See Fair, Speak Truth: Equitable Attention Improves Grounding and Reduces Hallucination in Vision-Language Alignment

Anonymous ACL submission

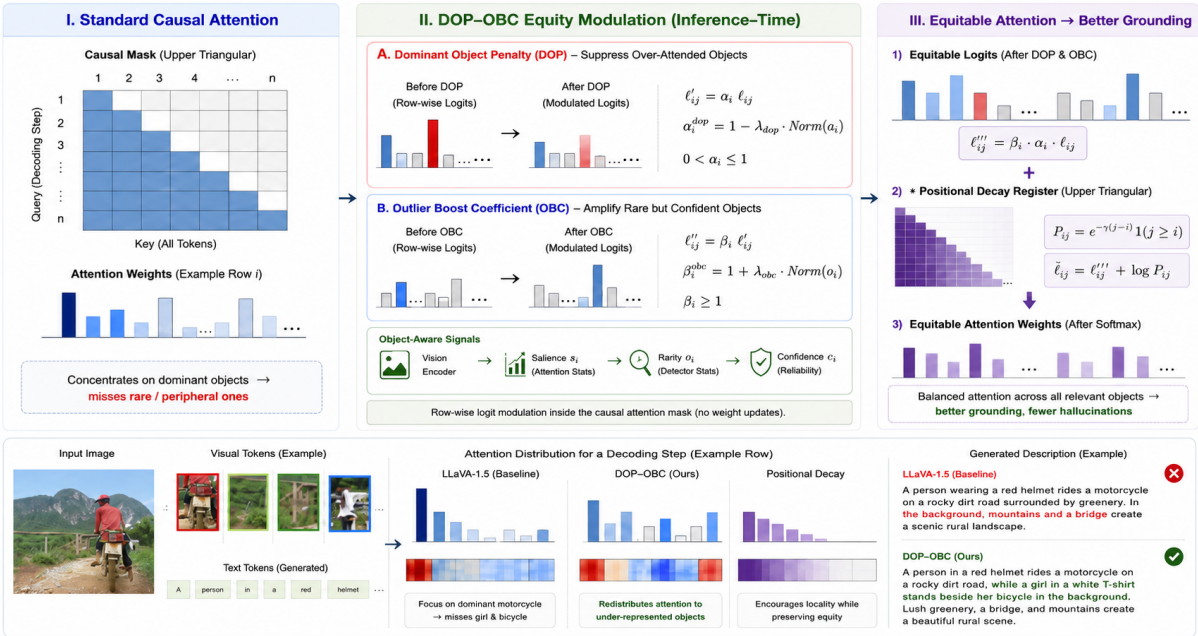


Figure 1: DOP–OBC equitable attention modulation during autoregressive decoding. DOP suppresses over-attended dominant objects, while OBC amplifies rare yet confident objects through row-wise logit modulation. Combined with an upper-triangular positional decay register, the method rebalances causal attention at inference time, improving grounding of under-represented visual entities without weight updates.

Abstract

Multimodal large language models (MLLMs) frequently hallucinate objects that are absent from the visual input, often because attention during decoding is disproportionately drawn to visually dominant or frequently occurring content. We observe that this inequity in attention allocation is a root cause of object hallucination: when rare, small, or contextually peripheral objects receive insufficient attention, the model fails to ground its generation in the full visual scene. We argue that every object in an image, regardless of its size, frequency or salience, deserves equal representational opportunity during decoding. To this end, we propose **DOP-OBC**, a training-free and architecture-agnostic decoding strategy built on the principle of *equitable attention*. Two complementary object-aware signals work in tandem: a **Dominant Object Penalty (DOP)** that softly suppresses attention over-concentration on visually dominant regions, and an **Outlier Boost Coefficient (OBC)** that amplifies attention toward rare yet confidently detected objects. These signals are injected as per-row logit modulations within the causal attention mask, requiring no weight updates and preserving autoregressive decoding properties. Extensive experiments across image and video MLLMs demonstrate consistent reductions in object hallucination on CHAIR and POPE benchmarks, alongside improvements in GPT-4o-assessed captioning quality across correctness, consistency, detail, context and temporal dimensions. DOP-OBC establishes that fairness in attention allocation is not merely a design principle but a practical

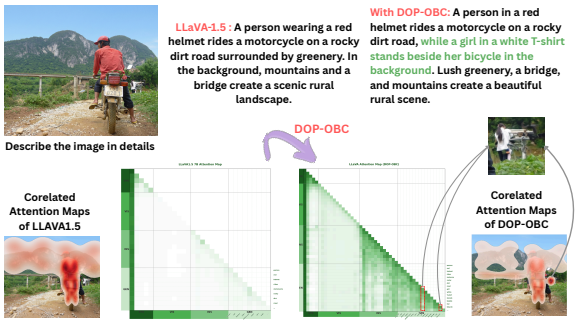


Figure 2: Example showing that DOP–OBC yields a more complete and grounded description than the LLaVA-1.5 baseline. The method recovers correct background entities, highlighted in green, while the attention maps show reallocation of attention toward previously neglected objects.

and effective path toward more faithful multimodal generation.

