SAKED: MITIGATING HALLUCINATION IN LARGE VISION-LANGUAGE MODELS VIA STABILITY-AWARE KNOWLEDGE ENHANCED DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucinations in Large Vision-Language Models (LVLMs) pose significant security and reliability risks in real-world applications. Inspired by the observation that humans are more error-prone when uncertain or hesitant, we investigate how instability in a model's internal knowledge contributes to LVLM hallucinations. We conduct extensive empirical analyses from three perspectives, namely attention heads, model layers, and decoding tokens, and identify three key hallucination patterns: (i) visual activation drift across attention heads, (ii) pronounced knowledge fluctuations across layers, and (iii) visual focus distraction between neighboring output tokens. Building on these findings, we propose **S**tability-**A**ware **K**nowledge-**E**nhanced **D**ecoding (**SAKED**), which introduces a layer-wise **K**nowledge **S**tability **S**core (**KSS**) to quantify knowledge stability throughout the model. By contrasting the most stability-aware and stability-agnostic layers, SAKED suppresses decoding noise and dynamically leverages the most reliable internal knowledge for faithful token generation. Moreover, SAKED is training-free and can be seamlessly integrated into different architectures. Extensive experiments demonstrate that SAKED achieves state-of-the-art performance for hallucination mitigation on various models, tasks, and benchmarks.

1 INTRODUCTION

In recent years, Large Vision–Language Models (LVLMs) have made substantial progress in cross-modal understanding and generation. By connecting a pretrained language model with a vision encoder and fine-tuning on instruction-based datasets, LVLMs achieve strong performance across a wide range of multimodal tasks OpenAI (2023); Chen et al. (2023); Yin et al. (2024); Zhao et al. (2024b). Nevertheless, LVLMs remain prone to hallucinations Rohrbach et al. (2018); Zhou et al. (2023); Li et al. (2023b), where models generate content that is not grounded in the input signals but instead relies on their internal knowledge Heiman et al. (2025). Such failures raise pressing safety and security concerns in high-stakes applications, such as autonomous driving, medical diagnosis, and financial analysis, etc Li et al. (2026); Zhang et al. (2023); Xia et al. (2026).

Prior work has investigated the causes of LVLM hallucinations, including language priors Wang et al. (2024a); Li et al. (2026), dataset bias Li et al. (2023b); Zhou et al. (2023), error accumulation Huang et al. (2024); Zhou et al. (2023), and suboptimal training policy Kalai et al. (2025); Yang et al. (2025). Accordingly, a variety of mitigation strategies have been proposed, such as contrastive decoding Leng et al. (2024); Favero et al. (2024); Chuang et al. (2023), fine-tuning with improved learning policies Jiang et al. (2024); Yang et al. (2025); Sarkar et al. (2025), early intervention Wang et al. (2024a); Li et al. (2026); Wang et al. (2025a), information steering Li et al. (2025); Park et al. (2025a), and representation editing Liu et al. (2024); Kang et al. (2025); Gong et al. (2024); Kong et al. (2025). Despite this progress, most existing methods do not explicitly model how internal knowledge fluctuations give rise to hallucinations. Without characterizing when and where internal knowledge becomes unreliable, mitigation methods can be brittle and may fail under distribution shifts Heiman et al. (2025); Lv et al. (2024). In contrast, a stability-aware perspective enables the model to dynamically leverage more reliable internal knowledge, thereby achieving more faithful decoding.

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.

Motivated by studies of metacognition, uncertainty, and error monitoring in human decision-making Yeung & Summerfield (2012); Fleming (2024), we investigate whether hallucinations in LVLMs are accompanied by observable fluctuations in their internal knowledge states. Our central intuition is that reliable visual evidence should lead different model components to form relatively consistent internal judgments, whereas insufficient, ambiguous, or conflicting evidence may result in unstable representations during generation. A key challenge, therefore, is to explicitly characterize and quantify such knowledge fluctuations.

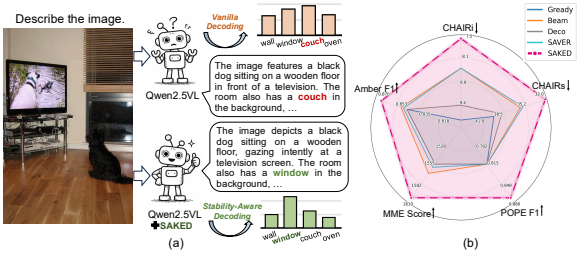

Figure 1: (a) SAKED mines internal stability-aware knowledge to enhance the decoding. (b) SAKED consistently achieves strong average performance across datasets.

To this end, we conduct extensive empirical analyses of internal knowledge stability from three complementary perspectives: attention heads, model layers, and consecutive decoding tokens. These analyses reveal three patterns associated with hallucinated generation. **Pattern #1**: Hallucinated tokens tend to exhibit greater visual activation drift across attention heads, whereas consistent head-wise activations indicate stronger visual grounding and more faithful decoding. **Pattern #2**: Large fluctuations in predictive knowledge across model layers are associated with a higher risk of hallucination, while relatively stable layers tend to preserve more reliable information for token prediction. **Pattern #3**: Hallucinated generation is also accompanied by stronger distraction in visual focus between neighboring decoding steps, suggesting that abrupt cross-token shifts in visual attention indicate unreliable internal knowledge. To verify that these observations extend beyond individual examples, we further conduct a statistical analysis of true and hallucinated object tokens.

