DAPE-BR: Distance-Aware Positional Encoding for Mitigating Object Hallucination in LVLMs

Mingrui Xie[1*] **Tianxiang Xu**[2*] **Qianhai Tang**[1] **Shanming Yao**[1]
Xiaofeng Zhang[3] **Junliang Du**[3†]
[1] China University of Geosciences
[2] Peking University
[3] Shanghai Jiao Tong University

Abstract

Large Vision Language Models (LVLMs) have garnered substantial interest owing to their impressive ability to interpret visual inputs and converse with users. Nevertheless, LVLMs still suffer from object hallucination — generating descriptions for objects that are absent from the image, which undermines their reliability and hinders real-world deployment. We propose DAPE-BR, a positional-alignment scheme that (i) preserves the pretrained weight order while globally aligning visualtext distances, (ii) embeds an isotropic fused patch-distance metric, and (iii) applies a patch-distance causal mask to enforce spatial causality. Extensive experiments on POPE, MMStar, and SQA show that DAPE-BR consistently reduces hallucinations and boosts overall performance.

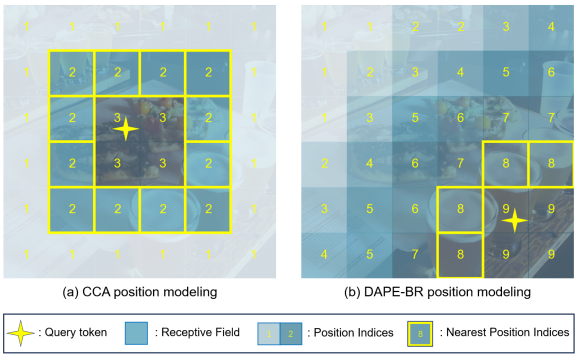


Figure 1: Long-term decay under two positional encodings. Attenuation follows the *downward numerical gradient* in each matrix-from larger to smaller values. (a) CCA exhibits concentric multi-directional decay; (b) DAPE-BR spreads decay more evenly in space. Dark cells mark weak decay, light cells strong decay. Example shows 36 image tokens.

1 Introduction

LVLMs(Liu et al., 2023b, 2024b; Bai et al., 2023; Cha et al., 2024; Ye et al., 2023; Zhu et al., 2023) excel at parsing images and conducting natural-language conversations, but they still hallucinate-describing objects that are missing or mis-characterised in the picture(Li et al., 2023; Rohrbach et al., 2018; Cui et al., 2023; Liu et al., 2024a). These factual slips erode user trust and impede reliable, real-world deployment, so reducing object hallucination has become a central research priority.

To tackle this challenge,Various strategies have been developed to mitigate object hallucination in LVLMs. A widely adopted approach involves employing post-hoc correction through revisor models(Yin et al., 2024; Zhou et al., 2023; Lee et al., 2023), which refine generated outputs to suppress hallucinated descriptions. Another promising avenue enhances supervised fine-tuning by enriching the diversity of instruction-tuning datasets(Liu et al., 2023a; Yu et al., 2024a). Although these methods effectively reduce object hallucination, they rely on high-quality annotations, which entail substantial manual labor and make them costly to deploy in practice. Recently, several studies have explored training-free techniques that mitigate object hallucination by directly correcting inaccuracies during the autoregressive decoding phase of LVLMs(Leng et al., 2024; Huo et al., 2024; Huang et al., 2024). While these training-free heuristics operate at the output layer, another line of work revisits the *input* sidenamely, the positional encoding itself.

A recent study (Xing et al., 2024) traces object hallucination in LVLMs to the problem of "long-term decay" in Rotary Position Embeddings (RoPE) (Su et al., 2024): as token indices increase, RoPE's sinusoidal phases shrink exponentially and thus cause tokens at higher positions to overshadow earlier visual context. The authors counter this by adding a lightweight positional-alignment module (Figure 1a) that rescales and recenters the phases

[*] Equal contribution.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 8638–8649
November 4-9, 2025

before decoding and then fine-tunes the model. This modification reduces hallucinations but still suffers from three issues. First, its concentric-ring encoding merges all patches on the same ring into a single index, slashing the positional vocabulary from $H \times W$ to $\mathcal{O}(\min\{H, W\})$ and causing under-utilization and aliasing (border patches dominate, and objects that cross rings split encodings). Second, because the algorithm scans patches from the periphery inward, the last visual token fed to the language model is an arbitrary border patch rather than the bottom-right patch used during pre-training. As a result, the learned visualtext offset is disrupted, and additional fine-tuning becomes necessaryhence the method is not truly training-free. Third, since the indices depend only on the radius, angular information is lost: patches that are opposite each other on a ring (approximately $2r$ apart) appear co-located, whereas adjacent patches in neighboring rings seem distant. This angular insensitivity introduces anisotropic distortions that hinder consistent 2-D spatial reasoning.

