	**145.12**$_{\pm 0.48}$	170.66$_{\pm 0.66}$	**669.76**$_{\pm 1.72}$

4.2 Evaluation on Hallucination Benchmarks

As shown in Table 1, existing training-free hallucination mitigation methods can be broadly categorized into two groups. The first group (OPERA (Huang et al., 2024), DOPRA (Wei & Zhang, 2024), DoLa (Chuang et al., 2023) VCD (Leng et al., 2024), AGLA (An et al., 2025), etc.) focuses on correcting the decoding process to reduce hallucinations at inference time, while the second group (TAME(Tang et al., 2025a), VAR(Kang et al., 2025), EAH (Zhang et al., 2024b), VHR (He et al., 2025), FarSight(Tang et al., 2025b), etc.) improves MLLMs' reliability by calibrating attention heads.

Our method belongs to the second group, but differs from prior attention head-based approaches by explicitly disentangling the internal information composition and dynamically calibrating modality equilibrium. On the MME and POPE benchmarks, our method achieves strong and consistent performance gains. Compared with EAH (Zhang et al., 2024b), HEAL reaches a higher recall and longer generation length on CHAIR. We attribute this to the fact that EAH mainly strengthens certain heads, whereas our method avoids over-emphasizing one modality and instead performs an equilibrium reallocation between visual and language information. TAME (Tang et al., 2025a) aggregates token-to-token attention scores but largely overlooks the role of visual information, while VAR (Kang et al., 2025) suppresses attention collapse by reinforcing visual information but tends to underweight textual signals. Consequently, both methods may degrade performance on long-form generation benchmarks like CHAIR. In contrast, our calibration strategy preserves the model's language fluency while improving the visual evidence in the generated output.

4.3 Evaluation on Comprehensive Benchmarks

As shown in Table 1, the results on the MME dataset show that HEAL consistently achieves higher scores across different evaluation categories. This suggests that our method is effective not only on hallucination-specific benchmarks, but also on a broader set of multimodal reasoning and perception tasks. In Tables 2 and 15, we further integrate HEAL as a plug-and-play module into several advanced MLLMs. These results show that our method consistently improves the hallucination-related and comprehensive metrics across different architectures, which demonstrates its strong generalization.

Table 2: MLLM performance with and without HEAL on the POPE, CHAIR, and LLaVA-Bench.

Method	Comprehensive Benchmark	Hallucination Benchmark				
	LLaVA-Bench ↑	$CHAIR_S$ ↓	$CHAIR_I$ ↓	POPE-R↑	POPE-F1↑	POPE-A↑
LLaVA-1.5-7B	72.5	51.0	15.2	87.0	85.4	84.0
+ HEAL	**75.2**	**36.9**	**10.7**	**89.1**	**87.8**	**88.5**
LLaVA-NeXT-7B	81.6	29.9	9.2	87.4	86.5	84.7
+ HEAL	**82.2**	**24.6**	**7.9**	**89.9**	**88.1**	**89.0**
Qwen2.5-VL-7B	76.8	27.2	9.0	80.4	87.4	88.4
+HEAL	**78.5**	**23.3**	**8.5**	**81.3**	**88.4**	**89.1**
Qwen2-VL-7B	75.6	25.0	7.3	79.1	86.6	87.6
+ HEAL	**78.0**	**23.1**	**6.2**	**81.9**	**88.1**	**88.9**
InternVL-7B	51.6	46.6	12.4	80.0	85.3	86.2
+ HEAL	**53.4**	**39.2**	**9.6**	**86.5**	**87.8**	**87.9**

4.4 Ablation Analysis

Update interval for head types. The head attribution distribution remains relatively stable within short generation windows, which motivates sparse updates instead of per-step recomputation. To quantify this effect, we vary the update interval from 2 to 20 and report the corresponding performance on the POPE dataset in Figure 4(a). While smaller intervals achieve marginally better results, they suffer from higher computational overhead due to more counterfactual analysis. In contrast, update intervals of 10-15 maintain comparable performance while significantly improving decoding efficiency. Additional visualizations of head distribution dynamics across different generation steps are displayed in Appendix D.3.

Equilibrium factor α. We next investigate the effect of the equilibrium factor α, which controls the trade-off between visual and language information in synergy heads. We evaluate both hallucination suppression and general generation quality under different settings of α. As shown in Figure 4(b), on CHAIR with LLaVA-1.5, the hallucination metric exhibits a U-shaped trend: performance first improves as α increases, but degrades when α becomes too large. This suggests that an intermediate calibration is necessary to reduce hallucination effectively; overly aggressive calibration suppresses language information and may instead harm generation quality.

We observe a similar pattern on LLaVA-Bench for both LLaVA-1.5 and Qwen2.5-VL in Table 3, although the optimal value of α differs across models and tasks. This is expected, as different model architectures and tasks impose different demands on visual evidence. For coarse-grained tasks, limited visual information may be sufficient, whereas fine-grained tasks involving spatial reasoning typically

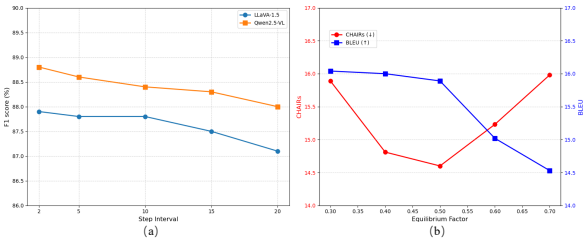


Figure 4: (a) Effect of update interval on model performance. (b) Effect of equilibrium factor on $CHAIR_s$ and BLEU (LLaVA-1.5).

Table 3: LLaVA-Bench scores under different equilibrium factors.

Equilibrium Factor	LLaVA-1.5	Qwen2.5-VL
0.3	72.1	75.6
0.4	73.6	78.3
0.5	74.6	78.5
0.6	75.2	78.2
0.7	74.8	77.9

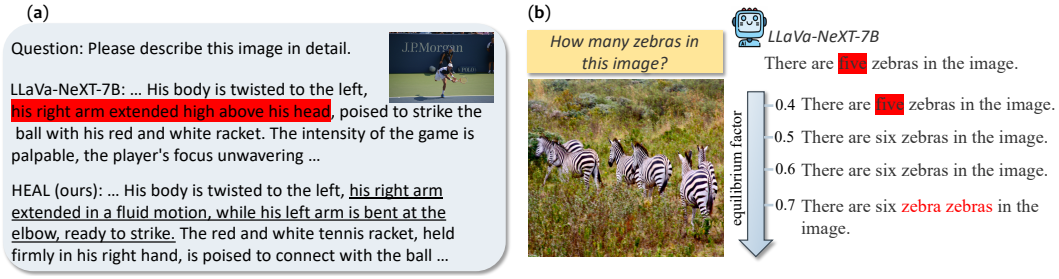


Figure 5: (a) An example of HEAL eliminating the hallucinated content. (b) The generated content under varying settings of the equilibrium factor.

benefit from stronger visual grounding and thus a larger α. Moreover, models with higher-quality visual representations can often achieve good performance with a smaller equilibrium factor.

Qualitative results. Figure 5 provides qualitative examples of hallucination mitigation. In particular, Figure 5(b) shows that hallucinations can only be effectively reduced once α reaches an appropriate range; if α is too large, the model may produce ungrammatical or degraded outputs, which harms the overall performance. We also note that some errors remain unresolved even with a large α. This is likely because the relevant visual information has already been lost during visual encoding or early fusion, and such cases are better addressed by improving the model architecture or training procedure.

4.5 Bidirectional Causal Analysis and Robustness Analysis

To support our claim that the information disequilibrium in synergy heads can cause hallucinations, we conduct a bidirectional intervention by directly manipulating the visual-language information ratio in Appendix D.1. Specifically, decreasing the visual information proportion can induce hallucinations in originally correct generations, whereas increasing it can alleviate hallucinations in originally hallucinated generations. These results provide stronger evidence that the visual-language information distribution is not merely correlated with hallucination behavior but can be causally intervened upon to change it. We further conduct comprehensive robustness analyses to examine whether the identified head taxonomy and observed information drift are sensitive to different counterfactual replacements, masking strategies, measurements, and threshold selections. Across these settings, the head assignments and drift patterns remain highly consistent, supporting the stability of our interpretation. Detailed experimental settings and results are provided in Appendix D.2.

5 Conclusion

In this paper, we reveal that hallucinations occur when information distribution drifts away from the equilibrium in synergy heads, and propose a dynamic information calibration strategy at inference time to mitigate hallucinations. Despite its effectiveness, the calibration factors and update interval are currently determined empirically and may vary across models and tasks. Moreover, hallucinations caused by early visual encoding failures or missing visual evidence cannot always be resolved by attention head calibration alone. These limitations suggest that future work should explore more adaptive calibration policies and tighter integration with model training.

