ORBIT: Overcoming Hallucination Risks via Bi-Subspace Interaction and Traction

Zhehan Kan[1,2†], Yanlin Liu[1†], Xiaochen Yang[3], Hongyang Yu[2], Qingmin Liao[1,2], Wenming Yang[1,2✉]

[1]Tsinghua University [2]Pengcheng Lab [3]University of Glasgow

Abstract. Despite the remarkable capabilities of Large Vision-Language Models (LVLMs), they remain highly susceptible to hallucinations. Existing inference-time mitigation strategies largely rely on visual attention as a diagnostic proxy. However, we demonstrate that this approach is insufficient in certain cases due to the "anchoring trap," where parameterized linguistic priors override correct visual grounding. In this paper, we pioneer a representation geometry perspective, identifying hidden states as a complementary diagnostic signal beyond visual attention. We model the evolution of hidden states across layers and generation steps as constrained sequences governed by a static visual subspace and a dynamic linguistic subspace within a high-dimensional Euclidean space. We reveal that hallucinations inherently stem from abrupt structural deviations, specifically, the spontaneous departure of latent representations from the visual subspace and their subsequent drift into the linguistic domain. To quantify this phenomenon, we propose the Geo-Semantic Turbulence Index (GSTI) to adaptively monitor visual misalignment and semantic volatility. Building upon GSTI, we introduce Orthogonal Residual-Based Intervention (ORBIT), an architecture-agnostic, plug-and-play mechanism that recalibrates divergent representation sequences by employing a closed-loop intervention encompassing residual extraction, visual re-anchoring, and norm-preserving injection of suppressed visual evidence. Extensive experiments across diverse LVLM architectures demonstrate that ORBIT significantly mitigates hallucinations and enhances general vision-centric performance with exceptional computational efficiency.

Keywords: LVLMs · Hallucination · Representation Geometry

1 Introduction

LVLMs have emerged as a scalable paradigm toward Artificial General Intelligence (AGI), leveraging visual encoders and modality connectors to ground language models in visual contexts [2, 6, 23, 28]. Despite their remarkable multimodal understanding capabilities, LVLMs suffer from a pervasive vulnerability to hallucinations, generating textual responses that contradict the provided visual evidence. This unfaithfulness introduces critical bottlenecks for deployment

† Equal contribution. [1,2] China; [3] United Kingdom. ✉ Corresponding author.

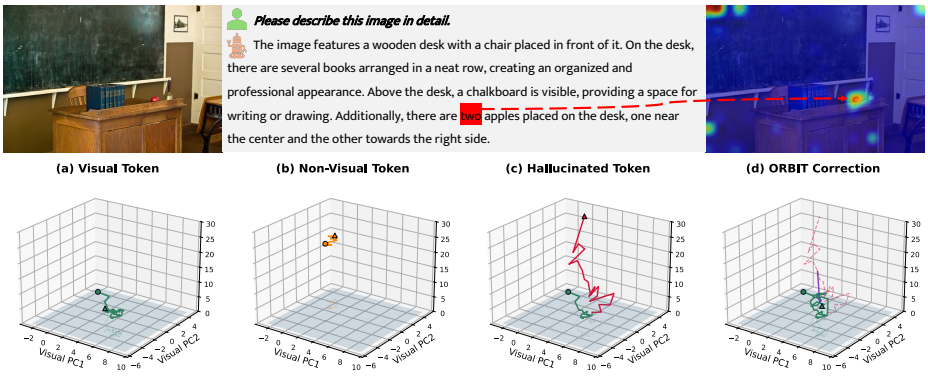

Fig. 1: Illustration of the "anchoring trap" and the representation geometry of LVLM hallucinations. Top: Despite the visual attention correctly localizing on the single apple, the model hallucinates the token "two". **Bottom: (a)-(b)** At a fixed generation step t, faithful visual and non-visual hidden states evolve stably across layers within their respective **geometric subspaces** (**visual subspace** and **linguistic subspace**) of the **Euclidean latent space**. **(c)** The hallucinated token exhibits an abrupt **structural deviation**, where its cross-layer sequence spontaneously detaches from the visual subspace (defined by the principal components of visual features) and **drifts** into the linguistic domain. **(d)** Our proposed ORBIT successfully recalibrates the divergent **representation sequence** back toward the visual subspace, restoring grounding and mitigating the hallucination.

in high-stakes environments, such as healthcare and autonomous systems [4, 8], where misaligned outputs can precipitate severe consequences. To mitigate these issues, early efforts predominantly adopted training-centric paradigms, such as curating high-fidelity instruction-tuning datasets or integrating negative feedback via preference optimization [12, 16–19, 22]. However, these retraining strategies incur prohibitive computational costs. More critically, they risk degrading the models' general capabilities, as distribution shifts in the curated data can interfere with previously internalized linguistic priors. Driven by these limitations, recent research has pivoted toward inference-time interventions, which offer a lightweight and effective alternative to enforce multimodal faithfulness without requiring any parameter updates [13, 21].

Existing inference-time interventions have evolved from black-box logit manipulations to interpretable attention-based mechanisms, positing that hallucinations arise from misaligned visual attention. However, as illustrated in Fig. 1, hallucinations persist even when visual attention is precisely localized on semantically relevant regions. This observation demonstrates that attention serves as an unreliable monitoring proxy. Sole reliance on attention maps induces an "anchoring trap," where the generated content is overridden by parameterized linguistic priors despite correct visual grounding. Consequently, **uncovering the fundamental mechanisms underlying hallucinations in LVLMs and developing principled solutions remain critical open challenges.**

In this work, we investigate hidden states as a complementary diagnostic signal beyond visual attention. From the perspective of representation geometry, we formulate the evolution of hidden states during LVLM inference as sequences within a high-dimensional Euclidean space. Specifically, the cross-layer progression of each token is characterized by its relationship with two functional geometric structures: a static visual subspace and a dynamic linguistic subspace. As illustrated in Fig. 1, hallucinations stem from abrupt structural deviations in these latent representation sequences. The hidden states spontaneously depart from the visual subspace, which provides stable perceptual grounding, and drift into the linguistic subspace, driven by statistical token co-occurrences. To quantify this phenomenon, we introduce the **Geo-Semantic Turbulence Index (GSTI)**, which adaptively monitors the cross-layer evolution of each token by measuring Visual Deviation (the extent of displacement from the visual subspace) and Semantic Volatility (the magnitude of semantic shifts). Building upon this metric, we propose **Orthogonal Residual-Based Intervention (ORBIT)** by employing a closed-loop intervention encompassing residual extraction, visual re-anchoring, and iso-norm injection to recalibrate divergent representation sequences and mitigate hallucinations. ORBIT is architecture-agnostic and functions as a plug-and-play module for various LVLMs. Extensive evaluations integrating ORBIT into Qwen2.5-VL [2] and LLaVA-1.5 [23] demonstrate significant performance gains over SOTA inference-time interventions across hallucination mitigation and numerous vision-centric benchmarks, while maintaining high computational efficiency.