Building on this analysis, we propose DAPE-BR (see Figure 1b), a two-stage positional-alignment module that first scans visual tokens from the bottom-right corner in a special raster order and rectifies the causal mask, then encodes each token with the mean of its Euclidean, Manhattan, and Chebyshev distances, thereby preserving two-dimensional continuity and markedly reducing hallucinations during LVLMs training; this scheme shortens global tokeninstruction paths without disturbing pretrained attention, introduces an isotropic fused-distance metric that removes directional bias, and applies a three-step pipelinedistance fusion, shell quantization, and reverse causal maskingto confine attention to truly causal regions.

Our key contributions include:

- Refines DAPE-BR positional encoding, enriching spatial indices and strengthening visual grounding.
- Introduces a fully training-free hallucination-mitigation strategy that plugs into existing LVLMs without changing their parameters, sharply reducing computation and annotation costs.
- Shows consistent hallucination reduction and overall performance gains across multiple public benchmarks(Li et al., 2023)(Lu et al., 2022)(Chen et al., 2024a)(+1.25% on Accuracy and +1.84% on F1 score, as compared to the state-of-the-art method (Zou et al., 2025) on POPE).

2 Related Works

Large Vision Language Models. BERT and its variants (Devlin et al., 2019; Lu et al., 2019; Chen et al., 2020) laid the groundwork for multimodal AI, which was later amplified by GPT-3, PaLM, T5, and LLaMA (Brown et al., 2020; Chowdhery et al., 2023; Raffel et al., 2020; Touvron et al., 2023a). In Vision Language learning, ViLBERT and LXMERT (Lu et al., 2019; Tan and Bansal, 2019) were soon surpassed by the contrastive giants CLIP and ALIGN (Radford et al., 2021; Jia et al., 2021). Modern LVLMs such as LLaVA, Gemini, and Qwen (Liu et al., 2023b; Team et al., 2023; Bai et al., 2023) attach a frozen LLM to a vision encoder through a slim projection head and are tuned with visual instructions, yet they still hallucinate objects absent from the image.

Object hallucination. Hallucinating nonexistent objects compromises model reliability (Cui et al., 2023; Li et al., 2023; Rohrbach et al., 2018; Liu et al., 2024a; Guan et al., 2024; Wang et al., 2024; Nie et al., 2024; An et al., 2024; Favero et al., 2024; Wang et al., 2023). Remedies include post-hoc grounding or post-hoc self-correction, both of which break end-to-end flow (Yin et al., 2024; Zhou et al., 2023; Lee et al., 2023; Liu et al., 2024c; Wu et al., 2024); additional human-annotated data for instruction tuning (Liu et al., 2023a; Yu et al., 2024b; Sun et al., 2023; Jiang et al., 2024a; Yue et al., 2024; Yu et al., 2024a); and training-free reranking, which increases inference latency (Huang et al., 2024; Leng et al., 2024; Chen et al., 2024b). Instead, we study how rotary position encoding (RoPE) influences this failure mode.

Position encoding in Transformers. Because self-attention is order-agnostic (Vaswani et al., 2017), researchers have proposed several positional-encoding schemes, including sinusoidal (Vaswani et al., 2017), learnable (Dosovitskiy et al., 2020), and relative approaches (Shaw et al., 2018; Ke et al., 2020; He et al., 2020; Huang et al., 2020), the last of which excels on variable-length inputs (Su et al., 2024; Peng et al., 2023). Rotary position encoding (RoPE) encodes positions by rotating embedding pairs (Su et al., 2024); this design boosts linear attention and enables large-scale pre-training in LLaMA (Touvron et al., 2023a,b), and it is now being explored for vision

tasks (Chu et al., 2024; Lu et al., 2024). In this work, we examine whether RoPEs long-term decay contributes to object hallucination and how to curb it without costly retraining.

3 Motivation

3.1 Rotary Position Encoding in LVLMs.

Modern LVLMs (e.g. LLaVA) adopt Rotary Position Encoding (RoPE) to model positional dependencies in the Transformer. RoPE encodes token position p by multiplying the token embedding with a position-specific rotation matrix $R(p)$. In practice, $R(p)$ is a block-diagonal matrix composed of 2D rotation submatrices for each pair of hidden dimensions: for example, one such 2×2 submatrix is $R_{\text{2D}}(\theta_p) = \begin{pmatrix} \cos\theta_p & -\sin\theta_p \\ \sin\theta_p & \cos\theta_p \end{pmatrix}$, where θ_p is determined by predefined sinusoidal functions of p. Applying RoPE to a query or key embedding $\mathbf{x}_p$ yields $\tilde{\mathbf{x}}_p = \mathbf{x}_p\, R(p)$. This rotation imprints the position into the embedding such that (i) the inner product between any two position-encoded vectors depends on their relative index difference, and (ii) it enables extrapolation beyond a fixed length by cyclically repeating positional phase patterns. In LVLMs architectures, the rotation $R(p)$ is applied to *all* query and key vectors across each self-attention layer, so that positional relationships are consistently encoded throughout the network. Formally, for a query at position i and a key at position j, the scaled dot-product attention score is:

$$
\begin{aligned}
s_{i,j} &= \frac{(\mathbf{q}_i R(i))(\mathbf{k}_j R(j))^\top}{\sqrt{d}} \\
&= \frac{\mathbf{q}_i \big(R(i)R(j)^\top\big)\mathbf{k}_j^\top}{\sqrt{d}}.
\end{aligned} \tag{1}
$$