AI USE STATEMENT

We used generative AI tools only for language editing and polishing, including improving grammar, clarity, and overall readability of the manuscript. All AI-assisted edits were carefully reviewed and verified by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We have made substantial efforts to ensure that the proposed method and experimental results are reproducible. The main paper provides detailed descriptions of the algorithm, model architecture, and experimental settings, while additional implementation and experimental details are presented in the appendix. We will release the source code on GitHub in a subsequent stage.

REFERENCES

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29915–29926, 2025. 3, 7

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. 1, 2, 7, 15, 25

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. 7

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 1, 3, 15

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. 18, 19

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. PerturboLLaVA: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*, 2025. 1, 3

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024a. 7

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b. 1, 7, 15

Zhiyuan Chen, Yuecong Min, Jie Zhang, Bei Yan, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. A survey of multimodal hallucination evaluation and detection. *International Journal of Computer Vision*, 134(3):131, 2026. 1, 15

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023. 7

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, 2024. 6

Luca Faes, Laura Sparacino, Gorana Mijatovic, Yuri Antonacci, Leonardo Ricci, Daniele Marinazzo, and Sebastiano Stramaglia. Partial information rate decomposition. *Physical Review Letters*, 135 (18):187401, 2025. 18

Wanlong Fang, Tianle Zhang, Wen Tao, and Alvin Chan. Towards understanding modality interaction in multimodal language models via partial information decomposition. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. 21

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 7

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in LVLMs with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 3488–3501, Vienna, Austria, 2025. Association for Computational Linguistics. 1, 3, 5, 7, 26

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024. 3, 7

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-Introspective Decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations*, 2025. 7

Robin AA Ince. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):318, 2017. 19

Jitesh Jain, Jianwei Yang, and Humphrey Shi. VCoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27992–28002, 2024. 1

Dongzhi JIANG, Jun He, Baichuan Zhou, Zilong Huang, Zhiyuan Yan, Hongsheng Li, Conghui He, Weijia Li, et al. Blink-twice: You see, but do you observe? a reasoning benchmark on visual perception. *Advances in Neural Information Processing Systems*, 38, 2026. 7, 25

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025. 1

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, 2025. 3, 7, 8

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. 6

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. 1, 3, 7

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a. 15

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023b. 1

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023c. 7, 15

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 35799–35819, 2025. 7

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023. 21

Hubert W Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967. 23

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, 2024a. 3

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. 15

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26286–26296, 2024b. 1, 7, 15

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. LLaVA Blog, January 2024c. 1, 2, 7, 15

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision*. Springer, 2024d. 3

OpenAI. GPT-4o system card. Technical report, OpenAI, 2024. 15

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. ConVis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6434–6442, 2025. 1, 3

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018. 7, 15

Yan Shu, Hangui Lin, Yexin Liu, Yan Zhang, Gangyan Zeng, Yan Li, Yu Zhou, Ser-Nam Lim, Harry Yang, and Nicu Sebe. When semantics mislead vision: Mitigating large multimodal models hallucinations in scene text spotting and understanding. *arXiv preprint arXiv:2506.05551*, 2025. 15

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110. Association for Computational Linguistics, 2024a. 3

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024b. 7, 25, 26

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 *International Conference on Learning Representations*, 2025a. 7, 8

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, Wenxue Li, Yulong Li, Wenxuan Song, Shiyan Su, Wei Feng, Jionglong Su, Minquan Lin, Yifan Peng, Xuelian Cheng, Imran Razzak, and Zongyuan Ge. Seeing far and clearly: Mitigating hallucinations in MLLMs with attention causal decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26147–26159, 2025b. 3, 7

Chao Tian and Shlomo Shamai. Broadcast channel cooperative gain: An operational interpretation of partial information decomposition. *Entropy*, 27(3):310, 2025. 18

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2

Zifu Wan, Ce Zhang, Silong Yong, Martin Q. Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. ONLY: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3225–3234, 2025. 7

Jiahui Wang, Zuyan Liu, Yongming Rao, and Jiwen Lu. Sparsemm: Head sparsity emerges from visual concept responses in mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23177–23187, 2025a. 5, 26

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 7

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. 7, 15, 25

Jinfeng Wei and Xiaofeng Zhang. DOPRA: Decoding over-accumulation penalization and re-allocation in specific weighting layer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7065–7074, 2024. 7

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010. 4, 18, 19

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 3

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13040–13051, 2024. 15

In-Kwon Yeo and Richard A Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000. 23

Liu Yu, Zhonghao Chen, Ping Kuang, Zhikun Feng, Fan Zhou, Lan Wang, and Gillian Dobbie. Causally-grounded dual-path attention intervention for object hallucination mitigation in LVLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36021–36029, 2026. 1, 3

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. HalluciDoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12944–12953, 2024. 1, 3

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019. 17

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pp. 196–213. Springer, 2024a. 1, 3

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in LVLMs. *arXiv preprint arXiv:2411.09968*, 2024b. 3, 7, 8

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Barrett Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination begins where saliency drops. In *International Conference on Learning Representations*, volume 2026, pp. 15062–15083, 2026. 7, 25

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing LVLMs through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. 1, 3

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, Huiyu Zhou, and Xuming Hu. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6193–6212, 2025. 15

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *International Conference on Learning Representations*, 2025. 3, 7

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024. 3

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, 2024. 15

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. IBD: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1624–1633, 2025. 1

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024. 7

A More Related Works

Multimodal Large Language Models. MLLMs extend large language models with visual perception, enabling them to encode visual inputs into language-aligned representations and generate instruction-following responses grounded in visual evidence (Li et al., 2023a; Bai et al., 2025a; Chen et al., 2024b; Liu et al., 2024c;b). The development of MLLMs has evolved from modular vision-language alignment to more general-purpose multimodal intelligence. Early models such as BLIP-2 (Li et al., 2023a) bridged frozen image encoders and large language models through lightweight alignment modules, while LLaVA demonstrated the effectiveness of visual instruction tuning (Liu et al., 2023). Subsequent representative models further expanded MLLM capabilities, including MiniGPT-4 (Zhu et al., 2024), mPLUG-Owl2 (Ye et al., 2024), LLaVA-NeXT (Liu et al., 2024c), Qwen3-VL (Bai et al., 2025a), and InternVL-3.5 (Wang et al., 2025b), showing strong performance on multimodal understanding, grounding, document parsing, etc. Despite these advances, recent studies show that MLLMs may generate fluent but visually inconsistent outputs, including object, attribute, relation and semantic hallucinations (Chen et al., 2026; Bai et al., 2024; Li et al., 2023c; Rohrbach et al., 2018; Shu et al., 2025; Zheng et al., 2025). These hallucinations can undermine user trust, reduce system reliability, and even lead to harmful downstream decisions in real-world applications (OpenAI, 2024).

B Proofs

B.1 Theorem 1

Proof. For an attention head, we can decompose its output into visual and language components by Equation 1:

$$\mathbf{o}^{(t,l,i)} = [\mathbf{A}_{vis}^{(t,l,i)}, \mathbf{A}_{lang}^{(t,l,i)}] \begin{bmatrix} \mathbf{V}_{vis}^{(t,l,i)} \\ \mathbf{V}_{lang}^{(t,l,i)} \end{bmatrix} = \mathbf{A}_{vis}^{(t,l,i)} \mathbf{V}_{vis}^{(t,l,i)} + \mathbf{A}_{lang}^{(t,l,i)} \mathbf{V}_{lang}^{(t,l,i)} = \mathbf{h}_{\mathrm{v}} + \mathbf{h}_{\mathrm{l}}, \tag{19}$$

where $\mathbf{A}_{vis}^{(t,l,i)}$ and $\mathbf{A}_{lang}^{(t,l,i)}$ denote the attention weights corresponding to the visual and language tokens, respectively. Let's assume the continuous information measure $\mathcal{I}(\cdot)$ is additive and positively homogeneous, i.e.,

$$\mathcal{I}(\lambda \mathbf{z}) = \lambda \mathcal{I}(\mathbf{z}), \qquad \lambda \geq 0. \tag{20}$$

Then we can compute the original visual-language ratio as

$$\alpha_{old} = \frac{\mathcal{I}(\mathbf{h}_{\mathrm{v}})}{\mathcal{I}(\mathbf{h}_{\mathrm{v}}) + \mathcal{I}(\mathbf{h}_{\mathrm{l}})}. \tag{21}$$

Given a target equilibrium factor $\alpha \in (0, 1)$, we set

$$\beta = \frac{\alpha}{\alpha_{old}}, \qquad \gamma = \frac{1-\alpha}{1-\alpha_{old}}. \tag{22}$$