In summary, our main contributions are detailed as follows: (1) We pioneer a fundamental investigation into LVLM hallucinations through the lens of representation geometry. We demonstrate that existing attention-based monitoring is unreliable due to the "anchoring trap" and reveal that hallucinations inherently stem from abrupt structural shifts in hidden states, specifically, the spontaneous departure of representation sequences from stable visual subspaces and their subsequent drift into linguistic subspaces. (2) We introduce the Geo-Semantic Turbulence Index (GSTI) to dynamically monitor cross-layer token evolution by quantifying visual deviation and semantic volatility. Building upon this metric, we propose Orthogonal Residual-Based Intervention (ORBIT), an architecture-agnostic, plug-and-play mechanism that recalibrates divergent sequences during inference by employing a closed-loop intervention encompassing residual extraction, visual re-anchoring, and iso-norm injection of suppressed visual evidence. (3) We extensively validate ORBIT by integrating it into leading architectures, including Qwen2.5-VL and LLaVA-1.5. Comprehensive evaluations across diverse benchmarks demonstrate that our method consistently outperforms SOTA inference-time interventions in both hallucination mitigation and general vision-centric tasks, achieving significant performance gains while maintaining exceptional computational efficiency.

2 Related Work

2.1 Hallucination in LVLMs

Although LVLMs have demonstrated impressive capabilities in multimodal understanding and generation, they remain susceptible to hallucination, defined as the generation of textual content inconsistent with the provided visual evidence [12, 22]. Early mitigation efforts primarily focused on *data- and training-centric* approaches. These methods involve the construction of higher-fidelity instruction-tuning corpora [23] and the incorporation of stronger negative feedback signals through preference optimization or RLHF-style alignment for multimodal outputs [26]. While effective in specific scenarios, retraining-based strategies are often computationally expensive to implement. Furthermore, they may compromise general capabilities due to distribution shifts in the curated data or interference with previously acquired linguistic behaviors. These limitations have motivated a growing body of research on *inference-time* interventions designed to enhance faithfulness without the need for parameter updates.

2.2 Inference-time Interventions

Logit-based Interventions. Recent inference-time approaches for hallucination mitigation can be broadly categorized into *logit-based* and *attention-based* interventions. One prevalent category of methods operates directly on next-token distributions. Representative techniques, such as VCD [21] and DoLa [7], suppress hallucinations through contrastive or layer-wise decoding. These methods typically involve the construction of an auxiliary distribution designed to capture hallucinatory tendencies, followed by a reweighting of the logits to favor more faithful continuations. Specific strategies include the injection of perturbations into visual inputs to induce uncertainty and the contrasting of the final layer with earlier layers that may encode more conservative predictions [27]. Despite empirical improvements, these strategies often rely on decoding heuristics, the mechanisms of which are difficult to interpret. Consequently, they may be brittle under distribution shifts and can degrade the quality of general generation when the contrastive signal is inaccurate.

Attention-based Interventions. Subsequent research focuses on internal attention patterns, operating under the hypothesis that hallucination stems from weak visual grounding. Methods such as VAR [20] and OPERA [13] utilize cross-attention maps as proxies for grounding and intervene during the decoding process to amplify attention weights on semantically relevant visual tokens. Compared to techniques that operate purely at the logit level, attention-based methods provide more direct control over the image-conditioned computation of the model. Additionally, they enhance interpretability by identifying which regions or visual tokens influence the output. However, visual attention serves as an unreliable diagnostic proxy due to the "anchoring trap." Departing from attention-centric paradigms, we model hidden states as representation sequences

within a high-dimensional Euclidean space. Recent latent-space projection or steering methods also intervene on internal representations to suppress undesirable generation patterns [24]. Nevertheless, our goal is not to introduce a generic projection or fixed steering direction. We reformulate hallucinations as abrupt structural deviations, where these sequences spontaneously depart from the visual subspace and drift into the dynamic linguistic domain. To quantify this divergence, we introduce the GSTI for adaptive monitoring of representation evolution across layers and generation steps. Guided by GSTI, our ORBIT mechanism employs orthogonal residuals and iso-norm injection to dynamically recalibrate divergent representation sequences, thereby restoring visual grounding and effectively mitigating hallucinations while preserving linguistic fluency.

3 Methodology

3.1 Motivation

As illustrated in Fig. 1, a critical observation of our work is that hallucinations persist even when visual attention mechanisms ostensibly align with semantically relevant regions. This observation underscores that while attention serves as a necessary condition for visual grounding, it is insufficient to guarantee faithfulness. Sole reliance on attention maps risks a "grounding trap," wherein the model attends to correct visual stimuli yet generates content overridden by parametric language priors. To transcend this limitation, we investigate hallucination through the lens of representation geometry. Specifically, for each generation step t, we formulate the cross-layer evolution of hidden states $\mathbf{h}_t^{(l)}$ as discrete sequences within a high-dimensional Euclidean space. We observe that hallucinatory errors manifest as abrupt structural deviations in these latent sequences: hidden states spontaneously decouple from a static visual subspace, characterized by stable perceptual anchoring, and drift into a dynamic linguistic subspace governed by statistical token co-occurrences. This geometric shift signifies the critical juncture where internal priors supersede external visual evidence. Motivated by this perspective, our framework is designed to detect these cross-layer representation anomalies and perform targeted inference-time rectification.

3.2 Geometric Formulation: The Bi-Subspace Hypothesis

Let an LVLM be characterized by a visual encoder $\mathcal{E}_v$, a projection module P, and an LLM backbone $\mathcal{D}$. The model maps a multimodal input (I, x) to a sequence of tokens $y_{1:t}$. We model the latent space of the LLM as a high-dimensional Euclidean space $\mathcal{H} = \mathbb{R}^d$. At any generation time step t and layer l, the hidden state $\mathbf{h}_t^{(l)} \in \mathcal{H}$ represents the semantic integration of visual perception and linguistic reasoning. We propose the **Bi-Subspace Hypothesis**, postulating that in a faithful generation process, the representation sequence of $\mathbf{h}_t^{(l)}$ evolves under the dual constraints of two distinct geometric structures: **The Static Visual Subspace ($\mathcal{S}_{vis}^{(l)}$):** A global, time-invariant but layer-specific

linear subspace spanned by the objective visual features. It serves as the geometric anchor, providing a structural constraint that ensures the representation remains aligned with the provided visual evidence. **The Dynamic Linguistic Subspace ($\mathcal{S}_{lang}^{(t,l)}$):** A local, time-variant structure representing the direction of autoregressive evolution. It encodes the syntactic momentum and linguistic priors of the current context.