We propose Knowledge Stability Score (KSS), which comprises (i) Cross-Head Stability Score (CHSS) to measure visual activation drift across attention heads, (ii) Cross-Layer Stability Score (CLSS) to quantify knowledge fluctuations across layers, and (iii) Cross-Token Stability Score (CTSS) to assess visual focus distraction between adjacent decoding steps. Building on KSS, we further introduce stability-aware contrastive decoding that contrasts logits from the most stability-aware layer and the most stability-agnostic layer to suppress decoding noise and improve faithfulness. As shown in Figure 1, SAKED achieves the best average performance across the evaluated models and benchmarks. In summary, our main contributions are as follows:

- We conduct extensive empirical and statistical analyses to examine when and where hallucinations arise by tracking model knowledge evolution from three perspectives: attention heads, model layers, and generated tokens. We find that unstable internal model knowledge is associated with a higher risk of hallucination.
- We propose a novel **S**tability-**A**ware **K**nowledge **E**nhanced **D**ecoding (**SAKED**) method that quantifies knowledge stability for each layer. We further improve decoding by contrasting the most stability-aware layer with the most stability-agnostic layer, thereby reducing hallucinations. SAKED is plug-and-play and can be seamlessly integrated into various LVLMs.
- Extensive experiments demonstrate that SAKED achieves state-of-the-art hallucination mitigation performance across a wide range of models, datasets, and tasks.

2 Related Work

2.1 Large Vision-Language Models (LVLMs)

Recent advances in LVLMs have yielded strong performance across a wide range of vision–language tasks. Most LVLMs consist of three components: a vision encoder that extracts visual features from the input image, a vision–language connector that aligns these features with the language representation space, and a pretrained LLM that consumes the resulting multimodal embeddings to generate responses. To reduce the modality gap between visual features and textual representations, prior work has explored diverse alignment strategies, including lightweight linear projection layers that

map visual features into the LLM token space (e.g., LLaVA Liu et al. (2023), MiniGPT-4 Zhu et al. (2023)), Q-former modules that learn query embeddings for multimodal grounding (e.g., BLIP-2 Li et al. (2023a), InstructBLIP Dai et al. (2023)), and cross-attention architectures that enable iterative interaction between visual tokens and text (e.g., Flamingo Alayrac et al. (2022), OpenFlamingo Awadalla et al. (2023)). More recent systems, such as GPT-5 OpenAI (2025) and Gemini-3 Google DeepMind (2025), further demonstrate enhanced capabilities on vision tasks requiring complex reasoning. Despite this progress, LVLMs remain susceptible to severe hallucinations, raising pressing safety and security concerns, especially in critical application scenarios (e.g., medical diagnosis).

2.2 Hallucination in LVLMs

Hallucination in LVLMs typically refers to outputs that are not grounded in the input signals and instead rely excessively on the model's internal knowledge Heiman et al. (2025). Recent tuning-based approaches usually improve their learning objectives or training policies Jiang et al. (2024); Yang et al. (2025); Sarkar et al. (2025); Park et al. (2025b) on high-quality curated datasets Wang et al. (2024b); Zhang et al. (2024); Jing & Du (2024). Representative efforts also explore architectural refinement for hallucination mitigation, such as HalluRNN Yu et al. (2025). In contrast, training-free methods are generally more computationally efficient. For example, contrastive decoding Leng et al. (2024); Favero et al. (2024); Chuang et al. (2023) compares output logits across paired inputs (e.g., different image conditions) to suppress ungrounded generations, while recent adaptive decoding methods such as Mixture of Decoding Chen et al. (2025) further improve inference-time mitigation. Early-intervention approaches Wang et al. (2024a); Li et al. (2026); Wang et al. (2025a); Chen et al. (2024b); Wu et al. (2024); Fang et al. (2025a); Wang et al. (2025b); Fang et al. (2025b) dynamically refine decoding by leveraging more reliable signals from earlier layers. In addition, representation-editing methods Liu et al. (2024); Kang et al. (2025); Gong et al. (2024); Zhao et al. (2024a); Zou et al. (2024); Chen et al. (2024a) reallocate attention to encourage the model to focus on relevant visual regions. Despite substantial progress, existing decoding methods generally select intermediate knowledge using token confidence, fixed early-layer signals, or visual-attention feedback, without explicitly measuring the stability of internal knowledge. Unlike these approaches, SAKED estimates token-specific reliability jointly through cross-head agreement, cross-layer predictive continuity, and cross-token visual consistency, and contrasts the most and least stable layers during a single decoding pass.

