MAP: Map-Level Evidence Aggregation for Faithful Decoding in Multimodal Large Language Models

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) often generate fluent but visually unsupported responses, raising reliability concerns for grounded multimodal generation. Recent training-free decoding methods mitigate such hallucinations by intervening along either the layer or token dimension, but they typically overlook factual evidence distributed across the full decoding trajectory. In this work, we study MLLM hallucination from a map-level perspective by treating hidden states across layers and positions as a two-dimensional semantic map. Through a systematic logit-lens analysis, we show that visually grounded evidence is dispersed across intermediate layer-position states. Motivated by this observation, we propose **M**ap-Level **A**ttention **P**rocessing (MAP), a training-free decoding method that aggregates dispersed grounded evidence for faithful generation. MAP refines prediction representations with Layer-Wise Criss-Cross Attention and further combines local and global evidence through logit-level fusion. Experiments on closed- and open-ended hallucination benchmarks show that MAP consistently improves factuality across diverse MLLM architectures with modest decoding overhead.

1 Introduction

Multimodal Large Language Models (MLLMs), particularly vision-language models, have become a central paradigm for grounded multimodal generation. Through large-scale image-text pretraining and multimodal alignment, these models acquire broad visual-semantic knowledge and exhibit strong capabilities in cross-modal reasoning, instruction following, and visual understanding (Sun et al., 2023a; Li et al., 2021, 2025; Shu et al., 2025). Such capabilities make MLLMs increasingly useful for tasks that require natural-language responses grounded in visual inputs, such as Visual Question Answering (VQA) (Huynh et al., 2025) and Visual Grounding (VG) (Yan et al., 2024).

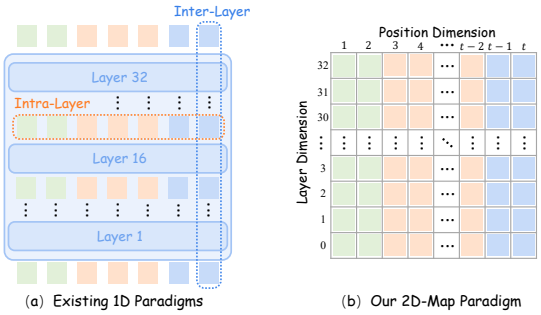

Figure 1: Comparison between single-dimensional decoding interventions and MAP. Existing methods mainly exploit either inter-layer or intra-layer information, while MAP treats hidden states across layers and positions as a semantic map to aggregate dispersed visually grounded evidence.

Despite these advances, hallucinations remain a major obstacle to reliable grounded generation in MLLMs. In vision-language settings, hallucinations refer to generated content that is fluent and plausible but inconsistent with the visual input, such as non-existent objects, incorrect attributes, erroneous spatial relations, or inaccurate quantities (Li et al., 2023; Zhao et al., 2025; Wang et al., 2023; Rawte et al., 2023). These visually unsupported responses are particularly problematic when MLLM outputs are used as factual descriptions of visual scenes, motivating efficient methods that improve the faithfulness of generation during inference.

Existing hallucination mitigation methods can be broadly categorized into training-based and training-free approaches. Training-based methods, including Supervised Fine-Tuning (SFT) (Jiang et al., 2024; Liu et al., 2023a), Reinforcement Learning from Human Feedback (RLHF) (Zhao et al., 2023; Jing and Du, 2024; Kim et al., 2024), and Retrieval-Augmented Generation (RAG) (Qu et al., 2024; Sarwar, 2025), can reduce hallucinations but usually require additional data, model retraining, or external knowledge. By contrast, training-free decoding methods intervene only dur-

