One Token, Two Fates: A Unified Framework via Vision Token Manipulation Against MLLMs Hallucination

Supplementary Material

This appendix provides supplementary material to accompany our main paper, offering more detail on our experimental setup, providing extended results and analyses, and presenting the full theoretical analysis of our framework. The structure is as follows:

First, we detail our experimental setup in Sec.1, including specifics of the datasets and benchmarks used, our evaluation protocols and algorithmic pseudocode for our Synergistic Visual Calibration (SVC) and Causal Representation Calibration (CRC) modules to ensure full reproducibility.

Second, we present extended experimental results across additional model architectures and more benchmarks in Sec.2, in order to further show generalizability of our unified framework.

Third, in Sec.3, we provide in-depth analyses and ablation studies on critical design choices, such as the choice of augmentation types, providing empirical justification for the configurations used in our main experiments. Furthermore, we delve deeper into the synergistic relationship between SVC and CRC.

Fourth, to offer further qualitative insights, we include additional case studies that visually demonstrate our framework's effectiveness in mitigating specific types of hallucinations and enhancing visual grounding in Sec.4.

Finally, we present a more comprehensive theoretical discussion in Sec.5, expanding upon the Structural Causal Model (SCM) introduced in the main text to provide a derivation of our causal claims.

1. More Experiment&Algorithm Setup

1.1. Dataset and Benchmark Details

We evaluate our method on standard benchmarks for hallucination and general MLLM capabilities.

POPE [7]. The Polling-based Object Probing Evaluation assesses object hallucination in a binary question-answering format. Given an image and a question "Is there a [object] in the image?", the model must answer "Yes" or "No". We evaluate on the three standard splits based on object occurrence frequency (random, popular, adversarial) using subsets derived from COCO [9], AOKVQA [12], and GQA [5]. We report the average Accuracy and F1 score across the three splits for each dataset, following the official evaluation protocol.

CHAIR [11]. Caption Hallucination Assessment with Image Relevance measures object hallucination in open-ended image descriptions. It compares objects mentioned in the generated caption against ground-truth object labels. We report the instance-level score (CHAIR_I, percentage of hallucinated object instances) and the sentence-level score (CHAIR_S, percentage of captions containing hallucinations), where lower scores indicate better performance. We evaluate on 500 randomly sampled images from the MSCOCO validation set, using the prompt "Please help me describe the image in detail" with a maximum generation length as specified in the main paper (e.g., 64 and 128 tokens).

MMHal-Bench [13]. This benchmark provides a rigorous test of hallucination in complex reasoning scenarios beyond simple object presence. It consists of 96 image-question pairs across 8 challenging categories (Object attribute, Adversarial object, Comparison, Counting, Spatial relation, Environment, Holistic description, Others). We follow the official protocol, using GPT-4 [1] to score the model's responses against ground-truth answers on a scale relevant to the benchmark (details in [13]). We report the score across all categories.

MME [19]. The MME benchmark offers a comprehensive evaluation of general MLLM capabilities across 14 tasks: Existence, Count, Position, Color, Poster, Celebrity, Scene, Landmark, Artwork, OCR, Commonsense Reasoning, Numerical Calculation, Text Translation, Code Reasoning, grouped into Perception and Cognition categories. It uses carefully curated image-question pairs requiring diverse skills like OCR, spatial reasoning, and commonsense understanding. We conduct the full-set evaluation and report the overall sum of Perception and Cognition categories scores following the official script.

1.2. Pseudocode of Our Unified Framework

Algorithm 1 outlines the integration of SVC and CRC into MLLM's autoregressive generation process at each step t.