3.3 Subspace Construction via Spectral Decomposition

To operationalize the Bi-Subspace Hypothesis, we employ spectral analysis to construct discrete approximations of $\mathcal{S}_{vis}^{(l)}$ and $\mathcal{S}_{lang}^{(t,l)}$. This process translates high-dimensional activations into quantifiable coordinate systems via singular value decomposition (SVD). To ensure inference efficiency, we strictly decouple the computational overhead into the *Prefill* and *Decoding* phases.

The Static Visual Subspace (Prefill Phase). While the input image I is static, its representation undergoes a hierarchical transformation through the network depth. Consequently, $\mathcal{S}_{vis}^{(l)}$ must be layer-specific to align with the semantic abstraction level of $\mathbf{h}_t^{(l)}$.

During the prefill phase, we isolate the activations corresponding to the visual tokens. Let $\mathbf{H}^{(l)} \in \mathbb{R}^{L_{seq} \times d}$ denote the full sequence activation at layer l. We extract the visual feature matrix $\mathbf{V}^{(l)}$ by slicing $\mathbf{H}^{(l)}$ at indices $\mathcal{I}_{img}$:

$$\mathbf{V}^{(l)} = \mathbf{H}^{(l)}[\mathcal{I}_{img}] \in \mathbb{R}^{N_v \times d}, \tag{1}$$

where N_v denotes the number of visual patches. To identify the principal directions of visual variation, we perform randomized SVD [11] on the centered feature matrix:

$$(\mathbf{V}^{(l)} - \boldsymbol{\mu}_{vis}^{(l)})^\top \approx \mathbf{U}^{(l)} \boldsymbol{\Sigma}^{(l)} (\mathbf{Q}^{(l)})^\top. \tag{2}$$

Here, the columns of $\mathbf{U}^{(l)} \in \mathbb{R}^{d \times d}$ constitute the orthonormal basis of the visual subspace. To suppress high-frequency noise, we select the minimal k such that the cumulative explained variance exceeds a threshold η (set to 0.95):

$$k = \min \left\{ r \;\middle|\; \frac{\sum_{i=1}^{r} (\sigma_i^{(l)})^2}{\sum_{j=1}^{d} (\sigma_j^{(l)})^2} \geq \eta \right\}. \tag{3}$$

The projection operator onto the static visual subspace is then defined as:

$$\mathbf{P}_{vis}^{(l)} = \mathbf{U}_{1:k}^{(l)} (\mathbf{U}_{1:k}^{(l)})^\top. \tag{4}$$

Geometrically, $\mathbf{P}_{vis}^{(l)}$ functions as a spectral filter, preserving components aligned with the visual ground truth while nullifying orthogonal perturbations associated with hallucination. Since $\mathbf{V}^{(l)}$ depends solely on the input image, $\mathbf{P}_{vis}^{(l)}$ is precomputed and cached, incurring zero latency during decoding.

The Dynamic Linguistic Subspace (Decoding Phase). In contrast to the visual anchor, linguistic context is non-stationary. We model the local linguistic subspace $\mathcal{S}_{lang}^{(t,l)}$ as the span of the representation sequence at generation step t.

We construct a sliding context window $\mathbf{H}_{ctx}^{(t,l)}$ containing the hidden states of the K most recent tokens. To explicitly isolate linguistic context from visual evidence, this window strictly excludes visual tokens and system prompts:

$$\mathbf{H}_{ctx}^{(t,l)} = \left[\mathbf{h}_{t-K}^{(l)}, \dots, \mathbf{h}_{t-1}^{(l)}\right]^\top \in \mathbb{R}^{K\times d}. \tag{5}$$

Here, K (empirically $K = 5$) defines the locality of the syntactic approximation. We compute the principal directions of this local sequence via a thin SVD on the centered window:

$$(\mathbf{H}_{ctx}^{(t,l)} - \boldsymbol{\mu}_{ctx}^{(t,l)})^\top \approx \mathbf{U}_{lang}^{(t,l)} \boldsymbol{\Sigma}_{lang} (\mathbf{Q}_{lang})^\top. \tag{6}$$

The dynamic projection operator $\mathbf{P}_{lang}^{(t,l)}$ is constructed using the top k_{lang} left singular vectors $\mathbf{U}_{lang,1:k_{lang}}^{(t,l)}$:

$$\mathbf{P}_{lang}^{(t,l)} = \mathbf{U}_{lang,1:k_{lang}}^{(t,l)} (\mathbf{U}_{lang,1:k_{lang}}^{(t,l)})^\top. \tag{7}$$

This operator captures the instantaneous direction of the linguistic flow, allowing us to mathematically disentangle syntax-driven momentum from novel semantic information.

3.4 Detecting Semantic Turbulence

Building upon the Bi-Subspace formulation, we analyze the projection residuals and evolution path of generated sequences relative to the static visual subspace. As visualized in Fig. 1, we categorize tokens into three regimes: (a) *Visual Tokens*, which necessitate strict alignment with visual representations; (b) *Non-Visual Tokens*, which rely primarily on contextual priors; and (c) *Hallucinated Tokens*, which require visual grounding but erroneously succumb to parametric overconfidence. Empirically, we observe that visual tokens adhere closely to the static visual subspace $\mathcal{S}_{vis}^{(l)}$. Non-visual tokens maintain a stable evolution at a higher vertical displacement relative to $\mathcal{S}_{vis}^{(l)}$ (i.e., along dimensions not spanned by the primary visual components). Crucially, hallucinated tokens initially track the visual subspace but undergo a sudden deviation in later layers, characterized by an abrupt detachment followed by erratic, high-frequency directional shifts. We formalize hallucination as a geometric shift: a critical point where the representation sequence decouples from $\mathcal{S}_{vis}^{(l)}$ and drifts into a high-volatility region. To quantify this, we introduce the **Geo-Semantic Turbulence Index (GSTI)**, designed to monitor the stability of the latent sequence by monitoring the concurrence of (1) **visual deviation** and (2) **semantic volatility**.

