ROT: Rotating Hidden States towards Contextual Vectors for Hallucination Mitigation in LVLMs

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) frequently suffer from object hallucination. Existing training-free interventions primarily manipulate attention weights, which indirectly affect the deep semantics reaching the final predictive layers. In this work, we shift our focus to the hidden state vectors extracted after self-attention and residual addition. Empirical analysis reveals that hallucinated tokens do not simply over-rely on linguistic priors; instead, they exhibit an anomalous contextual deviation, showing significantly lower similarities to both textual and visual contexts in intermediate layers. Motivated by this, we propose ROT, a layer-specific, training-free framework. ROT dynamically detects semantic deviation in the middle layers and applies a norm-preserving rotation to steer the hidden states back toward the grounded factual plane spanned by the contexts. For subsequent layers, a representational smoothing mechanism is introduced to stabilize the calibrated trajectory. Extensive experiments on multiple benchmarks demonstrate that ROT consistently reduces hallucinations across various model architectures and scales, offering an efficient, geometry-driven solution for grounded generation.

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated exceptional capabilities in integrating visual perception with linguistic reasoning. However, these models are persistently plagued by object hallucination, generating descriptions of objects or attributes that are inconsistent with the visual input (Li et al., 2023; Zhou et al., 2023). This fundamental flaw poses severe reliability risks for real-world applications. While extensive efforts have been made to mitigate hallucinations via instruction tuning, such training-based paradigms are often computationally prohibitive. Consequently, training-free inference-time interventions have emerged as an efficient alternative to dynamically calibrate model outputs.

Recent inference-time interventions have made notable strides in this direction. For instance, some approaches adjust visual information propagation in the middle layers (Jiang et al., 2025), selectively drop visual tokens based on epistemic uncertainty (Fang et al., 2024), or leverage gradient-aware saliency to filter ungrounded tokens (Zhang et al., 2026). Despite their effectiveness, these methods primarily manipulate attention weights or passively filter input contexts. We argue that solely adjusting attention weights does not directly determine the actual semantic content reaching the final predictive layers. To effectively rectify hallucinations, we shift our focus to the hidden state vectors after self-attention and before the feed-forward network.

To understand the dynamics of these hidden states during hallucination, we extract the normalized attention matrix at a given layer k to aggregate the hidden states of the previous layer, yielding a textual context vector C_T and a visual context vector C_V. Under the conventional language prior assumption, one would expect hallucinated tokens to over-attend to text, thereby exhibiting a high cosine similarity with C_T. Strikingly, our empirical analysis reveals the exact opposite: the hidden states of hallucinated tokens h_k consistently exhibit abnormally low similarities with both C_T and C_V. This observation resonates with recent findings that hallucinations often begin where contextual memory or saliency drops (Zhang et al., 2026). It indicates that hallucinated tokens do not simply over-rely on textual priors; rather, their representations diverge from the grounded text-vision subspace into erroneous directions.

Recognizing this behavior, we propose **ROT, a training-free framework featuring layer-specific hidden state intervention**. As shown in Figure 1, we formulate a deviation indicator E, which is derived from the complement of the cosine sim-

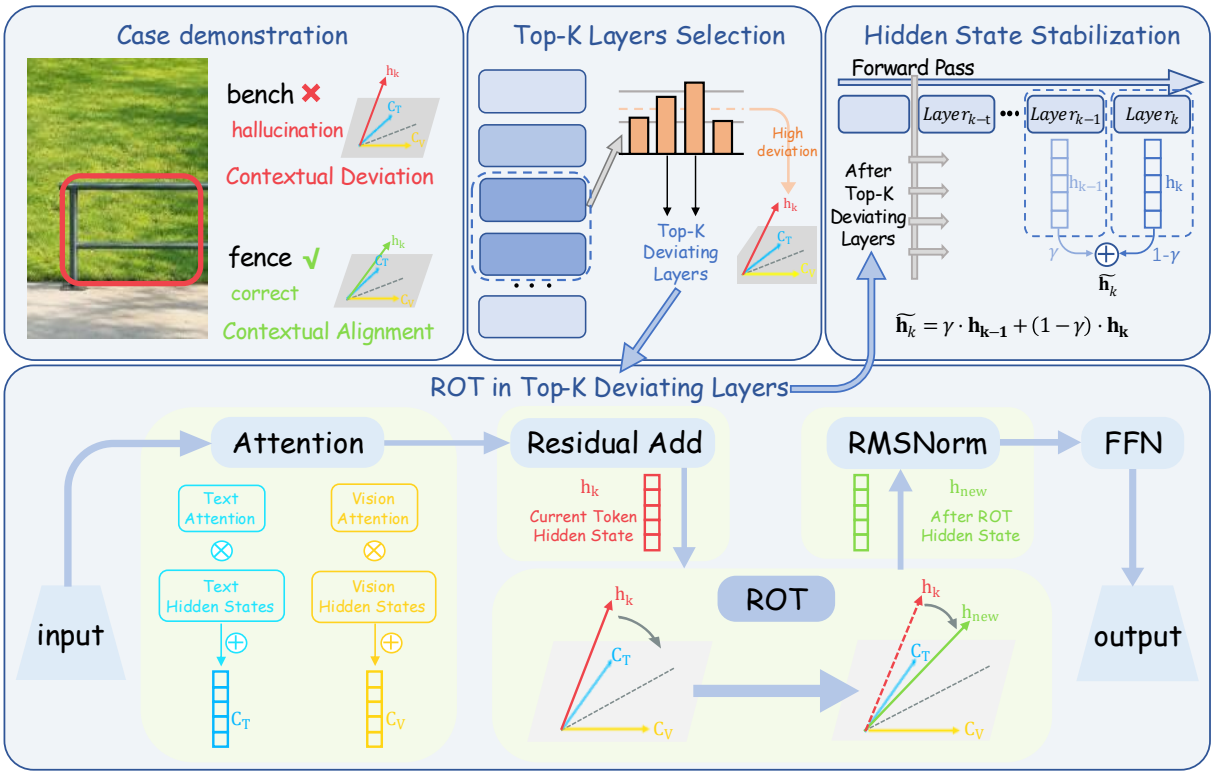

Figure 1: The overall framework of the proposed ROT method. The process consists of two stages: (1) Dynamic Hallucination Detection in the critical middle layers, which identifies contextual deviation based on the similarities between the hidden state and multimodal contexts; and (2) Layer-Specific Hidden State Mitigation, which applies a norm-preserving rotation in the intermediate layers to steer the deviating representation back toward the grounded factual plane, followed by a representational smoothing mechanism in subsequent layers to maintain stability during propagation.