2. Additional Experimental Results

2.1. More Architectures

To further demonstrate the broad applicability and robustness of our unified framework, we conducted additional experiments on two more diverse MLLM architectures:

Algorithm 1 Unified Framework: SVC + CRC during Autoregressive Generation Step t

Inputs:
Image I, Augmented Image I_{aug}, Query Tokens $\mathbf{Q}$, Previously generated tokens $\mathbf{y}_{<t}$, MLLM $\mathcal{M}$ (Encoder $\mathcal{E}_V$, Projector $\mathcal{P}$, Decoder $\mathcal{D}$ with L layers)
Outputs:
Next token probability distribution $p_\theta(y_t|\cdot)$
Hyperparameters:
SVC layer L_c, SVC strength λ_s, CRC strength λ_c, Num negative samples K, Num retained tokens N_h
Global: CRC_Cache $\leftarrow$ None ▷ Cached $\{\mathbf{v}_{\text{crc}}^{(l)}\}_{l=1}^{L_c}$, computed once at first token
$\mathbf{V} \leftarrow \mathcal{P}(\mathcal{E}_V(I))$; $\mathbf{V}_{\text{aug}} \leftarrow \mathcal{P}(\mathcal{E}_V(I_{\text{aug}}))$ ▷ Get vision tokens
$\mathbf{X}_{\text{context}} \leftarrow [\mathbf{V}; \mathbf{Q}]$
$\mathbf{H}_{\text{current}} \leftarrow \text{Embed}([\mathbf{X}_{\text{context}}; \mathbf{y}_{<t}])$ ▷ Initial embeddings
Store initial $\mathbf{H}_{t,\text{org}}^{(0)} \leftarrow \mathbf{H}_{\text{current}}$ ▷ Store uncalibrated state
CRC_Cache $\leftarrow$ PRECOMPUTE_ALL_CRC_VECTORS($\mathbf{H}_{t,\text{org}}^{(0)}, \mathbf{V}, \mathbf{Q}, \mathbf{y}_{<t}, L_c, K, N_h$)
for $l = 1$ to L **do** ▷ Main forward pass through decoder layers
  $\mathbf{H}_{\text{input}}^{(l)} \leftarrow \mathbf{H}_{\text{current}}$
  if $l \leq L_c$ **then**
    $\mathbf{v}_{\text{crc}}^{(l)} \leftarrow$ CRC_Cache[l] ▷ Get pre-computed CRC vector
    $\mathbf{H}_{\text{input}}^{(l)} \leftarrow \text{Norm-Calibrate}(\mathbf{H}_{\text{input}}^{(l)}, \mathbf{v}_{\text{crc}}^{(l)}, \lambda_c)$ ▷ Apply norm-preserving calibration (Eq.(11))
  end if
  if $l == L_c$ **then**
    $\mathbf{V}_{\text{syn}} \leftarrow [\mathbf{V}; \mathbf{V}_{\text{aug}}]$ ▷ Get synergistic tokens
    $\mathbf{C}_t \leftarrow \text{softmax}\left(\frac{\mathbf{H}_{\text{input}}^{(l)}(\mathbf{V}_{\text{syn}})^T}{\sqrt{d}}\right)\mathbf{V}_{\text{syn}}$ ▷ Compute context (Eq.(4))
    $\mathbf{H}_{\text{input}}^{(l)} \leftarrow (1-\lambda_s)\cdot\mathbf{H}_{\text{input}}^{(l)} + \lambda_s \cdot \mathbf{C}_t$ ▷ Apply interpolation (Eq.(5))
  end if
  $\mathbf{H}_{\text{current}} \leftarrow \text{TransformerBlock}^{(l)}(\mathbf{H}_{\text{input}}^{(l)})$
  Store uncalibrated $\mathbf{H}_{t,\text{org}}^{(l)} \leftarrow \mathbf{H}_{\text{current}}$
end for
$h_{t,\text{last}}^{(L)} \leftarrow \mathbf{H}_{\text{current}}$[last_token_index]
$p_\theta(y_t|\cdot) \leftarrow \text{softmax}(\text{LM\_Head}(h_{t,\text{last}}^{(L)}))$
return $p_\theta(y_t|\cdot)$

function PRECOMPUTE_ALL_CRC_VECTORS($\mathbf{H}_{\text{org}}^{(0)}, \mathbf{V}, \mathbf{Q}, \mathbf{y}_{<t}, L_c, K, N_h$)
  Run full forward pass of decoder up to layer L_c to obtain $\{\mathbf{H}_{\text{org}}^{(l)}\}_{l=1}^{L_c}$ from $\mathbf{H}_{\text{org}}^{(0)}$
  Initialize cache as empty dictionary
  for $l = 1$ to L_c **do**
    Compute $\mathbf{v}_{\text{crc}}^{(l)}$ as average difference between $\mathbf{H}_{\text{org}}^{(l)}$ and K negative passes at layer l
    cache[l] $\leftarrow \mathbf{v}_{\text{crc}}^{(l)}$
  end for
  return cache
end function