3 Methodology

In this section, we investigate how instability in internal model knowledge contributes to hallucinations in LVLMs. We begin by formalizing the generation process and introducing the notation used throughout the paper. We then examine the relationship between knowledge stability and hallucination from three complementary perspectives: attention heads, model layers, and decoding steps. Our empirical analysis reveals three characteristic instability patterns associated with hallucination: *(i)* visual activation drift across attention heads, *(ii)* pronounced knowledge fluctuation across layers, and *(iii)* visual focus distraction between adjacent decoding steps. Motivated by these observations, we propose **S**tability-**A**ware **K**nowledge **E**nhanced **D**ecoding (**SAKED**). Specifically, we introduce three stability measures, namely Cross-Head Stability Score (CHSS), Cross-Layer Stability Score (CLSS), and Cross-Token Stability Score (CTSS), and aggregate them into a unified Knowledge Stability Score (KSS). Based on KSS, we identify a stability-aware layer and a stability-agnostic layer, and incorporate their contrast into decoding so as to suppress noisy internal knowledge and improve generation faithfulness. Since SAKED operates entirely at inference time, it is training-free and can be seamlessly integrated into different LVLM architectures. The theoretical analysis is provided in the appendix section *Additional Theoretical Analysis*, where we show that under a monotone deviation-control assumption, our analysis interprets KSS as a surrogate for hallucination risk and provides sufficient conditions under which stability-aware contrastive correction favors grounded tokens.

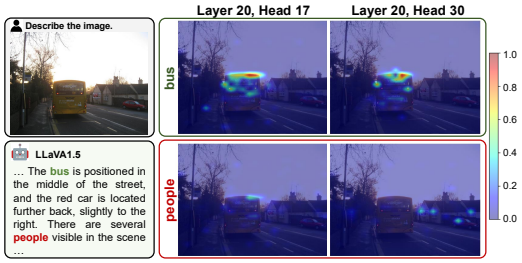

Figure 2: The left panel shows the input image and the output caption. The right panel visualizes the attention maps of a grounded token ("bus") and a hallucinated token ("people") across different attention heads.

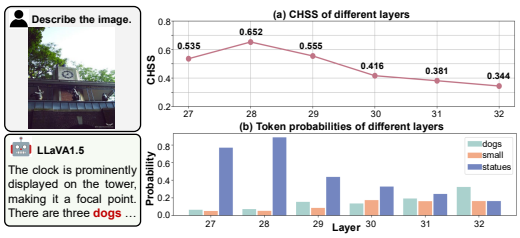

Figure 3: (a) Layer-wise CHSS. (b) Layer-wise token probabilities, showing the final layer's top three tokens.

3.1 Preliminaries

An LVLM typically consists of three main components: a vision encoder, a projection module, and an autoregressive language model. Given an input image, the vision encoder first converts it into a sequence of visual tokens $\mathbf{x}^I = \{\boldsymbol{x}_1^I, \boldsymbol{x}_2^I, \ldots, \boldsymbol{x}_m^I\}$. In parallel, a textual prompt is tokenized into $\mathbf{x}^P = \{\boldsymbol{x}_1^P, \boldsymbol{x}_2^P, \ldots, \boldsymbol{x}_n^P\}$. Here, m and n denote the numbers of visual and textual tokens, respectively. These embeddings are concatenated to form the multimodal input $\mathbf{x}$, which is processed by an autoregressive language model consisting of L stacked transformer decoder layers. At layer l, the model produces hidden states $\mathbf{h}^l = \{\boldsymbol{h}_0^l, \boldsymbol{h}_1^l, \ldots, \boldsymbol{h}_{m+n-1}^l\}$. During generation, the hidden state at the final position is projected by an affine mapping $\phi(\cdot)$ to produce a token distribution over the vocabulary $\mathcal{V}$:

$$\boldsymbol{y}_t \sim p_\theta(\boldsymbol{y}_t \mid \mathbf{x}^I, \mathbf{x}^P, \boldsymbol{y}_{<t}), \tag{1}$$

where θ denotes the model parameters.

3.2 Cross-Head Stability

We first examine the relationship between cross-head consistency and hallucination. In transformer architectures, multi-head self-attention aggregates multiple attention views to form richer and more robust representations. Figure 2 visualizes the visual activations of a grounded token ("bus") and a hallucinated token ("people") across attention heads within the same layer. The grounded token exhibits concentrated high-activation regions that align well with its semantic referent, whereas the hallucinated token shows more diffuse and higher-uncertainty activations. Moreover, the grounded token displays substantially stronger agreement across attention heads. These observations suggest that hallucinations are more likely to arise when visual evidence varies markedly across heads. To quantify the informativeness of the k-th attention head in layer l at decoding step t, we define the Visual Activation Score (VAS) as

$$\mathrm{VAS}_t^{(l,k)} = M_t^{(l,k)} + O_t^{(l,k)}, \tag{2}$$

where $M_t^{(l,k)}$ and $O_t^{(l,k)}$ denote the peak activation and the concentration statistic of the visual attention map $A_t^{(l,k)} \in \mathbb{R}^{a \times a}$, respectively. Heads with larger VAS are regarded as more informative and thus more suitable for characterizing visual grounding. For each layer l, we select the top-K representative heads with the highest VAS values from the full set of H heads. Denote these selected maps by $\hat{A}_t^{(l,k)}$, where $k \in \mathcal{K} = \{1, \ldots, K\}$. We then define the Cross-Head Stability Score (CHSS) as

$$\mathrm{CHSS}_t^l = \frac{2}{K(K-1)} \sum_{s,k \in \mathcal{K},\, s<k} \mathrm{SoftIoU}(\hat{A}_t^{(l,s)}, \hat{A}_t^{(l,k)}). \tag{3}$$

Here, SoftIoU measures the consistency between two selected attention maps:

$$\mathrm{SoftIoU}\left(\hat{A}^{(l,s)}, \hat{A}^{(l,k)}\right) = \frac{\left\|\min\left(\hat{A}^{(l,s)}, \hat{A}^{(l,k)}\right)\right\|_1}{\left\|\max\left(\hat{A}^{(l,s)}, \hat{A}^{(l,k)}\right)\right\|_1 + \varepsilon}, \tag{4}$$

where $\min(\cdot)$ and $\max(\cdot)$ are applied elementwise and $\varepsilon > 0$ is a small constant for numerical stability.