Here $R(i)R(j)^\top = R(j-i)$, meaning that the attention score depends on the *relative position* $j-i$. This property effectively introduces a distance-dependent attenuation in attention: as the relative index difference $|i-j|$ grows, the rotation $R(j-i)$ represents a larger phase shift, making the dot-product $s_{i,j}$ smaller on average (a *long-term decay*). In language modeling, such decay is desirable since distant words typically have weaker direct dependencies.

However, in multimodal sequences this decay can be harmful e.g. visual token v_1 (the first image patch) and a late instruction token w_{N_t} may correspond to the same object(see Figure 3(b)), yet RoPE attenuate their interaction simply because they are far apart in the sequence.

3.2 Limitations of CCA for Long-Term Decay

Underutilization. CCA adopts a coarsegrained concentric-ring encoding that merges every patch on the same ring into one positional index(seeFigure 3(c)). Let the ring index for patch coordinate (x, y) be $r(x, y)$. This collapses the number of distinct indices from $H \times W$ (for an $H \times W$ grid) to $\mathcal{O}(\max\{H/2, W/2\})$. Worse, the *population* of each index is highly imbalanced: the innermost ring contains only 4 patches, whereas the outermost ring contains $4H-4$ patches, so tokens concentrated near the borders dominate the positional vocabulary. Whenever $r(x_1, y_1) = r(x_2, y_2)$ the embeddings coincide, $\mathbf{p}_{(x_1,y_1)} = \mathbf{p}_{(x_2,y_2)}$, causing *positional aliasing*: fine objects that straddle two rings get split encodings, and distant, unrelated regions sharing a ring become indistinguishable.

Sequential-bias mismatch (not training-free). CCA traverses patches from the periphery toward the centre, so the visual token that immediately precedes the first text token is an *arbitrary border patch* rather than the bottom-right patch produced by standard raster scan. This shifts the visualtext 'bridge' index and perturbs the relative offset $d_{\text{vis}\to\text{text}}$ internalized during language-only pre-training, thereby distorting RoPEs learned attention phases and hindering visualinstruction fusion. Correcting this misalignment requires additional fine-tuning to re-establish cross-modal positional correlations, so CCA cannot serve as a truly training-free remedy

Anisotropy. Ring indices depend solely on radial distance; angular displacement is ignored. A radial step changes the index by $\Delta r = \pm 1$, whereas any tangential move along a ring leaves $\Delta r = 0$, even though the Euclidean displacement is comparable. Consequently, two patches on opposite sides of the same ringseparated by $\approx 2r$ grid unitsare treated as co-located, whereas two adjacent patches in neighbouring rings are treated as far apart. This direction-dependent (non-isotropic) notion of distance introduces inconsistent geometry and weakens the models ability to perform coherent 2-D spatial reasoning.

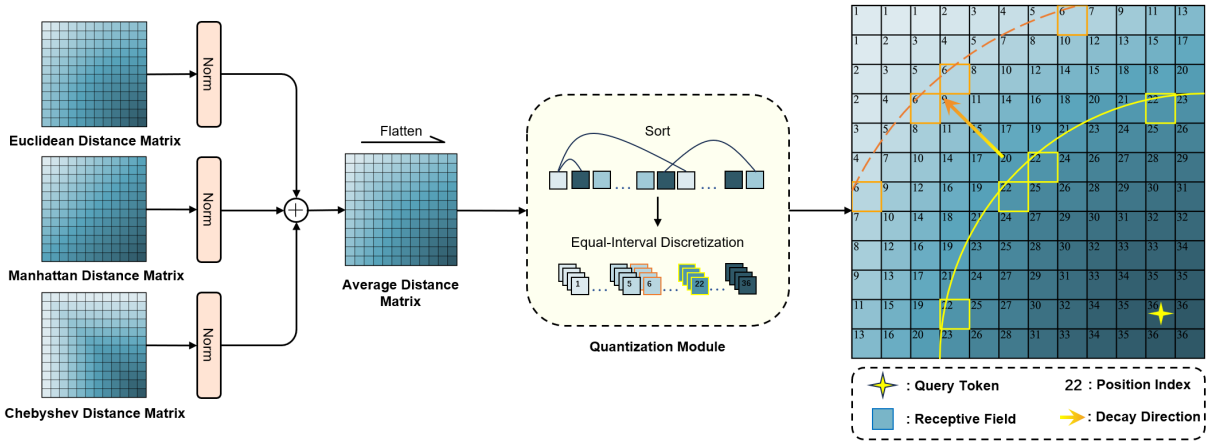


Figure 2: Overall workflow of the proposed **DAPE-BR**. We first normalise the Euclidean, Manhattan, and Chebyshev distance matrices, then average them to obtain the fused distance map. After flattening, the distances are sorted and discretised at equal intervals to generate shell indices, which are finally re-projected onto the 2-D grid (right).