1 Introduction

Multimodal Large Language Models (MLLMs) have made substantial progress in vision–language understanding, including image captioning, visual question answering, instruction following, and multimodal reasoning (Liu et al., 2024, 2023a; Dai et al., 2023). By coupling strong vision encoders with large language models, they can generate fluent and contextually coherent visual descriptions. However, their outputs are often only partially grounded in the image. A major failure mode is *object hallucination*, where models either omit visible objects or confidently mention objects absent from the scene (Rohrbach et al., 2018; Li et al., 2023). Its persistence across models and benchmarks indicates that the problem is not merely representational, but is closely tied to how multimodal evidence is integrated and decoded.

Figure 2 illustrates this imbalance. In the example, **LLaVA-1.5** correctly describes the dominant motorcycle and rider, but overlooks a clearly visible girl standing beside her bicycle. The corresponding attention maps reveal the cause: although the visual encoder activates over multiple regions, the decoder increasingly concentrates attention on the dominant foreground object. In the green attention-weight plots, most visual tokens receive uniformly low weights, while tokens associated with the motorcycle accumulate strong and persistent attention. This concentration intensifies during decoding, causing attention collapse and suppressing peripheral but visually present objects. Thus, the failure is not due to missing visual evidence in the encoder; rather, it arises during decoding, where attention allocation is biased toward large, salient, and frequent objects because of token abundance and language priors. Uniformly redistributing attention is insufficient, since equality assigns identical resources, whereas equity compensates for differences in object scale, salience, and frequency.

To mitigate this issue, we propose **DOP–OBC**, a training-free and architecture-agnostic decoding strategy that rebalances attention at inference time. As shown in Figure 1, DOP–OBC introduces two complementary object-aware controls into attention computation. The **Dominant Object Penalty (DOP)** softly suppresses over-attended objects by reducing their row-wise logit amplitude, while the **Outlier Boost Coefficient (OBC)** amplifies rare but confidently detected objects through calibrated scaling. These signals are combined with an upper-triangular positional decay register, preserving autoregressive causality while stabilizing long-range grounding. In Figure 2, this mechanism reduces excessive attention to the motorcycle and increases attention to underrepresented girl and bicycle. The resulting attention maps exhibit a more balanced and sustained distribution across decoding steps, enabling a more faithful description of all relevant objects. Crucially, DOP–OBC operates entirely at inference time and requires no retraining, auxiliary models, or architectural modification.

The main contributions of this paper are as :

- We identify a systematic attention imbalance in MLLMs: dominant objects consume excessive decoding capacity, suppressing small, rare, or peripheral objects and contributing to object hallucination.
- We propose **DOP–OBC**, a training-free, architecture-agnostic decoding method that rebalances attention using two plug-in signals: Dominant Object Penalty for over-attended objects and Outlier Boost Coefficient for rare but confidently detected objects, applied as per-row modulations within an upper-triangular attention register.
- We validate DOP–OBC across image and video MLLMs, reducing hallucination on CHAIR and POPE and improving all five GPT-4o-evaluated video captioning dimensions, demonstrating the effectiveness of equitable inference-time attention allocation.

2 Related Work

2.1 Vision-language models

Recent progress in multimodal large language models (MLLMs) has substantially advanced vision–language integration, improving multimodal reasoning, instruction following, and conversational capabilities (Liu et al., 2023a, 2024; Dai et al., 2023). LLaVA aligns a pretrained LLM with a visual encoder through instruction tuning, enabling open-ended image-conditioned generation (Liu et al., 2023a). Flamingo introduces an interleaved visual-text token architecture that supports few-shot multimodal reasoning (Alayrac et al., 2022). InstructBLIP extends BLIP-2 with instruction tuning to enhance image-grounded dialogue and general multimodal understanding (Dai et al., 2023). For video understanding, Video-LLaVA and related frameworks incorporate temporal modeling into LLaVA-style architectures to enable coherent video dialogue and reasoning (Lin et al., 2024). Despite these advances, object hallucination and grounding failures remain persistent challenges. Rohrbach et al. (Rohrbach et al., 2018) first systematically analyzed object hallucination in image captioning, showing that language priors often override visual evidence. POPE (Li et al., 2023) further demonstrates that modern MLLMs frequently generate objects unsupported by the input image, even when confidence is high. More recently, FarSight (Tang et al., 2025) identifies decoding-time attention collapse and positional drift as key contributors to hallucination in LLaVA-style models. These findings indicate that hallucination is not solely a representation issue but is closely tied to attention allocation during decoding, motivating inference-time attention rebalancing strategies such as ours.

2.2 Contrastive decoding and attention modification

Contrastive decoding is an inference-time strategy for reducing object hallucination by contrasting predictions from original and perturbed inputs. Visual Contrastive Decoding (VCD) contrasts logits from clean and distorted visual inputs to reduce language-prior bias and improve grounding without retraining (Leng et al., 2024a; Yin et al., 2025). Recent extensions steer attention, as in ASCD (Wang et al., 2025b), target hallucination-related frequency components, as in MFCD (Liu et al., 2025), or contrast image-focused intermediate layers, as in iTaD (Xu et al., 2025). Other methods amplify visual signals during modality fusion to counteract language bias (Yin et al., 2025). Unlike these approaches, DOP-OBC rebalances object-level attention through dominance suppression and rarity-aware boosting at inference time.