ilarities between the current hidden state h_k and the two context vectors C_T and C_V. Our layer-wise analysis reveals that the value of E is predominantly high in the middle transformer layers, indicating a high propensity for hidden states to experience severe semantic deviation in these intermediate stages. Consequently, we specifically target the middle layers for hallucination detection. When E exceeds a predefined threshold, it signifies an erroneous semantic deviation, thereby triggering the mitigation mechanism.

For mitigation within these middle layers, instead of manipulating attention weights, we introduce a Hidden State Rotation strategy. To explicitly correct the deviation, we rotate the hidden state h_k toward the plane spanned by C_T and C_V. Crucially, this rotation is norm-preserving, ensuring that the erroneous semantic direction is steered back to a reliable context without compromising the token's inherent representation strength. Furthermore, for the Transformer layers subsequent to the middle layers, we introduce a smoothing mechanism to maintain stability. For each layer in this later stage, we update the representation by performing a weighted interpolation between the current hidden state and the output from the preceding layer using a controllable smoothing factor. This ensures that the corrected semantic trajectory smoothly propagates to the final predictive layers.

Extensive experiments on widely adopted benchmarks (e.g., CHAIR, POPE, and MME) demonstrate the superiority of our proposed method across various LVLM architectures. In summary, our main contributions are as follows:

- We propose a novel perspective for mitigating LVLM hallucinations by shifting the intervention focus from attention weights to hidden state vectors, addressing the semantic deviation directly in the hidden space.
- We introduce a robust hallucination detection mechanism based on hidden state dynamics. By observing that the semantic deviation indicator E is predominantly high in the middle layers, we utilize a similarity-based threshold to precisely detect hallucinations in these

susceptible intermediate stages.

- We design a layer-specific mitigation strategy. For deviating hidden states in the middle layers, we perform a norm-preserving rotation towards the plane formed by textual and visual context vectors; for subsequent layers, we employ a smoothing mechanism to maintain representational stability.

2 Related Work

2.1 Large Vision-Language Models

The rapid evolution of Large Language Models has catalyzed the development of Large Vision-Language Models, extending their profound reasoning capabilities into the multimodal domain. Models such as LLaVA (Liu et al., 2023) bridge the gap between modalities by aligning visual encoders with pre-trained language models through visual instruction tuning. Subsequent advancements, such as Qwen-VL (Bai et al., 2023), have further enhanced these capabilities by introducing robust fine-grained perception and supporting multi-image interactions at varying resolutions. Despite their remarkable success in diverse applications ranging from visual question answering to complex reasoning, these models predominantly suffer from a critical vulnerability: they frequently generate plausible yet ungrounded text, a phenomenon widely recognized as hallucination.

2.2 Object Hallucination in LVLMs

Object hallucination in LVLMs refers to the generation of objects, attributes, or relationships that are completely inconsistent with the provided visual input. Previous studies indicate that this flaw primarily stems from an over-reliance on statistical language priors, inadequate visual-linguistic alignment, and the inherent uncertainty during autoregressive decoding. To systematically quantify this issue, several authoritative benchmarks have been established. The CHAIR benchmark (Rohrbach et al., 2018) evaluates hallucinations in open-ended image captioning by analyzing the proportion of generated ungrounded objects at both the instance and sentence levels. More recently, the POPE benchmark (Li et al., 2023) was introduced to assess the phenomenon in a more controlled, discriminative manner by framing the evaluation as a binary classification task regarding object existence. These metrics have become the standard testbed for evaluating the reliability of LVLMs.

2.3 Training-Free Inference Intervention

To mitigate hallucinations, researchers have explored both training-based and inference-time strategies. Given the prohibitive computational cost of retraining or RLHF alignment for massive LVLMs, training-free inference interventions have emerged as a highly efficient and scalable alternative. Recent methods primarily fall into three paradigms. Visual Contrastive Decoding (Leng et al., 2024) mitigates the over-reliance on language priors by contrasting the output distributions between original and distorted visual inputs, though it fundamentally doubles the inference latency. Attention manipulation approaches, such as OPERA (Huang et al., 2024), introduce an over-trust penalty during beam search to trigger rollbacks when the model exhibits abnormal attention patterns. Feature injection methods, like MemVR (Zou et al., 2024), attempt to combat visual amnesia by performing memory-space visual retracing, which overwrites intermediate hidden states with key-value representations.

While these methods have demonstrated considerable success, they primarily rely on passive logits contrast, indirect attention manipulation, or rigid state substitution. In contrast, our proposed ROT framework pioneers a fundamentally different perspective. By discovering that hallucinated tokens suffer from a severe directional deviation in the hidden semantic space, ROT is the first to introduce an active geometric intervention. Rather than merely penalizing weights or replacing tokens, ROT directly rectifies the generation trajectory via a norm-preserving rotation. This targeted directional correction effectively re-anchors the hidden state to the factual multimodal subspace while preserving the original representational intensity, offering a highly efficient and structurally elegant solution to LVLM hallucinations.

3 Motivation and Empirical Analysis

3.1 Preliminaries of Hidden States in LVLMs

Large Vision-Language Models typically adopt a Transformer-based architecture to process interleaved multimodal inputs. Formally, given an input sequence consisting of both textual and visual tokens, the hidden representations are iteratively updated through multiple Transformer layers. In this work, we focus on the internal representation dynamics during the auto-regressive decoding phase.