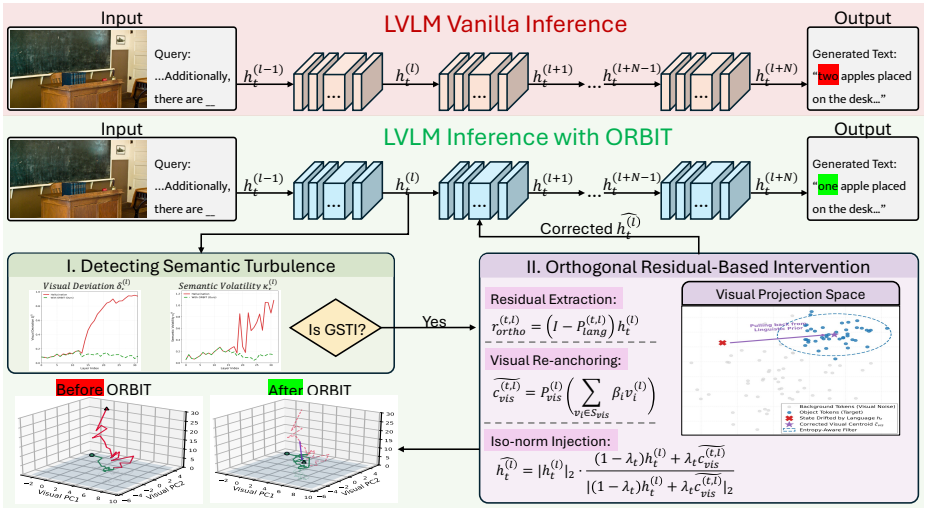

Fig. 2: Overview of ORBIT. While vanilla LVLM inference (top) is prone to hallucinations, ORBIT (bottom) intervenes dynamically during generation. **I. Detecting Semantic Turbulence:** The GSTI continuously monitors the cross-layer evolution of hidden states to identify abrupt structural deviations. **II. Orthogonal Residual-Based Intervention:** Upon detection, ORBIT suppresses parameterized linguistic priors via orthogonal projection, extracts faithful visual evidence, and injects the recalibrated state under an iso-norm constraint.

Visual Deviation ($\boldsymbol{\delta_t^{(l)}}$) We quantify the extent to which the hidden state $\mathbf{h}_t^{(l)}$ diverges into the orthogonal complement of the visual subspace. Utilizing the projection operator $\mathbf{P}_{vis}^{(l)}$ (Eq. 4), we define the orthogonal energy ratio:

$$\delta_t^{(l)} = \frac{\|(\mathbf{I} - \mathbf{P}_{vis}^{(l)})\mathbf{h}_t^{(l)}\|_2}{\|\mathbf{h}_t^{(l)}\|_2}, \tag{8}$$

where $\mathbf{I}$ denotes the identity matrix. Geometrically, $\delta_t^{(l)} \in [0, 1]$ measures the fraction of the activation energy orthogonal to the visual subspace. As $\delta_t^{(l)} \to 1$, the generation effectively decouples from visual constraints, becoming driven purely by the internal priors of the LLM.

Semantic Volatility ($\boldsymbol{\kappa_t^{(l)}}$) Hallucinations are frequently preceded by erratic shifts in the latent space, as the model vacillates between competing tokens without coherent visual grounding. We quantify this instability as the directional inconsistency of the evolution path. Let $\mathbf{v}_t^{(l)} = \mathbf{h}_t^{(l)} - \mathbf{h}_t^{(l-1)}$ denote the layer-wise velocity vector at step t. We define the local volatility $\kappa_t^{(l)}$ via the cosine distance between consecutive layer-wise updates:

$$\kappa_t^{(l)} = 1 - \frac{\langle \mathbf{v}_t^{(l)}, \mathbf{v}_t^{(l-1)} \rangle}{\|\mathbf{v}_t^{(l)}\|_2 \|\mathbf{v}_t^{(l-1)}\|_2}. \tag{9}$$

A low volatility ($\kappa \approx 0$) implies a consistent, directed evolution, whereas high volatility indicates stochastic, fragmented shifts in the representation.

Geo-Semantic Turbulence Index (GSTI) ($\Omega_t^{(l)}$) To capture the interaction between decoupling and instability, we define GSTI as a gated product governed by the statistical significance of the visual drift:

$$\Omega_t^{(l)} = \sigma\left(1.702 \cdot \frac{\delta_t^{(l)} - \mu_\delta}{\sigma_\delta}\right) \cdot \left(1 + \kappa_t^{(l)}\right), \tag{10}$$

where $\sigma(\cdot)$ denotes the standard Sigmoid function. The gating term approximates a Cumulative Distribution Function (CDF) of the Z-score of the visual drift, using the scaling factor 1.702 to align the Sigmoid curve with the normal CDF [3]. This provides an adaptive anchor mechanism:

- **Anchored State:** When the visual drift $\delta_t^{(l)}$ is low, the Sigmoid gate suppresses the volatility factor. This ensures that directional shifts are attenuated when the model remains grounded in the visual subspace.
- **Decoupled State:** As the drift significantly exceeds the statistical mean ($\delta_t^{(l)} \gg \mu_\delta$), the gate saturates. In this regime, the metric becomes fully sensitive to $\kappa_t^{(l)}$, interpreting directional inconsistency as a strong signal of hallucination.

Online Anomaly Detection We employ an online statistical trigger by maintaining a running estimate of the mean $\mu_\Omega^{(l)}$ and variance $(\sigma_\Omega^{(l)})^2$ of the GSTI stream. An intervention is triggered solely when the turbulence constitutes a statistical anomaly:

$$\mathbb{I}_{\text{trigger}}^{(t,l)} = \mathbb{1}\left[\Omega_t^{(l)} > \mu_\Omega^{(l)} + \alpha \cdot \sigma_\Omega^{(l)}\right], \tag{11}$$

where $\alpha = 3.0$ follows the 3σ rule for outlier detection. This adaptive mechanism ensures that our method intervenes only during significant deviations from the model's localized stability baseline, minimizing false positives in fluent but linguistically complex generations.