2.3 Externally guided decoding

Another line of work reduces hallucination by incorporating external guidance at inference time. CLIP-guided decoding uses pretrained contrastive vision–language embeddings to steer generation toward visually grounded token predictions, improving alignment between images and captions (Deng et al., 2024). Summary-guided decoding modifies the decoding process by using concise summaries to focus the model on image-relevant content, which can reduce language prior dominance (Min et al., 2025). Interaction guidance sampling explicitly guides the model to leverage multimodal interaction information during generation, reducing spurious predictions without additional training (Dong et al., 2025). Other approaches train auxiliary estimators, such as value models, to score candidate continuations and bias search toward low-hallucination outputs (Wang et al., 2025a; Deria et al., 2025). These externally guided strategies differ from our approach by relying on additional models or external signals; in contrast, we derive guidance solely from internal object proposals and attention statistics to provide targeted, object-level control at inference time.

3 Methodology

We propose a fully inference-time decoding strategy that improves attention allocation without any finetuning or architectural modification. The method operates directly within the attention computation and consists of three components. First, attention registers absorb surplus attention assigned to future positions, steering decoding toward contextually informative tokens (Sec. 3.1). Second, a progressively diminishing masking rate introduces absolute positional focus while preserving autoregressive causality (Sec. 3.2). Third, to ensure equitable allocation across long-tailed objects, we introduce two lightweight object-aware controls: the Dominant Object Penalty (DOP) and the Outlier Boost Coefficient (OBC), which rebalance attention between dominant and under-represented objects (Sec. 3.3).

3.1 Upper-Triangular Attention Registers

To mitigate attention collapse, we introduce an upper-triangular attention register that absorbs surplus mass assigned to future positions. Let $\boldsymbol{\omega} \in \mathbb{R}^{n\times n}$ denote the query–key score matrix and let $\mathbf{C} = \mathrm{tril}(\mathbf{1})$ be the standard lower-triangular causal mask. We modify the logits as

$$\mathbf{W} = \boldsymbol{\omega}\mathbf{C} + \mathbf{P}, \qquad \mathbf{P}_{i,j} = \begin{cases} -(j-i)\sigma, & j > i, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $\mathbf{P} \in \mathbb{R}^{n\times n}$ is an upper-triangular matrix with linear decay slope $\sigma > 0$. The term $\mathbf{P}$ provides a controlled reservoir for excess logits toward future tokens, preventing unintended redistribution over valid positions after normalization.

The masked attention is then computed as

$$\widetilde{\mathbf{W}} = \mathrm{SoftMax}(\mathbf{W})\mathbf{C}. \tag{2}$$

Remark. The mask $\mathbf{C}$ in Eq. (1) constrains logits before normalization, whereas the mask in Eq. (2) removes residual future mass after normalization, ensuring strictly causal decoding.

3.2 Positional Awareness Encoding

The linear decay in $\mathbf{P}$ produces a *progressively diminishing* allocation to future tokens across rows: as the row index increases, the normalized mass distributed over valid (past) tokens grows monotonically. This yields absolute positional awareness without altering model parameters. Practically, later tokens aggregate more historical context while future positions are softly discouraged, stabilizing long-range generation.

3.3 Equity-Aware Object Controls (DOP & OBC)

To reallocate decoding capacity toward small and rare-but-valid objects, we introduce two object-aware signals that act through the same mechanisms as the attention register: a row-wise amplitude α_i that scales pre-SoftMax logits and a row-wise decay σ_i that controls the register slope. Each attention row i is associated with an object proposal $o(i)$ via the same ROI/patch assignment used during routing; rows without an aligned object retain default parameters.

Dominant Object Penalty (DOP). For each object o, we compute a normalized dominance score

$$D(o) = w_1\hat{s}(o) + w_2\hat{L}(o) + w_3\hat{A}(o), \tag{3}$$

Algorithm 1 DOP–OBC Decoding (Single Decoder Layer)

Require: Hidden states $\mathbf{x}$; causal mask $\mathbf{C}$ (lower-triangular)
Require: Object set $\mathcal{O}$ with features $\{\text{size}, \text{len}, \text{attn}, \text{conf}\}$
Require: Row-to-object alignment map $\pi(i)$
Require: Hyperparameters $\alpha_0, \sigma_0, \mathbf{w} = [w_1, w_2, w_3], \lambda, \gamma, \tau_p, \tau_r, r_{\max}$
Ensure: Updated decoder output $\mathbf{y}$

1. QKV Projection and Raw Attention Scores
$(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \leftarrow \texttt{qkv_proj}(\mathbf{x})$
$\mathbf{S} \leftarrow (\mathbf{Q}\mathbf{K}^\top)/\sqrt{d_k}$

2. Dominant Object Penalty (DOP)
Normalize object statistics:

$$\hat{\mathbf{s}}, \hat{\mathbf{L}}, \hat{\mathbf{A}} \leftarrow \texttt{minmax}(\text{size}),\ \texttt{minmax}(\text{len}),\ \texttt{minmax}(\text{attn})$$

$\mathbf{D} \leftarrow w_1\hat{\mathbf{s}} + w_2\hat{\mathbf{L}} + w_3\hat{\mathbf{A}}$
$\mathbf{w}_{\text{pen}} \leftarrow \exp(-\lambda\mathbf{D})$

3. Outlier Boost Coefficient (OBC)
$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leftarrow \texttt{ema_gaussian}(\text{feat})$
$\mathbf{r} \leftarrow \texttt{clamp}\Big(\frac{\texttt{mahalanobis}(\text{feat}, \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\tau_r},\ 0,\ r_{\max}\Big)$
$\mathbf{b} \leftarrow \gamma\mathbf{r} \mathbb{1}[\text{conf} \geq \tau_p]$