4 DAPE-BR Method

Core contributions of the proposed DAPE-BR (Figure 2). Our approach can be summarized in three interconnected aspects:

Order-consistent re-indexing. Without any additional supervised fine-tuning, **DAPE-BR** globally *shrinks the relative distances from most visual tokens to the instruction token*, effectively counteracting the hallucination-prone long-range decay inherent to RoPE.

The first patch-distance metric and anisotropy removal. We are the first to *explicitly measure pairwise distances between image patches* and to encode them in the positional indices. This fused metric removes the directional anisotropy inherent in prior ring-based or raster layouts, enabling the model to perceive patch separations isotropically across the grid.

Three-stage pipeline. The method proceeds through (i) *fused distance computation*, (ii) *shell quantization* that converts the distance into discrete indices, and (iii) a *causal mask* that allows a query to attend only to keys in the same or inner shells. In the following subsections,We elaborate on each stage in the following subsections, starting with the definition of the fused distance.

4.1 Fused distance and Shell quantization.

Consider an image feature map (Figure 3(d)) of height H and width W, yielding $v = H \times W$ visual tokens arranged on a 2D grid. We denote coordinates on this grid as (r, c) with $0 \le r < H$ and $0 \le c < W$, where $(0, 0)$ corresponds to the top-left corner. Set the *anchor* point (r_a, c_a) be the bottom-right corner $(H-1,\ W-1)$. We construct three distance matrices capturing standard distance metrics from the anchor:

$$D_E[r, c] = \sqrt{(r_a - r)^2 + (c_a - c)^2}, \quad (2)$$

$$D_M[r, c] = |r_a - r| + |c_a - c|, \quad (3)$$

$$D_C[r, c] = \max\{\, |r_a - r|,\ |c_a - c| \,\}. \quad (4)$$

Each matrix $D_* \in \mathbb{R}^{H \times W}$ encodes the distance of token (r, c) from the anchor under the specified metric. We then define the **fused distance matrix** as the elementwise average of the above:

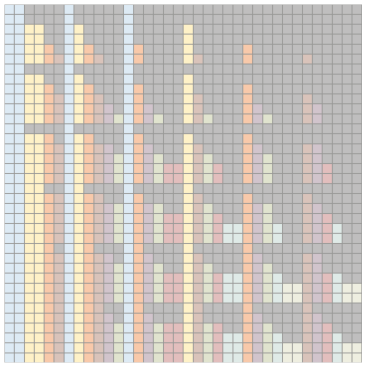
$$D_F[r, c] = \tfrac{1}{3}\big(D_E[r, c] + D_M[r, c] + D_C[r, c]\big). \quad (5)$$

which blends the geometric perspectives of Euclidean, Manhattan, and Chebyshev distances into a single scalar field D_F.

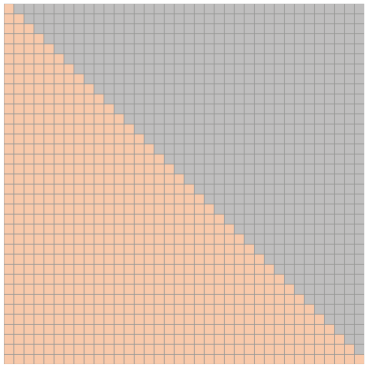
Using D_F, we induce an ordering of all v image tokens by their fused distance to the anchor. Let $\rho(r, c)$ denote the *rank* index of token (r, c) in ascending order of D_F (i.e., $\rho(r_a, c_a) = v$ for the anchor itself, and $\rho(r, c) = 1$ for the farthest token). We then quantize these ranks into discrete concentric bands, or **shells**, by applying a floor division with a fixed shell width Δ (in number of tokens):

$$s(r, c) \ = \ \left\lfloor \frac{\rho(r, c)}{\Delta} \right\rfloor. \quad (6)$$

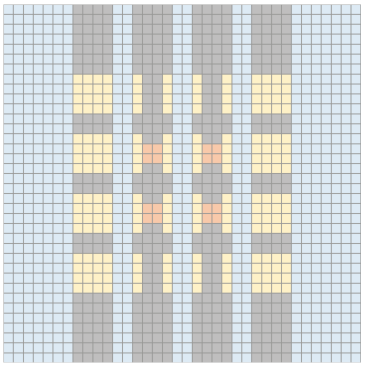
where $s(r, c) \in \{0, 1, \ldots, \lfloor v/\Delta \rfloor\}$ is the *shell index* assigned to token (r, c). All tokens that shar-



(a) Our DAPE-BR pattern



(b) LLaVA raster pattern



(c) CCA receptive pattern

$$\begin{bmatrix} 1 & 1 & \cdots & \cdots & k \\ 1 & 2 & \cdots & \cdots & k{+}1 \\ \vdots & \vdots & \lfloor \frac{\mathrm{rank}(\delta_{ij})}{4} \rfloor & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \ell & \ell{+}1 & \cdots & \cdots & \frac{v}{4} \end{bmatrix}$$