3.5 Orthogonal Residual-Based Intervention

Once the GSTI triggers an alarm ($\mathbb{I}_{\text{trigger}}^{(t,l)} = 1$, see Sec. 3.4), Orthogonal Residual-Based Intervention (ORBIT) initiates a targeted geometric rectification. The objective is to realign the divergent representation by re-injecting suppressed visual evidence. However, a naive additive injection of visual features risks perturbing the feature distribution required for coherent autoregression, potentially degrading linguistic fluency. To resolve this, we propose a three-step intervention protocol: *Residual extraction, Visual re-anchoring, and Iso-norm injection.*

Step 1: Residual Extraction. First, we isolate the visual signal component that the language model has failed to predict or has actively suppressed. We posit that a hallucinated hidden state is dominated by a strong linguistic prior that overshadows the visual signal. Using the linguistic projection operator $\mathbf{P}_{lang}^{(t,l)}$, we project the current hidden state $\mathbf{h}_t^{(l)}$ onto the orthogonal complement of the local linguistic subspace:

$$\mathbf{r}_{ortho}^{(t,l)} = (\mathbf{I} - \mathbf{P}_{lang}^{(t,l)})\mathbf{h}_t^{(l)}. \tag{12}$$

The vector $\mathbf{r}_{ortho}^{(t,l)}$ represents the residual information in the current state that is geometrically perpendicular to the local linguistic context. Ideally, this residual encodes the suppressed visual details; however, in practice, it also contains high-frequency noise and irrelevant artifacts.

Step 2: Visual Re-anchoring. To filter out noise and construct a valid correction vector, we verify $\mathbf{r}_{ortho}^{(t,l)}$ against the ground-truth visual subspace $\mathcal{S}_{vis}^{(l)}$. We treat $\mathbf{r}_{ortho}^{(t,l)}$ as a query vector to retrieve support from the cached visual features $\mathbf{V}^{(l)}$. We compute the relevance scores via dot-product attention:

$$s_i = \frac{\langle \mathbf{r}_{ortho}^{(t,l)}, \mathbf{v}_i^{(l)} \rangle}{\sqrt{d}}, \quad \boldsymbol{\beta} = \mathrm{Softmax}(\mathbf{s}), \tag{13}$$

where $\mathbf{v}_i^{(l)}$ denotes the rows of $\mathbf{V}^{(l)}$. To ensure we only inject high-confidence features, we apply Nucleus Sampling (Top-p) on $\boldsymbol{\beta}$, selecting the minimal set of indices $S_{vis}^{(t,l)}$ such that the aggregated probability mass $\sum_{i \in S_{vis}^{(t,l)}} \beta_i \geq \rho$ (set to 0.9). The Corrected Visual Centroid is then computed as the weighted sum of this support set, strictly re-anchored onto the visual subspace:

$$\tilde{\mathbf{c}}_{vis}^{(t,l)} = \mathbf{P}_{vis}^{(l)} \left(\sum_{i \in S_{vis}^{(t,l)}} \beta_i \mathbf{v}_i^{(l)} \right). \tag{14}$$

This step guarantees that the correction vector $\tilde{\mathbf{c}}_{vis}^{(t,l)}$ is confined within the valid visual subspace $\mathcal{S}_{vis}^{(l)}$, effectively denoising the residual.

Step 3: Iso-norm Injection. Finally, we fuse the corrected visual centroid $\tilde{\mathbf{c}}_{vis}^{(t,l)}$ back into the hidden state. In deep Transformer architectures, the Euclidean norm $\|\mathbf{h}_t^{(l)}\|_2$ significantly influences LayerNorm statistics and downstream attention saturation. To mitigate representational collapse, we impose an Iso-norm Constraint, ensuring the intervention modifies the direction (semantics) rather than the energy of the state:

$$\hat{\mathbf{h}}_t^{(l)} = \|\mathbf{h}_t^{(l)}\|_2 \cdot \frac{(1-\lambda_t)\mathbf{h}_t^{(l)} + \lambda_t \tilde{\mathbf{c}}_{vis}^{(t,l)}}{\|(1-\lambda_t)\mathbf{h}_t^{(l)} + \lambda_t \tilde{\mathbf{c}}_{vis}^{(t,l)}\|_2}, \tag{15}$$

where λ_t denotes the injection strength. Rather than employing a fixed value, λ_t is adaptively coupled to the severity of the detected turbulence $\Omega_t^{(l)}$ via a calibrated hyperbolic tangent function:

$$\lambda_t = \left(1 - \frac{H(\beta)}{\log |S_{vis}^{(t,l)}|}\right) \cdot \tanh(\Omega_t^{(l)}), \tag{16}$$

where $|S_{vis}^{(t,l)}|$ denotes the cardinality of the selected visual token set from Nucleus Sampling, and $H(\beta)$ represents the entropy of the attention distribution over this set. This formulation incorporates an adaptive modulation: the confidence term $(1 - H(\beta)/\log |S_{vis}^{(t,l)}|)$ attenuates the injection when the visual correspondence is diffuse. Consequently, ORBIT applies a potent corrective force only during severe subspace detachment, while remaining subtle during minor fluctuations to preserve linguistic fluency.

4 Experiments

4.1 Experimental Setups

Evaluated LVLMs. To assess the generalizability of our framework, we integrate ORBIT into a diverse array of advanced LVLMs, spanning different parameter scales and architectural designs. Specifically, our evaluation suite includes the LLaVA series (v1.5-7B/13B) [23] and the Qwen series (Qwen2.5-VL-7B/32B) [2], representing the current SOTA in open-source multimodal understanding.

Evaluation Benchmarks. We conduct a comprehensive evaluation across three distinct categories of benchmarks to ensure a holistic assessment: (1) **Hallucination Evaluation:** We employ POPE [23] and CHAIR [5] to strictly measure object existence hallucinations and description faithfulness. (2) **General Visual Question Answering:** We use MME [10] and MMBench [25] to evaluate broad multimodal reasoning capabilities. (3) **Fine-Grained Perception:** We utilize CV-Bench [33] and CLEVR [15] to test precise spatial reasoning and counting abilities, which are particularly sensitive to visual grounding. To further assess robustness under distribution shifts, we additionally evaluate ORBIT on MME-RealWorld [32] for high-resolution real-world visual understanding and GEOBench-VLM [9] for satellite-imagery-oriented geometric reasoning.

Implementation Details. In this work, we adhere to the default query format for the input data used in both LLaVA-1.5 and Qwen2.5-VL. Additionally, we set the linguistic context window size to $K = 5$ (Eq. 5). Crucially, ORBIT operates as a plug-and-play inference-time method, requiring neither parameter tuning nor retraining of the backbone models.

Table 1: Results on hallucination benchmarks.

