Causal Head Gating: Separating Visual Evidence and Knowledge Priors for Hallucination Mitigation in Multimodal LLMs

Anonymous ACL submission

Abstract

Multimodal large language models excel at vision-language tasks but suffer from hallucinations that undermine their reliability. Existing mitigation methods apply uniform heuristics at the output or attention level, overlooking the heterogeneous causal roles of attention heads and their dynamic behaviors during inference. We revisit hallucination from a causal perspective and identify it as competition between two pathways: a direct path from visual evidence to outputs, and an indirect path dominated by knowledge priors. When visual evidence weakens, MLLMs increasingly rely on parametric knowledge priors—such as spurious correlations learned during pretraining. This over-reliance causes attention heads responsible for visual grounding to drift toward language-driven patterns, resulting in generations consistent with the language model but unfaithful to the image. We propose Causal Head Gating, an inference-time framework that separates visual evidence from knowledge priors by modeling their causal contributions at the attention-head level. Under this framework, we introduce **CHERI** (Causal Head Evidence Reweighting and Intervention), which identifies attention heads whose visual contributions weaken under hallucination and dynamically restores their influence. Specifically, **CHERI** reinforces heads that propagate visual evidence while suppressing those dominated by knowledge priors, strengthening the direct causal effect of vision on generation and attenuating spurious indirect effects. Experiments across multiple benchmarks show that **CHERI** consistently reduces hallucination rates across diverse task formats while preserving overall generation quality and task performance.

1 Introduction

Multimodal Large Language Models (MLLMs) have become the de facto standard for vision-language tasks. Yet, a persistent issue undermines their reliability: hallucination—the generation of text that is consistent with the language context but unfaithful to the image. This disconnect poses a severe risk for real-world deployment.

The community's approach to this problem has evolved significantly. Initially, solutions leaned heavily on training, utilizing Supervised Fine-Tuning (SFT) or Reinforcement Learning (RLHF) to align models. While effective, these methods impose heavy computational burdens—often requiring hundreds of GPU hours—and risk catastrophic forgetting. Consequently, the field has pivoted toward inference-time intervention, which corrects model behavior on the fly without parameter updates. However, current inference strategies, such as Contrastive Decoding, suffer from two major flaws. First, they operate at a coarse granularity (e.g., shifting entire logits or attention maps), which often inadvertently suppresses valid visual details. Second, they are computationally expensive at runtime; many require multiple forward passes per token, making them impractical for latency-sensitive applications.

We argue that resolving this dilemma requires a microscopic view of the model's internal mechanism: the attention heads. We reframe hallucination through a causal lens, modeling it as a conflict between two distinct flows: a *Direct Visual Pathway* grounded in image features, and an *Indirect Knowledge Pathway* driven by parametric priors and spurious correlations. Our analysis reveals a distinct layer-wise functional division: early layers act as a perception stage focused on visual extraction, whereas deeper layers form a reasoning stage where strong knowledge priors often override weak visual signals.

Guided by this insight, we propose Causal Head Gating, a fine-grained, inference-time framework. Unlike static methods that apply blanket penalties, our approach uses lightweight probes to intervene

only when necessary. We design a Visual Perception Probe (VPP) to boost heads with weak visual grounding during perception, and a Knowledge Prior Probe (KPP) to dampen heads dominated by spurious biases during reasoning.

Our method, **CHERI** (Causal Head Evidence Reweighting and Intervention), distinguishes itself through surgical precision. By regulating information flow at the head level, **CHERI** effectively mitigates hallucinations without compromising general generation quality. Our contributions are:

- **Fine-grained Head-Level Control:** We move beyond coarse-grained interventions, targeting specific attention heads to fix hallucinations with minimal impact on correct tokens.
- **Dynamic Gating Mechanism:** Instead of a "one-size-fits-all" approach, our lightweight probes dynamically determine when and where to intervene, ensuring efficiency.
- **Causal-Based Architecture:** By modeling hallucination as a competition between Knowledge Prior and Visual Pathways, we design stage-specific interventions. This ensures that amplification and suppression are applied in accordance with the model's layer-wise specialization.
- **Inference Efficiency:** **CHERI** avoids the high latency of contrastive methods. It requires no additional forward passes and no expensive fine-tuning, offering a seamless plug-and-play solution.

2 Related Work

Hallucination in Multimodal Large Language Models. Despite the remarkable capabilities of Multimodal Large Language Models (MLLMs) in vision-language tasks (Liu et al., 2024c; Bai et al., 2023), they are prone to hallucinations—generating text that conflicts with visual content (Liu et al., 2024b; Bai et al., 2024). Unlike unimodal hallucinations, MLLM errors often arise from a misalignment between visual perception and linguistic generation (Li et al., 2023b). Research indicates that these failures are frequently driven by object co-occurrence biases (Zhou et al., 2024; Zhong et al., 2024) and the "snowballing" of language priors, where models rely on parametric knowledge rather than visual grounding (Zhong et al., 2024; Jiang et al., 2023). Mechanistic interpretability studies have further localized these issues, identifying specific network components, such as middle layers (Jiang et al., 2024) or attention heads (Gong et al., 2024; Arif et al., 2025), as critical stages where hallucinations emerge. Moreover, recent works suggest that internal model states (Chen et al., 2024a; Zhang et al., 2025a) and cross-modal information flows (Zhang et al., 2025b) retain detectable signals of these errors. Our work leverages these causal insights to distinguish between valid visual evidence and spurious knowledge priors at the attention head level.

Training-based Mitigation Strategies. A standard mitigation paradigm involves refining models through high-quality data interventions. Approaches range from robust instruction tuning (Liu et al., 2024a; Zhang et al., 2024) to constructing datasets that specifically penalize toxic visual instructions (Yu et al., 2023a; Yin et al., 2024b). Reinforcement learning from human feedback (RLHF) and Direct Preference Optimization (DPO) have also been adapted to align MLLMs with factual visual grounding (Sun et al., 2024; Yu et al., 2023b; Yang et al., 2025b; Fu et al., 2025). Additionally, some methods employ logic-closed loops (Wu et al., 2024) or cross-level interventions (Chen et al., 2025) to uncover and correct inconsistencies. Post-hoc correction frameworks like Woodpecker (Yin et al., 2024a) further refine outputs using external tools. While effective, training-based methods incur significant computational costs and risk catastrophic forgetting of general capabilities. In contrast, our approach focuses on inference-time intervention, offering a lightweight, plug-and-play solution without parameter updates.