At a specific layer k, let X^{k-1} denote the se-

quence of hidden states output from the preceding layer $k-1$. During the generation of the current token, the self-attention mechanism computes the attention scores between the current query and all preceding key representations. To analyze the information routed from different modalities, we extract the raw attention weights assigned to all past tokens and partition them into two distinct sets based on token modality: the textual attention weights and the visual attention weights. To accurately capture the aggregated semantic influence from each modality, we independently normalize these partitioned attention weights, yielding $\tilde{A}_{t,j}$ for the j-th text token and $\tilde{A}_{v,i}$ for the i-th vision token.

Using these normalized attention weights, we isolate the multimodal context by computing the weighted sum of the corresponding hidden states from X^{k-1}. Specifically, we construct a textual context vector C_T and a visual context vector C_V, which are formulated as follows:

$$C_T = \sum_{j \in T} \tilde{A}_{t,j} X_j^{k-1}, \quad C_V = \sum_{i \in V} \tilde{A}_{v,i} X_i^{k-1} \quad (1)$$

where T and V represent the sets of indices for the preceding textual and visual tokens, respectively. The resulting vectors, C_T and C_V, represent the isolated semantic subspaces grounded strictly in the textual and visual contexts.

Furthermore, we define h_k as the hidden state vector of the currently generated token at layer k, extracted after the self-attention mechanism and the subsequent residual addition, but prior to the layer normalization operation. Because h_k aggregates all contextual information and directly dictates the subsequent predictive trajectory through the Feed-Forward Network, investigating its geometric relationship with the modality-specific context vectors (C_T and C_V) provides a fundamental lens into the semantic dynamics of the model, particularly when hallucinations occur.

3.2 The Contextual Deviation of Hallucinated Tokens

To systematically investigate the representation dynamics of hallucination, we conduct an extensive empirical study. We utilize the widely adopted LLaVA-1.5-7B model and decode 3,000 evaluation instances from the standard CHAIR benchmark. Following the official CHAIR evaluation protocol, we meticulously categorize the generated tokens into two distinct groups: normal tokens and hallucinated tokens. For each token at a given layer k, we compute its cosine similarity with the textual context vector ($S_T = \cos(h_k, C_T)$) and the visual context vector ($S_V = \cos(h_k, C_V)$). We then average these similarity scores across the two token groups for comparative visualization.

Under the prevailing language prior hypothesis, one might naturally expect hallucinated tokens to exhibit an over-reliance on linguistic context, which would mathematically manifest as an anomalously high similarity S_T between the hidden state h_k and the textual context C_T. However, our analysis reveals a striking, counter-intuitive phenomenon.

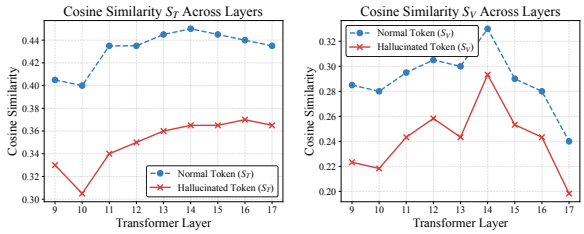

Figure 2: Average cosine similarities of the hidden state h_k with the textual context (S_T) and visual context (S_V). Hallucinated tokens consistently exhibit significantly lower similarities to both contexts compared to normal tokens.

As illustrated in Figure 2, the curves demonstrate a persistent and significant margin between normal and hallucinated tokens. Crucially, the representations of hallucinated tokens exhibit substantially lower cosine similarities with both the textual context C_T and the visual context C_V compared to their factual counterparts.

This observation profoundly refutes the simplistic language prior assumption. The simultaneous and sharp decline in both S_T and S_V indicates that the hidden state h_k is semantically derailed. Rather than over-fitting to one modality, the hallucinated representation completely diverges from the grounded multimodal subspace spanned by C_T and C_V, straying into an erroneous direction in the high-dimensional vector space.

This geometric divergence serves as the fundamental motivation for our method. It implies that passively scaling attention weights is insufficient; instead, an active geometric intervention is required to explicitly steer the diverging hidden states back to the grounded contextual subspace.

4 Methodology

4.1 Dynamic Hallucination Detection

Building upon the empirical observations in Section 3.2, we propose a dynamic detection mechanism to identify hallucinated tokens during the inference phase. Our goal is to quantify the semantic deviation of the hidden state h_k relative to the grounded contextual subspace. To this end, we define a contextual deviation indicator, denoted as E, which aggregates the cosine distances between the current hidden state and the modality-specific context vectors:

$$E = (1 - S_T) + (1 - S_V) = 2 - S_T - S_V \quad (2)$$

where S_T and S_V represent the cosine similarities of h_k with the textual context C_T and visual context C_V, respectively. Geometrically, E represents the total angular displacement of the hidden state from the grounded semantic anchors. A higher value of E indicates that the representation h_k has significantly diverged into an erroneous direction that is unanchored to either textual or visual evidence.

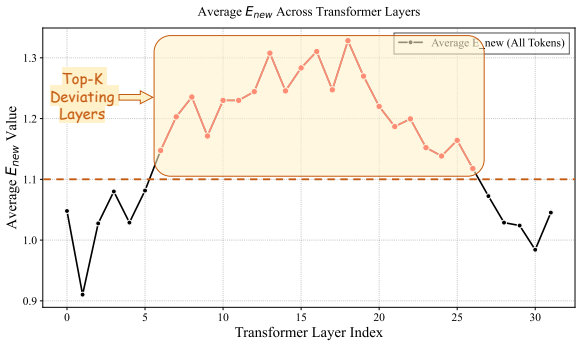

Figure 3: Distribution of the contextual deviation indicator E across different transformer layers for LLaVA-1.5-7B. The deviation is significantly more pronounced in the intermediate layers.

As illustrated in Figure 3, we quantify the average E values across all layers using the LLaVA-1.5-7B model on the CHAIR benchmark. We observe a clear hierarchical aggregation effect: the semantic deviation is not uniformly distributed but is instead concentrated within specific depths of the transformer stack. Specifically, the indicator E remains relatively low in the early and final layers but exhibits a significant peak in the intermediate stages. This suggests that the middle layers serve as a critical transition zone where multimodal features are fused, and it is within this zone that the hidden states are most susceptible to derailment.