Qwen-VL-Chat [2] and mPLUG-Owl2 [17]. Since both models utilize a 32-layer language backbone, we applied the same default hyperparameter settings as used in our main experiments (SVC at $L_c = 16$, $\lambda_s = 0.06$; CRC up to $L_c = 16$, $N_h = 5$, $K = 3$, $\lambda_c = 0.1$).

Table 1 presents the performance comparison on the POPE benchmark (averaged across COCO splits). Our framework consistently improves performance over the vanilla baseline for both models, showcasing its effectiveness across different vision-language alignment strategies.

Table 1. Additional results on the POPE benchmark (averaged over COCO splits). Our method demonstrates consistent improvements on diverse MLLM architectures using default settings. Best results are in **bold**.

Model	Method	Avg. Accuracy (%) ↑	Avg. F1 (%) ↑
Qwen-VL-Chat (7B)	Vanilla	84.23	85.10
	VCD	84.91	85.21
	Ours	**85.91**	**87.21**
mPLUG-Owl2 (7B)	Vanilla	76.22	79.73
	VCD	75.43	79.21
	Ours	**78.40**	**79.82**

2.2. Extended Performance Comparison on Additional Benchmarks

We present an extended quantitative comparison of different methods on LLaVA-1.5 (the core backbone in our main experiments) in Fig. 1, covering both object-level hallucination (POPE) and attribute-level hallucination (AMBER[15] subset) benchmarks, with three key metrics: POPE Accuracy, POPE F1-score, and AMBER Attribute Accuracy.

Results show that single-modality baselines (e.g., MemVR [21], VAF[18]) only bring marginal gains over Vanilla LLaVA-1.5. Our method achieves the best performance on all three metrics, including the AMBER subset for attribute hallucination.

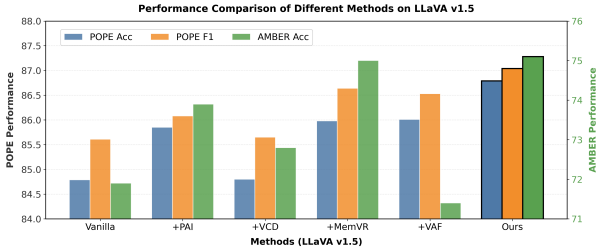

Figure 1. Performance comparison of different methods on LLaVA-1.5 for POPE (COCO) and AMBER (including an attribute hallucination subset) benchmarks. The left y-axis denotes POPE Accuracy/F1-score, and the right y-axis denotes AMBER Attribute Accuracy for numerical range adaptation. Our method outperforms all baselines on all metrics, including the AMBER subset for attribute hallucination verification.

2.3. Validating the Hallucination Direction

A core premise of our Causal Representation Calibration (CRC) module is that the computed calibration vector $\mathbf{v}_{\text{crc}}$ accurately captures the "hallucination direction" within the model's representation space. To validate this, we conduct an experiment where we apply the calibration vector not only in the intended suppressive direction (i.e., subtracting it from the original representation, corresponding to positive λ_c in Eq.10) but also in the reverse direction (i.e., adding it to the original representation, corresponding to negative λ_c).

Figure 2 visually conceptualizes this experiment. If $\mathbf{v}_{\text{crc}}$ truly represents the hallucination direction, then subtracting it should steer the representation towards a more visually grounded state, improving performance. Conversely, adding it should amplify the hallucination tendency, degrading performance.

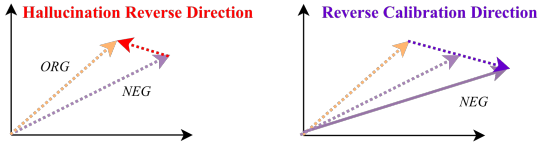

Figure 2. Conceptual illustration of validating the hallucination direction. **(Left)** Our CRC method subtracts the estimated hallucination direction ($v_{\text{crc}} \approx \text{ORG} - \text{NEG}$) from the original representation (ORG) to obtain a calibrated state (POS). **(Right)** Applying the vector in the reverse direction (adding v_{crc}) should push the representation further towards hallucination.