Method	LLaVA-v1.5-7B					Qwen2.5-VL-7B				
	POPE		CHAIR		Len	POPE		CHAIR		Len
	F1↑	Acc↑	C_S ↓	C_I ↓		F1↑	Acc↑	C_S ↓	C_I ↓	
Baseline	85.3	84.1	51.2	15.3	102.0	88.4	89.0	26.5	7.2	98.4
DoLa [7]	80.4	83.0	56.8	15.1	97.4	86.8	87.5	29.1	8.5	95.2
VCD [21]	85.2	85.1	50.8	14.8	102.1	88.5	89.0	26.8	7.4	100.1
OPERA [13]	84.1	85.3	47.2	14.5	95.5	88.5	88.9	26.2	7.3	94.8
HALC [5]	84.0	83.9	50.4	12.3	97.0	87.5	87.8	28.7	8.0	96.5
RITUAL [29]	85.1	84.2	45.4	13.1	99.1	88.6	89.1	26.4	7.2	98.2
SID [14]	85.7	85.9	44.0	12.1	99.5	88.9	89.3	26.1	7.8	98.8
VAR [31]	86.1	86.4	44.2	14.6	103.2	89.1	89.5	25.5	7.5	101.5
DeGF [30]	85.2	84.6	46.5	12.9	97.3	88.4	88.9	26.8	8.1	96.9
AGLA [1]	84.7	85.4	43.2	14.0	98.6	87.7	88.1	28.9	8.3	97.5
ORBIT (Ours)	**87.5**	**88.0**	**38.1**	**8.5**	100.3	**90.3**	**90.8**	**18.4**	**5.3**	98.5

4.2 Quantitative Results

OOD Robustness. Table 3 shows that ORBIT also improves robustness beyond the in-distribution hallucination benchmarks. On both MME-RealWorld and GEOBench-VLM, ORBIT achieves the best accuracy across LLaVA-v1.5-7B and Qwen2.5-VL-7B. Compared with SID and VAR, which provide marginal gains or degrade under some OOD settings, ORBIT shows more consistent improvements, suggesting that image-conditioned hidden-state rectification is effective under realistic visual shifts.

Performance on Hallucination Benchmarks. Table 1 summarizes the comparative performance on the POPE and CHAIR benchmarks. ORBIT consistently outperforms existing SOTA methods across all model architectures by significant margins. A critical concern in hallucination mitigation is the "refusal trap," where models achieve artificially high scores by generating overly brief or evasive responses. Notably, ORBIT maintains an average generation length (100.3 tokens) comparable to the baseline, confirming that the performance gains stem from substantive semantic correction rather than defensive brevity.

Furthermore, the efficacy of ORBIT is particularly pronounced on stronger baselines such as Qwen2.5-VL-7B. High-performance models pose a rigorous challenge for intervention methods; prior approaches like AGLA often exhibit diminishing returns or performance degradation (e.g., AGLA causes a drop in F1 from 88.4 to 87.7) because their heuristic thresholds conflict with the robust priors of advanced models. In contrast, ORBIT successfully achieves visual-semantic decoupling even within this highly optimized latent space, boosting the POPE F1 score to 90.3 and reducing the $CHAIR_S$ metric by a relative 30.5% (26.5 $\rightarrow$ 18.4). This validates our Bi-Subspace Hypothesis: even sophisticated models are susceptible to geometric misalignment, suggesting that subspace-aware rectification is a more robust paradigm than heuristic logit suppression.

Table 2: Results on general VQA benchmarks.

Backbone	Method	Visual Hallucination			General VQA		Vision-Centric	
		POPE ↑	C_S↓	C_I↓	MME ↑	MMB ↑	CV-B ↑	CLEVR ↑
LLaVA-v1.5-7B	Baseline	85.3	51.2	15.3	1495.5	64.2	56.6	43.6
	AGLA [1]	84.7	43.2	14.0	1491.0	64.0	56.5	43.4
	SID [14]	85.7	44.0	12.1	1505.5	64.9	56.9	44.2
	VAR [31]	86.1	44.2	14.6	1512.2	65.1	56.8	44.6
	ORBIT (Ours)	**87.5**	**38.1**	**8.5**	**1525.8**	**66.2**	**58.0**	**46.4**
LLaVA-v1.5-13B	Baseline	85.8	24.6	8.2	1501.2	67.7	58.2	68.0
	AGLA [1]	85.5	23.9	8.4	1498.5	67.5	57.9	67.5
	SID [14]	86.1	21.1	7.0	1510.0	67.9	58.6	68.3
	VAR [31]	86.3	20.5	7.5	1520.5	68.0	58.6	68.1
	ORBIT (Ours)	**88.2**	**16.4**	**6.1**	**1542.5**	**69.1**	**60.2**	**70.1**
Qwen2.5-VL-7B	Baseline	88.4	26.5	7.2	1973.3	82.1	73.5	74.3
	AGLA [1]	87.7	28.9	8.3	1965.0	82.0	73.1	74.6
	SID [14]	88.9	26.1	7.8	1970.5	82.2	73.4	74.3
	VAR [31]	89.1	25.5	7.5	1968.0	82.4	73.6	74.2
	ORBIT (Ours)	**90.3**	**18.4**	**5.3**	**1988.2**	**83.8**	**74.6**	**75.5**
Qwen2.5-VL-32B	Baseline	90.5	18.5	4.5	2046.5	83.5	77.0	78.2
	AGLA [1]	90.3	18.9	5.4	2035.0	83.3	76.0	77.7
	SID [14]	90.4	18.3	4.3	2045.8	83.6	77.1	78.0
	VAR [31]	90.7	18.0	4.3	2058.2	83.7	77.3	78.6
	ORBIT (Ours)	**91.2**	**15.2**	**4.1**	**2080.4**	**84.5**	**78.1**	**79.4**

Table 3: OOD evaluation on high-resolution real-world images and satellite imagery.

Method	MME-RealWorld Acc. ↑		GEOBench-VLM Acc. ↑	
	LLaVA-v1.5-7B	Qwen2.5-VL-7B	LLaVA-v1.5-7B	Qwen2.5-VL-7B
Baseline	26.4	56.9	25.1	33.2
SID	26.7	57.1	23.7	32.5
VAR	27.0	57.3	25.2	32.8
ORBIT	**28.2**	**58.2**	**26.1**	**34.0**

Table 4: latency as the number of visual tokens increases.