4.1.1 Top-K Deviating Layers Selection

Based on the observed layer-wise distribution of E, we implement a selective intervention strategy rather than applying corrections to the entire network. While this process involves identifying a subset of layers, our approach is fundamentally driven by the magnitude of semantic divergence rather than a fixed layer count. Empirically, for models like LLaVA-1.5-7B, we identify the critical middle layers where the average deviation E exceeds a specific sensitivity threshold. In our experimental setup, this criterion corresponds to layers 6 through 26, which we designate as the active execution zone for our proposed intervention.

Within these selected intermediate layers, we further refine the intervention to be token-adaptive by introducing a trigger threshold τ. For each generated token, the Norm-Preserving Rotation is activated if and only if the current deviation indicator satisfies $E \geq \tau$. This dual-level selection mechanism ensures that the intervention is both spatially targeted at susceptible layers and dynamically triggered for specific tokens, effectively rectifying the generation trajectory only when a significant semantic derailment is detected. Such a design maintains the model's original representational integrity for factual tokens while precisely capturing and correcting those undergoing contextual deviation.

4.2 Layer-Specific Hidden State Mitigation

4.2.1 Norm-Preserving Rotation (Top-K Deviating Layers)

For the intermediate layers identified in the selection phase, we execute a geometric intervention termed Norm-Preserving Rotation to explicitly rectify the deviating hidden states. Our objective is to steer the derailed representation h_k back toward the grounded contextual subspace while preserving its inherent representational intensity.

First, we define a grounded semantic plane $\mathcal{P}$ formed by the textual and visual context vectors:

$$\mathcal{P} = \text{span}\{C_T, C_V\} \quad (3)$$

This plane characterizes a factual subspace within the high-dimensional hidden space, providing a grounded reference for rectifying the deviating h_k through rotation.Within this coplanar space, we establish a reference target direction, C_{target}, which represents the ideal grounded orientation for the current token. The orientation of C_{target} is dynamically determined by balancing the textual and

visual anchors:

$$C_{target} = \text{Norm}((1-\alpha) \cdot C_T + \alpha \cdot C_V) \quad (4)$$

where $\text{Norm}(\cdot)$ denotes the vector normalization operation, and $\alpha \in [0, 1]$ is a hyperparameter that adjusts the relative proximity of the target direction toward the visual versus textual context.

Subsequently, we perform a rotation to move the hidden state h_k toward C_{target}. To ensure that the intervention only modifies the semantic direction without disrupting the model's activation magnitude, we implement a norm-preserving rotation. The updated hidden state h_{new} is formulated using the principle of spherical linear interpolation:

$$h_{new} = \|h_k\| \cdot \text{Norm}\left(\text{Slerp}(h_k, C_{target}; \beta)\right) \quad (5)$$

where $\beta \in [0, 1]$ controls the rotation intensity. A larger β signifies a stronger correction that pulls h_k closer to the grounded plane. By maintaining $\|h_{new}\| = \|h_k\|$, we guarantee that the calibration does not inadvertently weaken the feature's signal strength before it enters the LayerNorm and the subsequent FFN. This precise directional correction ensures that the generation trajectory is re-anchored to the factual multimodal context.

4.2.2 Representational Smoothing (Subsequent Layers)

While the rotation mechanism effectively rectifies peak semantic deviations in the middle layers, applying aggressive directional changes in the final layers of the Transformer stack may interfere with the stabilization of the probability distribution over the vocabulary. Therefore, for the layers subsequent to the active execution zone, we transition from rotation to a Representational Smoothing strategy.

In this stage, rather than forcing a geometric redirection, we aim to enhance the coherence of the corrected trajectory by incorporating historical semantic information. For each layer k in the later stages, the hidden state is updated via a weighted interpolation with the representation from the preceding layer:

$$h_k = \gamma h_{k-1} + (1-\gamma) h_k \quad (6)$$

where $\gamma \in [0, 1]$ is the smoothing factor. This operation facilitates a more consistent representation update, allowing the calibrated semantic signals from the intermediate layers to propagate reliably to the final prediction head. This dual-stage intervention, which combines rotation for correction and smoothing for stability, ensures a robust and grounded generation process throughout the entire network depth.

5 Experiments

5.1 Experimental Setups

Models. To comprehensively evaluate the effectiveness and generalizability of our proposed ROT framework, we conduct experiments across multiple mainstream Large Vision-Language Models with varying architectures and parameter scales. Specifically, we select the widely adopted LLaVA-1.5 series, including LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2024c). Furthermore, to validate our method on state-of-the-art models, we incorporate the Qwen-VL family, utilizing both Qwen2-VL-7B (Wang et al., 2024b) and the highly capable Qwen2.5-VL-32B (Bai et al., 2025).

Baselines. We compare ROT against standard Greedy Decoding and a comprehensive suite of strong inference baselines. These include contrastive-based methods such as VCD (Leng et al., 2024), attention-penalty methods like OPERA (Huang et al., 2024), and recent advanced mitigation strategies including RITUAL (Woo et al., 2024), Vissink (Kang et al., 2025), SID (Huo et al., 2024), and TAME (Tang et al., 2025). Additionally, we compare our method with LVLMs-Saliency (Zhang et al., 2026), a recent gradient-aware sampling approach.

Benchmarks and Metrics. To evaluate the effectiveness of our approach in mitigating object hallucinations, we conduct experiments on three widely adopted benchmarks: CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023), and MME(Fu et al., 2026). Following standard evaluation protocols, we primarily evaluate the hallucination rate at the sentence level C_S and instance level C_I for the CHAIR benchmark, where lower scores indicate better hallucination mitigation.