Inference-time Intervention and Decoding. Inference-time strategies dynamically rectify model behavior during generation, avoiding the need for retraining. A prominent direction involves Contrastive Decoding (CD) (Li et al., 2023a), which has been adapted for MLLMs to penalize statistical biases. For instance, VCD (Leng et al., 2024) contrasts logits from original and visually distorted inputs, while IBD (Zhu et al., 2024) and M3ID (Favero et al., 2024) leverage image-biased or instruction-level contrasts. Other variants include ICD, and CATCH (Kan et al., 2024), which further refine contrastive objectives. Beyond contrastive

methods, attention- and logit-based interventions have gained traction. OPERA (Huang et al., 2024) introduces an over-trust penalty and a retrospection-allocation mechanism to disrupt hallucination patterns in self-attention. Recent advancements include DeCo (Wang et al., 2025) , which employs dynamic correction decoding, and ADHH (Yang et al., 2025a), which utilizes modular attribution to guide intervention. Other notable approaches include reweighting attention to image-centric tokens (Liu et al., 2024d; Xing et al., 2024), applying classifier-free guidance (Zhao et al., 2024; Chen et al., 2024b), or utilizing coarse-to-fine feedback. Our method, CHERI, advances this frontier by introducing a fine-grained, causal gating mechanism that specifically targets the attention heads responsible for the conflict between visual evidence and knowledge priors.

3 Methods

3.1 Empirical Analysis: When Visual Evidence Meets Knowledge Priors

To understand how hallucinations emerge in MLLMs, we examine attention patterns across different transformer layers during generation. Our preliminary experiments on MLLMs uncover two key observations.

Functional specialization across layers. The model's internal layers exhibit clear functional division. Layers 1-16 serve as the perception stage, where the model extracts visual signals from image tokens. At this stage, attention concentrates on visual grounding, though semantic composition remains limited. Layers 17-32 constitute the reasoning stage, where the model integrates learned knowledge to produce fluent language. This stage-wise specialization suggests that different layers require tailored intervention strategies.

Knowledge priors override visual evidence. The example in Figure 1 shows that even when the model attends to the suitcase in the background and generates partial tokens like "suit" in layers 25-28, the final output becomes "handbag" in later layers. This happens because the co-occurrence of "restaurant" and "girl" in training data creates a strong bias toward "handbag," overwhelming the visual evidence for "suitcase." A similar pattern appears with "cup"—while the model generates this correct token, attention maps show it focuses on wrong spatial regions. The description contains errors in both quantity and location, indicating the token came from parametric knowledge rather than visual grounding. In other words, this is getting the right word for the wrong reasons.

Insufficient visual attention during hallucination. Figure 2 compares attention heatmaps for hallucinated tokens (e.g., "handbag") and non-hallucinated tokens (e.g., "sandwich") across all layers and heads. An obvious phenomenon emerges: when hallucination occurs, attention to image tokens remains notably weak throughout the perception stage (layers 1-16). This insufficient visual attention lets knowledge priors dominate generation, as the model lacks strong visual evidence to counter linguistic biases. The heatmap shows sparse, diffuse attention to visual tokens when generating "handbag," while "sandwich" triggers concentrated attention in corresponding visual regions.

These findings point to an imbalance between two causal paths: a direct path from visual evidence to outputs, and an indirect path driven by knowledge priors. When visual evidence weakens—through image ambiguity or inadequate attention—the model falls back on prior-driven generation. This motivates our design: amplify visual evidence during perception and suppress spurious priors during reasoning. Detailed analysis appears in the Appendix.

3.2 Causal Head Gating

Building on the empirical observations in Section 3.1, we introduce Causal Head Gating, a framework that separates visual evidence from knowledge priors at the attention head level. Each attention head is equipped with a lightweight probe to quantify its reliance on either visual evidence or knowledge priors. We design two types of probes tailored to different functional stages: Visual Perception Probes (VPP) for the 512 heads in the perception stage (layers 1-16), and Knowledge Prior Probes (KPP) for the 512 heads in the reasoning stage (layers 17-32). Here we use LLaVA-1.5-7B as an illustrative example.

Visual Perception Probe (VPP). VPP estimates the probability that a head suffers from insufficient visual evidence. We construct training pairs by pairing each clean image with its noisy counterpart, where noise is added following the DDPM formulation with 30% Gaussian noise injection. For each head (l, n) at layer l and head index n, we train a binary classifier parameterized

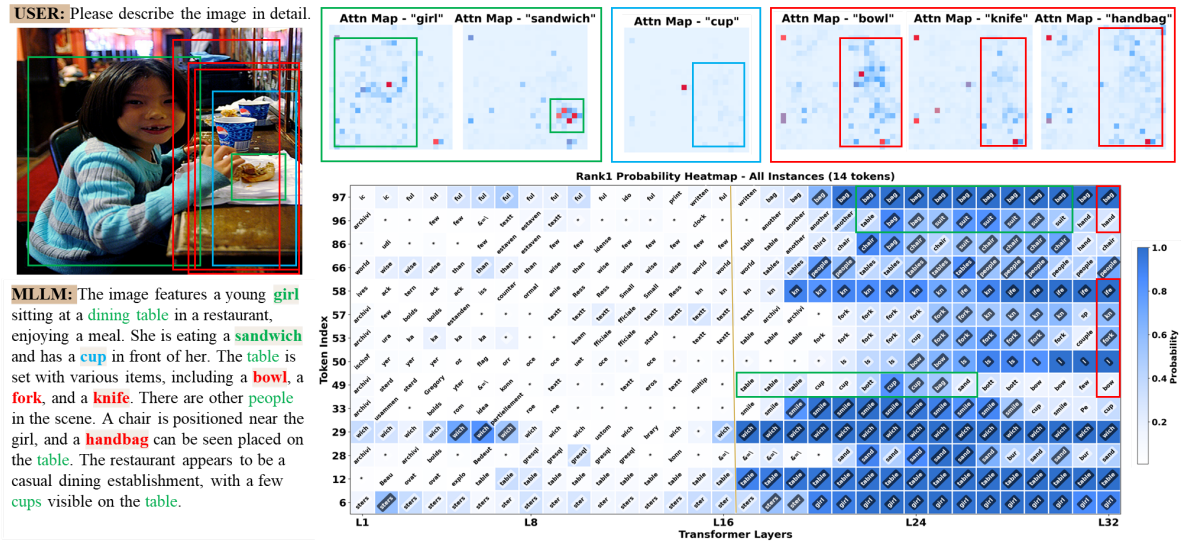


Figure 1: Focuses on the "Knowledge Priors override Visual Evidence" aspect. (The Shift from "suit" to "handbag")