4. Row-wise Modulation and Register Construction
Initialize $\boldsymbol{\alpha} \leftarrow \alpha_0\mathbf{1}, \quad \boldsymbol{\sigma} \leftarrow \sigma_0\mathbf{1}$
for $i = 1$ to n **do**
  $o \leftarrow \pi(i)$
  if $o \geq 0$ **then**
    $\alpha_i \leftarrow \alpha_i \cdot \mathbf{w}_{\text{pen}}[o] \cdot (1 + b_o)$
    $\sigma_i \leftarrow \sigma_i/(1 + b_o)$
  end if
end for
$\mathbf{P} \leftarrow \texttt{build_register}(n, \boldsymbol{\sigma})$

5. Attention Composition and Output
$\mathbf{S} \leftarrow (\mathrm{diag}(\boldsymbol{\alpha})\,\mathbf{S}) \odot \mathbf{C} + \mathbf{P}$
$\mathbf{A} \leftarrow \mathrm{SoftMax}(\mathbf{S}) \odot \mathbf{C}$
$\mathbf{y} \leftarrow \texttt{wo}(\mathbf{A}\mathbf{V})$
return $\mathbf{y}$

where $w_i \geq 0$ and $\sum_i w_i = 1$. Here, $\hat{s}(o)$ denotes normalized object size, $\hat{L}(o)$ its temporal persistence, and $\hat{A}(o)$ the attention share already received, each min–max normalized within the current window. A soft penalty

$$w_{\text{pen}}(o) = \exp\big(-\lambda D(o)\big), \quad \lambda \geq 0, \tag{4}$$

attenuates attention allocated to dominant objects, freeing representational capacity for less prominent instances.

Outlier Boost Coefficient (OBC). To promote rare yet reliable objects, we estimate a rarity score $r(o)$ using a clipped, temperature-scaled distance between object features and an exponential moving average scene distribution. The boost is gated by posterior confidence $p(o)$:

$$b(o) = \gamma\, r(o)\, \mathbf{1}[p(o) \geq \tau_p]\,, \tag{5}$$

so that only rare and confident objects receive amplification.

Integration into attention. For each row i linked to object $o(i)$, the controls are applied as

$$\alpha_i \leftarrow \alpha_0\, w_{\text{pen}}(o(i))\,\big(1 + b(o(i))\big), \qquad \sigma_i \leftarrow \frac{\sigma_0}{1 + b(o(i))}. \tag{6}$$

We replace $\boldsymbol{\omega}$ with $\mathrm{Diag}(\boldsymbol{\alpha})\,\boldsymbol{\omega}$ in Eq. (1) and construct the register $\mathbf{P}$ using per-row slopes $\{\sigma_i\}$, while Eq. (2) remains unchanged.

In effect, DOP dampens attention to already dominant objects, and OBC amplifies and prolongs attention to rare but credible ones, ensuring that long-tail objects are less likely to be suppressed during decoding.

3.4 Decoder Layer Composition

Each decoder layer applies the following sequence: (i) QKV projection; (ii) row-wise logit scaling via $\mathrm{Diag}(\boldsymbol{\alpha})$; (iii) register construction with per-row decay; (iv) masked SoftMax normalization; and (v) value aggregation and projection, as summarized in Alg. 1. An optional pre-routing step rescales proposal scores using the same DOP and OBC signals to align selection with attention reallocation.

3.5 Practical Notes and Defaults

Per-window normalization uses a small ε for stability. EMA density is vectorizable; kNN can be substituted. Default knobs: $w{=}\langle 0.5, 0.25, 0.25\rangle$, $\lambda{=}1.0$, $\tau_p{=}0.5$, $\gamma{=}0.3$, $r_{\max}{=}2.0$. Safeguards: clip b, floor α_i, and bound $\sigma_i \in [\sigma_{\min}, \sigma_0]$. The added cost is $O(|\mathcal{P}_t| + n)$; the attention complexity is unchanged.

4 Experiments

We evaluate our equity-aware decoding framework on both *image* and *video* benchmarks, following established protocols for hallucination mitigation and instruction-tuned MLLMs (Liu et al., 2023a; Rohrbach et al., 2018; Li et al., 2023; Tang et al., 2025). Unless otherwise specified, we adopt dataset splits and prompts, and we retain each base model's default decoding settings, including temperature. Our method operates as a training-free, drop-in plug-in applied solely at inference time.

4.1 Implementation Details

We keep all model weights frozen and apply our plug-ins at decode time. Object proposals/tracks and confidences come from each model's standard vision stack; our row-wise amplitude/decay are computed on-the-fly and vectorized. Default knobs (λ, γ, τ_p etc.) are as in Sec. 3. No extra training data or finetuning is used.

4.2 Benchmarks and Metrics

Image Benchmarks. We evaluate general multimodal reasoning and hallucination behavior on widely adopted image benchmarks, with quantitative results reported in Table 2 and cross-benchmark improvements illustrated in Figure 3. The evaluation suite includes MMBench (Liu et al., 2023b), LLaVA-Bench-in-the-Wild (LLaVA-W) (Liu et al., 2023a), MM-Vet (Yu et al., 2023), VizWiz (Gurari et al., 2018), ScienceQA (SQA) (Lu et al., 2022), and the hallucination-focused CHAIR and POPE datasets (Rohrbach et al., 2018; Li et al., 2023). For CHAIR, we report CHAIRS ↓ and CHAIRI ↓, where lower values indicate fewer hallucinations. For POPE, we report Recall, Precision, and Accuracy (R/P/A) ↑, where higher values reflect stronger visual grounding and reduced object-level hallucination.