After calibrating the V vectors, we can get

$$\mathbf{V}_{vis}^{'} \leftarrow \beta \mathbf{V}_{vis}, \qquad \mathbf{V}_{lang}^{'} \leftarrow \gamma \mathbf{V}_{lang}, \tag{23}$$

the visual and language components in Equation 19 become

$$\mathbf{h}'_{\mathrm{v}} = \beta \mathbf{h}_{\mathrm{v}}, \qquad \mathbf{h}'_{\mathrm{l}} = \gamma \mathbf{h}_{\mathrm{l}}. \tag{24}$$

The new information ratio is

$$\begin{aligned} \alpha_{\text{new}} &= \frac{\mathcal{I}(\mathbf{h}'_{\mathrm{v}})}{\mathcal{I}(\mathbf{h}'_{\mathrm{v}}) + \mathcal{I}(\mathbf{h}'_{\mathrm{l}})} \\ &= \frac{\beta \mathcal{I}(\mathbf{h}_{\mathrm{v}})}{\beta \mathcal{I}(\mathbf{h}_{\mathrm{v}}) + \gamma \mathcal{I}(\mathbf{h}_{\mathrm{l}})} \\ &= \alpha. \end{aligned} \tag{25}$$

Therefore, calibrating the value vectors is equivalent to steering the modality information distribution toward the target factor. In our implementation, $\mathcal{I}(\cdot)$ is instantiated by the counterfactual contribution score estimated via difference-in-differences, so the theorem holds up to a local additive approximation of the head output. □

B.2 THEOREM 2

Proof. According to Equation 19 and Equation 24, the calibrated output of an attention head can be decomposed as

$$\mathbf{h}(\alpha) = \beta(\alpha)\mathbf{h}_{\mathrm{v}} + \gamma(\alpha)\mathbf{h}_{\mathrm{l}}. \tag{26}$$

The cosine alignment between the calibrated head output and the visual component is

$$A_v(\alpha) = \cos\big(\mathbf{h}(\alpha), \mathbf{h}_{\mathrm{v}}\big) = \frac{\langle \mathbf{h}(\alpha), \mathbf{h}_{\mathrm{v}} \rangle}{\|\mathbf{h}(\alpha)\|_2 \|\mathbf{h}_{\mathrm{v}}\|_2}. \tag{27}$$

Let $c = \frac{\langle \mathbf{h}_{\mathrm{v}}, \mathbf{h}_{\mathrm{l}} \rangle}{\|\mathbf{h}_{\mathrm{v}}\|_2 \|\mathbf{h}_{\mathrm{l}}\|_2}$, then $-1 \le c \le 1$. Using Equation 26, the numerator of Equation 27 becomes

$$\langle \mathbf{h}(\alpha), \mathbf{h}_{\mathrm{v}} \rangle = \beta \|\mathbf{h}_{\mathrm{v}}\|_2^2 + \gamma \langle \mathbf{h}_{\mathrm{l}}, \mathbf{h}_{\mathrm{v}} \rangle = \beta \|\mathbf{h}_{\mathrm{v}}\|_2^2 + \gamma \|\mathbf{h}_{\mathrm{v}}\|_2 \|\mathbf{h}_{\mathrm{l}}\|_2 c. \tag{28}$$

Similarly,

$$\begin{aligned} \|\mathbf{h}(\alpha)\|_2^2 &= \|\beta \mathbf{h}_{\mathrm{v}} + \gamma \mathbf{h}_{\mathrm{l}}\|_2^2 \\ &= \beta^2 \|\mathbf{h}_{\mathrm{v}}\|_2^2 + \gamma^2 \|\mathbf{h}_{\mathrm{l}}\|_2^2 + 2\beta\gamma \|\mathbf{h}_{\mathrm{v}}\|_2 \|\mathbf{h}_{\mathrm{l}}\|_2 c. \end{aligned} \tag{29}$$

Substituting Equation 28 and Equation 29 into Equation 27 yields

$$\begin{aligned} A_v(\alpha) &= \frac{\beta \|\mathbf{h}_{\mathrm{v}}\|_2 + \gamma \|\mathbf{h}_{\mathrm{l}}\|_2 c}{\sqrt{\beta^2 \|\mathbf{h}_{\mathrm{v}}\|_2^2 + \gamma^2 \|\mathbf{h}_{\mathrm{l}}\|_2^2 + 2\beta\gamma \|\mathbf{h}_{\mathrm{v}}\|_2 \|\mathbf{h}_{\mathrm{l}}\|_2 c}} \\ &= \frac{1 + \tau(\alpha) c}{\sqrt{1 + \tau(\alpha)^2 + 2\tau(\alpha) c}}, \end{aligned} \tag{30}$$

where $\tau(\alpha) = \frac{\gamma \|\mathbf{h}_{\mathrm{l}}\|_2}{\beta \|\mathbf{h}_{\mathrm{v}}\|_2}$. Substituting the explicit forms of $\beta(\alpha)$ and $\gamma(\alpha)$ gives

$$\frac{\partial \tau(\alpha)}{\partial \alpha} = -\frac{\alpha_{vis}}{\alpha_{lang} \alpha^2} \frac{\|\mathbf{h}_{\mathrm{l}}\|_2}{\|\mathbf{h}_{\mathrm{v}}\|_2} < 0. \tag{31}$$

Hence, increasing the equilibrium factor α strictly decreases the relative magnitude of the language component with respect to the visual component.

It remains to determine how the cosine alignment changes as τ decreases. Differentiating $A_v(\alpha)$ with respect to τ gets

$$\begin{aligned} \frac{\partial A_v(\alpha)}{\partial \tau} &= c \left(1 + \tau^2 + 2\tau c\right)^{-1/2} \\ &\quad - (1 + \tau c) \left(1 + \tau^2 + 2\tau c\right)^{-3/2} (\tau + c) \\ &= \frac{c(1 + \tau^2 + 2\tau c) - (1 + \tau c)(\tau + c)}{(1 + \tau^2 + 2\tau c)^{3/2}} \\ &= -\frac{\tau(1 - c^2)}{(1 + \tau^2 + 2\tau c)^{3/2}} \le 0, \end{aligned} \tag{32}$$

because $\tau > 0, 1 - c^2 \ge 0$. Combining Equation 31 and Equation 32, the chain rule gives

$$\begin{aligned} \frac{\partial A_v(\alpha)}{\partial \alpha} &= \frac{\partial f(\tau)}{\partial \tau} \frac{\partial \tau(\alpha)}{\partial \alpha} \\ &= \frac{\tau(\alpha)(1 - c^2)}{[1 + \tau(\alpha)^2 + 2\tau(\alpha) c]^{3/2}} \frac{\alpha_{vis}}{\alpha_{lang} \alpha^2} \frac{\|\mathbf{h}_{\mathrm{l}}\|_2}{\|\mathbf{h}_{\mathrm{v}}\|_2} \ge 0. \end{aligned} \tag{33}$$

Hence, the visual alignment of the calibrated head output is monotonically non-decreasing with respect to the equilibrium factor. Moreover, because $\tau(\alpha) > 0$ for $\alpha \in (0, 1)$, the derivative is strictly positive whenever $1 - c^2 > 0 \iff |c| < 1$. Then, unless the visual and language components are collinear, increasing α strictly increases the cosine alignment with the visual component.

Extension to residual connection and RMSNorm (Zhang & Sennrich, 2019). Let the residual representation after the calibrated attention head be

$$\mathbf{r}(\alpha) = \mathbf{x} + \mathbf{h}(\alpha) = \mathbf{x} + \beta(\alpha)\mathbf{h}_{\mathrm{v}} + \gamma(\alpha)\mathbf{h}_{\mathrm{l}}, \tag{34}$$

and the output of the subsequent RMSNorm layer is

$$\mathbf{z}(\alpha) = \mathrm{RMSNorm}(\mathbf{r}(\alpha)) = \frac{G\mathbf{r}(\alpha)}{\rho(\mathbf{r}(\alpha))}, \tag{35}$$

where

$$G = \mathrm{diag}(\mathbf{g}), \qquad \rho(\mathbf{r}) = \sqrt{\frac{1}{d}\|\mathbf{r}\|_2^2 + \epsilon} > 0.$$

We define the gain-transformed representations

$$G\mathbf{r}(\alpha) = \bar{\mathbf{x}} + \beta(\alpha)\bar{\mathbf{h}}_{\mathrm{v}} + \gamma(\alpha)\bar{\mathbf{h}}_{\mathrm{l}}. \tag{36}$$