We test this on LLaVA-1.5 using the POPE-COCO benchmark, varying the calibration strength λ_c. Positive values correspond to our standard CRC calibration, while negative values simulate amplifying the hallucination. The results are presented in Table 2.

Table 2. Effect of calibration direction on POPE Accuracy (%). Applying CRC calibration (positive λ_c) improves accuracy over the baseline ($\lambda_c = 0$). Reversing the direction (negative λ_c) degrades accuracy, validating that $\mathbf{v}_{\text{crc}}$ captures the hallucination direction.

Model	LLaVA-1.5				
Calibration Strength (λ_c)	-0.1	-0.06	0 (Vanilla)	+0.06	+0.1
POPE Accuracy (%)	84.31	84.45	84.79	85.31	**86.11**

Results strongly support our observation: CRC with positive λ_c improves POPE accuracy over the Vanilla baseline (84.79% $\rightarrow$ 86.11% at $\lambda_c = 0.1$), while negative λ_c degrades it (e.g., 84.31% at $\lambda_c = -0.1$). The bidirectional, magnitude-dependent effect confirms that $\mathbf{v}_{\text{crc}}$ captures a hallucination-related direction in latent space, and our method effectively keep representations away from it.

3. More Analysis

3.1. Analysis on Layer Selection (L_c)

The choice of the intervention layer L_c (where SVC injects visual context and up to which CRC performs calibration) is critical [3]. We compare our static layer selection strategy against dynamic approaches and various static layer choices. Dynamic methods, such as those proposed by

[16, 21], aim to identify the most relevant layer at runtime based on intermediate representational states. Static choices involve intervening consistently at a fixed layer index.

Table 3. Ablation study on the intervention layer L_c. We report POPE Accuracy (%) and F1 (%) on COCO, and the relative inference latency compared to Vanilla. Dynamic methods incur significant overhead.

Intervention Layer (L_c)	POPE		Cost
	Acc (%) ↑	F1 (%) ↑	Relative Latency ↓
Dynamic Method 1 [16]	86.8	86.6	×1.56
Dynamic Method 2 [21]	86.3	86.4	×1.78
Static Layer 12	86.4	86.5	
Static Layer 14	86.6	**87.1**	
Static Layer 16 (Ours)	**86.8**	87.0	**×1.06 (Ours)**
Static Layer 18	86.7	86.8	
Static Layer 20	85.9	86.6	
Static Layer 24	85.1	86.1	

As shown in Table 3, dynamic layer selection methods, while conceptually appealing, introduce substantial inference latency and do not yield superior performance compared to well-chosen static layers. For static choices, we observe that intervening in the middle layers (around layer 16 for the 32-layer LLaMA-style architecture used in LLaVA-1.5) provides the most robust and highest performance. Intervening too early (e.g., layer 12) may not capture sufficient semantic context, while intervening too late (e.g., layer 24) risks instability and diminishing returns, potentially interfering with the final output formation. Considering the balance between performance, robustness, and efficiency, we select $L_c = 16$ as our default, consistent with findings in prior work on critical layers for semantic processing [16].

3.2. Semantic Retention Analysis of Local Context Representation

We analyze the semantic retention of local context representation (3×3 token neighbors) on 100 COCO samples (Fig. 3), comparing our pruning-based method with traditional masking-based methods via cosine similarity (quantifying semantic consistency between manipulated/original tokens).

Results show our pruning-based method achieves a median cosine similarity of 0.69, while masking only reaches 0.31. This gap confirms our method preserves original visual feature semantics and avoids contextual destruction from masking (which degrades features into OOD noise and induces hallucination-related bias).

3.3. Analysis on Augmentation Type for SVC

The choice of augmentation I_{aug} used to construct the synergistic visual context $\mathbf{V}_{\text{syn}}$ in SVC can influence performance. We compare several augmentation strategies by evaluating the performance of our framework with only the SVC module enabled (+SVC (V_{syn})).