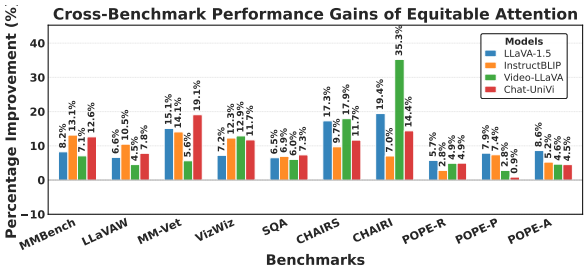


Figure 3: Cross-benchmark percentage improvement of DOP–OBC over base models across general multimodal benchmarks and hallucination metrics, showing consistent gains in task performance and reductions in CHAIR/POPE errors.

Video benchmarks. For video question answering, we use MSVD-QA (Xu et al., 2017) and ActivityNet-QA (Yu et al., 2019). For open-ended video captioning, we follow recent practice and employ GPT-4o as an automatic evaluator, reporting five dimensions: Correctness (Cr.), Consistency (Cs.), Detail (De.), Context (Ct.), and Temporal understanding (Te.), with higher scores indicating better performance (see Tab. 1).

4.3 Backbone Models

Image MLLMs: We evaluate representative instruction-tuned image MLLMs. LLaVA-1.5 (Liu et al., 2024) extends LLaVA with improved visual instruction tuning based on CLIP–Vicuna, but relies on standard RoPE positional encoding and can exhibit long-context drift and attention collapse,

Table 1: Video benchmarks. For Video-Based Text Generation we report GPT-4o–assessed metrics: Correctness (Cr.), Consistency (Cs.), Detail (De.), Context (Ct.), Temporal (Te.). Higher is better.

Base Model	Method	MSVD-QA		ActivityNet-QA		Video-Based Text Generation				
		Accuracy ↑	Score ↑	Accuracy ↑	Score ↑	Cr. ↑	Cs. ↑	De. ↑	Ct. ↑	Te. ↑
Char-UniVi (Jin et al., 2023)	Base	64.6	3.6	43.1	3.2	2.84	2.93	2.55	3.16	2.43
	DOP-OBC	**69.2**$^{(+4.6)}$	**4.0**$^{(+0.4)}$	43.0$^{(-0.1)}$	**3.6**$^{(+0.4)}$	**2.90**$^{(+0.06)}$	**3.00**$^{(+0.07)}$	**2.62**$^{(+0.07)}$	**3.30**$^{(+0.14)}$	**2.56**$^{(+0.13)}$
Video-LLaVA (Lin et al., 2024)	Base	64.8	3.7	41.5	3.3	2.32	2.34	2.65	2.75	2.09
	DOP-OBC	**72.2**$^{(+7.4)}$	**4.1**$^{(+0.4)}$	**44.2**$^{(+2.7)}$	**3.8**$^{(+0.5)}$	**2.97**$^{(+0.65)}$	**2.65**$^{(+0.31)}$	**3.12**$^{(+0.47)}$	**3.21**$^{(+0.46)}$	**2.47**$^{(+0.38)}$
VILA (Lin et al., 2023a)	Base	72.6	4.0	50.2	3.3	3.14	3.40	2.71	3.43	2.58
	DOP-OBC	**76.9**$^{(+4.3)}$	**4.7**$^{(+0.7)}$	**53.2**$^{(+3.0)}$	**3.9**$^{(+0.6)}$	**3.23**$^{(+0.09)}$	**3.76**$^{(+0.36)}$	**2.83**$^{(+0.12)}$	**3.77**$^{(+0.34)}$	**2.78**$^{(+0.20)}$
Video-LLAMA2 (Cheng et al., 2024)	Base	70.9	3.8	49.9	3.3	3.13	3.23	2.70	3.42	2.45
	DOP-OBC	**74.3**$^{(+3.4)}$	**4.5**$^{(+0.7)}$	**52.2**$^{(+2.3)}$	**4.0**$^{(+0.7)}$	**3.35**$^{(+0.22)}$	**3.42**$^{(+0.19)}$	**3.37**$^{(+0.67)}$	**3.89**$^{(+0.47)}$	**2.56**$^{(+0.11)}$

Table 2: Image benchmarks. Lower is better for **CHAIRS/CHAIRI**; higher is better for all others.