Since $\rho(\mathbf{r}(\alpha))$ is a positive scalar, it does not affect cosine alignment. Therefore, the visual alignment of the final RMSNorm output can be written as

$$A_{\mathrm{v}}^{\mathrm{out}}(\alpha) = \cos\left(\mathbf{z}(\alpha), \bar{\mathbf{h}}_{\mathrm{v}}\right) = \cos\left(G\mathbf{r}(\alpha), \bar{\mathbf{h}}_{\mathrm{v}}\right) = \frac{\langle G\mathbf{r}(\alpha), \bar{\mathbf{h}}_{\mathrm{v}}\rangle}{\|G\mathbf{r}(\alpha)\|_2 \|\bar{\mathbf{h}}_{\mathrm{v}}\|_2}. \tag{37}$$

Because $\beta'(\alpha) = \frac{1}{\alpha_{\mathrm{vis}}}, \gamma'(\alpha) = -\frac{1}{\alpha_{\mathrm{lang}}}$, we have

$$[G\mathbf{r}(\alpha)]' = \mathbf{w} = \frac{\bar{\mathbf{h}}_{\mathrm{v}}}{\alpha_{\mathrm{vis}}} - \frac{\bar{\mathbf{h}}_{\mathrm{l}}}{\alpha_{\mathrm{lang}}}. \tag{38}$$

Differentiating 37 with respect to α gives

$$\frac{\partial A_{\mathrm{v}}^{\mathrm{out}}(\alpha)}{\partial \alpha} = \frac{\langle \mathbf{w}, \bar{\mathbf{h}}_{\mathrm{v}}\rangle \|G\mathbf{r}(\alpha)\|_2^2 - \langle G\mathbf{r}(\alpha), \bar{\mathbf{h}}_{\mathrm{v}}\rangle \langle G\mathbf{r}(\alpha), \mathbf{w}\rangle}{\|G\mathbf{r}(\alpha)\|_2^3 \|\bar{\mathbf{h}}_{\mathrm{v}}\|_2}. \tag{39}$$

To characterize the sign of this derivative, decompose $G\mathbf{r}(\alpha)$ and $\mathbf{w}$ into components parallel and orthogonal to the visual direction $\bar{\mathbf{h}}_{\mathrm{v}}$. Let $\mathbf{u}_{\mathrm{v}} = \frac{\bar{\mathbf{h}}_{\mathrm{v}}}{\|\bar{\mathbf{h}}_{\mathrm{v}}\|_2}$, and write

$$\mathbf{G}\mathbf{r}(\alpha) = s(\alpha)\mathbf{u}_{\mathrm{v}} + \mathbf{p}(\alpha), \qquad \mathbf{p}(\alpha) \perp \mathbf{u}_{\mathrm{v}}, \tag{40}$$

and

$$\mathbf{w} = t\mathbf{u}_{\mathrm{v}} + \mathbf{q}, \qquad \mathbf{q} \perp \mathbf{u}_{\mathrm{v}}, \tag{41}$$

where $s(\alpha) = \langle G\mathbf{r}(\alpha), \mathbf{u}_{\mathrm{v}}\rangle, t = \langle \mathbf{w}, \mathbf{u}_{\mathrm{v}}\rangle$. So $s'(\alpha) = t, \mathbf{p}'(\alpha) = \mathbf{q}$. Substituting these decompositions into Equation 37 and Equation 39, we can get

$$\begin{aligned} A_{\mathrm{v}}^{\mathrm{out}}(\alpha) &= \frac{s(\alpha)}{\sqrt{s(\alpha)^2 + \|\mathbf{p}(\alpha)\|_2^2}}, \\ \frac{\partial A_{\mathrm{v}}^{\mathrm{out}}(\alpha)}{\partial \alpha} &= \frac{t\|\mathbf{p}(\alpha)\|_2^2 - s(\alpha)\langle \mathbf{p}(\alpha), \mathbf{q}\rangle}{\left(s(\alpha)^2 + \|\mathbf{p}(\alpha)\|_2^2\right)^{3/2}}. \end{aligned} \tag{42}$$

If the calibrated representation after RMSNorm satisfies

$$\mathcal{M}(\alpha) = t\|\mathbf{p}(\alpha)\|_2^2 - s(\alpha)\langle \mathbf{p}(\alpha), \mathbf{q}\rangle \geq 0, \tag{43}$$

then

$$\frac{\partial A_{\mathrm{v}}^{\mathrm{out}}(\alpha)}{\partial \alpha} \geq 0.$$

The residual compatibility margin $\mathcal{M}(\alpha)$ has a simple geometric interpretation. In Equation 43, t measures the α-induced representation change along the visual direction, while $\mathbf{p}(\alpha)$ denotes the non-visual component orthogonal to the visual. The vector $\mathbf{q}$ describes how this orthogonal component changes with α. The first term $t\|\mathbf{p}(\alpha)\|_2^2$ represents the visual-alignment gain induced by increasing α, whereas the second term $s(\alpha)\langle \mathbf{p}(\alpha), \mathbf{q}\rangle$ measures the opposing angular perturbation introduced by the residual branch. Therefore, $\mathcal{M}(\alpha) \geq 0$ characterizes the regime in which increasing α moves the residual representation toward the visual direction without introducing a stronger opposing angular change. In particular, the condition is naturally satisfied when $t \geq 0$ and the orthogonal residual component does not move further in its current direction, i.e., $\langle \mathbf{p}(\alpha), \mathbf{q}\rangle \leq 0$. Additionally, RMSNorm only introduces the positive scalar factor $\rho(\mathbf{r}(\alpha))^{-1}$ after the fixed gain transformation G, so it does not lead to any additional angular variation. The above monotonicity analysis holds up to the visual alignment after both the residual connection and RMSNorm. □

C Theoretical Analysis of HEAL

C.1 Theoretical Contributions

Our theoretical results include:

- First, value vector calibration is equivalent to adjusting the information distribution between visual and language components within synergy heads (Theorem 1).
- Second, the calibration changes the representation geometry predictably, rather than acting as an empirical heuristic (Theorem 2).

Beyond these theorems, HEAL also provides a potential information decomposition framework for Partial Information Decomposition (PID) Theory (Williams & Beer, 2010; Faes et al., 2025; Tian & Shamai, 2025), where multimodal information can be interpreted as visual-specific, language-specific, shared (synergistic), and redundant components. Strict PID decomposition generally requires strong mathematical formulations that are difficult to directly instantiate in modern neural representations. HEAL provides a practical decomposition instantiation for multimodal transformers.

C.2 Discussions between PID Theory and HEAL

Our decomposition framework is closely related to the classical PID theory (Williams & Beer, 2010), which characterizes the information that multiple sources provide about a target in terms of redundant, unique, and synergistic components. Let X_V and X_L denote the visual and language sources, and Y a target variable associated with the information represented by a multimodal attention head. PID decomposes the joint mutual information as

$$\begin{aligned} I(Y; X_V, X_L) &= \int_{\mathcal{Y}} \int_{\mathcal{X}_V} \int_{\mathcal{X}_L} p(y, x_V, x_L) \log \frac{p(y, x_V, x_L)}{p(y)p(x_V, x_L)} \, \mathrm{d}x_L \, \mathrm{d}x_V \, \mathrm{d}y \\ &= R(Y; X_V, X_L) + U_V(Y; X_V|X_L) + U_L(Y; X_L|X_V) + S(Y; X_V, X_L). \end{aligned} \tag{44}$$

where R denotes redundant information shared by the two sources, U_V and U_L denote source-specific information, and S denotes information that emerges only from the joint availability of both sources. Equivalently, using conditional entropy,

$$\begin{aligned} I(Y; X_V, X_L) &= H(Y) - H(Y \mid X_V, X_L) \\ &= \int p(y, x_V, x_L) \log \frac{p(y \mid x_V, x_L)}{p(y)} \, \mathrm{d}y \, \mathrm{d}x_V \, \mathrm{d}x_L. \end{aligned} \tag{45}$$

$I(Y; X_V, X_L)$ measures the uncertainty reduction about Y obtained by jointly observing the visual and language sources. In PID theory, marginal mutual information contains both redundant and unique information, whereas conditional mutual information contains unique and synergistic information. Thus, we have

$$I(Y; X_V) = \int_{\mathcal{Y}} \int_{\mathcal{X}_V} p(y, x_V) \log \frac{p(y \mid x_V)}{p(y)} \, \mathrm{d}x_V \, \mathrm{d}y = R + U_V, \tag{46}$$

$$I(Y; X_L) = \int_{\mathcal{Y}} \int_{\mathcal{X}_L} p(y, x_L) \log \frac{p(y \mid x_L)}{p(y)} \, \mathrm{d}x_L \, \mathrm{d}y = R + U_L, \tag{47}$$

$$\begin{aligned} I(Y; X_V \mid X_L) &= \int p(y, x_V, x_L) \log \frac{p(y \mid x_V, x_L)}{p(y \mid x_L)} \, \mathrm{d}y \, \mathrm{d}x_V \, \mathrm{d}x_L \\ &= H(Y \mid X_L) - H(Y \mid X_V, X_L) = U_V + S, \end{aligned} \tag{48}$$

$$\begin{aligned} I(Y; X_L \mid X_V) &= \int p(y, x_V, x_L) \log \frac{p(y \mid x_V, x_L)}{p(y \mid x_V)} \, \mathrm{d}y \, \mathrm{d}x_V \, \mathrm{d}x_L \\ &= H(Y \mid X_V) - H(Y \mid X_V, X_L) = U_L + S. \end{aligned} \tag{49}$$