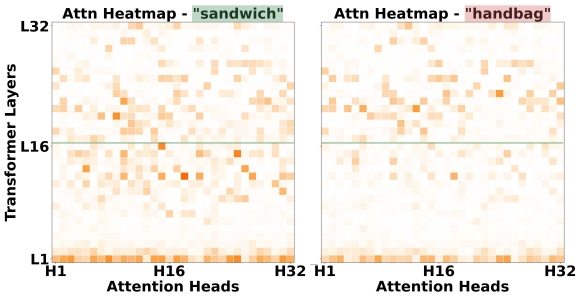


Figure 2: Focuses on the "Insufficient Visual Evidence" aspect. (The Attention Map comparison between "sandwich" and "handbag")

by $\theta_{(l,n)}$ using standard cross-entropy loss. The probe outputs a probability:

$$p^v = \text{sigmoid}(\langle \theta_{(l,n)}, H_{(l,n)} \rangle) \tag{1}$$

where $H_{(l,n)}$ denotes the head's output representation. A higher p^v indicates that the head operates with weak visual evidence, similar to processing a degraded image.

Knowledge Prior Probe (KPP). KPP quantifies how strongly a head is driven by knowledge priors rather than visual input. We adopt a log probability increase metric to measure each head's causal influence on token prediction. Specifically, we compute:

$$\Delta \log P_{(l,n)}(B) = \log P(B \mid h_{(l-1)} + H_{(l,n)}) - \log P(B \mid h_{(l-1)}) \tag{2}$$

This measures how much head (l, n) shifts the probability toward a specific token B. To isolate heads influenced by knowledge priors, we need to distinguish between heads that promote correct tokens and those that push toward hallucinated ones. We construct two token sets B^+ and B^- as follows:

For yes/no questions, B^+ contains the ground-truth token. For caption tasks, we generate multiple captions and partition tokens into grounded set G and hallucinated set H via automated filtering. At each generation step t, we extract the top-k logits to form candidate set C_t. If the current token is grounded, B^+ is the current token and $B^- = C_t \cap H$. If the current token is hallucinated, B^- is the current token and $B^+ = C_t \cap G$. Neutral tokens are excluded from training.

The final head-level score is:

$$s_{(l,n)} = \max\big(0, \Delta \log P_{(l,n)}(B^-) - \Delta \log P_{(l,n)}(B^+)\big) \tag{3}$$

We then train a probe to predict this score via regression:

$$p^k = \text{sigmoid}(s_{(l,n)}) \tag{4}$$

A higher p^k indicates the head strongly promotes hallucinated tokens driven by knowledge priors.

Training details. Both probes are trained on 10,000 samples from MS-COCO that do not overlap with our evaluation datasets. The probes remain frozen during inference and serve as diagnostic tools to identify when and where intervention is needed.

3.3 CHERI : Causal Head Evidence Reweighting and Intervention

With probes p^v and p^k quantifying visual evidence insufficiency and knowledge prior dominance, we now describe how **CHERI** performs targeted intervention at each stage.

Perception stage: Visual evidence reweighting. When VPP detects weak visual evidence ($p^v > 0.5$), we amplify the head's attention to visual tokens. Let v_s and v_e denote the start and end indices of visual tokens. We first compute a normalized attention distribution over visual tokens:

$$\mu_{(l,n)} = \left(\frac{A_{(l,n)}[:, v_s : v_e]}{\sum_{j=v_s}^{v_e} A_{(l,n)}[:, j] + \epsilon} \right) \cdot V_{(n)}[v_s : v_e] \tag{5}$$

where $A_{(l,n)}$ is the attention matrix of head (l, n), $V_{(n)}$ is the value matrix, and ϵ prevents division by zero. This renormalized representation $\mu_{(l,n)}$ focuses solely on visual content. We then interpolate between the original head output $H_{(l,n)}$ and $\mu_{(l,n)}$ using an adaptive scaling factor:

$$\alpha = \alpha_0 \cdot p^v \tag{6}$$

$$H^{\text{new}}_{(l,n)} = (1 - \alpha) \cdot H_{(l,n)} + \alpha \cdot \mu_{(l,n)} \tag{7}$$

The scaling factor α grows with p^v, ensuring stronger intervention when visual evidence is weaker. This reweighting pushes the head toward visual grounding without discarding its original function entirely.

Reasoning stage: Knowledge prior suppression. In the reasoning stage, we take a different approach. As shown in Figure 4 (left), most heads exhibit low p^k values, indicating that only a small subset strongly relies on knowledge priors. Aggressive intervention across all heads risks degrading linguistic coherence and performance on tasks requiring factual knowledge (e.g., MME benchmarks). We therefore apply selective suppression only when $p^k > 0.5$:

$$\beta = \beta_0 \cdot p^k, \quad H^{\text{new}}_{(l,n)} = (1 - \beta) \cdot H_{(l,n)} \tag{8}$$

This dampens the influence of heads dominated by knowledge priors while leaving other heads intact. Figure 4 (right) visualizes the resulting intervention pattern: only a sparse set of heads receives substantial suppression, preserving the model's overall reasoning capacity.