(d) DAPE-BR ($A[r, c]$)

$$\begin{bmatrix} 1 & 2 & \cdots & \sqrt{v}-1 & \sqrt{v} \\ \sqrt{v}+1 & \sqrt{v}+2 & \cdots & 2\sqrt{v}-1 & 2\sqrt{v} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ v-\sqrt{v}+1 & v-\sqrt{v}+2 & \cdots & v-1 & v \end{bmatrix}$$

(e) Raster-scan ($R[r, c]$)

$$\begin{bmatrix} 1 & 1 & \cdots & \cdots & 1 & 1 \\ 1 & 2 & \cdots & \cdots & 2 & 1 \\ & \ddots & & & \iddots & \\ \vdots & \vdots & \frac{\sqrt{v}}{2} & \frac{\sqrt{v}}{2} & \vdots & \vdots \\ \vdots & \vdots & \frac{\sqrt{v}}{2} & \frac{\sqrt{v}}{2} & \vdots & \vdots \\ & \iddots & & & \ddots & \\ 1 & 2 & \cdots & \cdots & 2 & 1 \\ 1 & 1 & \cdots & \cdots & 1 & 1 \end{bmatrix}$$

(f) Concentric CCA ($C[r, c]$)

Figure 3: **Bottom row:** position-index matrices produced by three ordering strategies. Each cells value is the positional index assigned to that location. **Top row:** query-patch receptive fields under the same three schemes. Compared with the strip-like raster of LLaVA (e) and the ring-shaped CCA (f), our DAPE-BR (d) keeps 2-D locality while shortening the distance to text tokens, thereby suppressing hallucination.

ing the same integer shell index k lie within the k-th concentric shell around the bottom-right anchor. By appropriate choice of Δ, one can control the granularity of positional grouping: smaller Δ yields finer-grained shells (approaching the fully distinct positions of a raster scan), whereas larger Δ yields coarser shells (approaching the highly compressed concentric grouping of CCA).

4.2 Causal mask.

Finally, we apply a causal mask to ensure that each position only attends to preceding positions in the sequence order. Formally, for any query position i and key position j (where positions are indexed in the chosen sequence order):

$$C(i,j) = \begin{cases} 1, & \text{if } j \le i\,. \\ 0, & \text{if } j > i\,. \end{cases} \tag{7}$$

i.e., $C(i, j) = 1$ only when j is not a future position relative to i. We use $C(i, j)$ to mask out any d_{ij} values for which $j > i$, thereby that position j (a future position) does not influence the computation for position i. The resulting distance matrix that is strictly lower-triangular, so no information flows from future to past in the attention process.

By integrating the fused distance measure with shell quantization and the causal mask, DAPE-BR constructs a positional index matrix that preserves 2D spatial locality while respecting the sequences causality. Consequently,this indexed representation is then used to inform the models attention mechanism, allowing it to capture relative positional relationships more effectively than the standard raster-scan or the CCA-based approaches.

5 Experiments

All experiments were implemented in PyTorch and run on NVIDIA A100 80GB GPU hardware. We evaluate DAPE-BR on the unified LLaVA-v1.5-7B(Liu et al., 2023b) foundation model across three complementary benchmarks, each chosen to stress a different aspect of multimodal grounding. Specifically, we use the **POPE**(Li et al., 2023) dataset to test object hallucination suppression, **MMStar**(Chen et al., 2024a)dataset to assess

Table 1: Comparison of decoding and reranking methods on three promptselection settings.

Model	Method	Citation	Random		Popular		Adversarial		Average	
			Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
LLaVA-1.5	Sampling		84.27	82.38	81.32	79.54	78.81	77.37	81.46	79.76
	Beam Search		87.00	85.40	85.80	84.27	83.50	82.15	85.43	83.94
	DOLA	ICLR 24	84.78	84.19	79.75	80.61	76.32	76.16	80.20	80.32
	VCD	CVPR 24	87.20	86.63	84.83	84.55	80.76	81.18	84.26	84.12
	OPERA	CVPR 24	87.02	85.42	85.79	84.28	83.51	82.15	85.44	83.95
	HALC	ICLR 24	87.26	87.21	84.06	84.19	79.23	81.02	83.51	84.14
	CCA	NeurIPS 24	87.70	86.68	86.87	85.64	85.70	84.46	86.86	84.54
	CCA (Training Free)	NeurIPS 24	88.62	88.09	87.53	86.75	83.36	83.06	86.50	85.97
	AGLA	CVPR 25	88.54	87.71	85.14	84.68	81.13	81.36	84.93	84.58
	TAME	ICLR 25	–	–	–	–	–	–	85.40	85.70
	MEMVR	ICML 25	–	–	–	–	–	–	87.00	85.87
	RITUAL	arXiv2025	88.87	88.81	85.83	86.17	78.80	80.54	84.40	85.17
	DAPE-BR(Training Free)	(Ours)	**90.07**	**89.40**	**88.87**	**88.26**	**85.80**	**85.50**	**88.25**	**87.71**

Table 2: Performance of different methods on the *MMStar* benchmark.