Importantly, these mutual information quantities in 46-49 provide only three independent equations for the four unknown PID atoms. Therefore, the redundancy term R cannot be uniquely determined from Shannon mutual information alone. A specific redundancy function must be introduced (Bertschinger

et al., 2014; Ince, 2017). For example, under the original $I_{\min}$ construction of Williams and Beer (Williams & Beer, 2010), the pointwise specific information of a source X about an outcome y is

$$I_{\text{spec}}(y; X) = D_{\text{KL}}\left[p(x \mid y) \| p(x)\right] = \int p(x \mid y) \log \frac{p(x \mid y)}{p(x)} \, \mathrm{d}x. \tag{50}$$

The redundancy between the visual and language sources is then

$$R_{\min}(Y; X_V, X_L) = \int_{\mathcal{Y}} p(y) \min\left\{I_{\text{spec}}(y; X_V), I_{\text{spec}}(y; X_L)\right\} \, \mathrm{d}y. \tag{51}$$

Once a redundancy function $R^\star$ is selected, the remaining PID atoms follow algebraically:

$$U_V^\star = I(Y; X_V) - R^\star, U_L^\star = I(Y; X_L) - R^\star,$$

$$S^\star = I(Y; X_V, X_L) - I(Y; X_V) - I(Y; X_L) + R^\star. \tag{52}$$

Equation 52 is particularly relevant to Equation 11 in HEAL. It makes explicit that the synergistic atom is not simply the difference between joint and marginal information: the redundancy term must also be restored. Different choices of $R^\star$ lead to different PID decompositions. For example, BROJA defines unique and synergistic information through an optimization over distributions having fixed source-target marginals (Bertschinger et al., 2014), while the CCS approach defines redundancy by identifying common pointwise changes in surprisal (Ince, 2017). The existence of these alternatives emphasizes that PID provides a decomposition principle rather than a unique redundancy estimator.

HEAL does not estimate $p(y, x_V, x_L)$ explicitly. Instead, it observes the change induced in an attention head representation after selectively removing visual and language inputs. For two hidden states a and b, HEAL defines the representation discrepancy as

$$\begin{aligned} D(a, b) &= 1 - \text{Sim}(a, b) \\ &= \frac{1}{2}\left[1 - \frac{\langle a, b\rangle}{\|a\|_2 \|b\|_2}\right]. \end{aligned} \tag{53}$$

Rather than interpreting $D(a, b)$ as a Shannon mutual information, we regard it as a task-relevant measure of the information contribution associated with the corresponding counterfactual intervention. In fact, for a distribution of inputs (V, T), the expected counterfactual information is written as

$$\begin{aligned} \mathcal{I} &= \mathbb{E}_{(V,T)\sim p(V,T)}\left[D\left(H(V, T), H^{\text{cf}}(V, T)\right)\right] \\ &= \int_{\mathcal{V}} \int_{\mathcal{T}} p(v, t) \frac{1}{2}\left[1 - \frac{\langle H(v, t), H^{\text{cf}}(v, t)\rangle}{\|H(v, t)\|_2 \|H^{\text{cf}}(v, t)\|_2}\right] \mathrm{d}v \, \mathrm{d}t. \end{aligned} \tag{54}$$

This expectation-level formulation makes clear the analogy with mutual information: Shannon information averages a pointwise log-likelihood ratio over samples, whereas HEAL averages a pointwise representation discrepancy over counterfactual interventions. The two quantities have different statistical definitions, but they share the same operational structure of measuring information contribution through changes induced by conditioning or intervention.

In HEAL framework (Section 3.2), we define

$$D_{00} = D(H_{11}, H_{00}), \qquad D_{10} = D(H_{11}, H_{10}), \qquad D_{01} = D(H_{11}, H_{01}). \tag{55}$$

Thus, the total information contribution of the two modalities is

$$I_{\text{total}} = D_{00} = \frac{1}{2}\left[1 - \frac{\langle H_{11}, H_{00}\rangle}{\|H_{11}\|_2 \|H_{00}\|_2}\right]. \tag{56}$$

The visual contribution is obtained by subtracting the discrepancy that remains after masking language:

$$I_{\text{vis}} = D_{00} - D_{10} = \frac{1}{2}\left[\frac{\langle H_{11}, H_{10}\rangle}{\|H_{11}\|_2 \|H_{10}\|_2} - \frac{\langle H_{11}, H_{00}\rangle}{\|H_{11}\|_2 \|H_{00}\|_2}\right]. \tag{57}$$

Similarly,

$$I_{\text{lang}} = D_{00} - D_{01} = \frac{1}{2}\left[\frac{\langle H_{11}, H_{01}\rangle}{\|H_{11}\|_2 \|H_{01}\|_2} - \frac{\langle H_{11}, H_{00}\rangle}{\|H_{11}\|_2 \|H_{00}\|_2}\right]. \tag{58}$$

The remaining term is obtained by

$$\begin{aligned} I_{\text{syn}} &= I_{\text{total}} - I_{\text{vis}} - I_{\text{lang}} \\ &= D_{00} - (D_{00} - D_{10}) - (D_{00} - D_{01}) \\ &= D_{10} + D_{01} - D_{00}. \end{aligned} \tag{59}$$

The relationship to PID becomes explicit under an information-faithfulness assumption. Suppose that, in expectation over the data distribution, the counterfactual representation discrepancies approximate the corresponding Shannon information quantities:

$$\begin{aligned} \mathbb{E}[I_{\text{total}}] &\approx I(Y; X_V, X_L), \\ \mathbb{E}[I_{\text{vis}}] &\approx I(Y; X_V), \\ \mathbb{E}[I_{\text{lang}}] &\approx I(Y; X_L). \end{aligned} \tag{60}$$

Substituting the PID identities into the HEAL interaction term gives

$$\begin{aligned} I_{\text{syn}} &\approx I(Y; X_V, X_L) - I(Y; X_V) - I(Y; X_L) \\ &= (R + U_V + U_L + S) - (R + U_V) - (R + U_L) \\ &= S - R. \end{aligned} \tag{61}$$

At the same time,

$$\begin{aligned} I_{\text{vis}} &\approx R + U_V, \\ I_{\text{lang}} &\approx R + U_L, \\ I_{\text{total}} &\approx R + U_V + U_L + S. \end{aligned} \tag{62}$$

Thus, the modality-specific scores of HEAL should be understood as modality-attributable information rather than strictly unique information. Let $R^\star$ be any admissible redundancy function; then

$$\boxed{\begin{aligned} U_V^\star &\approx I_{\text{vis}} - R^\star, \\ U_L^\star &\approx I_{\text{lang}} - R^\star, \\ S^\star &\approx I_{\text{syn}} + R^\star. \end{aligned}} \tag{63}$$

Indeed,

$$\begin{aligned} R^\star + U_V^\star + U_L^\star + S^\star &\approx R^\star + (I_{\text{vis}} - R^\star) + (I_{\text{lang}} - R^\star) + (I_{\text{syn}} + R^\star) \\ &= I_{\text{vis}} + I_{\text{lang}} + I_{\text{syn}} \\ &= I_{\text{total}}, \end{aligned} \tag{64}$$

which demonstrates that an explicit redundancy estimate provides the missing degree of freedom required to transform the HEAL decomposition into a full PID decomposition.

The above derivation also explains a characteristic property of the HEAL synergy score. In strict PID, the synergistic atom S is defined after explicitly separating the redundancy R. In HEAL, this redundancy is not separately estimated and is implicitly counted in both I_{vis} and I_{lang}. As a result, $I_{\text{syn}}^{\text{HEAL}} \approx S - R$, rather than S itself. Therefore,

$$\boxed{\begin{aligned} I_{\text{syn}} > 0 &\iff S > R, \\ I_{\text{syn}} = 0 &\iff S = R, \\ I_{\text{syn}} < 0 &\iff S < R. \end{aligned}} \tag{65}$$

A negative HEAL synergy score should not be interpreted as negative Shannon synergy. It indicates that shared information is sufficiently large to dominate the net joint interaction. This is also why the signed nature of I_{syn} is not inconsistent with the non-negativity of the atoms in the classical PID formulation. The original PID was introduced partly to separate the non-negative synergy and redundancy atoms that become conflated in conventional interaction information.