Adaptive gating mechanism. The key advantage of **CHERI** lies in its adaptive gating, which addresses both *when* and *how* to intervene. Unlike static methods that uniformly modify all heads regardless of context, **CHERI** uses runtime diagnostics (p^v and p^k) to determine whether intervention is necessary for each head at each generation step. This case-by-case adaptation is crucial because attention heads exhibit context-dependent behavior—a head's function shifts with different prompts and generation stages. For instance, a head that encodes spatial relationships in one context may contribute to linguistic fluency in another. Treating all heads uniformly ignores this functional heterogeneity.

Moreover, attention heads operate in a coupled manner, where one head's behavior depends on others. Aggressive intervention across all heads disrupts these intricate dependencies, leading to degraded internal representations. This over-intervention problem manifests as incoherent language or reduced performance on knowledge-intensive tasks (e.g., MME benchmarks). By coupling intervention strength with probe confidence, **CHERI** selectively modifies only the heads that exhibit problematic patterns—those with weak visual evidence or strong prior dominance—while leaving others intact. This selective approach preserves the model's overall reasoning capacity and linguistic coherence.

4 Experiments

4.1 Experimental Setup

Evaluation Benchmarks. To comprehensively validate the effectiveness of CHERI, we conduct experiments on three hallucination evaluation benchmarks: POPE, CHAIR, and MME. These benchmarks assess hallucination mitigation from different perspectives. Full benchmark details are in Appendix.

4.2 Hallucination Mitigation Performance

We compare CHERI against four mainstream inference-time intervention methods across all benchmarks. Results demonstrate that CHERI consistently achieves the best performance while generalizing robustly across two multimodal architectures.

POPE Results. Table 1 presents POPE evaluation results for CHERI on LLaVA-1.5 and Qwen-VL-Chat, respectively. On LLaVA-1.5, CHERI achieves the best performance under all three sampling strategies. Specifically, under the Random

USER: "Please help me describe the image in detail."

LLaVA-v1.5-7b: "The image features a **woman** standing on a **train**, holding a blue **suitcase** in her hand. She is wearing a black dress and appears to be posing for the camera. The **train** is red and green, and the **woman** is positioned near the door of the **train car**. In addition to the woman and the **suitcase**, there is a **handbag** placed on the ground near her. The scene captures the essence of travel and the excitement of embarking on a journey."

Ours: "The image features a **woman** standing on a **train**, holding a blue **suitcase** in her hand. She is wearing a white t-shirt and a black skirt, and she has **glasses** on her face. The **woman** has tattoos visible on her arms and appears to be posing for the camera. The **train** behind her is painted in red and green colors. The **woman** is standing near the door of the **train**. Her unique appearance and the train setting create an interesting and captivating scene."

Figure 3: A qualitative example where **CHERI** eliminates hallucinated tokens while enriching visual details. After applying **CHERI**, the generated caption no longer contains fabricated objects. More importantly, the model now captures fine-grained visual elements that were previously missed—it correctly identifies the person's tattoos and glasses, both of which are small and easily overlooked. The intervention also prevents spurious associations: the baseline model hallucinates "car" when it sees a train, and produces "handbag" when a suitcase appears in the image. **CHERI** stops these errors by strengthening visual grounding in the perception stage and suppressing prior-driven predictions in the reasoning stage. Additional qualitative results are provided in the Appendix.

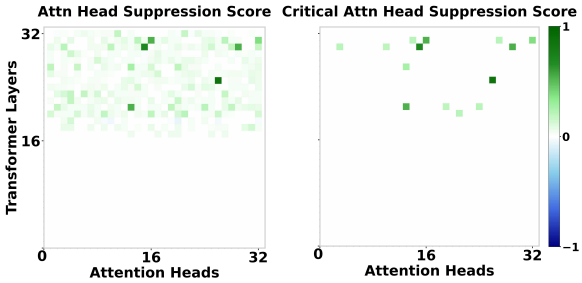


Figure 4: Intervention pattern visualization. Left: Distribution of p^k values across heads in the reasoning stage, showing most heads have low knowledge prior dominance. Right: Resulting intervention strength, where only sparse heads with $p^k > 0.5$ receive substantial suppression.

setting, CHERI attains 90.74% accuracy and 88.72 F1 score, improving over the baseline by 7.23 and 6.43 percentage points, respectively. Under the more challenging Popular setting, CHERI achieves 87.27% accuracy and 86.48 F1 score, surpassing the second-best method ADHH by 2.52 and 2.06 percentage points. Under the most difficult Adversarial setting, CHERI maintains its lead with 85.66% accuracy and 84.07 F1 score, representing gains of 9.62 and 7.80 percentage points over the baseline. On average, CHERI achieves 87.89% accuracy and 86.42 F1 score, demonstrating stable hallucination detection capability.

On Qwen-VL-Chat, CHERI also obtains the best results. Under the Random setting, it achieves 90.15% accuracy and 89.03 F1 score; under the Popular setting, 88.71% accuracy and 87.58 F1 score; and under the Adversarial setting, 86.17% accuracy and 84.73 F1 score. It is worth noting that CHERI achieves consistent performance improvements on both models with different architectures, which validates the generalization capability of our method. Compared with logits-based intervention methods (VCD) and attention-based intervention methods (OPERA, DeCO, ADHH), CHERI maintains more stable performance advantages across different difficulty settings through its head-level dynamic gating mechanism.

CHAIR Results. Table 2 presents object hallucination evaluation results on the image captioning task. On LLaVA-1.5, CHERI reduces the sentence-level hallucination rate CHAIR_S to 26.6, a 46.9% reduction from the baseline's 50.1. The instance-level hallucination rate CHAIR_I is reduced to 7.2, a 53.2% reduction from the baseline's 15.4. Compared with the second-best method ADHH, CHERI achieves improvements of 6.6 and 2.2 percentage points on CHAIR_S and CHAIR_I, respectively. On Qwen-VL-Chat, CHERI also performs well, achieving 27.1 on CHAIR_S and 7.3 on CHAIR_I, both the lowest among all methods. These results indicate that CHERI not only accurately identifies object hallucinations but also effectively suppresses the generation of false objects in open-ended generation

Table 1: Experimental Results Comparison, Qwen-VL-Chat and LLaVA-v1.5-7B in POPE Benchmark.