N_v	Latency ↓	Relative Cost ↓
576	0.051s	×1.0
1024	0.096s	×1.9
2304	0.235s	×4.6
4096	0.421s	×8.3

Impact on General Capabilities. A common in intervention methods is the "over-correction" effect, where suppressing hallucinations inadvertently impairs the model's general reasoning or grounding capabilities. Table 2 extends our evaluation to general VQA (MME, MMBench) and vision-centric tasks (CV-Bench, CLEVR). ORBIT consistently improves performance across these benchmarks, indicating that our method preserves the linguistic fluency and reasoning backbone of the LLM. On tasks demanding precise spatial localization and counting (CLEVR) or fine-grained perception (CV-Bench), ORBIT achieves new SOTA results. This confirms that ORBIT does not merely suppress the model's output but actively re-injects lost visual details, effectively transforming the intervention into a constructive feature enhancement mechanism.

4.3 Ablation Studies and Efficiency Analysis

Ablation Studies. ORBIT is purposefully designed with minimal heuristic tuning. Parameters such as $\eta = 0.95$, $\rho = 0.9$, and $\alpha = 3.0$ strictly adhere to

Table 5: Ablation study on context window K.

K	Accuracy				Latency ↓
	POPE↑	CHAIR↓	MME↑	CV-B↑	
3	87.2	38.8	1523.4	57.6	5.60
5	87.5	38.1	1525.8	58.0	6.27
10	87.6	38.1	1526.4	57.9	8.92
20	87.8	37.9	1529.9	58.4	13.22

Table 6: Efficiency comparison.

Method	CHAIR ↓	Latency ↓
Baseline	51.2	4.21s (×1.0)
VCD	50.8	8.88s (×2.1)
OPERA	47.2	29.2s (×6.9)
HALC	50.4	26.6s (×6.3)
SID	44.0	6.40s (×1.5)
VAR	44.2	7.11s (×1.7)
DEGF	46.5	18.2s (×4.3)
AGLA	43.2	11.2s (×2.6)
Ours	38.1	7.73s (×1.8)

Table 7: Component ablations on LLaVA-v1.5-7B.

Variant	δ	κ	Res.	Re-anchor	Iso-norm	POPE F1 ↑	CHAIR_S ↓	CHAIR_I ↓
Baseline	–	–	–	–	–	85.3	51.2	15.3
w/o GSTI	–	–	✓	✓	✓	86.9	41.3	10.1
w/o Visual Deviation	–	✓	✓	✓	✓	87.0	40.7	9.8
w/o Semantic Volatility	✓	–	✓	✓	✓	87.2	39.8	9.4
w/o Residual Extraction	✓	✓	–	✓	✓	86.0	45.8	12.4
w/o Re-anchoring	✓	✓	✓	–	✓	86.3	44.2	11.6
w/o Iso-norm Injection	✓	✓	✓	✓	–	86.6	42.7	10.8
Full ORBIT	✓	✓	✓	✓	✓	**87.5**	**38.1**	**8.5**

established statistical conventions (the PCA energy threshold, nucleus sampling prior, and the 3σ outlier rule, respectively), ensuring robust generalization across diverse architectures. Therefore, our ablation study strategically isolates K, the context window size. As the core structural variable, K defines the receptive field of the linguistic subspace and governs the critical trade-off between syntactic approximation fidelity and computational overhead. Table 5 details this ablation. The empirical results demonstrate that performance remains fundamentally robust across various configurations of K. We adopt $K = 5$ as the optimal default to maximize accuracy while curbing inference latency. Further ablation experiments are provided in supplemental materials.

Component Ablations. Table 7 isolates the contribution of GSTI and ORBIT. Applying the correction without GSTI still improves over the baseline but underperforms full ORBIT, indicating the benefit of adaptive anomaly triggering. Removing either visual deviation or semantic volatility hurts performance, suggesting that the two signals provide complementary evidence. Among the correction modules, residual extraction is particularly important, while visual re-anchoring and iso-norm injection further reduce hallucination by filtering noisy residuals and preserving hidden-state scale.

Computational Efficiency. ORBIT decouples its overhead between the prefill and decoding phases. The visual subspace is computed once during prefill and cached. For a rank-k approximation with N_v visual tokens and hidden dimension

d, the SVD costs $O(N_v dk + (N_v + d)k^2)$. During decoding, projection onto the cached visual basis costs only $O(dk)$ per triggered layer, while the linguistic SVD uses a small context window and costs $O(dK^2)$. Therefore, the decoding-time overhead is lightweight and incurred only for triggered layers.

Table 4 reports the prefill SVD latency under different visual-token counts. Although the cost increases with N_v, this operation is performed only once per input and the resulting visual basis is cached before decoding. Thus, the decoding stage only requires lightweight projections on triggered layers.

4.4 Qualitative Results

Fig. 3 illustrates a representative case demonstrating the efficacy of ORBIT. Crucially, the baseline model exhibits hallucinatory behavior even when its visual attention is accurately localized on the corresponding target region. By integrating ORBIT, such underlying semantic deviations are successfully detected and dynamically rectified, thereby yielding factually accurate outputs.

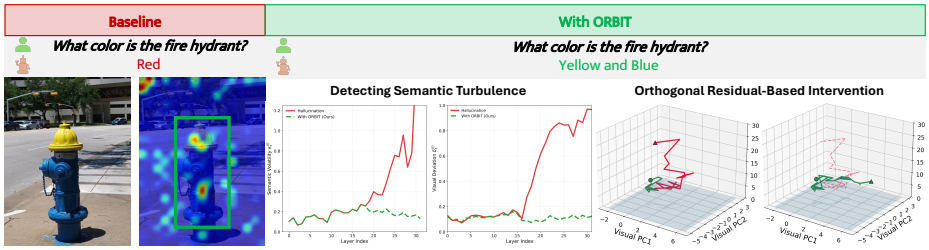

Fig. 3: Qualitative case study of ORBIT intervention. Left: The baseline model hallucinates the color "Red" due to strong linguistic priors, despite accurate visual attention on the yellow and blue fire hydrant. **Middle & Right:** ORBIT successfully detects the abrupt topological shift via spikes in GSTI metrics (semantic turbulence) and dynamically recalibrates the divergent latent trajectory. This intervention restores visual grounding, enabling the model to correctly output "Yellow and Blue."

5 Conclusion

In this work, we present a fundamental rethinking of hallucination mitigation in LVLMs through the lens of representation geometry. Moving beyond the unreliable proxy of visual attention and the pervasive "anchoring trap," we establish that hallucinations are intrinsically driven by abrupt geometric deviations, where latent trajectories spontaneously decouple from stable visual subspaces and drift into parameterized linguistic priors. To counter this, we introduce the Geo-Semantic Turbulence Index (GSTI) for precise trajectory monitoring, alongside Orthogonal Residual-Based Intervention (ORBIT), a plug-and-play mechanism that dynamically recalibrates divergent states by re-anchoring suppressed visual evidence. Comprehensive evaluations confirm that ORBIT significantly suppresses hallucinations and bolsters general multimodal capabilities with exceptional computational efficiency.