Base Model	Method	MMBench ↑	LLaVAW ↑	MM-Vet ↑	VizWiz ↑	SQA ↑	CHAIRS ↓	CHAIRI ↓	POPE-R ↑	POPE-P ↑	POPE-A ↑
LLaVA-1.5 (Liu et al., 2024)	Base	64.3	72.5	30.5	48.5	64.5	48.0	13.9	87.0	82.8	76.6
	ICD	63.1	69.7	30.4	46.9	62.8	47.7	13.6	87.9	84.0	80.2
	VCD	63.9	70.9	29.5	43.4	63.3	46.8	13.2	87.0	83.5	78.1
	OPERA	64.4	72.0	31.4	50.0	64.9	45.2	12.7	88.8	82.8	79.2
	FarSight	66.0	74.7	32.5	50.8	67.4	41.6	13.2	90.5	86.1	80.4
	DOP-OBC	**69.6**$^{(+5.3)}$	**77.3**$^{(+4.8)}$	**35.1**$^{(+4.6)}$	**52.0**$^{(+3.5)}$	**68.7**$^{(+4.2)}$	**39.7**$^{(+8.3)}$	**11.2**$^{(+2.7)}$	**92.0**$^{(+5.0)}$	**89.3**$^{(+6.5)}$	**83.2**$^{(+6.6)}$
Video-LLaVA (Lin et al., 2024)	Base	60.9	73.1	32.0	48.1	64.6	50.2	15.6	81.6	85.3	86.2
	DOP-OBC	**65.2**$^{(+4.3)}$	**76.4**$^{(+3.3)}$	**33.8**$^{(+1.8)}$	**54.3**$^{(+6.2)}$	**68.5**$^{(+3.9)}$	**41.2**$^{(+9.0)}$	**10.1**$^{(+5.5)}$	**85.6**$^{(+4.0)}$	**87.7**$^{(+2.4)}$	**90.2**$^{(+4.0)}$
Chat-UniVi(Jin et al., 2023)	Base	56.3	70.4	28.3	46.9	59.9	52.3	16.7	85.1	69.5	64.4
	DOP-OBC	**63.4**$^{(+7.1)}$	**75.9**$^{(+5.5)}$	**33.7**$^{(+5.4)}$	**52.4**$^{(+5.5)}$	**64.3**$^{(+4.4)}$	**46.2**$^{(+6.1)}$	**14.3**$^{(+2.4)}$	**89.3**$^{(+4.2)}$	**70.1**$^{(+0.6)}$	**67.3**$^{(+2.9)}$
InstructBLIP (Dai et al., 2023)	Base	43.4	58.2	25.6	33.4	62.1	55.6	24.2	88.7	81.3	74.4
	DOP-OBC	**49.1**$^{(+5.7)}$	**64.3**$^{(+6.1)}$	**29.2**$^{(+3.6)}$	**37.5**$^{(+4.1)}$	**66.4**$^{(+4.3)}$	**50.2**$^{(+5.4)}$	**22.5**$^{(+1.7)}$	**91.2**$^{(+2.5)}$	**87.3**$^{(+6.0)}$	**78.3**$^{(+3.9)}$

which contribute to object hallucination. InstructBLIP (Dai et al., 2023) combines a frozen vision encoder with a Q-Former and an instruction-tuned LLM. While broadly capable, it inherits common LLM issues such as hallucination and miscalibration, as noted by its authors (Dai et al., 2023).

Video MLLMs: For video, we consider Video-LLaVA (Lin et al., 2024), VILA (Lin et al., 2023b), Video-LLaMA2 (Cheng et al., 2024), and Chat-UniVi (Jin et al., 2023). Video-LLaVA adapts LLaVA to video via frame sampling and temporal adapters but inherits similar positional and attention biases. VILA emphasizes unified vision–language pretraining improvements but does not address inference-time attention allocation. Video-LLaMA2 strengthens visual backbones and alignment for video understanding, yet relies on standard positional schemes that can drift over long temporal spans. Chat-UniVi unifies multimodal instruction following for video, but like prior MLLMs, lacks decoding-time mechanisms to balance attention across dominant and long-tail objects. As a result, visually salient content can monopolize decoding capacity while smaller or rarer objects remain underrepresented.

Decoding-time baselines: We additionally compare against training-free inference controls designed to improve grounding. ICD (Wang et al., 2024) biases decoding toward image evidence, VCD (Leng et al., 2024b) integrates visual features through contrastive decoding, and OPERA (Huang et al., 2024) optimizes retrieval of visual evidence during generation. While these approaches enhance faithfulness, they do not explicitly redistribute attention mass across objects, nor do they address attention collapse or positional decay effects identified by FarSight (Tang et al., 2025).

Table 3: CHAIR/POPE comparison. C_S and C_I denote $CHAIR_S$ and $CHAIR_I$; P-R and P-P denote POPE-Random and POPE-Popular. ↓ lower is better; ↑ higher is better.

Model	Method	C_S ↓	C_I ↓	P-R ↑	P-P ↑
LLaVA-1.5	RoPE	48.0	13.9	87.0	82.8
	FixVPE	47.3	13.4	87.5	84.7
	EDVT	46.8	14.5	87.8	85.4
	FarSight	41.6	13.2	90.5	86.1
	DOP–OBC	**39.7**$_{-8.3}$	**11.2**$_{-2.7}$	**92.0**$_{+5.0}$	**89.3**$_{+6.5}$
Video-LLaVA	RoPE	50.2	15.6	81.6	85.3
	FixVPE	48.5	14.9	81.9	85.2
	EDVT	46.8	13.7	82.5	84.7
	FarSight	44.8	12.9	83.2	85.8
	DOP–OBC	**41.2**$_{-9.0}$	**10.1**$_{-5.5}$	**85.6**$_{+4.0}$	**87.7**$_{+2.4}$

4.4 Quantitative Results

Hallucination Diagnostics (CHAIR & POPE): Table 3 isolates CHAIR/ POPE under different positional schemes. Relative to RoPE and prior

fixes (FixVPE/EDVT), our method achieves reductions in **CHAIRS/CHAIRI** and increases in **POPE-R/POPE-P** on both LLaVA-1.5 and Video-LLaVA. The decreases in CHAIRI (instance-level) are notable, aligning with our goal of reallocating attention to small/rare objects rather than merely suppressing text length.