Methods	Random		Popular		Adversarial		Average	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
LLaVA-1.5	83.51	82.29	79.99	79.35	76.04	76.27	79.85	79.30
+VCD	86.85	86.25	82.66	83.39	77.33	79.31	82.28	82.98
+OPERA	87.55	86.47	84.23	83.50	80.91	80.70	84.23	83.56
+DeCo	88.67	87.32	84.73	84.92	81.40	81.77	84.93	84.67
+ADHH	88.40	87.61	84.75	84.42	81.54	81.91	84.90	84.65
+CHERI	**90.74**	**88.72**	**87.27**	**86.48**	**85.66**	**84.07**	**87.89**	**86.42**
QWen-VL-Chat	85.47	83.22	83.43	81.40	80.53	79.96	83.14	81.53
+VCD	86.51	84.62	85.30	83.37	83.01	81.24	84.94	83.08
+OPERA	87.43	86.52	85.77	84.38	83.24	81.70	85.48	84.20
+DeCo	88.99	87.54	85.23	85.10	83.99	82.07	86.07	84.90
+ADHH	88.73	87.68	86.64	85.83	84.12	82.86	86.50	85.46
+CHERI	**90.15**	**89.03**	**88.71**	**87.58**	**86.17**	**84.73**	**88.34**	**87.11**

Table 2: CHAIR Metrics Comparison across Different Methods.Lower is better (↓).

Methods	LLaVa						QWen					
	Vanilla	+VCD	+OPERA	+DeCo	+ADHH	**+CHERI**	Vanilla	+VCD	+OPERA	+DeCo	+ADHH	**+CHERI**
$CHAIR_S$↓	50.1	47.8	42.4	37.8	33.2	**26.6**	46.0	44.8	42.8	38.2	33.3	**27.1**
$CHAIR_I$↓	15.4	14.6	11.6	11.1	9.4	**7.2**	13.5	13.3	12.6	10.7	9.0	**7.3**

tasks.

MME Results. Table 3 shows CHERI's performance on the comprehensive MME benchmark. CHERI scores 1519.32 on perception tasks, 370.03 on cognition tasks, and achieves a total score of 1889.35, all the highest among all methods. Compared with the baseline model, CHERI improves by 10.32 points on perception tasks and 14.33 points on cognition tasks. Notably, methods like OPERA often compromise cognitive reasoning capabilities when suppressing hallucinations, achieving only 310.73 on cognition tasks. In contrast, CHERI maintains cognitive reasoning ability while enhancing visual perception through its stage-wise selective intervention strategy. This result validates our design principle proposed in Section 3.3: reinforcing visual evidence during the perception stage and selectively suppressing knowledge priors during the reasoning stage achieves a better balance between hallucination mitigation and task performance.

4.3 Ablation Study

We validate the effectiveness of the two core components, VPP and KPP, through ablation experiments. As shown in Table 4, on LLaVA-1.5, using VPP alone improves the average F1 score from the baseline's 79.30 to 83.26, an improvement of 3.96 points. Using KPP alone yields a more substantial gain, achieving 84.72 F1 score with an improvement of 5.42 points. This indicates that knowledge prior suppression in the reasoning stage contributes slightly more to hallucination mitigation than visual evidence amplification in the perception stage. When both components are enabled, the F1 score further improves to 86.42, a gain of 1.70 points, validating the complementary nature of the two-stage intervention strategy. The same

Table 3: Performance Comparison across Different Metrics in MME. Higher is better (↑).

Methods	**Perception↑**	**Cognition↑**	**Total↑**
Vanilla	1509.0	355.7	1864.7
+VCD	1514.0	358.9	1872.9
+OPERA	1473.6	310.7	1784.3
+DeCo	1495.7	316.4	1812.1
+ADHH	1493.6	318.7	1812.4
+CHERI	**1519.3**	**370.0**	**1889.4**

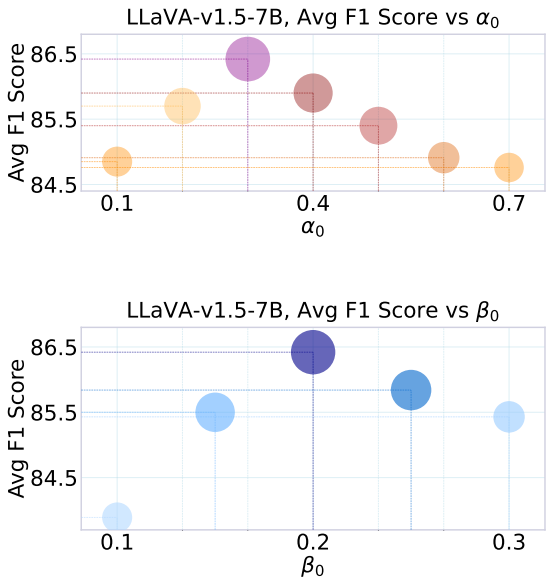


Figure 5: Parameter sensitivity analysis. The upper panel shows the impact of visual evidence amplification factor α_0 in the perception stage on average F1 score, while the lower panel shows the impact of knowledge prior suppression factor β_0 in the reasoning stage. Experiments are conducted on LLaVA-1.5-7B, with the average F1 score across three POPE settings as the evaluation metric. Results indicate that optimal performance is achieved at $\alpha_0 = 0.3$ and $\beta_0 = 0.2$.

trend is observed on Qwen-VL-Chat, where the complete model improves by 2.86 and 1.97 points over using VPP or KPP alone, respectively.

Figure 5 presents the sensitivity analysis for hyperparameters α_0 and β_0. For the perception stage amplification factor α_0, performance peaks at 0.3, with both smaller and larger values leading to performance degradation. When α_0 is too small, visual enhancement is insufficient; when too large, it may overly suppress language modeling capabilities. For the reasoning stage suppression factor β_0, the optimal value is 0.2. Excessive suppression ($\beta_0 > 0.2$) harms the model's reasoning ability, which aligns with our selective intervention principle proposed in Section 3.3—suppressing only the small subset of heads dominated by knowledge priors while preserving the normal function of most heads.