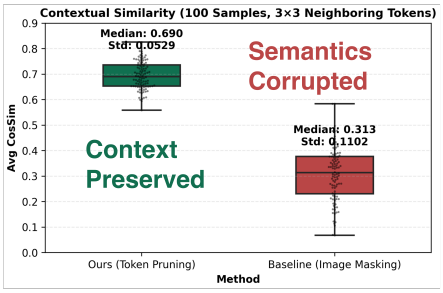

Figure 3. Local context representation (3×3 token neighbors) analysis on 100 COCO samples. Cosine similarity quantifies semantic consistency between manipulated (pruning/masking) and original tokens. Our method (Median = 0.69) retains far higher semantic identity than masking (Median = 0.31), avoiding OOD noise-induced bias.

Table 4. Ablation study on augmentation type for SVC (V_{syn}). We report POPE Accuracy (%) on COCO using only the SVC module.

Augmentation Type for I_{aug}	POPE Accuracy (%) ↑
None (Using V_{ori} only)	85.04
Gaussian Blur Only	85.25
Random Noise Only	85.20
Random Flip Only	85.18
Gaussian Blur + Random Noise & Flip (Ours)	**85.55**
Gaussian Blur + Random Flip	85.28

Table 4 shows the results. While individual augmentations offer slight improvements over using only V_{ori}, combining Gaussian Blur with Random Noise yields the best performance. We hypothesize that this combination provides the most effective semantic complementarity (F2): Gaussian blur forces the model to capture coarser, global features by removing fine details, while random noise might disrupt local texture patterns, potentially encouraging attention to different structural aspects compared to the original image. Random flips provide geometric variance but seem less effective in generating complementary semantic cues for this task. Thus, we adopt Gaussian Blur + Random Noise & Flip as our default augmentation strategy.

3.4. Analysis on Module Contribution and Synergy

Our main ablation study in paper reveals that CRC alone yields stronger improvements in hallucination mitigation (POPE Acc) than SVC alone, although both are effective. This observation warrants further discussion.

Why is CRC more impactful alone? We attribute CRC's stronger individual performance to its direct targeting of the *source* of the text inertia. By intervening in shallow lay-