5.2 Main Results

Mitigation of Object Hallucination. We first evaluate the hallucination mitigation performance on the LLaVA-1.5-7B model, as detailed in Table 1. On the CHAIR benchmark, our ROT framework achieves remarkable quantitative improvements, establishing a new standard for training-free

Table 1: Performance comparison of ROT against various state-of-the-art training-free baselines on LLaVA-1.5-7B across CHAIR, POPE, and MME benchmarks.

Method	CHAIR				POPE		MME			
	$C_S \downarrow$	$C_I \downarrow$	Recall↑	length↑	Acc↑	F1↑	Exist.↑	Pos.↑	Color↑	Total↑
Greedy	54.4	15.7	75.2	82.7	85.5	85.9	175.67	114.00	151.00	440.67
VCD (Leng et al., 2024)	51.1	14.7	75.2	82.3	85.0	85.3	184.67	128.67	153.00	466.34
OPERA (Huang et al., 2024)	47.0	14.6	74.5	77.1	85.2	84.2	180.67	111.67	123.33	415.67
RITUAL (Woo et al., 2024)	45.2	13.2	74.3	80.3	84.3	85.2	187.50	125.00	164.17	476.67
Vissink (Kang et al., 2025)	52.4	14.5	75.1	83.4	86.5	86.0	190.00	138.33	155.00	483.33
SID (Huo et al., 2024)	44.2	12.8	73.0	80.4	85.8	85.6	183.90	127.80	155.90	467.60
TAME (Tang et al., 2025)	41.3	12.4	74.4	79.9	85.7	85.4	193.00	139.00	164.67	496.67
LVLMs-Saliency (Zhang et al., 2026)	35.6	8.2	75.4	79.5	**87.5**	**87.0**	190.00	133.33	**175.00**	498.33
ROT (Ours)	**33.0**	**6.7**	**75.6**	**85.1**	86.6	86.3	**195.00**	**143.33**	168.33	**506.66**

Table 2: Generalization performance of ROT across various LVLM architectures and parameter scales on CHAIR and the POPE benchmark (including its adversarial, popular, and random subsets).

Method	CHAIR		POPE					
			Adversarial		Popular		Random	
	$C_S \downarrow$	$C_I \downarrow$	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
Base Model: LLaVA-1.5-7B								
Greedy	54.4	15.7	80.4	81.7	86.2	86.4	89.9	89.6
VCD	51.1	14.7	81.2	82.1	85.7	85.9	88.0	88.1
MemVR	50.4	14.3	79.8	81.8	86.3	85.9	89.4	89.6
ROT	**33.0**	**6.7**	**82.3**	**82.6**	**87.2**	**86.6**	**90.2**	**89.8**
Base Model: LLaVA-1.5-13B								
Greedy	41.4	12.7	83.8	84.4	86.1	86.8	87.5	88.4
VCD	44.7	13.1	81.9	81.8	85.7	85.1	87.4	86.6
MemVR	42.2	11.5	81.2	82.4	87.1	85.3	88.2	88.5
ROT	**30.8**	**4.3**	**85.0**	**84.5**	**87.7**	**86.9**	**89.5**	**88.6**
Base Model: Qwen2-VL-7B								
Greedy	24.2	8.0	85.9	84.8	87.6	86.4	89.3	88.6
ROT	**21.4**	**4.9**	**87.3**	**86.2**	**89.2**	**88.5**	**90.5**	**89.9**
Base Model: Qwen2.5-VL-32B								
Greedy	43.6	9.5	85.8	84.6	86.9	85.6	88.5	87.1
ROT	**30.9**	**8.7**	**86.5**	**85.5**	**87.4**	**86.4**	**89.3**	**88.1**

interventions. Compared to the standard greedy decoding, ROT drastically reduces the sentence-level hallucination metric C_S from 54.4 to 33.0 and the instance-level metric C_I from 15.7 to 6.7. Even when compared to the highly competitive, second-best method, LVLMs-Saliency, our geometric intervention proves superior, further depressing C_S by an absolute margin of 2.6 points and C_I by 1.5 points. Transitioning to the POPE benchmark, ROT continues to deliver robust quantitative gains. It elevates the Accuracy from the greedy baseline of 85.5 to 86.6 and boosts the F1-score from 85.9 to 86.3, decisively outperforming widely adopted baseline strategies such as OPERA, VCD, and TAME. On the comprehensive MME benchmark, ROT surpasses all evaluated methods and achieves a state-of-the-art total score of 506.66. This validates that correcting the semantic trajectory directly in the hidden space is a highly effective mechanism for anchoring generation to factual visual evidence across diverse evaluation protocols.

Generalization Across Architectures and Scales. To demonstrate the universal applicability and scalability of ROT, we extend our evaluation across diverse model families and parameter sizes, encompassing LLaVA-1.5-7B, LLaVA-1.5-13B, Qwen2-VL-7B, and Qwen2.5-VL-32B, as presented in Table 2. Across all tested architectures, ROT consistently and significantly suppresses hallucinated outputs on the CHAIR benchmark. Quantitatively, the intervention achieves an average relative reduction of 26.4% in C_S and an impressive 42.7% in C_I across the four models. Beyond CHAIR, we comprehensively assess factual grounding using the POPE benchmark. Across the four diverse models and all testing subsets, our method guarantees steady improvements, yielding consistent Accuracy lifts of up to 2.0 absolute points relative to their respective greedy decoding baselines. Exhibiting robust efficacy even on massive, highly optimized models like Qwen2.5-VL-32B underscores that the semantic deviation of hidden states is a fundamental issue across LVLMs, and our geometric intervention provides a scalable, architecture-agnostic solution.

5.3 Ablation Study

To validate the efficacy and necessity of the core components within our proposed framework, we conduct a comprehensive ablation study on the CHAIR benchmark. We isolate the two primary mechanisms of our approach: the Norm-Preserving Rotation in the intermediate layers (denoted as w/o Rotation) and the Representational Smoothing in the subsequent layers (denoted as w/o Smoothing). The evaluation is performed on both LLaVA-1.5-

7B and Qwen2-VL-7B architectures, with the quantitative results presented in Table 3.