The preceding results suggest a natural two-stage interpretation of HEAL. First, the counterfactual Difference-in-Differences procedure identifies modality-attributable information directly in the representation space:

$$\{H_{11}, H_{10}, H_{01}, H_{00}\} \longrightarrow \{I_{\text{vis}}, I_{\text{lang}}, I_{\text{syn}}\}. \tag{66}$$

Second, if a redundancy functional $R^\star$ is estimated from the underlying joint distribution, the corresponding PID atoms can be obtained through

$$\{I_{\text{vis}}, I_{\text{lang}}, I_{\text{syn}}, R^\star\} \longrightarrow \{U_V^\star, U_L^\star, R^\star, S^\star\}\,. \tag{67}$$

This establishes HEAL as a counterfactual, representation-level analogue of PID rather than an exact replacement for a probability-based PID estimator. The distinction is important: the HEAL discrepancy in Equation 53 is based on cosine geometry of hidden representations, whereas Shannon mutual information is based on a log-density ratio. Nevertheless, both constructions quantify the contribution of a source through the change in information available about a target, and the counterfactual factorial design in HEAL provides a natural operational analogue of marginal and joint source access. This connection is also complementary to existing multimodal information decomposition approaches, which estimate redundancy, uniqueness, and synergy primarily at the modality level (Fang et al., 2026; Liang et al., 2023). In contrast, HEAL applies the decomposition locally to individual attention heads and generation steps, allowing modality interaction to be analyzed and subsequently intervened upon at the level of internal computation.

D More Results and Visualization

D.1 Bidirectional Causal Analysis

Table 4: Bidirectional causal intervention in LLaVA-1.5-7B and Qwen2.5-VL-7B by manipulating the visual-language ratio. Performance scores are evaluated on the LLaVA-Bench.

Ratio	**0.3**	**base (no intervening)**	**0.4**	**0.5**	**0.6**	**0.7**
LLaVA-1.5-7B	72.1	72.5	73.6	74.6	75.2	74.8
Qwen2.5-VL-7B	75.6	76.8	78.3	78.5	78.2	77.9

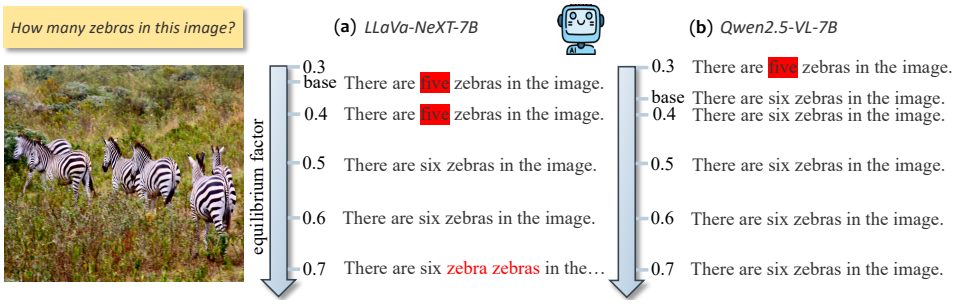


Figure 6: Bidirectional causal intervention in LLaVA-1.5-7B and Qwen2.5-VL-7B replies by manipulating the visual-language information distribution (equilibrium factor). *base* means no intervention.

We conduct a bidirectional causal intervention analysis by directly manipulating the visual-language information ratio. As shown in Figure 6 and Table 4, the results demonstrate two causal directions:

- When the model originally produces a correct response, artificially decreasing the visual information proportion can induce hallucination.
- When the model originally produces hallucination, increasing the visual information proportion can alleviate hallucination.

Therefore, the observed relationship between hallucination and visual-language disequilibrium is not merely correlational. The intervention results show that manipulating this distribution leads to predictable changes in hallucination behavior, providing stronger evidence for our causal claim.

D.2 Robustness Results and Details

The credibility of our interpretation depends on whether the proposed head taxonomy and the observed drift remain stable under different design choices. To support the central claim, we conduct a comprehensive robustness analysis.

Experiment setting. We sampled 50 image-caption pairs generated by LLaVA-Next-7B and Qwen2.5-VL-7B on the CHAIR benchmark, covering both correct and various hallucinated tokens (approximately 9,000 tokens in total). For each token, every attention head is assigned a label, and we further computed the visual-language information ratio within the identified synergy heads for both correct and hallucinated tokens. The reported Head Assignment Agreement is defined as the percentage of head labels that remain identical to those obtained using the default setting across all tokens.

Table 5: Head type assignment agreement under different masking strategies.

Masking strategy	**Gaussian masking (base)**	**one-point (zero) masking**	**uniform masking**	**swapping actual tokens**
Head Assignment Agreement (%)	100.00	92.13	93.21	95.36

Table 6: Robustness to different masking strategies. We report the visual-language information ratios of correct and hallucinated tokens in synergy heads and LLaVA-1.5-7B performance on POPE.

Masking strategy	**Gaussian masking (base)**	**one-point (zero) masking**	**uniform masking**	**swapping actual tokens**
F1 score ↑	87.84	87.12	86.93	87.53
correct tokens	0.43:0.54	0.41:0.53	0.39:0.48	0.41:0.56
hallucinated tokens	0.28:0.62	0.21:0.52	0.23:0.57	0.31:0.66

(1) Robustness to masking strategies. To validate whether our conclusions depend on the specific masking strategy, we further consider zero masking, uniform masking, and swapping the actual visual/language tokens from another image or text. As shown in Table 5, the resulting head taxonomy is highly similar across all masking strategies, achieving 92.13%-95.36% head assignment agreement. More importantly, from Table 6, the central claim that the visual-language information drift between correct and hallucinated tokens, remains consistently observable under every masking strategy. Additionally, all masking strategies yield comparable downstream performance on POPE, with Gaussian masking achieving the best. These results indicate that our taxonomy and the observed drift are properties of the model itself rather than artifacts of a particular masking strategy.

Table 7: Robustness to threshold selection. We report the visual-language information ratios of correct and hallucinated tokens in synergy heads and LLaVA-1.5-7B performance on POPE.

(a) σ_{total} coefficient in Equation 12

σ_{total} coefficient	**1**	**2**	**3**	∞
correct tokens	0.45:0.51	0.40:0.49	0.43:0.54	0.45:0.52
hallucinated tokens	0.24:0.63	0.23:0.61	0.28:0.62	0.29:0.66
F1 score ↑	87.06	87.43	**87.84**	87.65

(b) MAD coefficient

MAD coefficient (λ)	**1.4826**	**2.9652**	**4.4478**	∞
correct tokens	0.43:0.57	0.43:0.54	0.47:0.52	0.38:0.49
hallucinated tokens	0.19:0.56	0.28:0.62	0.20:0.66	0.23:0.65
F1 score ↑	86.95	**87.84**	87.51	86.23

Note. We use MAD to robustly estimate dispersion, where $1.4826 \approx 1/\Phi^{-1}(0.75)$ is the standard consistency factor under normality. Thus, 1.4826, 2.9652, and 4.4478 correspond approximately to the conventional 1σ, 2σ, and 3σ thresholds, respectively.

(2) Robustness to threshold selection. The $3\sigma_{total}$ criterion and λ·MAD modality-ratio threshold indeed affect the distribution of head types because they explicitly define the classification boundaries. As shown in Tables 7a and 7b, although the proportion of heads assigned to each category changes moderately, the fundamental observation remains unchanged: hallucinated tokens consistently exhibit a significant distribution shift toward language representations in synergy heads. The POPE performance is also stable, with the best performance achieved around the standard $3\sigma_{total}$ information redundancy criterion and $\lambda = 2.9652$. Overall, these experiments show that while exact head assignments naturally vary with different decision boundaries, the existence of synergy heads, the distribution disequilibrium, and the effectiveness of HEAL remain remarkably stable.

(3) Robustness to measurement metrics. We further evaluate alternative head scoring metrics, including normalized L_2 norm and a hybrid cosine+L_2 metric. In Table 8, all metrics produce highly similar head assignments, suggesting that the proposed taxonomy is largely independent of the specific metric. We adopt cosine similarity because it is most consistent with the geometric interpretation of HEAL. As illustrated in Figure 3 and theoretically supported by Theorem 2, HEAL performs a directional calibration of the output representation by adjusting the equilibrium between visual and language components. Cosine similarity directly measures this angular change, providing a more interpretable geometric characterization than magnitude-based metrics such as L_2 norm.

Table 8: Head assignment agreement under different measurement metrics.