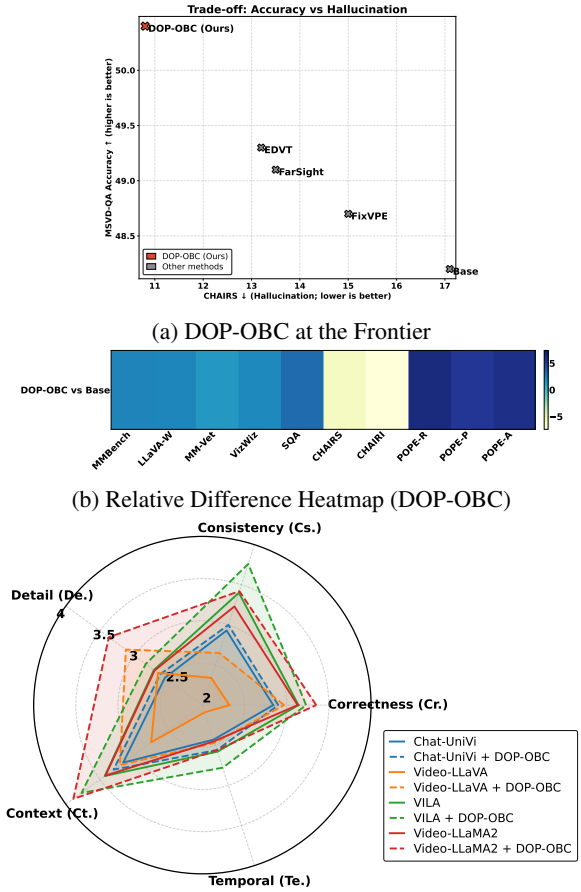
Video QA and Open-Ended Generation: Table 1 reports video QA accuracy/score on MSVD-QA and ActivityNet-QA, and GPT-4o-assessed text quality. Our method improves accuracy and open-ended scores for all four video MLLMs. Gains on **Cr./Cs.** suggest fewer factual inconsistencies. Improvements on **Te.** indicate better temporal grounding, consistent with our per-row decay that sustains evidence across steps.

Image Benchmarks: Table 2 summarizes image-level results. Across general-ability suites (MMBench, LLaVA-W, MM-Vet, VizWiz, SQA) our method improves over the base models, while substantially reducing object-hallucination scores (CHAIR/CHAIR$_I$) and improving POPE. For example, on LLaVA-1.5 we observe gains on hallucination diagnostics (e.g., ↓CHAIRS and ↓CHAIRI) alongside increases on POPE-R/P/A, indicating stronger grounding rather than over-conservative outputs. InstructBLIP and Video-LLaVA (image mode) show similar trends, confirming the plug-and-play nature of our approach.

Holistic Performance Analysis: While Tables 2–1 report individual metrics, Figure 4 provides a consolidated view of DOP–OBC's impact. As shown in Figure 4(a), our method defines a new empirical frontier in the accuracy–hallucination plane, achieving higher accuracy with lower CHAIR scores compared to prior decoding fixes. Figure 4(b) illustrates consistent relative improvements across ten benchmarks, while Figure 4(c) shows uniform gains across correctness, consistency, detail, context, and temporal dimensions in video captioning. These trends indicate that equitable attention reallocation improves grounding without trading off general performance.

4.5 Qualitative Results

Figure 5 compares **LLaVA-1.5 (base)**, **FarSight** and **DOP–OBC** on a complex plaza scene containing both dominant (arch, food truck) and peripheral objects (traffic lights, street signs, red vehicle). LLaVA-1.5 captures the main structures but



(a) DOP-OBC at the Frontier

(b) Relative Difference Heatmap (DOP-OBC)

(c) Equity-Aware Quality Growth

Figure 4: **Comprehensive analysis of DOP–OBC's impact on multimodal performance.** (a) Accuracy (↑) vs. CHAIRS (↓): DOP–OBC defines the empirical frontier, achieving higher accuracy with lower hallucination. (b) Relative Difference Heatmap (DOP–OBC | Base): consistent gains across ten benchmarks, improving general metrics and POPE while reducing CHAIRS/CHAIRI. (c) Equity-Aware Quality Growth: outward shifts across Correctness, Consistency, Detail, Context, and Temporal dimensions indicate uniform improvements in video caption quality.

exhibits early attention collapse, leading to omission of peripheral yet clearly visible objects. FarSight improves object coverage but remains unstable across decoding steps, resulting in incomplete grounding. In contrast, **DOP–OBC** maintains stable image attention throughout generation and correctly grounds all GT-mentioned objects, including peripheral elements. These qualitative results confirm that equity-aware attention reallocation mitigates attention collapse and improves completeness and visual grounding without retraining or architectural modification.

Figure 5: **Qualitative comparison and attention dynamics during decoding. LLaVA-1.5 (base model)** exhibits early attention collapse, leading to omission of peripheral objects (e.g., traffic lights, street signs, red vehicle). FarSight partially mitigates this effect but remains unstable. **DOP–OBC** maintains stable image attention throughout generation, correctly grounding all GT-mentioned objects without architectural modification or retraining.