Figure 3 presents the layer-wise CHSS values together with the corresponding token distributions. Layers with larger CHSS, such as layers 27 and 28, are associated with more accurate token predictions, indicating that stronger cross-head agreement yields more reliable grounding. In contrast, the final layer exhibits the lowest CHSS and produces a hallucinated token. This observation motivates using stability to identify layers that contain more trustworthy internal knowledge.

3.3 Cross-Layer Stability

We next investigate how knowledge evolves across layers and where instability emerges. Figure 4 visualizes the Jensen–Shannon divergence (JSD) of layer-wise token distributions. For each generated token $\boldsymbol{y}_t$, we apply LogitLens to the hidden state $\boldsymbol{h}_t^l$ at each layer and obtain

$$\phi(\boldsymbol{h}_t^l) = \mathrm{softmax}(\mathrm{LogitLens}(\boldsymbol{h}_t^l)), \quad l \in \{1, \ldots, L\}. \tag{5}$$

LogitLens maps intermediate representations into the vocabulary space, thereby making it possible to inspect how token preferences evolve throughout the network.

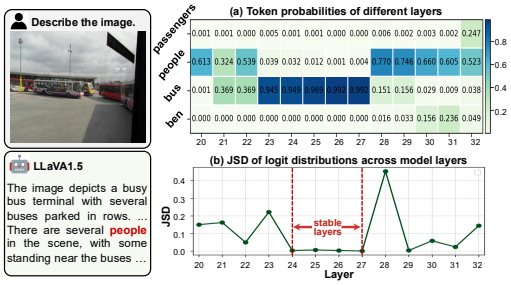

Figure 4: (a) Layer-wise token probabilities for "people", showing the final layer's top four tokens. (b) JSD between adjacent-layer logit distributions; lower values indicate greater transition stability.

We define the Cross-Layer Stability Score (CLSS) as

$$\mathrm{CLSS}_t^l = 1 - \mathrm{JSD}\big(\phi(\boldsymbol{h}_t^l), \phi(\boldsymbol{h}_t^{l-1})\big). \tag{6}$$

A larger CLSS indicates that the token distribution changes smoothly from layer $l-1$ to layer l, whereas a smaller CLSS indicates a more abrupt fluctuation in internal knowledge.

Figure 4(a) illustrates how the token distribution of the hallucinated token "people" changes across layers. Figure 4(b) shows that the JSD rises sharply around layer 28, suggesting a substantial disruption in the knowledge trajectory. In such cases, token preferences become unstable and may drift toward hallucinated outputs. By contrast, smoother transitions in intermediate layers are more strongly associated with faithful generations.

3.4 Cross-Token Stability

We further study how instability manifests across consecutive decoding steps. Neighboring tokens often refer to the same visual entity. For instance, in the phrase "yellow bus", both "yellow" and "bus" describe the same object. This suggests that faithful decoding should maintain similar visual focus across adjacent tokens. To quantify this phenomenon, we define Visual Focus Distraction

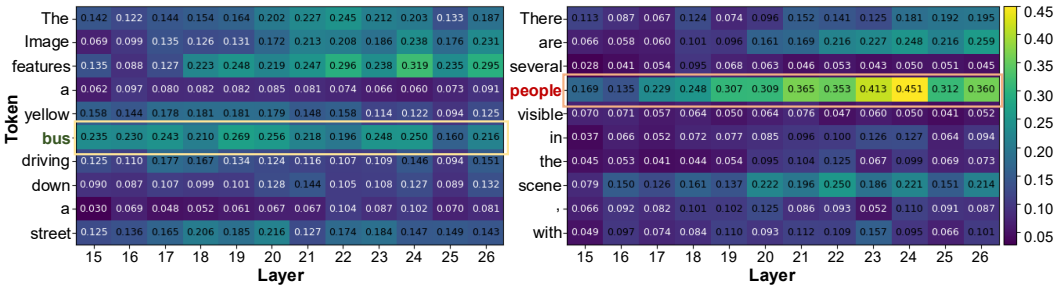

Figure 5: Layer-wise VFD between adjacent tokens. Grounded "bus" and hallucinated "people" are highlighted in green and red, respectively; the hallucinated token shows substantially greater visual focus distraction.

(VFD) as the average JSD between head-wise visual attention maps at two consecutive steps:

$$\mathrm{VFD}_t^l = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \mathrm{JSD}(A_t^{(l,h)}, A_{t-1}^{(l,h)}). \tag{7}$$

Figure 5 shows that the grounded token "bus" maintains consistently low VFD across layers, indicating stable visual grounding over successive steps. Its neighboring tokens, such as "yellow" and "driving", also exhibit low VFD, consistent with their shared referent. In contrast, the hallucinated token "people" exhibits substantially larger VFD, especially in deeper layers, indicating that the model's visual focus deviates from that of the previous step. This suggests that cross-step distraction is another useful indicator of hallucination risk. We therefore define the Cross-Token Stability Score (CTSS) as

$$\mathrm{CTSS}_t^l = 1 - \mathrm{VFD}_t^l. \tag{8}$$

3.5 Statistical Validation of Internal Instability

The preceding examples qualitatively illustrate the relationship between internal instability and hallucination. To determine whether this pattern generalizes beyond individual cases, we conduct an instance-level analysis using the CHAIR annotation pipeline. Generated object tokens are categorized as either true or hallucinated according to whether the corresponding objects are supported by the input image.