Metric	normalized cosine similarity	L_2 norm + Sigmoid	cosine similarity + L_2 norm + Sigmoid
Head Assignment Agreement (%)	100.00 (base)	95.73	97.98

Table 9: Robustness to different replacement distributions in Section 3.1. We report the Head Assignment Agreement metric and LLaVA-1.5-7B performance on POPE.

	no intervention	Gaussian (base)	one-point (zero)	uniform	Cauchy (extreme outliers)	swapping actual outputs
Head Assignment Agreement (%)	-	100.00	95.25	96.31	97.46	96.97
F1 score ↑	86.89	87.84	87.73	87.69	87.81	87.75

Note. Every attention head is only assigned a binary label (redundant or non-redundant) to compute Head Assignment Agreement.

(4) Robustness to replacement distributions. In Section 3.1, our method does not assume that attention head outputs follow a Gaussian distribution. The Gaussian distribution is only used as a replacement noise during the causal intervention, rather than as a probabilistic model of the underlying activations. The objective of this intervention is simply to remove the instance-specific information carried by a head so that its causal contribution can be estimated.

To verify whether the choice of replacement affects causally redundant head identification, we further replace Gaussian noise with several different alternatives, including a one-point distribution (zero replacement), a uniform distribution, a heavy-tailed Cauchy distribution containing extreme outliers, and swapping the head outputs from another sample. As reported in Table 9, all replacements have highly consistent classifications. These results indicate that redundant head identification is largely distribution-independent, and our causal intervention is robust even under heavy-tailed perturbations or real activation replacement. More importantly, the hallucination mitigation performance also remains stable. Replacing the Gaussian distribution with these alternatives produces very similar POPE F1 scores, while all intervention strategies consistently outperform the variant without causal intervention. This further demonstrates that the success of this stage stems from the intervention itself rather than the specific choice of the replacement distribution.

Regarding swapping the actual head outputs, it may provide a meaningful alternative for estimating head importance. However, such a strategy requires an additional reference sample during inference, making it impractical for the inference-time setting. In contrast, Gaussian replacement is computationally efficient, low-order moment-preserving, and can be performed directly on the current input without introducing external samples.

Table 10: Redundant head assignment agreement before and after Yeo-Johnson transformation.

Metric	before Yeo-Johnson transformation (base)	after Yeo-Johnson transformation
Head Assignment Agreement (%)	100.00	98.25

(5) Why $3\sigma(\mathbf{I}^{(\mathbf{t},\mathbf{l},:)})$ criterion. The Gaussian assumption in Section 3.1 is not imposed on the intervention noise itself, but on the distribution of the estimated head contribution scores, which is required for applying the 3σ criterion to identify causally redundant heads.

To verify this assumption, we conduct an additional statistical analysis. For each token, we computed the contribution score of every attention head and performed a Lilliefors normality test (Lilliefors, 1967), which tests the null hypothesis that a sample is drawn from a normal distribution when the mean and variance are unknown. In almost all cases, the obtained p-values are greater than 0.05, indicating that we cannot reject the normality hypothesis for the contribution scores.

We also applied a Yeo-Johnson transformation (Yeo & Johnson, 2000) to further Gaussianize the score distribution and repeated the redundant-head identification. As shown in Table 10, the head assignment agreement before and after the transformation reaches 98.25%, indicating that the identified redundant heads are almost unchanged. In Table 11, the F1 score is also stable under 1σ, 2σ, and 3σ settings, with the best achieved around the standard 3σ causal redundancy criterion. This demonstrates that the causally redundant head identification is robust and that the 3σ criterion is well justified in practice.

Table 11: Robustness analysis on the $\sigma(I^{(t,l,:)})$ coefficient. We report LLaVA-1.5-7B performance on POPE.

$\sigma(I^{(t,l,:)})$ coefficient	**1**	**2**	**3**	∞
F1 score ↑	86.97	87.21	**87.84**	87.89

From a statistical perspective, each head contribution is an aggregate effect of numerous independent factors (e.g., token context, modality interaction). Such aggregated quantities are often empirically well approximated by a Gaussian distribution, which is also consistent with the normality tests.

D.3 Head Distribution Dynamics in an Update Interval



Figure 7: Head distribution in an update interval of 10 steps. We observe that the macroscopic distribution of head attributes shows negligible variation, justifying our periodic update strategy.

D.4 Empirical Analysis on the Equilibrium Factor and Update Interval

At present, we cannot provide a closed-form theoretical rule for determining the optimal values of these hyperparameters. However, our empirical analysis reveals that they exhibit stable operating ranges across different MLLM families and tasks, making parameter selection easier in practice.

Table 12: LLaVA-Bench score and C_S (CHAIR) under different equilibrium factors.

Equilibrium Factor	**LLaVA-Bench (LLaVA-1.5-7B)** ↑	**LLaVA-Bench (Qwen2.5-VL-7B)** ↑	C_S **(LLaVA-1.5-7B)** ↓	C_S **(Qwen2.5-VL-7B)** ↓
0.3	72.1	75.6	39.0	23.8
0.4	73.6	78.3	37.1	**23.3**
0.5	74.6	**78.5**	**36.9**	23.5
0.6	**75.2**	78.2	37.4	24.1
0.7	74.8	77.9	38.1	24.9

Equilibrium factor. As shown in Table 12 together with Figure 4(b), the optimal equilibrium factor consistently falls within **0.4-0.6** across both the LLaVA and Qwen model families. Although the exact optimum varies slightly across datasets, the performance remains stable within this interval.

Moreover, we observe an interesting empirical trend: models with stronger visual understanding (e.g., the Qwen series) generally achieve their best performance with a relatively smaller equilibrium factor, suggesting that such models require less additional visual calibration because their visual representations are already more reliable. In contrast, models with relatively weaker visual grounding (e.g., LLaVA) benefit from slightly larger values of α. Therefore, rather than requiring exhaustive tuning, we recommend selecting α within 0.4-0.6, followed by only minor adjustments according to the visual capability of the target model.

Table 13: LLaVA-Bench score and POPE-F1 under different update intervals.

Update Interval	LLaVA-Bench (LLaVA-1.5-7B) ↑	LLaVA-Bench (Qwen2.5-VL-7B) ↑	POPE-F1 (LLaVA-1.5-7B) ↑	POPE-F1 (Qwen2.5-VL-7B) ↑
2	75.5	78.7	87.9	88.8
5	75.3	78.7	87.8	88.6
10	**75.2**	**78.5**	**87.8**	**88.4**
15	75.0	78.2	87.5	88.3
20	74.7	77.8	87.1	88.0

Update interval. Similarly, Figure 4(a) together with Table 13 shows that the performance is highly similar when the update interval is chosen between **5 and 15** decoding steps, while we use 10 as the default setting throughout the paper. Importantly, unlike the equilibrium factor, we find that the update interval transfers well across different tasks and datasets, indicating that it is considerably less sensitive to downstream applications.

Finally, we would like to note that these parameters are analogous to commonly used inference-time hyperparameters such as temperature or top-p, which are intentionally exposed to users to flexibly control the decoding behavior. In our case, the equilibrium factor provides an interpretable control over the balance between visual evidence and language information. As theoretically supported by Theorem 2, increasing α continuously shifts the output representation toward the visual direction, allowing users to explicitly control how strongly the model relies on visual evidence.

Table 14: Intern-VL performance comparison with our and other methods.

Method	LLaVA-Bench↑	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	POPE-R↑	POPE-F1↑	POPE-A↑
InternVL-7B	51.6	46.6	12.4	80.0	85.3	86.2
+LocoRE (Zhang et al., 2026)	52.8	40.2	10.5	85.8	87.2	87.3
+HEAL	**53.4**	**39.2**	**9.6**	**86.5**	**87.8**	**87.9**

To further verify the transferability of these practical guidelines, we follow the above strategy for the other models in our experiments. For example, Intern-VL adopts the same parameter setting as the LLaVA family due to their similar visual capability. Although we do not separately tune these hyperparameters for Intern-VL, HEAL still consistently improves all evaluation metrics (Table 14), suggesting that the proposed parameter ranges generalize well to other MLLMs.

D.5 Performance on Recent Benchmarks and MLLMs

Table 15: HEAL performance on recent benchmarks and MLLMs.

Method	MMHal-Bench		BLINK-Twice				
	Halluc. Rate↓	Score↑	No-Acc	Yes-Acc	Q-Acc	I-Acc	G-Acc
Qwen3-VL-8B	17.5	4.82	0.476	0.653	0.493	0.336	0.140
+HEAL	16.6	4.91	0.557	0.714	0.536	0.353	0.213
InternVL3.5-8B	19.4	4.53	0.355	0.588	0.475	0.264	0.109
+HEAL	18.3	4.71	0.463	0.624	0.583	0.334	0.158

We conducted additional experiments on Qwen3-VL-8B (Bai et al., 2025a) and InternVL3.5-8B (Wang et al., 2025b), two recent open-source MLLMs, and evaluated them on MMHal-Bench (Sun et al., 2024b) (ACL 2024) and BLINK-Twice (JIANG et al., 2026) (NeurIPS 2026), which provide more challenging evaluations of multimodal hallucination and visual reasoning than earlier benchmarks.