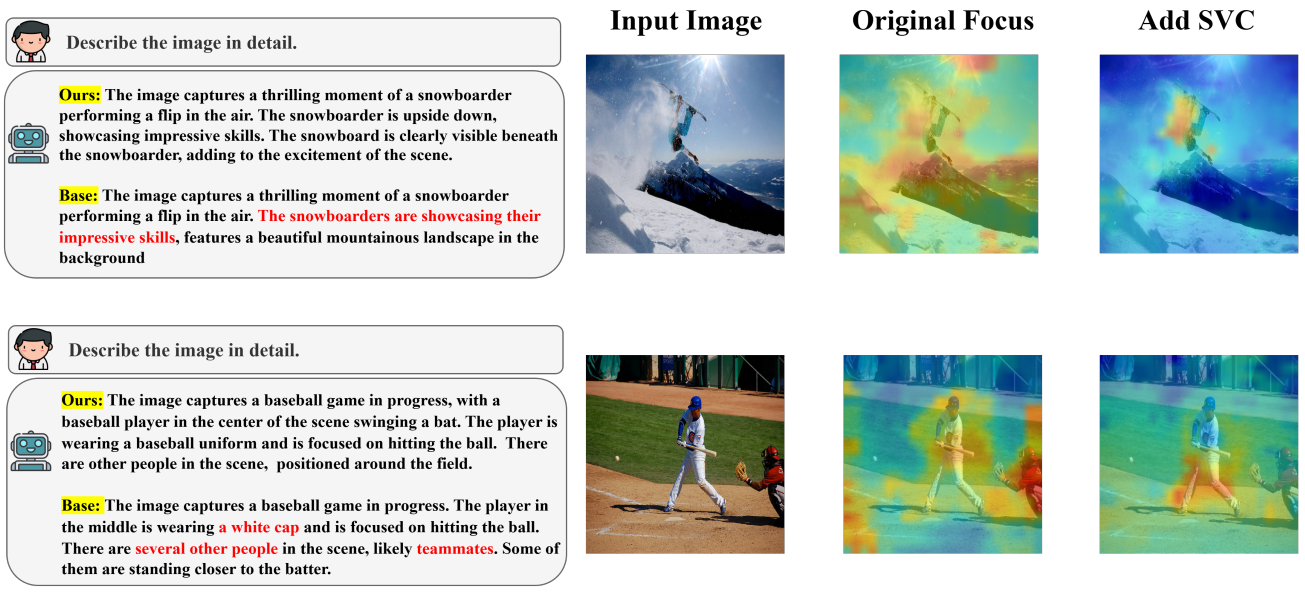

Figure 4. **Qualitative Case Studies.** We compare text outputs and visual focus (via TAM) between a baseline MLLM (Base) and our method (Ours). **(Top)** Our method correctly identifies a single snowboarder, unlike the baseline which hallucinates multiple. SVC sharpens focus on the subject. **(Bottom)** Our method avoids hallucinating details like a "white cap" and extra "teammates" seen in the baseline output. SVC helps concentrate attention on relevant scene elements. Red text highlights hallucinations in baseline responses.

ers, CRC purifies the representations before text inertia can strongly take hold. It directly counteracts the overpowering linguistic priors, which is arguably the dominant factor in the vision-language imbalance leading to hallucination. SVC, while crucial, enhances the *counteracting* visual signal, which might still struggle against a deeply ingrained bias if applied alone, especially in later generation stages where visual fading (F1) is most pronounced.

Complementarity and Synergy of SVC and CRC. Despite CRC's stronger individual effect, the full model (SVC+CRC) consistently outperforms either module alone, demonstrating clear synergy. We posit that SVC and CRC address different aspects of the vision-language imbalance in a complementary manner:

- **CRC creates a cleaner foundation:** By purifying representations in shallow layers, CRC removes the initial dominance of linguistic bias, allowing subsequent layers to process visual information more effectively.
- **SVC provides a robust visual anchor:** With the initial bias suppressed by CRC, the enhanced, synergistic visual context injected by SVC in the middle layer (L_c) acts as a much stronger and more reliable anchor. It guides the generation process towards visually grounded outputs, preventing the model from reverting to linguistic priors even when CRC's early intervention effect might naturally decay in deeper layers.

Essentially, CRC cleans the representation, and SVC writes the correct visual information onto it more effectively.

4. Case Studies

To provide qualitative insights into how our unified framework mitigates hallucinations and enhances visual grounding, we present two case studies in Figure 4. We compare the outputs generated by our method (Ours) against a baseline MLLM (Base) and visualize the impact of our SVC module on the model's visual focus using Token Activation Mapping (TAM) [8].

These case studies qualitatively illustrate how our unified framework, through synergistic visual calibration and causal representation calibration, leads to more visually grounded and factually accurate descriptions by restoring a better vision-language balance.

5. Detailed Theoretical Derivations for CRC

This appendix provides a detailed theoretical derivation for our Causal Representation Calibration (CRC) module, grounding it in the framework of Structural Causal Models (SCM).

5.1. A Causal Framework for Hallucination

We posit that hallucination in MLLMs arises from a spurious causal pathway dominated by the model's intrinsic biases, confounding the true visual pathway. We formalize this using an SCM, depicted in Figure 4.

Variables of the Causal Graph:

- V: Variable representing the ground-truth visual facts in image I.
- Q: Variable representing the user's textual query T.
- B: Variable representing the model's **Intrinsic Bias** (including noise, linguistic priors, architectural artifacts).
- $H_t^{(l)}$: Variable representing the latent hidden state at layer l and generation step t.
- Y: Endogenous variable representing the final generated text sequence $\mathbf{y}$.

Causal Paths: The latent representation $H_t^{(l)}$ is influenced by three primary paths:

1. **True Visual Path**: $V \to H_t^{(l)} \to Y$.
2. **Query Path**: $Q \to H_t^{(l)} \to Y$.
3. **Spurious Bias Path**: $B \to H_t^{(l)} \to Y$.

Hallucination occurs when the influence of the spurious path ($B \to H_t^{(l)}$) dominates the true visual path ($V \to H_t^{(l)}$). The objective of CRC is to estimate and counteract the effect channeled through this spurious path.