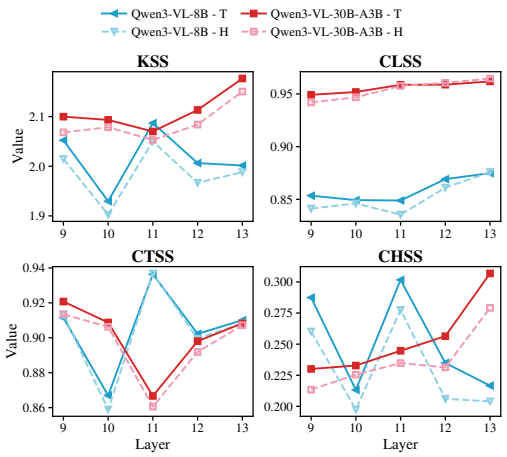

Figure 6: Layer-wise KSS, CLSS, CTSS, and CHSS values for true (T) and hallucinated (H) object tokens. Higher values indicate more stable internal knowledge.

As shown in Figure 6, hallucinated object tokens generally obtain lower scores than true object tokens, indicating weaker internal consistency and, consequently, greater internal instability. This pattern is consistently observed for KSS, CTSS, and CHSS across both models and nearly all examined layers. CLSS exhibits a similar tendency, although the separation is smaller and the true and hallucinated curves partially overlap at several layers. The magnitude of the difference also varies with model size and network depth, suggesting that different instability components may become more informative at different stages of generation. Overall, these results provide population-level evidence that hallucinated object generation is systematically associated with greater internal instability. Nevertheless, the association remains statistical rather than deterministic: a low consistency score alone does not necessarily imply that a generated object is hallucinated.

3.6 Stability-Aware Knowledge Enhanced Decoding (SAKED)

We define the overall Knowledge Stability Score (KSS) by aggregating the three stability measures:

$$\mathrm{KSS}_t^l = \lambda_1 \mathrm{CHSS}_t^l + \lambda_2 \mathrm{CLSS}_t^l + \lambda_3 \mathrm{CTSS}_t^l, \tag{9}$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$ control the relative contributions of cross-head, cross-layer, and cross-token stability, respectively.

Based on KSS, we select the most stability-aware layer l^+ and the most stability-agnostic layer l^-:

$$l^+ = \arg\max_{l \in L_c} \mathrm{KSS}_t^l, \qquad l^- = \arg\min_{l \in L_c} \mathrm{KSS}_t^l. \tag{10}$$

We then construct a stability-aware contrastive correction from their logits:

$$\boldsymbol{z}_t^{\mathrm{cont}} = (1+\alpha)\,\mathrm{logit}(\boldsymbol{h}_t^{l^+}) - \alpha\,\mathrm{logit}(\boldsymbol{h}_t^{l^-}), \tag{11}$$

Table 1: Hallucination evaluation results on CHAIR across five models. Ci and Cs represent CHAIRi and CHAIRs, respectively. (OPERA cannot be applied to Qwen2.5VL and InternVL3 due to architectural constraints)

Model	LLaVA1.5		InstructBLIP		MiniGPT-4		InternVL3		Qwen2.5VL		Average	
Method	Ci ↓	Cs ↓	Ci ↓	Cs ↓	Ci ↓	Cs ↓	Ci ↓	Cs ↓	Ci ↓	Cs ↓	Ci ↓	Cs ↓
Greedy	14.1	49.2	**13.3**	48.4	9.9	45.0	7.8	31.6	9.8	40.0	11.0	42.8
Beam	14.2	51.8	14.3	54.6	13.9	48.8	7.6	31.6	8.4	36.0	11.7	44.6
Dola	19.3	61.2	26.2	73.2	13.8	47.8	10.4	41.6	11.0	41.2	16.1	53.0
OPERA	12.8	44.6	14.2	46.4	12.8	44.6	-	-	-	-	-	-
Deco	12.5	46.4	13.5	46.4	9.4	32.2	8.0	33.2	9.4	38.8	10.6	39.4
SAVER	12.8	40.4	13.6	47.8	10.0	32.8	8.3	32.8	8.4	35.6	10.6	37.9
SAKED	**11.6**	**39.6**	14.0	**42.2**	**7.7**	**30.8**	**7.5**	**30.4**	**7.6**	**32.2**	**9.7**	**35.0**

where α controls the contrast strength.

Finally, we revise the final-layer logits $\boldsymbol{z}_t$ using this contrastive correction:

$$\hat{\boldsymbol{z}}_t = \boldsymbol{z}_t + \beta \boldsymbol{z}_t^{\text{cont}}, \qquad \text{s.t. } \hat{\boldsymbol{z}}_t(v) = -\infty,\ \forall v \notin \mathcal{V}_{\text{candidate}}, \tag{12}$$

and decode the next token by

$$\hat{\boldsymbol{y}}_t = \text{argmax}_{v \in \mathcal{V}_{\text{candidate}}} \hat{\boldsymbol{z}}_t(v). \tag{13}$$

Here, β is the revision coefficient, and $\mathcal{V}_{\text{candidate}}$ denotes the top-q (default: $q = 20$) candidate tokens under the original distribution. Restricting the revision to $\mathcal{V}_{\text{candidate}}$ prevents implausible low-probability tokens from being spuriously promoted while preserving the grounded token with high probability. The pipeline of SAKED is summarized in Algorithm 1.