As shown in Table 15, HEAL consistently improves performance across both models and both benchmarks. On Qwen3-VL-8B, HEAL reduces the hallucination rate from 17.5 to 16.6 while improving the MMHal score from 4.82 to 4.91. On InternVL3.5-8B, the hallucination rate decreases from 19.4 to 18.3, and the MMHal score improves from 4.53 to 4.71. On BLINK-Twice, HEAL consistently improves all evaluation metrics, including No-Acc, Yes-Acc, Q-Acc, I-Acc, and G-Acc, showing that the proposed dynamic information calibration generalizes well to recent MLLMs and challenging visual reasoning scenarios.

E FURTHER DISCUSSIONS ON HEAL

(1) Do not resolve hallucinations that stem from early visual encoding failures or a fundamental lack of visual evidence in the input.

HEAL is an inference-time calibration framework that operates on the internal attention representations during decoding. Its objective is to calibrate the contributions of visual and language information when both sources of information are already available in the model. Consequently, if the visual encoder fails to extract reliable visual features, or if the input itself lacks sufficient visual evidence, there is no additional visual information that can be recovered through attention calibration alone.

(2) HEAL applies a single, global equilibrium factor to all synergy heads. Do all synergy heads truly behave uniformly?

In fact, our empirical observations show that almost every synergy head exhibits its own attention pattern and modality preference. The key motivation is that HEAL is designed to correct a global decoding imbalance rather than optimize each head independently. For each token, the final prediction is jointly determined by the aggregation of all attention heads. Consequently, the objective of HEAL is to regulate the *collective* distribution of visual and language information at the token level, making a global equilibrium factor a natural design choice.

Although a few synergy heads may exhibit highly specialized behaviors, applying an identical equilibrium factor does not require every head to respond identically. Instead, it provides a consistent global bias that shifts the overall attention representation toward a better visual-language balance. While the contribution of a small number of individual heads may become slightly suboptimal, this effect is compensated by the remaining heads during multi-head aggregation. Since the decoder prediction depends on the *collective* output of all heads rather than any individual head, the overall generation quality is preserved, and hallucinations are effectively reduced.

Our empirical results support this design. As shown in the hyperparameter analysis (Figure 4(b), Tables 12 and 14), a single equilibrium factor consistently improves hallucination-related metrics across different model families, including LLaVA, Qwen, and Intern-VL. These results suggest that although synergy heads are heterogeneous at the individual level, their *collective* imbalance can be effectively corrected through a shared global calibration, which is sufficient to achieve consistent performance improvements across diverse MLLMs.

(3) The method updates head types every 10 steps based on temporal locality. However, complex queries often shift abruptly between visual description and linguistic reasoning. Does this fixed interval delay adaptation? Could it cause localized hallucinations during these sharp transitions?

The fixed update interval inevitably introduces a certain delay in tracking the dynamic changes of attention head attributes. This is an intentional design choice that balances adaptation accuracy and inference efficiency. Updating the head taxonomy at every decoding step would provide the most up-to-date estimation, but it would also incur a substantially higher computational cost, making the method much less practical for inference-time deployment. Importantly, our empirical observations suggest that this delay has only a limited impact on the final decoding behavior. As discussed above, although the types of synergy heads evolve dynamically during token generation, the visual heads remain highly sparse and remarkably stable, while most transitions occur between language-oriented and synergy heads. Similar observations have also been reported in recent studies on attention head dynamics in MLLMs (Wang et al., 2025a; He et al., 2025). Consequently, a moderate update interval is sufficient to capture the dominant head organization throughout decoding.

Furthermore, our hyperparameter analysis (Figure 4(a) and Table 13) shows that update intervals between 5 and 15 decoding steps consistently achieve very similar performance, indicating that HEAL is not particularly sensitive to moderate delays in taxonomy updates. We therefore use 10 decoding steps as the default setting, which provides a favorable balance between computational overhead and hallucination mitigation. Regarding the concern about rapidly switching reasoning patterns (e.g., alternating between visual grounding and language reasoning), we acknowledge that existing benchmarks do not explicitly isolate this phenomenon, making it difficult to quantify its impact directly. Nevertheless, we further evaluate HEAL on more challenging benchmarks, like MMHal-Bench (Sun et al., 2024b), involving complex multimodal reasoning and localization, where

frequent transitions are more likely to occur. The consistent improvements in Table 15 suggest that the delayed updates introduce only limited degradation in practice.

(4) About the two redundancy definitions in the paper.

In Section 3.1, redundancy is defined from the perspective of causal contribution. Specifically, after replacing one head output with its counterfactual counterpart, we measure the change in the final multi-head attention representation. If a head has negligible influence on the final representation, it is considered a *causally redundant head*. This filtering step is necessary because some heads may have very limited contribution after the output projection layer, and retaining these heads can introduce noise into the subsequent head taxonomy.

In contrast, Section 3.2 focuses on information composition within each remaining head. Here, an *information redundant head* refers to a head whose information content does not change significantly after introducing visual or language tokens. In other words, these heads contain negligible modality-specific information and therefore cannot provide meaningful visual-language decomposition.

Both filtering stages are necessary. As shown in Table 9, removing the causal redundancy filtering affects the subsequent head taxonomy and final performance. Meanwhile, without the information redundancy filtering, the modality decomposition becomes unreliable because heads with almost zero information variation cannot provide meaningful visual-language information ratios.

(5) The difference between the MoE and dense models in the design of the equilibrium factor. For the MoE model with different activated experts in use, how to compute the value of the equilibrium factor α?

HEAL operates entirely at the Multi-Head Attention (MHA) level rather than the FFN/MLP level. Specifically, our causal intervention, Difference-in-Differences analysis, head taxonomy, and dynamic information calibration are all performed on the attention outputs and value vectors (see Figure 3 and Theorem 1). In contrast, sparse MoE architectures replace the feed-forward (MLP/FFN) sublayer with routed experts, while the attention module remains unchanged in mainstream MoE MLLMs. Therefore, the computational procedure of HEAL is identical for dense and MoE models, and is independent of which experts are activated during inference.

Regarding the equilibrium factor, α is a model-level hyperparameter rather than a quantity computed from the activated experts. Its role is to control the target balance between visual and language information during calibration. Once α is determined, the calibration factors are computed solely from the visual/language information proportions estimated for each attention head, which are independent of the downstream MoE routing. Consequently, the computation of both the equilibrium factor and the calibration factors is the same for dense and MoE models. In practice, α is selected using the same procedure as in dense models, since models within the same family (e.g., Qwen-VL dense and sparse variants) exhibit similar multimodal information characteristics.

(6) Does HEAL employ the counterfactual mechanism for all the attention layers?

HEAL performs the counterfactual analysis for all attention layers. The reason is that our central finding is global rather than layer-specific: hallucinations are causally associated with the disequilibrium between visual and language information within synergy heads distributed throughout the entire network, rather than being dominated by a few specific layers. Therefore, identifying synergy heads requires performing the proposed DiD analysis across all attention layers. Based on the identified head taxonomy, HEAL dynamically calibrates only the synergy heads, which account for a small portion of all attention heads, while the remaining heads are left unchanged.

Although the head taxonomy is obtained from all attention layers, the computational overhead during inference is greatly reduced by the optimizations described in Section 3.4. First, head attributes are updated periodically rather than at every decoding step, leveraging the temporal locality of head behaviors. Second, the counterfactual analysis is implemented in a batched and parallelized manner. During intermediate decoding steps without taxonomy updates, HEAL only computes the calibration factors for the identified synergy heads instead of repeating the full analysis.

(7) Geometric interpretation of HEAL.

It is important to distinguish the geometric representation space from the information decomposition space in HEAL. In the representation space, the output of a multimodal attention head can be locally

viewed as a combination of visual and language components together with a residual component,

$$h = r + v + l,$$

where v and l denote the visual and language representation components, and r collects residual components that are not explicitly attributed to either modality.

In contrast, the quantities I_{vis}, I_{lang}, and I_{syn} computed in Section 3.2 characterize the information structure induced by counterfactual interventions. In particular, $I_{\text{syn}} = I_{\text{total}} - I_{\text{vis}} - I_{\text{lang}}$ represents the interaction between visual and language information that cannot be accounted for by their individual contributions. Thus, I_{syn} should not be interpreted as an additional Euclidean vector added independently to v and l. Rather, it quantifies a cross-modal interaction in the information space. This distinction explains the geometric illustration in Figure 3. The figure depicts the representation space effect of calibration rather than a vector decomposition of all information atoms.