Model	Method	MMStar						
		Average	Coarse Percep↑	Fine Percep↑	Inst Reason↑	Log Reason↑	Math↑	Sci & Tech↑
LLaVA-1.5	baseline	30.00	–	–	–	–	–	–
	CCA	34.08	64.03	32.55	38.46	30.10	22.41	16.96
	CCA(Training Free)	32.73	54.20	23.89	35.26	30.70	27.55	24.79
	DAPE-BR(Training Free)	**34.92**	**61.29**	**24.23**	**37.87**	**30.57**	**28.30**	**27.28**

Table 3: Performance of different methods on the *SQA* benchmark.

Model	Method	SQA ↑
LLaVA-1.5	baseline	66.80
	CCA	**69.86**
	CCA(Training Free)	53.54
	DAPE-BR(Training Free)	68.32

image-grounded reasoning accuracy, and **SQA**(Lu et al., 2022)dataset to evaluate multi-turn dialog consistency. This evaluation design ensures that DAPE-BR is validated across these critical angles, demonstrating its ability to improve visual grounding and reduce spurious object generation.

5.1 Models and Baselines

We compare our method against standard baselines and the latest hallucination-mitigation techniques. These include DoLa (Chuang et al., 2024), a layer-contrastive decoding strategy; VCD (Leng et al., 2024), which introduces visual contrastive decoding; and OPERA (Huang et al., 2024), a retrospection-based self-correction strategy. We also compare against HALC (Jiang et al., 2024b), which uses hallucination-augmented contrastive learning, and CCA (Xing et al., 2024), the concentric causal attention method (we evaluate both its fine-tuned and training-free variants). Finally, we include several recent approaches: AGLA (An et al., 2025), a plug-and-play global/local attention assembly; TAME (Tang et al., 2025), a decoding method based on dynamically intervening token propagation; MemVR (Zou et al., 2025), a memory-space visual retracing mechanism; and RITUAL (Woo et al., 2025), which applies random image transformations during decoding. By comparing DAPE-BR with all the above methods, we demonstrate that our training-free positional alignment yields competitive or superior results in mitigating object hallucination.

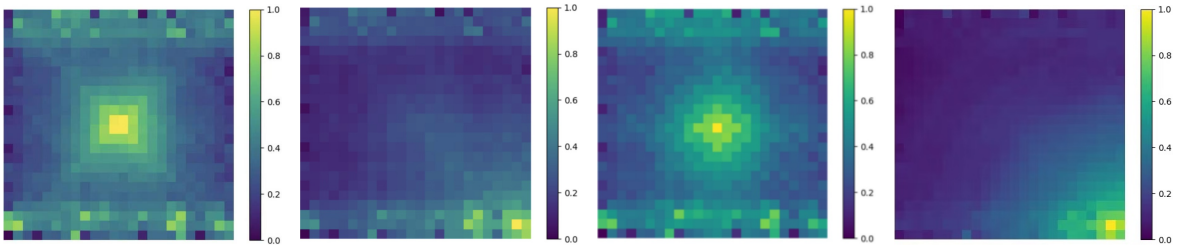
5.2 Results

Overall, DAPE-BR outperforms all training-free baselines (including the prior state-of-the-art CCA approach) across all three benchmarks. The improvements are consistent and substantial, indicating that improved positional encoding and anchoring directly translate to reduced hallucinations and more accurate visual grounding. Compared to CCA and other methods, DAPE-BR yields higher accuracy and fewer incorrect details in its answers, demonstrating the generality of our approach for mitigating hallucinations across diverse tasks.

POPE. DAPE-BR achieves state-of-the-art results on the POPE benchmark, surpassing every previously published methodboth training-free and

Table 4: Ablation experiment on POPE

Model	Method	Random		Popular		Adversarial		Average	
		Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
LLaVA-1.5	Sampling	84.27	82.38	81.32	79.54	78.81	77.37	81.46	79.76
	CCA	87.70	86.68	86.87	85.64	85.70	84.46	86.86	84.54
	CCA(Training Free)	88.62	88.09	87.53	86.75	83.36	83.06	86.50	85.97
	CCA(Training Free)+BR	89.34	88.84	88.17	87.43	84.13	83.84	87.21	86.70
	CCA(Training Free)+DAPE	89.27	88.58	88.23	87.62	83.50	83.48	86.97	86.56
	DAPE-BR(Training Free)	**90.07**	**89.40**	**88.87**	**88.26**	**85.80**	**85.50**	**88.25**	**87.71**



(a) CCA (SFT) (b) CCA (TF)+BR (c) CCA (TF)+DAPE (d) DAPE-BR

Figure 4: For every positional-alignment variant we run **LLaVA-1.5-7B** on 3 000 imagequery pairs from the *Adversarial* split of POPE. From the first decoder layer we extract the self-attention values that flow *from visual tokens to the instruction token*, then (1) average them over attention heads, visual tokens, and images, (2) reshape the resulting 24×24 patch grid, and (3) min-max normalise to $[0, 1]$ for visualisation (brighter stronger information flow). **(a)** CCA(SFT) shows a concentric focal peak caused by ring indices; **(b)** adding the bottom-right anchor (BR) shifts the focus towards the anchor; **(c)** adding distance-aware encoding (DAPE) spreads attention more isotropically; **(d)** combining DAPE and BR yields the most uniform map, indicating the greatest suppression of long-range decay and thus the lowest tendency to hallucinate objects.