4.4 Inference Latency Measurements

As shown in Table 5, CHERI requires only 0.2 GPU hours for probe training—100× less than POVID—while achieving the best AVG POPE F1 (86.42). At inference, CHERI adds no extra forward passes, running at 46.1ms (×1.1), versus VCD at 109.6ms (×2.6), OPERA at 89.8ms (×2.1), and DeCo at 52.4ms (×1.2). This demonstrates a uniquely favorable trade-off among training cost, inference speed, and hallucination mitigation.

Table 4: Ablation study results. We evaluate the contributions of the Visual Perception Probe (VPP) and Knowledge Prior Probe (KPP) to CHERI's performance. Experiments are conducted on LLaVA-1.5 and Qwen-VL-Chat, reporting average F1 scores on the POPE benchmark. Results show that both components contribute significantly to performance improvement, with the best results achieved when using both.

Components		Model	
VPP	KPP	LLaVA-1.5	Qwen-VL-Chat
✗	✗	79.30	81.53
✓	✗	83.26	84.25
✗	✓	84.72	85.14
✓	✓	86.42	87.11

Table 5: Training and Inference Cost Comparison.

Method	Training	Inference	AVG POPE F1
VCD	–	109.6ms (×2.6)	82.98
OPERA	–	89.8ms (×2.1)	83.56
DeCo	–	52.4ms (×1.2)	84.67
POVID	20.0 hours	**42.6ms** (×1.0)	86.12
V-DPO	2.3 hours	**42.6ms** (×1.0)	85.89
CHERI	**0.2 hours**	46.1ms (×1.1)	**86.42**

4.5 Generalizability

Further generalization results are deferred to the Appendix.

5 Conclusion

In this work, we trace the root of hallucinations to the causal imbalance between visual evidence and knowledge priors. Crucially, our analysis reveals that this imbalance is not random but follows a clear functional progression: early layers handle perception, while deeper layers drive reasoning. We leverage this natural division to design **CHERI**. It selectively intervenes on specific attention heads—amplifying visual signals when perception is weak and suppressing spurious priors when reasoning goes astray. We believe that such a causally-aware, low-latency approach offers a more viable path for deploying reliable multimodal systems in real-world scenarios.

Limitations

White-box model access. CHERI requires direct access to the internal attention states of the model—specifically, the per-head output representations $H_{(l,n)}$ and attention matrices $A_{(l,n)}$—to run the VPP/KPP probes and perform head-level interventions. This requirement precludes its application to closed-source, API-only models such as GPT-4V or Gemini. We note, however, that this is a shared constraint among virtually all attention-based inference-time intervention methods (e.g., OPERA, ADHH, DeCo), and does not limit applicability to the broad class of open-weight MLLMs.

References

Kazi Hasan Ibn Arif, Sajib Acharjee Dip, Khizar Hussain, Lang Zhang, and Chris Thomas. 2025. Fixing imbalanced attention to mitigate in-context hallucination of large vision-language model. *arXiv preprint arXiv:2501.12206*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zechen Bai, Pichao Wang, Tianjun Xiao, and 1 others. 2024. Hallucination of multimodal large language models: A survey. *CoRR*, abs/2404.18930.

Chao Chen, Kai Liu, Ze Chen, and 1 others. 2024a. INSIDE: llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024*.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. 2025. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4209–4221.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, and 1 others. 2024b. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14303–14312.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, and 1 others. 2025. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, and 1 others. 2024. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CVPR*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2023. Hallucination augmented contrastive learning for multimodal large language model. *Preprint*, arXiv:2312.06968.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2024. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *arXiv preprint arXiv:2411.16724*.

Zhehan Kan, Ce Zhang, Zihan Liao, and 1 others. 2024. Catch: Complementary adaptive token-level contrastive decoding to mitigate hallucinations in lvlms. *arXiv preprint arXiv:2411.12713*.

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

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023a. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025. Mllm can see? dynamic correction decoding for hallucination mitigation. *Preprint*, arXiv:2410.11779.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024. Mitigating object hallucination via concentric causal attention. *Advances in Neural Information Processing Systems*, 37:92012–92035.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025a. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024a. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Zhenfei Yin, Jiong Wang, Jianjian Cao, and 1 others. 2024b. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2023a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *Preprint*, arXiv:2311.13614.

Tianyu Yu, Yuan Yao, Haoye Zhang, and 1 others. 2023b. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *CoRR*, abs/2312.00849.

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. 2024. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2407.11422*.

Yudong Zhang, Ruobing Xie, Xingwu Sun, and 1 others. 2025a. Dhcp: Detecting hallucinations by cross-modal attention pattern in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025*, pages 3555–3564.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. 2025b. Cross-modal information flow in multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pages 19781–19791.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. *arXiv preprint arXiv:2407.00569*.

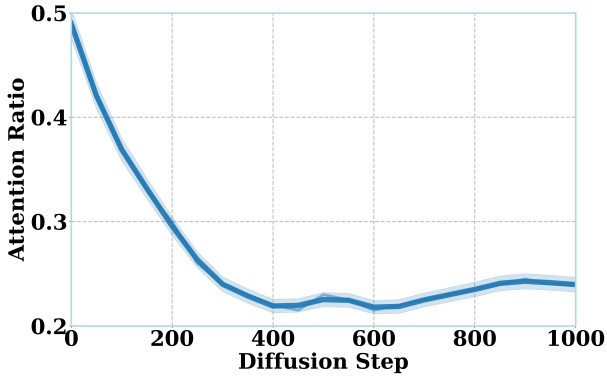
Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*.

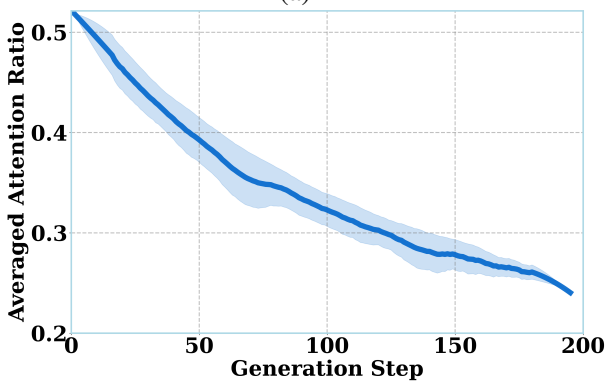
Use of Large Language Models

This article was written with the assistance of Large Language Models (LLMs) for text polishing and expression optimization to improve readability and fluency. All core arguments, ideas, and content are original work by the author(s).