5.2. Derivation of the CRC Vector Identity

Our derivation relies on two key assumptions about the local behavior of the MLLM's latent space, supported implicitly by recent interpretability findings [4, 6, 10, 14, 20]:

Assumption 1 (Effect Combination)**.** *Within a local region of the latent space, the hidden state $H_t^{(l)}$ can be approximated as a linear superposition of the causal effects from its direct parents V, Q, B:*

$$H_t^{(l)} \approx \mathcal{E}_V(V) + \mathcal{E}_Q(Q) + \mathcal{E}_B(B) + \epsilon^{(l)} \tag{1}$$

where $\mathcal{E}_V, \mathcal{E}_Q, \mathcal{E}_B$ are approximately linear functions mapping the causal variables to their effects in the latent space $\mathcal{H}^{(l)}$, and $\epsilon^{(l)}$ represents residual noise or interactions.

Assumption 2 (Approximate Subspace Separability)**.** *The causal effects from the visual input ($\mathcal{E}_V$) and the shared non-visual components ($\mathcal{E}_Q + \mathcal{E}_B$) predominantly reside in distinct subspaces within $\mathcal{H}^{(l)}$. We can thus group the non-visual effects:*

$$\mathcal{E}_{shared}(Q, B) = \mathcal{E}_Q(Q) + \mathcal{E}_B(B) \tag{2}$$

leading to the decomposition:

$$H_t^{(l)} \approx \mathcal{E}_V(V) + \mathcal{E}_{shared}(Q, B) + \epsilon^{(l)} \tag{3}$$

Now, consider the computation performed by CRC. We have the original hidden state $H_{t,\text{org}}^{(l)}$ derived from the true visual input V, and the negative hidden state $H_{t,\text{neg}}^{(l)}$ derived from the degraded visual input V_{neg} (obtained via latent token pruning). Applying our assumptions:

$$H_{t,\text{org}}^{(l)} \approx \mathcal{E}_V(V) + \mathcal{E}_{\text{shared}}(Q, B) + \epsilon_{\text{org}}^{(l)} \tag{4}$$

$$H_{t,\text{neg}}^{(l)} \approx \mathcal{E}_V(V_{\text{neg}}) + \mathcal{E}_{\text{shared}}(Q, B) + \epsilon_{\text{neg}}^{(l)} \tag{5}$$

Crucially, both computations share the same query Q and are subject to the same intrinsic bias B within the model θ. Therefore, their contribution $\mathcal{E}_{\text{shared}}(Q, B)$ is identical in both equations. The noise terms $\epsilon_{\text{org}}^{(l)}$ and $\epsilon_{\text{neg}}^{(l)}$ may differ slightly but are assumed to be small.

We compute the difference between these two hidden states:

$$\begin{aligned}
\Delta H_t^{(l)} &= H_{t,\text{org}}^{(l)} - H_{t,\text{neg}}^{(l)} && (6)\\
&\approx (\mathcal{E}_V(V) + \mathcal{E}_{\text{shared}}(Q, B) + \epsilon_{\text{org}}^{(l)}) \\
&\quad - (\mathcal{E}_V(V_{\text{neg}}) + \mathcal{E}_{\text{shared}}(Q, B) + \epsilon_{\text{neg}}^{(l)}) && (7)\\
&\approx \mathcal{E}_V(V) - \mathcal{E}_V(V_{\text{neg}}) + (\epsilon_{\text{org}}^{(l)} - \epsilon_{\text{neg}}^{(l)}) && (8)
\end{aligned}$$

Due to the assumed linearity of $\mathcal{E}_V$, we have $\mathcal{E}_V(V) - \mathcal{E}_V(V_{\text{neg}}) \approx \mathcal{E}_V(V - V_{\text{neg}})$.