SFT-based. In particular, our model reaches object-presence accuracy / F1 scores of 90.07 / 89.40 on Random prompts, 88.87 / 88.26 on Popular prompts, and 85.80 / 85.50 on the challenging Adversarial prompts.These numbers show that DAPE-BR says Yes when the object is truly in the image and No when it is noteven when the prompt tries to trick the model with scene-related distractors.Crucially, the accuracy drop from Random to Adversarial prompts is only 4.3 pp for DAPE-BR, compared with 5.3 pp for the strongest training-free baseline (CCA). This narrower gap highlights DAPE-BRs superior robustness against hallucination-inducing prompt biases. Under adversarial conditions it still posts an F1 of 85.50 , markedly higher than CCAs 83.06 , proving that DAPE-BR resists being induced into hallucinating objects that are not present.

MMStar and SQA. On the vision-indispensable benchmark MMStar (NeurIPS 2024), researchers first sampled 22401 items from eight mainstream multimodal datasets (MMMU, MMBench, ScienceQA, AI2D, SEED, MathVista, etc.(Chen et al., 2024a)). Automated filtering reduced this pool to 11 607 candidates, from which 1 500 examples that genuinely require visual information were manually curated to cover six core competencies and 18 fine-grained skills, eliminating pure-language shortcuts and training leakage at their root. Compared with the strongest training-free baseline CCA, DAPE-BR lifts the overall score from 32.73 to 34.92 (+2.2 pp); on the object-presence-critical Coarse Perception subtask it climbs from 54.20 to 61.29 (+7.1 pp). Other higher-order reasoning categories such as Instance Reasoning and Sci & Tech gain a steady 23 pp, while logic reasoning and fine-grained perception remain on par. These results show that, in the most challenging settings that readily expose hallucinations, DAPE-BR can more precisely locate the real objects in an image and reason about them. For the SQA task, created by Microsoft Research, 2022 complex table-QA items from WikiTableQuestions were split into 6 066 dialog sequences containing 17553 inter-linked questions, requiring a model to carry context across turns, avoid self-contradiction, and consistently refer back to previous answers. On this multi-turn consistency benchmark, DAPE-BR raises accuracy from 53.54 to 68.32 (+14.8 pp), drastically reduc-

ing cross-turn forgetting and hallucinated entities and demonstrating a clear advantage in long-range context tracking and entity consistency.

5.3 Ablation Study

To understand the contributions of each component in DAPE-BR, we perform an ablation study on the POPE benchmark (results in Table 4). Recall that DAPE-BR combines two key innovations: Distance-Aware Positional Encoding (DAPE) and Bottom-Right anchoring (BR). We compare four model variants: a baseline with neither DAPE nor BR (using a standard positional encoding), a model with DAPE only(on CCA), a model with BR anchoring only(on CCA), and the full DAPE-BR.Their qualitative attention patterns are visualised in Figure 4. And as shown in Table 4 both components independently improve performance on object presence queries each alone reduces hallucination rates compared to the no-DAPE/BR baseline.

Complementary effect. In particular, using distance-aware encoding (without BR) already yields higher accuracy on POPE(**Accuracy +0.71% F1 Score +0.73%**), indicating that encoding relative spatial distances helps the model distinguish objects and avoid confusion. Similarly, applying the bottom-right anchoring (without DAPE) provides a boost(**Accuracy +0.47% F1 Score +0.59%**), which suggests that changing the coordinate reference frame can make the positional indices more informative for the model.Most importantly, the combination of DAPE + BR achieves the best results, outperforming either component alone across all POPE metrics (**Accuracy +1.75% F1 Score +1.74%**).

Analysis of results. This demonstrates that DAPE and BR complement each other: distance-aware encoding and anchor-shifted coordinates together provide the model with a richer and more distinct positional signal, leading to the largest reduction in hallucinations. We hypothesize that the DAPE component improves positional index separability i.e., it ensures that each objects positional embedding carries unique distance-based information, making it easier for the model to tell objects apart and not hallucinate one for another. Meanwhile, the BR anchoring aligns the coordinate system with the models internal RoPE (Rotary Position Embedding) representation, which can simplify the geometric learning problem for the transformer. By anchoring positions at the bottom-right, the spatial embeddings may better synchronize with how RoPE encodes angles and distances, thus enhancing the models ability to attend to the correct regions. Together, these effects explain why DAPE-BR yields the lowest object hallucination: it provides a more discriminative and well-aligned positional encoding scheme, enabling the LVLMs to stay grounded in the actual image content.