Appendix



(a)



(b)

Figure 6: Impact of visual evidence strength and knowledge priors on attention allocation. (a) Under DDPM noise injection, the attention ratio to visual tokens in layer-0 decreases continuously from 0.50 to approximately 0.23 as diffusion steps increase, then stabilizes (mean standard deviation $\sigma = 0.0125$). The shaded area indicates the variation range of attention ratios across different samples. (b) During caption generation on the CHAIR dataset, the averaged attention ratio in layer-0 exhibits a monotonic decline from 0.54 to 0.35 across generation steps (mean standard deviation $\sigma = 0.0254$ over 195 generation steps). The shaded area represents attention fluctuation across samples at each generation step. Both experiments are conducted on 1,000 samples, validating the influence of weakened visual evidence and dominant knowledge priors on the model's attention mechanism.

A Quantitative Analysis: Dynamic Interplay Between Visual Evidence and Knowledge Priors

To further quantify the relationship between visual evidence and knowledge priors, we conduct controlled experiments on 1,000 samples, focusing on attention pattern variations in layer-0. We analyze this from two dimensions:

First, we examine how visual evidence strength affects attention allocation. Figure 6(a) demonstrates that under DDPM noise injection, the attention ratio to visual tokens decreases continuously from an initial 0.50 to approximately 0.23 as the diffusion step increases, then stabilizes. This pattern indicates that when objects in images are small or occluded—that is, when visual evidence weakens—the model's reliance on visual drops significantly. Generated descriptions in such cases lack detail, with many accessories and small objects overlooked, leading to increased hallucination rates. This result directly validates the critical role of the Direct Visual Pathway ($V \rightarrow Y$): sufficient visual evidence is a necessary condition for hallucination suppression.

Second, we analyze the mechanism by which knowledge priors operate during generation. Figure 6(b) presents the evolution of averaged attention ratio in layer-0 across generation steps on the CHAIR dataset. The results show a monotonic decrease from 0.54 to approximately 0.35. This declining trend reveals an important phenomenon: although visual evidence for "suit" exists in the image, the proximity of "restaurant" in the sequence activates knowledge priors that dominate the generation process. This validates the existence of the Indirect Prior Pathway ($K \rightarrow Y$), where K is jointly determined by visual input V, textual instruction X, generation history Y_{t-1}, and model-inherent bias C. Here we specifically illustrate how Y_{t-1} triggers knowledge priors, demonstrating that co-occurrence patterns in context can override current visual evidence.

Synthesizing both dimensions, we observe a dynamic equilibrium between visual evidence and knowledge priors: when visual signals are strong, the model relies on the direct pathway for accurate visual grounding; when visual signals weaken or context introduces strong linguistic biases, the indirect pathway dominates, causing the model to generate content consistent with linguistic statistics but unfaithful to the image.

B Benchmark Details

POPE (Polling-based Object Probing Evaluation). POPE reformulates hallucination detection as a binary classification task by querying the model with questions such as "Is there an X in the image?" The benchmark employs three negative sampling strategies with varying difficulty levels. The *Random* strategy selects non-existent objects randomly from the dataset as negative samples. The *Popular* strategy chooses objects that frequently appear in the dataset but are absent in the current image, such as "person" or "tree," based on statistical bias. The *Adversarial* strategy selects objects that co-occur with existing objects in the image but are not present in the current instance, such as "chopsticks" or "bowl" in a restaurant scene. Since the Adversarial strategy exploits statistical co-occurrence patterns in the training data, it presents the highest difficulty, while Random is the easiest. We evaluate model performance under all three settings and report Accuracy and F1 Score.

CHAIR (Caption Hallucination Assessment with Image Relevance). CHAIR evaluates object hallucinations in image captioning tasks. Objects mentioned in generated captions but absent from the image are considered hallucinations. The benchmark provides evaluation at two granularities. CHAIR_S measures sentence-level hallucination rate, computed as the proportion of sentences containing hallucinated objects over all sentences. CHAIR_I measures instance-level hallucination rate, computed as the proportion of hallucinated objects over all mentioned objects. These two metrics offer complementary insights into the reliability of generated descriptions.

MME (Multimodal LLM Evaluation). MME is a comprehensive benchmark comprising 14 subtasks designed for multimodal large language models. These tasks are divided into two categories: perception and cognition. Perception tasks assess fine-grained visual understanding, including existence, count, position, color, poster recognition, celebrity recognition, scene recognition, landmark recognition, artwork recognition, and optical character recognition. Cognition tasks evaluate high-level reasoning abilities, including commonsense reasoning, numerical calculation, text translation, and code reasoning. All questions are formulated as yes/no binary classification. Following the official implementation, we report the sum of Accuracy and Accuracy+ as the evaluation metric.

C About VPP/KPP

We report per-head probe performance on a held-out MS-COCO validation set. As expected, the degree of detectability varies across heads, reflecting the functional heterogeneity of attention heads:

- **VPP (binary classification):** Classification accuracy across heads ranges from 71.2% to 87.6%.
- **KPP (regression):** The Pearson correlation coefficient between predicted and ground-truth $\Delta \log P$ scores ranges from $r = 0.43$ to $r = 0.69$.

Ultimately, probe accuracy is not the end goal—the consistent gains across three benchmarks and two architectures demonstrate that even heads with moderate probe accuracy contribute effectively under selective intervention.

D About Generalizability

To further validate the generalizability of CHERI beyond the primary evaluation settings, we conduct two additional experiments: cross-domain generalization and cross-architecture generalization.