Removing the rotation mechanism causes severe performance degradation. On LLaVA-1.5-7B, omitting rotation drastically increases C_I from 6.7 to 12.4 and C_S from 33.0 to 40.5. This confirms that merely applying passive smoothing is insufficient; actively redirecting the hidden state toward the grounded factual plane is essential to fundamentally correct semantic deviations.

Similarly, disabling the smoothing mechanism notably increases hallucination rates, degrading LLaVA-1.5-7B's C_I from 6.7 to 8.3. This indicates that while rotation effectively anchors the semantic direction, the absence of subsequent smoothing allows the calibrated hidden states to destabilize in deeper layers, triggering secondary ungrounded fluctuations.

Ultimately, the full ROT framework significantly outperforms both incomplete variants. The norm-preserving rotation provides crucial geometric correction at peak deviation, while smoothing reliably preserves this structural integrity to the output head, confirming their strong synergy in mitigating multimodal hallucinations.

Table 3: Ablation study of ROT components on LLaVA-1.5-7B and Qwen2-VL-7B models across the CHAIR benchmark.

Method	LLaVA-1.5-7B		Qwen2-VL-7B	
	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
Greedy	54.5	15.7	24.2	8.0
ROT	**33.0**	**6.7**	**21.4**	**4.9**
w/o Rotation	40.5	12.4	23.4	6.6
w/o Smoothing	38.2	8.3	21.9	5.2

5.4 Case Study

To intuitively demonstrate the mechanism and effectiveness of our proposed method, we present a qualitative case study from the CHAIR benchmark using the LLaVA-1.5-7B model, as illustrated in Figure 4.

In this example, given an image of a person skateboarding, the baseline model generates a description containing ungrounded objects, specifically hallucinating a bench situated near the ramp and a clock on the wall. By analyzing the internal representations at the moment of generation, we find that these hallucinated tokens exhibit notably low contextual similarities. Specifically, the token bench presents an S_T of 0.41 and S_V of 0.24, while the

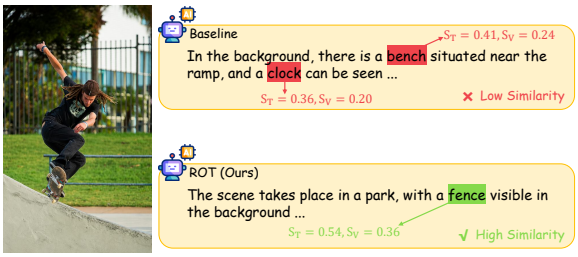

Figure 4: Qualitative comparison and hidden state similarities. The baseline model hallucinates a bench and a clock, which exhibit distinctly lower S_T and S_V compared to the factual token fence generated by ROT.

token clock further drops to an S_T of 0.36 and S_V of 0.20. These values indicate that the hidden states of these tokens have deviated significantly from the grounded textual and visual context.

In contrast, our ROT framework dynamically detects this deviation and applies the norm-preserving rotation to steer the representation back toward the grounded semantic plane. As a result, the model successfully identifies the factual object in the scene, generating the token fence with significantly higher similarities ($S_T = 0.54$, $S_V = 0.36$). This case demonstrates that ROT effectively re-anchors the generation trajectory to the visual evidence, replacing ungrounded hallucinations with accurate, grounded descriptions.

6 Conclusion

In this paper, we introduce ROT, a novel training-free framework designed to mitigate object hallucinations in LVLMs. Our approach is motivated by the empirical discovery that hallucinated tokens exhibit a significant directional deviation in the hidden semantic space, rather than a simple over-reliance on language priors. To address this, we propose a layer-specific intervention strategy consisting of two key components: a norm-preserving rotation that explicitly corrects the semantic trajectory in the most deviating intermediate layers, and a representational smoothing mechanism that ensures stability in subsequent layers. Extensive experiments across multiple benchmarks, including CHAIR, POPE, and MME, demonstrate that ROT consistently reduces hallucination rates and outperforms various state-of-the-art baselines while maintaining the model's fundamental multimodal capabilities. Our findings underscore the importance of geometric calibration in the hidden space as a robust and efficient solution for grounded generation.

7 Limitations

Despite the effectiveness of the ROT framework, several limitations remain to be addressed in future work. First, the construction of the grounded semantic plane is derived solely from the local textual and visual contexts provided within the model's attention mechanism. While effective, this approach does not incorporate external structured knowledge or global scene graphs, which could potentially offer a more comprehensive reference for resolving complex hallucinations. Second, the identification of the critical deviating layers and the selection of the trigger threshold τ are currently based on empirical observations from representative model families. While ROT demonstrates strong generalization across different scales (7B to 32B), these configurations might require adaptive fine-tuning or automated search strategies when applied to ultra-large-scale models with significantly different depth-wise representational dynamics.

References

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, et al. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1381–1390.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. 2025. Multimodal language models see better when they look shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6688–6706.

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

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

Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. 2024. Enhancing vision-language model reliability with uncertainty-guided dropout decoding. *arXiv preprint arXiv:2412.06474*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2026. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. 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*, pages 13418–13427.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

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

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25004–25014.

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

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

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Quanjiang Li, Zhiming Liu, Wei Luo, Tingjin Luo, and Chenping Hou. 2026. Correcting visual blur induced by attention distraction to reduce hallucinations: Algorithm and theory. *Preprint*, arXiv:2605.24602.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, volume 2024, pages 57689–57733.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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

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

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2024. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 37–58.

Xiaoye Qu, Qiyuan Chen, Wei Wei, Jiashuo Sun, Daizong Liu, and Jianfeng Dong. 2025. Alleviating hallucination in large vision-language models with active retrieval augmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(9):1–22.

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

Barrett Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. 2025. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *International Conference on Learning Representations*, volume 2025, pages 27745–27776.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2026. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *International Journal of Computer Vision*, 134(1):22.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024a. Docllm: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8529–8548.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

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

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024c. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. 2024. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024. Logical closed loop: Uncovering object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6944–6962.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024. Mitigating object hallucination via concentric causal attention. *Advances in neural information processing systems*, 37:92012–92035.