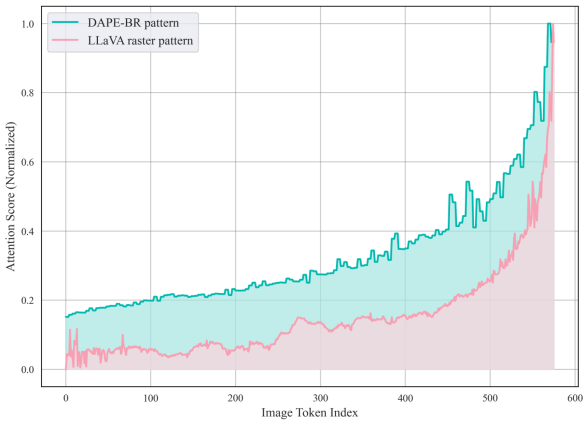


Figure 5: We run LLaVA-1.5-7B and DAPE-BR on 3 000 imagequery pairs from POPE (Adversarial split), extract the self-attention flowing *from every image token to the instruction token* in the first decoder layer, average those values across heads, queries, and images, then linearise the resulting 24×24 patch grid in standard raster order and min-max normalise the scores to $[0, 1]$.

Scale selection. In the following Figure 5,DAPE-BR keeps just $H \times W/4$ positional IDs the amount that, on LLaVAs pink curve (which uses the full $H \times W$ IDs), attain an attention score above 0.2. These IDs are then evenly re-allocated across all image tokens, yielding the turquoise curve. In this way we avoid over-focusing attention on patches adjacent to the instruction token, retain as much token discriminability as possible, and at the same time lessen RoPEs long-term decay.

6 Conclusion

We propose DAPE-BR, a positional-alignment scheme that mitigates object hallucination in LVLMs. It adds three light componentsorder-consistent re-indexing, fused distance-aware patch encoding, and reverse causal maskingwithout changing pretrained weights. Across POPE, MM-Star, and SQA it surpasses prior methods; each module contributes, and their combination yields the largest gains. Thus, precise positional realign-

ment,rather than extra data or fine-tuning, markedly improves grounding and is expected to extend to deeper layers, longer sequences, and other multimodal models.

Limitations

Although **DAPE-BR** markedly suppresses object hallucination, the technique also unveils several *opportunities for future research*. We frame these not as flaws, but as natural extensions that could amplify the methods impact:

- **Scalability to longer contexts.** In principle, orderconsistent re-indexing should extend to deeper transformer stacks and longer token streams, yet rigorous tests on lengthy imagedialog sequences remain to be carried out.
- **Finer or adaptive patch indexing.** Allowing the shell width Δ or the anchor position to adapt dynamically could yield even more precise spatial grounding for objects that span multiple patches.
- **Robustness in edgecase layouts.** Extremely atypical visual compositions may still challenge the current alignment assumptions. Developing additional safeguards or theoretical guarantees is especially important for high-stakes deployments.
- **Data diversity and fairness.** Our evaluation relies on popular benchmarks; validating DAPE-BR on larger, more heterogeneous image corpora will help reveal any hidden biases and verify generalizability.
- **Complementary enhancements.** DAPE-BR tackles positional alignment but not fluency or fine-grained grounding. Combining it with explicit regionword linking or caption-quality refinements could produce fully grounded and eloquent outputs.

Taken together, these directions highlight how *precise positional realignment* can serve as a foundation for continued progressscaling to longer sequences, broader models, and richer taskswithout the heavy cost of additional data or fine-tuning.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. 2024. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2024. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. 2024. Visionllama: A unified llama backbone for vision tasks. In *European Conference on Computer Vision*, pages 1–18. Springer.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

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.

Zhiheng Huang, Davis Liang, Peng Xu, and Bing Xiang. 2020. Improve transformer models with better relative position embeddings. *arXiv preprint arXiv:2009.13658*.

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*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024a. 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.

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

Guolin Ke, Di He, and Tie-Yan Liu. 2020. Rethinking positional encoding in language pre-training. *arXiv preprint arXiv:2006.15595*.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.

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.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. 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. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. 2024. Fit: Flexible vision transformer for diffusion model. *arXiv preprint arXiv:2402.12376*.

Jiahao Nie, Gongjie Zhang, Wenbin An, Yap-Peng Tan, Alex C Kot, and Shijian Lu. 2024. Mmrel: A relation understanding dataset and benchmark in the mllm era. *arXiv preprint arXiv:2406.09121*.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

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

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Feilong 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 *Proceedings of the International Conference on Learning Representations (ICLR)*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, and 1 others. 2024. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. 2025. RITUAL: Random image transformations as a universal anti-hallucination lever in LVLMs. *arXiv preprint arXiv:2405.17821*.

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

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

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

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

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

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*.

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*.

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

Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kening Zheng, Junkai Chen, Chang Tang, and Xuming Hu. 2025. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.