Cross-domain Generalization. Table 6 reports zero-shot cross-domain performance on A-OKVQA and GQA, where all models are trained on COCO only and evaluated directly without any domain-specific fine-tuning. On LLaVA-1.5, CHERI achieves 84.02% accuracy and 83.75 F1 on A-OKVQA, and 84.96% accuracy and 84.01 F1 on GQA, outperforming the second-best method DeCo by 3.38 and 1.88 points on A-OKVQA accuracy and F1, and by 2.21 and 1.62 points on GQA accuracy and F1, respectively. On Qwen-VL, CHERI similarly achieves the best performance across all four metrics (84.21/83.60 on A-OKVQA, 85.16/84.18 on GQA), surpassing DeCo by 2.33 and 1.64 points on A-OKVQA, and by 3.35 and 2.01 points on GQA. These results demonstrate that CHERI's causal head gating mechanism generalizes robustly across datasets without requiring any task-specific adaptation, confirming that the learned probes capture

Table 6: Zero-shot cross-domain generalization on A-OKVQA and GQA. Networks are trained on COCO only and evaluated directly without domain-specific fine-tuning.

Model	Method	A-OKVQA Acc	A-OKVQA F1	GQA Acc	GQA F1
LLaVA-1.5	Vanilla	78.13	78.10	77.89	78.13
	+VCD	79.99	81.30	80.16	81.67
	+DeCo	80.64	81.87	82.75	82.39
	+CHERI	**84.02**	**83.75**	**84.96**	**84.01**
Qwen-VL	Vanilla	78.31	79.18	78.15	79.22
	+VCD	80.07	80.96	80.65	81.43
	+DeCo	81.88	81.96	81.81	82.17
	+CHERI	**84.21**	**83.60**	**85.16**	**84.18**

Table 7: More mllms results

Method	InstructBLIP		MiniGPT-4	
	$\text{CHAIR}_\text{S} \downarrow$	$\text{CHAIR}_\text{I} \downarrow$	$\text{CHAIR}_\text{S} \downarrow$	$\text{CHAIR}_\text{I} \downarrow$
+DeCo	39.3	12.6	32.4	9.6
+CHERI	27.4	7.9	23.7	6.8

generalizable patterns of visual evidence insufficiency and knowledge prior dominance rather than dataset-specific artifacts.

Cross-architecture Generalization. Table 7 presents CHAIR hallucination results on two additional MLLM architectures: InstructBLIP and MiniGPT-4. CHERI consistently outperforms DeCo on both architectures and both metrics. On InstructBLIP, CHERI reduces CHAIR_S from 39.3 to 27.4 and CHAIR_I from 12.6 to 7.9, achieving reductions of 11.9 and 4.7 points respectively. On MiniGPT-4, CHERI reduces CHAIR_S from 32.4 to 23.7 and CHAIR_I from 9.6 to 6.8, corresponding to reductions of 8.7 and 2.8 points. These consistent improvements across architecturally distinct models further confirm that CHERI's head-level intervention strategy is not tailored to a specific model design, but rather targets a general mechanism of hallucination that manifests across diverse multimodal architectures.

E Other examples

Additional examples are presented below.

USER: "Please help me describe the image in detail."

LLaVA-v1.5-7b: "The image depicts a **man** and a young **boy** riding a **horse** in a grassy field. The man is sitting on the **horse**, **while the boy is sitting on his lap.** They are surrounded by a few **cows** in the field, with one cow located to the left of the **horse** and another cow further to the right. The man and the boy are carrying bags, with the man holding a **backpack** and the boy having a **handbag**. The scene appears to be a peaceful and leisurely outdoor activity, with the man and the boy enjoying their time together on the horse.",

Ours: "The image depicts a **man** and a young **boy** riding a **horse** in a grassy field. The man is sitting on the horse, while **the boy is sitting behind him.** They are surrounded by a few **cows** in the field, with one cow located to the left of the horse and another cow further to the right. The man and the boy are both wearing light-colored clothing, with the man wearing **a white hat** and the boy wearing **blue pants**. There are some **blue-green bags** visible on the horse. The scene looks peaceful and leisurely, with the man and the boy enjoying their outdoor time together on the horse."

USER: "Please help me describe the image in detail."

LLaVA-v1.5-7b: "The image features a **man** and a **woman** standing next to each other, with the **woman** helping the **man** tie his **tie**. The **man** is wearing a **suit** and **tie**, while the **woman** is dressed in **a black shirt**. They appear to be in a room with a **dining table** in the background. There are several **books** scattered around the room, with some placed on the **dining table** and others on the floor. **A cell phone can also be seen on the table**, possibly belonging to one of the individuals."

Ours: "The image features a **man** and a **woman** standing next to each other, with the **woman** helping the **man** tie his **tie**. The **man** is wearing a **suit**, **holding a cell phone in his hand**, while the **woman** is dressed in a **black dress**. In the background, there is a **red-framed glass** with an **exit sign** on it. A **trash can** is also visible on the floor near them. The scene suggests they are preparing for a formal event or gathering."

USER: "Please help me describe the image in detail."

LLaVA-v1.5-7b: "The image features a young **girl** sitting at a **dining table** in a restaurant, enjoying a meal. She is eating a **sandwich** and has **a cup in front of her.** The **table** is set with various items, including a **bowl**, a **fork**, and a **knife**. There are other **people** in the scene. A chair is positioned near the girl, and a **handbag** can be seen placed on the **table**. The restaurant appears to be a casual dining establishment, with a few **cups** visible on the **table**."

Ours: "The image features a young **girl** sitting at a **dining table** in a restaurant, enjoying a meal. She is wearing a **blue striped sweater** and is eating a **sandwich**. There are **two cups on the table in front of her**, with **one cup near her and another cup further away.** In the background, there are other **people** visible in the restaurant, and the setting appears to be a casual dining establishment. The restaurant has a warm atmosphere."

Figure 7: Additional qualitative examples comparing LLaVA-1.5-7b (baseline) and our method **CHERI** across diverse scenes. In each case, the baseline generates hallucinated objects driven by language priors (highlighted in red), such as confusing positional relationships or fabricating non-existent items. **CHERI** corrects these errors and produces descriptions that are more faithful to the visual content (highlighted in green), demonstrating consistent improvements in both hallucination suppression and fine-grained visual grounding.