Algorithm 1 Stability-Aware Knowledge Enhanced Decoding (SAKED)

Require: Inputs $\{\boldsymbol{x}^I, \boldsymbol{x}^P\}$; layers L_c; weights α, β, λ_1, λ_2, λ_3.
for every decoding step t **do**
  for $l \in L_c$ **do**
    Compute CHSS_t^l using Eq. 3.
    Compute CLSS_t^l using Eq. 6.
    Compute CTSS_t^l using Eq. 8.
    $\text{KSS}_t^l = \lambda_1 \text{CHSS}_t^l + \lambda_2 \text{CLSS}_t^l + \lambda_3 \text{CTSS}_t^l$.
  end for
  $l^+ = \arg\max_{l \in L_c} \text{KSS}_t^l$.
  $l^- = \arg\min_{l \in L_c} \text{KSS}_t^l$.
  $\boldsymbol{z}_t^{\text{cont}} = (1+\alpha)\text{logit}(\boldsymbol{h}_t^{l^+}) - \alpha \text{logit}(\boldsymbol{h}_t^{l^-})$.
  $\hat{\boldsymbol{y}}_t = \arg\max_{v \in \mathcal{V}_{\text{candidate}}} \big(\boldsymbol{z}_t(v) + \beta \boldsymbol{z}_t^{\text{cont}}(v)\big)$.
end for

4 Experiments

4.1 Implementation Details

We evaluate SAKED on five representative LVLMs: LLaVA-1.5 Liu et al. (2023), InstructBLIP Dai et al. (2023), MiniGPT-4 Zhu et al. (2023), InternVL3 Zhu et al. (2025), and Qwen2.5-VL Bai et al. (2025). We use deterministic decoding and compare against greedy decoding, beam search, Dola Chuang et al. (2023), OPERA Huang et al. (2024), Deco Wang et al. (2024a), and SAVER Li et al. (2026). Experiments cover CHAIR Rohrbach et al. (2018), POPE Li et al. (2023b), MME Fu et al. (2024), and AMBER Wang et al. (2023). Unless otherwise specified, models use 7B-scale language backbones, and experiments run on one NVIDIA GPU with 48 GB of memory. Full decoding settings, baseline descriptions, and benchmark protocols are provided in the appendix subsection *Implementation Details*; further experiments are reported in the appendix section *Additional Experimental Results*.

Table 2: Average POPE F1 scores across the popular, random, and adversarial settings.

Model	Greedy	Beam	Dola	Deco	SAVER	**SAKED**
LLaVA1.5	0.8220	0.8490	0.8320	0.8320	0.8489	**0.8536**
InstructBLIP	0.8000	0.8440	0.8340	0.8492	0.8486	**0.8551**
MiniGPT-4	0.5850	0.7030	0.7280	**0.7714**	0.7680	0.7639
InternVL3	**0.9074**	0.9021	0.8657	0.9069	0.9063	0.9027
Qwen2.5VL	0.8136	0.8138	0.7489	0.8136	0.8137	**0.8721**
Average	0.7856	0.8224	0.8017	0.8346	0.8371	**0.8495**

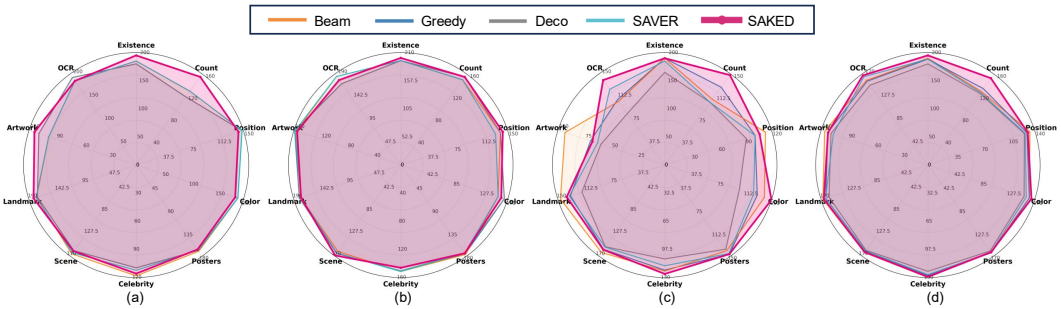

Figure 7: Detailed MME evaluation results on 10 subsets on (a) Qwen2.5VL, (b) InternVL3, (c) LLaVA1.5, and (d) Average results on the three models.