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

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

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, et al. 2026. Hallucination begins where saliency drops. *arXiv preprint arXiv:2601.20279*.

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

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pages 56969–56998.

Beier Zhu, Yulei Niu, Saeil Lee, Minhoe Hur, and Hanwang Zhang. 2023. Debiased fine-tuning for vision-language models by prompt regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3834–3842.

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

A Algorithm of the ROT Framework

To facilitate a deeper understanding and ensure the reproducibility of our proposed method, we provide the detailed pseudocode for the ROT framework in Algorithm 1.

The algorithm operates during the autoregressive decoding phase at each generation step. It takes the layer-wise hidden states and attention distributions as inputs, dynamically computes the modality-specific contexts, and applies the dual-stage intervention (Norm-Preserving Rotation and Representational Smoothing) based on the layer depth and the contextual deviation indicator E.

B Evaluation Metrics Definitions

To comprehensively evaluate the hallucination mitigation performance, we employ three widely used benchmarks. The detailed formulations of their respective metrics are defined as follows:

CHAIR Metrics. The Caption Hash Algorithm for Image Reporting (CHAIR) benchmark quantifies object hallucination in open-ended image captioning at two different granularities: the instance level (C_I) and the sentence level (C_S). The instance-level metric C_I calculates the proportion of hallucinated object instances among all objects mentioned by the model:

$$C_I = \frac{\text{\# of hallucinated objects}}{\text{\# of all objects mentioned}} \tag{7}$$

The sentence-level metric C_S measures the proportion of sentences that contain at least one hallucinated object:

$$C_S = \frac{\text{\# of sentences with hallucinated objects}}{\text{\# of all sentences}} \tag{8}$$

For both metrics, a lower value indicates a reduced degree of object hallucination.

POPE Metrics. The Polling-based Object Probing Evaluation (POPE) frames hallucination evaluation as a binary visual question answering task. We report the standard Accuracy (Acc) and F1-score:

- **Accuracy (Acc):** The ratio of correctly answered probing questions (both existing and non-existing objects) to the total number of questions.
- **F1-score:** The harmonic mean of Precision and Recall. It provides a balanced measure

Algorithm 1 Dynamic Hallucination Detection and Mitigation via ROT

Input: Current token hidden states X; Normalized attention weights $\tilde{A}$; Textual token indices T; Visual token indices V.
Parameters: Rotation execution zone (layers 1 to L_{mid}); Smoothing zone (layers $L_{mid}+1$ to L); Trigger threshold τ; Balance factor α; Rotation intensity β; Smoothing factor γ.
Output: Calibrated hidden states for vocabulary prediction.

Initialize h_0 as the input embedding of the current token.
for each transformer layer $k = 1$ to L **do**
  // Forward pass through Self-Attention
  Attn_Output, $\tilde{A} \leftarrow$ Self-Attention(h_{k-1}, X)
  $h_k \leftarrow h_{k-1} +$ Attn_Output ▷ Extract post-residual hidden state
  // Context Aggregation
  $C_T \leftarrow \sum_{j \in T} \tilde{A}_{t,j} X_j^{k-1}$ ▷ Textual context
  $C_V \leftarrow \sum_{i \in V} \tilde{A}_{v,i} X_i^{k-1}$ ▷ Visual context
  // Layer-Specific Interventions
  if $k \leq L_{mid}$ **then** ▷ Active execution zone for Rotation
    $S_T \leftarrow \cos(h_k, C_T)$
    $S_V \leftarrow \cos(h_k, C_V)$
    $E \leftarrow 2 - S_T - S_V$ ▷ Calculate contextual deviation
    if $E \geq \tau$ **then** ▷ Trigger Norm-Preserving Rotation
      $C_{target} \leftarrow \text{Norm}(\alpha \cdot C_V + (1-\alpha) \cdot C_T)$
      $h_{new} \leftarrow \|h_k\| \cdot \text{Norm}\left(\text{Slerp}(h_k, C_{target}; \beta)\right)$
      $h_k \leftarrow h_{new}$ ▷ Rectify semantic trajectory
    end if
  else ▷ Subsequent layers for Smoothing
    $h_k \leftarrow \gamma \cdot h_{k-1}^{(prev)} + (1-\gamma) \cdot h_k$ ▷ Representational smoothing
  end if
  $h_{k-1}^{(prev)} \leftarrow h_k$ ▷ Cache current state for the next layer's smoothing
  // Forward pass through FFN
  $h_k \leftarrow \text{LayerNorm}(h_k)$
  $h_k \leftarrow h_k + \text{FFN}(h_k)$
end for
return h_L ▷ Final hidden state for language modeling head

of the model's ability to correctly identify existing objects without falsely confirming non-existent ones.

MME Metrics. For the MME benchmark, we focus on the foundational perception subtasks (e.g., Existence, Position, Color). Following the official MME scoring protocol, the score for each subtask is derived from a combination of basic accuracy and pair-wise accuracy. Specifically, the evaluation utilizes pairs of positive and negative questions. The maximum possible score for each subtask is 200, where 100 points correspond to the basic accuracy across all individual questions, and the remaining 100 points are awarded based on strict pair-wise accuracy (i.e., both questions regarding a specific object must be answered correctly).

C Experimental Details

In this section, we elaborate on the concrete hardware environment and the specific hyperparameter configurations employed throughout our empirical evaluations to guarantee reproducibility. All experiments, encompassing both the standard greedy decoding baselines and our proposed ROT framework, are consistently executed on a unified computing node equipped with an NVIDIA A800 (80GB) GPU.

The dual-stage intervention mechanism of ROT is governed by three key hyperparameters: the context balance factor α, the rotation intensity β, and the representational smoothing factor γ. Taking CHAIR as an example, for the LLaVA model family, the hyperparameters for the LLaVA-1.5-7B model are established at $\alpha = 0.44$, $\beta = 0.062$, and $\gamma = 0.51$. When scaling up to the larger LLaVA-1.5-13B architecture, the configurations are tuned to $\alpha = 0.45$, $\beta = 0.050$, and $\gamma = 0.55$.On the Qwen2-VL-7B model, we configure the intervention framework with $\alpha = 0.50$, $\beta = 0.048$, and $\gamma = 0.43$. For the Qwen2.5-VL-32B model, the parameters are fine-tuned to $\alpha = 0.46$, $\beta = 0.024$, and $\gamma = 0.41$.