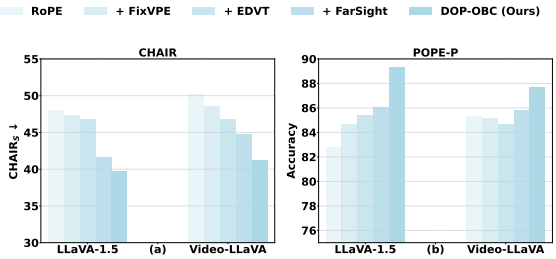


Figure 6: **CHAIR and POPE-P comparisons on LLaVA-1.5 and Video-LLaVA.** We compare CHAIR$_S$ (lower is better) and POPE-P (higher is better) across RoPE, FixVPE, EDVT, FarSight, and **DOP-OBC(Ours)**. **DOP-OBC(Ours)** yields the strongest gains on both backbones.

5 Discussion

Our results suggest that object hallucination in MLLMs is largely driven by decoding-time attention imbalance rather than missing visual perception. Across image and video backbones, **DOP–OBC** consistently reduces CHAIR/POPE hallucinations while improving VQA and captioning metrics, as shown in Table 2, Table 1, and Figure 6. This supports our claim that equitable attention reallocation improves grounding without degrading fluency or general reasoning. DOP–OBC works by reshaping decoder attention. DOP prevents dominant objects from monopolizing attention, while OBC preserves rare but confident object evidence across decoding steps. The positional decay register further stabilizes long-range grounding while maintaining causal masking. As shown qualitatively in Figure 5, this leads to more stable and evenly distributed attention than prior baselines. DOP–OBC operates entirely at inference time, requiring no retraining, auxiliary supervision, or architectural changes. However, its effectiveness depends on object proposal quality. In crowded scenes, suppression may become conservative and under-represent mid-frequency objects, while noisy proposals may occasionally amplify spurious content despite confidence gating. Future work should therefore improve proposal calibration and uncertainty-aware attention control. Overall, our findings show that better allocation of decoding capacity is a lightweight and effective route to more faithful multimodal generation.

6 Conclusions

We presented **DOP–OBC**, a lightweight, training-free decoding strategy that alleviates object hallucination in multimodal large language models by directly addressing structural imbalances in decoder attention. Instead of altering model weights, adding auxiliary supervision, or redesigning architectures, DOP–OBC rebalances attention at inference time via soft suppression of over-attended objects and calibrated amplification of rare yet confident object signals within a causal attention register. Extensive evaluation across diverse image and video models shows that DOP–OBC consistently reduces CHAIR and POPE hallucination rates while enhancing performance on standard captioning and VQA benchmarks, indicating improved visual grounding without sacrificing descriptive quality. These results support the view that hallucination in current MLLMs arises in large part from capacity allocation during decoding. By reallocating representational resources toward under-served visual entities, equity-aware attention control offers a simple and effective means to achieve more faithful and comprehensive multimodal generation.

7 Limitations

DOP–OBC relies on lightweight rarity and confidence signals, but does not model full uncertainty or proposal reliability. In crowded scenes with similar object confidence, dominance cues may weaken, and noisy detections may affect rarity-based boosts despite confidence gating. Since our method is limited to inference-time control, future work can improve uncertainty modeling and proposal calibration, extend equity-aware decoding to multi-image, long-video, and dialogue settings, and explore lightweight training objectives for multimodal grounding.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Wenliang Dai, Junnan Li, Dongxu Li, and 1 others. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*.

Ankan Deria, Adinath Madhavrao Dukre, Feilong Tang, Sara Atito, Sudipta Roy, Muhammad Awais, Muhammad Haris Khan, and Imran Razzak. 2025. Dual-stage value-guided inference with margin-based reward adjustment for fast and faithful vlm captioning. *arXiv preprint arXiv:2506.15649*.

Xin Dong, Shichao Dong, Jin Wang, Jing Huang, Li Zhou, Zenghui Sun, Lihua Jing, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. 2025. Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2534–2544.

Danna Gurari and 1 others. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*.

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 (CVPR)*, pages 13418–13427.

Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2023. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024a. 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.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024b. 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 (CVPR)*, 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. In *EMNLP*. POPE benchmark.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *EMNLP*.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023a. VILA: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023b. VILA: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*.

Bingqian Liu, Fu Zhang, Guoqing Chen, and Jingwei Cheng. 2025. Multi-frequency contrastive decoding: Alleviating hallucinations for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28556–28572.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *CVPR*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv:2304.08485*.

Haotian Liu and 1 others. 2023b. MMBench: Is your multi-modal language model getting better? *arXiv preprint arXiv:2307.06281*.

Pan Lu and 1 others. 2022. Learn to explain: Multi-modal reasoning via thought chains for science QA. In *NeurIPS*.

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2025. Mitigating hallucinations in large vision-language models via summary-guided decoding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4183–4198.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *EMNLP*.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zelin Peng, Zhiwei Yang, and 1 others. 2025. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *CVPR*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*.

Xiyao Wang, Zhengyuan Yang, Linjie Li, Hongjin Lu, Yuancheng Xu, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025a. Scaling inference-time search with vision value model for improved visual comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1184.

Yujun Wang, Jinhe Bi, Soeren Pirk, Yunpu Ma, and 1 others. 2025b. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. *arXiv preprint arXiv:2506.14766*.

Jun Xu and 1 others. 2017. MSVD-QA: A large-scale video QA dataset for open-ended understanding. In *ICCV*.

Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. 2025. Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1571–1590.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zhou Yu and 1 others. 2019. Activitynet-QA: A dataset for video question answering. In *CVPR*.