4.2 Experiments on CHAIR

Table 1 reports hallucination performance on the CHAIR benchmark across five representative LVLMs, measured by CHAIRi (Ci) and CHAIRs (Cs), where lower values indicate fewer hallucinated objects and sentences. Greedy and Beam decoding exhibit consistently high hallucination rates across models. Language-based method Dola ignores the visual information, thereby significantly degrading performance on various models. Hallucination-aware decoding methods (e.g., OPERA, Deco, SAVER) enhance models' visual perception and show partial improvements on some models. However, they fail to generalize consistently across architectures. Moreover, OPERA cannot be applied to all models due to architectural constraints. In contrast, our method SAKED achieves the lowest hallucination rates on almost all evaluated models, yielding the best average performance with Ci = 9.7 and Cs = 35.0, outperforming all baselines. These results provide evidence that internal knowledge instability is broadly associated with hallucination across different LVLM architectures. By dynamically leveraging the most stable internal knowledge without additional training, SAKED achieves state-of-the-art hallucination mitigation performance and can be flexibly adapted to different architectures.

4.3 Experiments on POPE

Table 2 summarizes hallucination mitigation performance on POPE using the average F1 score across the random, popular, and adversarial settings. Prior methods such as Deco and SAVER improve visual perception, yet they remain inferior to our method. SAKED achieves the highest average performance across the evaluated backbones, with an average F1 score of 0.8495. This consistency across diverse LVLM architectures suggests that hallucinations are not solely due to limited visual perception, but are also closely related to unstable knowledge aggregation.

4.4 Experiments on MME

Detailed results on a wide range of visual tasks are shown in Figure 7, where our method consistently achieves outstanding evaluation scores on Qwen2.5VL, InternVL3 and LLaVA1.5 across different tasks. Figure 8 in appendix shows the overall scores across these tasks. SAKED consistently achieves the highest performance on all evaluated models in both perception-oriented and cognition-oriented tasks, yielding an average score of 1544, surpassing both the vanilla decoding strategy and previous hallucination mitigation methods.

Table 3: Comparison of decoding efficiency on LLaVA1.5.

Method	Lat./Tok. (ms)↓	Lat./Sample (s)↓	Tok./s↑	Peak Mem. (MB)↓	Ci↓	Cs↓
Greedy	19.8±1.5	2.13±0.16	51.0±3.2	13914	14.1	49.2
Beam	28.9±0.8	3.34±0.09	35.0±0.8	15639	14.2	51.8
DoLa	23.5±0.1	2.57±0.12	43.6±1.6	14054	19.3	61.2
OPERA	68.9±0.2	7.69±0.03	16.8±0.1	23702	12.8	44.6
DECO	23.3±0.8	2.73±0.09	43.1±1.4	14305	12.5	46.4
SAVER	29.3±1.3	3.52±0.16	34.4±1.4	16417	12.8	40.4
SAKED	23.9±0.2	2.69±0.07	41.9±0.3	15358	**11.6**	**39.6**

Table 4: Ablation results on CHAIR, reported as Ci/Cs (↓).

Model	w/o CHSS	w/o CLSS	w/o CTSS	w/o SACD	**SAKED**
LLaVA1.5	13.5/44.8	12.1/41.0	12.4/42.4	12.0/43.6	**11.6/39.6**
InstructBLIP	14.1/42.4	17.5/48.4	15.1/47.6	15.9/44.8	**14.0/42.2**
InternVL3	8.0/33.0	8.3/32.4	7.8/**30.2**	8.1/31.0	**7.5**/30.4
Qwen2.5VL	**7.0**/32.4	7.5/32.4	8.0/34.2	7.5/34.8	7.6/**32.2**
Average	10.7/38.1	11.4/38.6	10.8/38.6	10.9/38.6	**10.2/36.1**

4.5 Inference Efficiency Analysis

Table 3 compares SAKED with baselines in latency, throughput, memory usage, and hallucination mitigation. Results are averaged over five runs on one NVIDIA Pro6000 GPU. SAKED achieves the best hallucination performance with modest overhead over DECO: latency increases from 23.3 to 23.9 ms/token, while memory usage rises only marginally. Meanwhile, Ci decreases from 12.5 to 11.6 and Cs from 46.4 to 39.6. This efficiency stems from using intermediate decoding stability for dynamic token selection, avoiding repeated forward passes or extra decoding. On Qwen2.5-VL-32B, SAKED similarly increases latency only from 59.2 to 61.1 ms/token and peak memory from 64.2 to 65.7 GB, while maintaining comparable throughput. Overall, SAKED provides a favorable efficiency and effectiveness trade-off and scales well to larger backbones.

4.6 Ablation Studies

Table 4 and Table 9 in Appendix A.6 evaluate the effectiveness and efficiency of SAKED's components. Removing CHSS, CLSS, CTSS, or SACD consistently degrades CHAIR performance, whereas the complete method achieves the best average results (Ci=10.0 and Cs=35.7). Single-score variants reduce computation, with CTSS-only providing the highest throughput and lowest per-token latency and memory usage. The full method nevertheless maintains a favorable trade-off, requiring 23.9 ms per token and achieving 41.9 tokens/s; further efficiency results are reported in Table 3. We also tune the KSS weights with $\lambda_1 = 1.0$. Table 10 in appendix shows that $\lambda_2 = \lambda_3 = 1.0$ performs best, yielding Ci=10.2 and Cs=36.1 on average. Ablations of α, β, and the candidate layer set L_c are provided in the appendix subsection *Additional Ablation Studies*.