Acknowledgments

This work was partly supported by the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen(No.KJZD20231023094700001) and the Shenzhen-Tsinghua Special Project for Fundamental & Frontier Research in Artificial Intelligence(No.AI2026018).

References

1. An, W., Tian, F., Leng, S., Nie, J., Lin, H., Wang, Q., Chen, P., Zhang, X., Lu, S.: Mitigating object hallucinations in large vision-language models with assembly of global and local attention (2025), https://arxiv.org/abs/2406.12718
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023), https://arxiv.org/abs/2308.12966
3. Camilli, G.: Teacher's corner: origin of the scaling constant d= 1.7 in item response theory. Journal of educational and behavioral statistics **19**(3), 293–295 (1994)
4. Chen, J., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., Yu, G., et al.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280 (2024)
5. Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., Zhou, J.: Halc: Object hallucination reduction via adaptive focal-contrast decoding (2024), https://arxiv.org/abs/2403.00425
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks (2024), https://arxiv.org/abs/2312.14238
7. Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J., He, P.: Dola: Decoding by contrasting layers improves factuality in large language models (2024), https://arxiv.org/abs/2309.03883
8. Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.D., et al.: A survey on multimodal large language models for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 958–979 (2024)
9. Danish, M., Munir, M.A., Shah, S.R.A., Kuckreja, K., Khan, F.S., Fraccaro, P., Lacoste, A., Khan, S.: Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7132–7142 (2025)
10. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2025), https://arxiv.org/abs/2306.13394
11. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review **53**(2), 217–288 (2011)
12. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (Jan 2025). https://doi.org/10.1145/3703155, http://dx.doi.org/10.1145/3703155

13. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation (2024), https://arxiv.org/abs/2311.17911
14. Huo, F., Xu, W., Zhang, Z., Wang, H., Chen, Z., Zhao, P.: Self-introspective decoding: Alleviating hallucinations for large vision-language models (2025), https://arxiv.org/abs/2408.02032
15. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning (2016), https://arxiv.org/abs/1612.06890
16. Kan, Z., Jiang, X., Liu, Y., Yang, X., Wei, Z., Liu, S., Zhu, Y., Liao, Q., Yang, W., Li, X., Liu, Y., Jiang, D., Sun, X.: Uvu: Improving multimodal understanding via vision-language unified autoregressive paradigm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26230–26239 (June 2026)
17. Kan, Z., Li, X., Liu, Y., Yang, X., Jiang, X., Liu, Y., Jiang, D., Sun, X., Liao, Q., Yang, W.: RAR: Reversing visual attention re-sinking for unlocking potential in multimodal large language models. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=xTbFeDhdRG
18. Kan, Z., Liu, Y., Yin, K., Jiang, X., Li, X., Cao, H., Liu, Y., Jiang, D., Sun, X., Liao, Q., Yang, W.: Taco: Think-answer consistency for optimized long-chain reasoning and efficient data learning via reinforcement learning in lvlms (2025), https://arxiv.org/abs/2505.20777
19. Kan, Z., Zhang, C., Liao, Z., Tian, Y., Yang, W., Xiao, J., Li, X., Jiang, D., Wang, Y., Liao, Q.: Catch: Complementary adaptive token-level contrastive decoding to mitigate hallucinations in lvlms (2024), https://arxiv.org/abs/2411.12713
20. Kang, S., Kim, J., Kim, J., Hwang, S.J.: See what you are told: Visual attention sink in large multimodal models (2025), https://arxiv.org/abs/2503.03321
21. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding (2023), https://arxiv.org/abs/2311.16922
22. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models (2024), https://arxiv.org/abs/2402.00253
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), https://arxiv.org/abs/2304.08485
24. Liu, S., Ye, H., Xing, L., Zou, J.: Reducing hallucinations in vision-language models via latent space steering (2024), https://arxiv.org/abs/2410.15778
25. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? (2024), https://arxiv.org/abs/2307.06281
26. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.Y., Wang, Y.X., Yang, Y., Keutzer, K., Darrell, T.: Aligning large multimodal models with factually augmented rlhf (2023), https://arxiv.org/abs/2309.14525
27. Wang, C., Chen, X., Zhang, N., Tian, B., Xu, H., Deng, S., Chen, H.: Mllm can see? dynamic correction decoding for hallucination mitigation (2025), https://arxiv.org/abs/2410.11779
28. Wei, Z., Li, Y., Kan, Z., Jiang, X., Long, Z., Liu, S., Shen, H., Liu, W., Tan, X., Lin, H., Zhu, Y., Li, Q., Yin, D., Cao, H., Gu, W., Li, X., Liu, Y., Jiang, D., Sun,

X., Wu, Y., Tang, M., Liu, S., Tang, L., Lin, H., Lu, J., Qin, J., Qiao, L., Qiao, R., Ke, B., He, J., Li, K., Li, Y., Shen, Y., Zhang, M., Chen, P., Yin, K., Liu, B., Wu, Y., Chen, H., Cai, Z., Li, X.: Youtu-vl: Unleashing visual potential via unified vision-language supervision (2026), https://arxiv.org/abs/2601.19798

29. Woo, S., Jang, J., Kim, D., Choi, Y., Kim, C.: Ritual: Random image transformations as a universal anti-hallucination lever in large vision language models (2024), https://arxiv.org/abs/2405.17821
30. Zhang, C., Wan, Z., Kan, Z., Ma, M.Q., Stepputtis, S., Ramanan, D., Salakhutdinov, R., Morency, L.P., Sycara, K., Xie, Y.: Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models (2025), https://arxiv.org/abs/2502.06130
31. Zhang, X., Quan, Y., Gu, C., Shen, C., Yuan, X., Yan, S., Cheng, H., Wu, K., Ye, J.: Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms (2024), https://arxiv.org/abs/2411.09968
32. Zhang, Y., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., et al.: Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In: International Conference on Learning Representations. vol. 2025, pp. 89655–89701 (2025)
33. Zhu, N., Dong, Y., Wang, T., Li, X., Deng, S., Wang, Y., Hong, Z., Geng, T., Niu, G., Huang, H., Yao, X., Jiao, S.: Cvbench: Benchmarking cross-video synergies for complex multimodal reasoning (2026), https://arxiv.org/abs/2508.19542