D System Prompts

In our experiments, we design specific system prompts to guide the Large Vision-Language Models to generate appropriate responses tailored to the evaluation formats of different benchmarks. To ensure fair and rigorous comparisons, these instruction templates are kept strictly identical across the baseline methods and our proposed ROT framework.

For the open-ended image captioning task in the **CHAIR** benchmark, we utilize a standard descriptive prompt to encourage comprehensive textual generation:

```
"Please describe this image in detail."
```

For the probing-based tasks in the **POPE** and **MME** benchmarks, we instruct the model to first briefly analyze the image content and then output the final binary classification result. This step-by-step instruction effectively standardizes the output format for automated exact-match evaluation:

```
"Please describe the image content concisely, and then answer with yes or no at the end."
```

E Extensive Qualitative Results

To further demonstrate the robust effectiveness of the ROT framework in mitigating object hallucinations, we present additional qualitative case studies in this section. These examples illustrate how our geometric intervention correctly anchors the generation trajectory across diverse and complex visual scenes.

As illustrated in Figure 5, the standard greedy decoding hallucinates non-existent objects, specifically fabricating a "bench" and a "clock". Our proposed ROT method effectively corrects this contextual deviation, completely eliminating these hallucinations while successfully identifying and generating the correct factual token "fence".

In Figure 6, the baseline model erroneously generates a "handbag" that does not exist in the visual input. The ROT framework not only removes this hallucinated object but also demonstrates a strong capability to capture global visual attributes, accurately describing the overall scene as "black and white" to match the monochrome nature of the image.

As shown in Figure 7, the greedy strategy suffers from object hallucination by fabricating a "bench" in the background. ROT successfully intercepts this ungrounded semantic shift in the intermediate layers, yielding a concise and strictly factual description without introducing any non-existent instances.

Figure 8 demonstrates a more complex scenario where the baseline model suffers from severe

multi-object hallucinations, incorrectly fabricating a “truck”, “traffic lights”, and “handbags”. Conversely, ROT effectively suppresses this clustered hallucination and precisely anchors the generation to the actual visual evidence, accurately identifying the “bridge”.

Baseline

The image captures a young man skillfully riding a skateboard down a ramp. He is in the middle of performing a trick, with his arms outstretched for balance. The skateboarder is wearing a black shirt and appears to be enjoying the thrill of the ride.\n\nIn the background, there is a bench situated near the ramp, and a clock can be seen hanging on the wall. The scene is set in an outdoor area, possibly a park or a skate park.

ROT (Ours)

The image features a young man performing a skateboarding trick. He is in the air, jumping, and is in the middle of a skateboarding trick. The skateboarder is wearing a black shirt and a blue wristband. The skateboarder is in the air, and his skateboard is visible.\n\nThe scene takes place in a park, with a fence visible in the background. The skateboarder is performing his trick in a park, and the skateboard is visible in the air.

Figure 5: Qualitative Example 1. ROT successfully removes the hallucinated “bench” and “clock”, and accurately describes the “fence”.

Baseline

The image features a man riding a bicycle on a sidewalk. He is wearing a hat and appears to be enjoying his ride. The bicycle is positioned towards the left side of the scene.\n\nIn the background, there is a building with a window, and a car is parked nearby. The scene also includes a handbag placed on the ground, possibly belonging to the man riding the bicycle.

ROT (Ours)

The image features a man riding a bicycle, standing on the street. He is wearing a hat and is standing next to a wall. The bicycle is positioned near the left side of the image. The man appears to be looking at the camera, possibly taking a picture. The scene is black and white, giving it a vintage feel.

Figure 6: Qualitative Example 2. ROT eliminates the hallucinated “handbag” and correctly identifies the “black and white” attribute of the image.

Baseline

The image captures a man riding a skateboard down a sidewalk. He is wearing a blue jacket and a green hat, adding a touch of color to the scene. The man appears to be enjoying his ride, possibly performing a trick or simply cruising down the street.\n\nIn the background, there are a few cars parked along the sidewalk, and a bench can be seen further down the path. The overall atmosphere of the scene is lively and energetic, with the skateboarder being the main focus

ROT (Ours)

The image features a man riding a skateboard, wearing a hat. He is riding on a street, possibly in a park. The skateboarder is wearing a blue jacket and a hat. The skateboarder is in motion, moving down the street. \n\nIn the background, there is a car visible, adding to the urban setting. The scene is captured in a yellow filter, giving it a warm, vintage feel.

Figure 7: Qualitative Example 3. ROT intercepts the semantic deviation to prevent the hallucination of a “bench”.

Baseline

The image captures a busy city street at dusk, with a large skyscraper towering over the surrounding buildings. The street is filled with traffic, including multiple cars and a truck, as well as a bus. The traffic lights are visible at various points along the street, ensuring the smooth flow of traffic.\n\nThe scene is bustling with pedestrians, with at least 12 people walking or standing on the sidewalks. Some of them are carrying handbags, adding to the urban atmosphere. The combination of the tall building, the traffic, and the people creates a lively and dynamic cityscape.

ROT (Ours)

The image is a nighttime city street, featuring a large building, possibly a skyscraper, towering over the city. The building is surrounded by a city street, with a bridge visible in the background. The street is busy, with several cars driving on the road. There are a few cars in the foreground, and a few more in the background.\n\nThe city is bustling with activity, with a few people walking around, and a few cars driving. The city is illuminated by the city lights, creating a vibrant cityscape. The city is surrounded by a tall building, which stands out in the cityscape.

Figure 8: Qualitative Example 4. ROT effectively suppresses clustered hallucinations (truck, traffic lights, handbags) and correctly recognizes the “bridge”.