5 Conclusion

In this paper, we introduce Stability-Aware Knowledge Enhanced Decoding (SAKED), a training-free, plug-and-play method for mitigating hallucinations in LVLMs. We show that hallucinations are closely linked to knowledge instability across attention heads, layers, and adjacent tokens. Based on these findings, we propose the Knowledge Stability Score (KSS), which integrates CHSS, CLSS, and CTSS to identify stable and unstable layers and contrast their logits during decoding. Experiments across diverse tasks, benchmarks, model scales, and architectures demonstrate that SAKED consistently improves faithfulness without external information or fine-tuning. Future work will explore more efficient and flexible ways to incorporate these stability signals into generation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with data analysis. We did not use generative AI tools to design or provide feedback on the research methodology or experiments, propose or refine hypotheses. Other required disclosure tasks were not applicable to this work.

Additionally, we used generative AI tools to assist with code checking and verification, and to edit the manuscript for formatting, language clarity, and readability. All AI-assisted analyses, code-related suggestions, and manuscript edits were reviewed by the authors. AI-assisted manuscript edits were limited to presentation, formatting, and language improvement and did not alter the scientific conclusions of the work.

We take responsibility for the final content of this work, including all text, claims, analyses, code, and other artifacts produced with the aid of generative AI. (This section is **required** and does not count toward the page limit.)

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

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*, 2025.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *arXiv preprint arXiv:2403.01548*, 2024a.

Xinlong Chen, Yuanxing Zhang, Qiang Liu, Junfei Wu, Fuzheng Zhang, and Tieniu Tan. Mixture of decoding: An attention-inspired adaptive decoding strategy to mitigate hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8525–8542, 2025.

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*, 2024b.

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.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. *arXiv preprint arXiv:2505.19678*, 2025a.

Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. Enhancing vision-language model reliability with uncertainty-guided dropout decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

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

Stephen M Fleming. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75(1):241–268, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL https://arxiv.org/abs/2306.13394.

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

Google DeepMind. Introducing gemini 3. https://blog.google/products/gemini/gemini-3/, 2025. Accessed: 2025-12.

Alice Heiman, Xiaoman Zhang, Emma Chen, Sung Eun Kim, and Pranav Rajpurkar. Factchexcker: Mitigating measurement hallucinations in chest x-ray report generation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30787–30796, 2025.

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.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27036–27046, 2024.

Liqiang Jing and Xinya Du. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*, 2024.

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

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 2025.

Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C Kot. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *IEEE Transactions on Dependable and Secure Computing*, 2025.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

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*, 2023b.

Zhaoxu Li, Chenqi Kong, Yi Yu, Qiangqiang Wu, Xinghao Jiang, Ngai-Man Cheung, Bihan Wen, Alex Kot, and Xudong Jiang. Saver: Mitigating hallucinations in large vision-language models via style-aware visual early revision. *The Association for the Advancement of Artificial Intelligence (AAAI)*, 2026.

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. *arXiv preprint arXiv:2502.03628*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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*, pp. 125–140. Springer, 2024.

Qitan Lv, Jie Wang, Hanzhu Chen, Bin Li, Yongdong Zhang, and Feng Wu. Coarse-to-fine highlighting: Reducing knowledge hallucination in large language models. *arXiv preprint arXiv:2410.15116*, 2024.

OpenAI. Gpt-4 technical report. `https://openai.com/research/gpt-4`, 2023. Accessed: 2024-10-18.

OpenAI. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`, 2025. Accessed: 2025-12-20.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. Steer llm latents for hallucination detection. *arXiv preprint arXiv:2503.01917*, 2025a.

Woohyeon Park, Woojin Kim, Jaeik Kim, and Jaeyoung Do. Second: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding. *arXiv preprint arXiv:2506.08391*, 2025b.

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

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. *International Conference on Learning Representations*, 2025.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*, 2024a.

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.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pp. 32–45. Springer, 2024b.

Sudong Wang, Yunjian Zhang, Yao Zhu, Enci Liu, Jianing Li, Yanwei Liu, and Xiangyang Ji. Shift: Smoothing hallucinations by information flow tuning for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3639–3649, 2025a.

Weixing Wang, Zifeng Ding, Jindong Gu, Rui Cao, Christoph Meinel, Gerard de Melo, and Haojin Yang. Image tokens matter: Mitigating hallucination in discrete tokenizer-based large vision-language models via latent editing. *arXiv preprint arXiv:2505.21547*, 2025b.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in large vision-language models. *arXiv preprint arXiv:2406.16449*, 2024.

Song Xia, Meiwen Ding, Chenqi Kong, Wenhan Yang, and Xudong Jiang. Provable robustness in multimodal large language models via feature space smoothing. *arXiv preprint arXiv:2601.16200*, 2026.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10610–10620, 2025.

Nick Yeung and Christopher Summerfield. Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1310–1321, 2012.

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.

Le Yu, Kaishen Wang, Jianlong Xiong, Yue Cao, Lei Zhang, and Zhang Yi Tao He. Hallurnn: Mitigating hallucinations via recurrent cross-layer reasoning in large vision-language models. *arXiv preprint arXiv:2506.17587*, 2025.

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, 2024.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023.

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

Zhen Zhao, Jingqun Tang, Binghong Wu, Chunhui Lin, Shu Wei, Hao Liu, Xin Tan, Zhizhong Zhang, Can Huang, and Yuan Xie. Harmonizing visual text comprehension and generation. *arXiv preprint arXiv:2407.16364*, 2024b.

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. *arXiv preprint arXiv:2310.00754*, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

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. *arXiv preprint arXiv:2410.03577*, 2024.