The CRC vector $\mathbf{v}_{\text{crc}}^{(l)}$ is obtained by averaging this difference over K independent negative samples $V_{\text{neg}}^{(k)}$. Assuming the residual noise differences $(\epsilon_{\text{org}}^{(l)} - \epsilon_{\text{neg}}^{(l,k)})$ have zero mean, the averaging process further reduces their influence:

$$\begin{aligned}
\mathbf{v}_{\text{crc}}^{(l)} &= \frac{1}{K}\sum_{k=1}^{K}(H_{t,\text{org}}^{(l)} - H_{t,\text{neg}}^{(l,k)}) && (9)\\
&\approx \frac{1}{K}\sum_{k=1}^{K}\left(\mathcal{E}_V(V - V_{\text{neg}}^{(k)}) + (\epsilon_{\text{org}}^{(l)} - \epsilon_{\text{neg}}^{(l,k)})\right) && (10)\\
&\approx \mathbb{E}_{V_{\text{neg}}}[\mathcal{E}_V(V - V_{\text{neg}})] && (11)\\
&\approx \mathcal{E}_V(V - \mathbb{E}[V_{\text{neg}}]) \quad \text{(by } \textbf{Assumption 1}\text{)} && (12)
\end{aligned}$$

Since V_{neg} represents a heavily pruned version of V, the difference $(V - \mathbb{E}[V_{\text{neg}}])$ primarily captures the *missing* visual information content. Therefore, Eq. 12 formally justifies our claim: $\mathbf{v}_{\text{crc}}^{(l)}$ serves as a robust estimate of the causal effect originating purely from the visual information difference induced by our pruning intervention, effectively disentangled from the shared query and intrinsic bias components.

Empirical Verification of CRC Assumptions. As illustrated in main paper Fig.7–8, we observe that the visual token distributions under augmentation exhibit quasi-linear displacement in the latent space, supporting the approximate local linearity assumption used in main paper Eq.(11). The observed partial overlap between visual and shared clusters further motivates our subtraction-based calibration rather than an ideal orthogonal decomposition.

Connection to Counterfactual Calibration. The final calibration step in CRC, $H^{(l)}_{t,\text{pos}} = H^{(l)}_{t,\text{org}} - \lambda_c \cdot \mathbf{v}^{(l)}_{\text{crc}}$, can be interpreted as a counterfactual adjustment. We estimate the deviation component $\mathbf{v}^{(l)}_{\text{crc}}$ that arises due to weakened visual input (allowing bias B to exert more relative influence) and subtract it from the original representation. This steers the hidden state away from the bias-induced direction and towards a state that would have occurred had the visual signal remained strong relative to the bias, thus promoting a more faithful, visually-grounded output.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[3] Haoran Chen, Junyan Lin, Xinhao Chen, Yue Fan, Xin Jin, Hui Su, Jianfeng Dong, Jinlan Fu, and Xiaoyu Shen. Rethinking visual layer selection in multimodal llms. *arXiv preprint arXiv:2504.21447*, 2025. 3

[4] Woody Haosheng Gan, Deqing Fu, Julian Asilis, Ollie Liu, Dani Yogatama, Vatsal Sharan, Robin Jia, and Willie Neiswanger. Textual steering vectors can improve visual understanding in multimodal large language models. *arXiv preprint arXiv:2505.14071*, 2025. 6

[5] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1

[6] Jaewoo Lee, Keyang Xuan, Chanakya Ekbote, Sandeep Polisetty, Yi R Fung, and Paul Pu Liang. Tamp: Token-adaptive layerwise pruning in multimodal large language models. *arXiv preprint arXiv:2504.09897*, 2025. 6

[7] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

[8] Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. Token activation map to visually explain multimodal llms. *arXiv preprint arXiv:2506.23270*, 2025. 5

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[10] Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Grains: Gradient-based attribution for inference-time steering of llms and vlms. *arXiv preprint arXiv:2507.18043*, 2025. 6

[11] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 1

[12] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 1

[13] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 1

[14] Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. Cross-modal projection in multimodal llms doesn't really project visual attributes to textual space. *arXiv preprint arXiv:2402.16832*, 2024. 6

[15] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 3

[16] Sudong Wang, Yunjian Zhang, Yao Zhu, Jianing Li, Zizhe Wang, Yanwei Liu, and Xiangyang Ji. Towards understanding how knowledge evolves in large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29858–29868, 2025. 4

[17] 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*, pages 13040–13051, 2024. 2

[18] Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14625–14634, 2025. 3

[19] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), 2024. 1

[20] 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 *The Thirteenth International Conference on Learning Representations*. 6

[21] Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *Forty-second International Conference on Machine Learning*. 3, 4