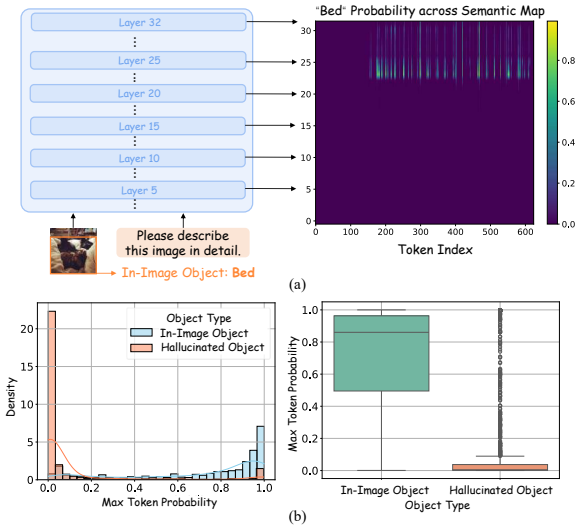

Figure 2: Logit-lens evidence for dispersed visually grounded signals. (a) Intermediate layer-position states can assign high vocabulary probabilities to objects present in the image. (b) Across 3,000 POPE images, in-image object words receive higher maximum logit-lens confidence than hallucinated object words, suggesting that factual evidence is dispersed across hidden states.

ing inference, making them more efficient and easier to deploy. Recent examples include contrastive decoding (Wang et al., 2024b; Chuang et al., 2023; Leng et al., 2024a), guided decoding (Wang et al., 2024a), anchor-token reallocation (Huang et al., 2024), visual attention amplification (Sarkar et al., 2025a), and layer-wise consistency (Wang et al., 2025; Tang et al., 2025; Yu et al., 2025).

While these training-free methods have shown promising results, they typically exploit only a partial view of the decoding trajectory. As illustrated in Figure 1, inter-layer methods mainly compare or aggregate representations across transformer layers, whereas intra-layer methods refine attention or token representations within a single layer. Both paradigms implicitly assume that useful factual evidence can be sufficiently captured from one dominant dimension of the hidden-state space. However, autoregressive generation in MLLMs produces a richer trajectory of hidden states spanning both layers and token positions. This raises a natural question: *Is visually grounded evidence localized within a single dimension, or is it dispersed across the full layer-position hidden-state map?*

To answer this question, we conduct a logit-lens analysis using a vanilla LLaVA-1.5 model (Liu et al., 2023b) and 3,000 images from the random split of POPE (Li et al., 2023). We project hidden states from different layers and positions into the vocabulary space and compare their confidence scores for in-image and hallucinated object words. As shown in Figure 2, factual visual evidence is not confined to final-layer or token-local representations. Instead, intermediate layer-position states often assign high probabilities to objects that are present in the image, while hallucinated objects receive lower maximum confidence. This suggests that hallucination mitigation can be viewed as recovering dispersed grounded evidence from the decoding trajectory.

Motivated by this finding, we propose **M**ap-Level **A**ttention **P**rocessing (`MAP`), a training-free decoding method for faithful generation in MLLMs. `MAP` treats hidden states across layers and positions as a two-dimensional semantic map, and aggregates dispersed grounded evidence during decoding. Specifically, Layer-Wise Criss-Cross Attention gathers evidence along both the layer and position dimensions, while Global-Local Logit Fusion combines local and global evidence at the logit level. Experiments on POPE (Li et al., 2023), MME (Fu et al., 2024), and MMHal-Bench (Sun et al., 2023b) show that `MAP` consistently improves factuality across multiple MLLM architectures with modest decoding overhead.

Our contributions are summarized as follows:

- We provide a logit-lens analysis showing that visually grounded factual evidence is dispersed across layer-position hidden states.
- We formulate hallucination mitigation as map-level evidence aggregation and propose `MAP`, a training-free decoding method that integrates inter-layer and intra-layer cues.
- We validate `MAP` on closed- and open-ended hallucination benchmarks across multiple MLLM architectures.

2 Related Work

2.1 Hallucinations in MLLMs

Hallucination is a widely studied problem in Large Language Models (LLMs), commonly defined as the generation of text that is factually inaccurate or unsupported by the input context (Rawte et al., 2023; Sriramanan et al., 2024; Huang et al., 2025). In Multimodal Large Language Models (MLLMs), hallucinations refer to outputs that conflict with the visual content, such as mentioning non-existent objects or misrepresenting visual attributes (Li et al.,

2023). Although the exact causes of hallucinations remain unclear, several factors are believed to contribute to their occurrence. Some studies (Liu et al., 2024; Jiang et al., 2025; Sarkar et al., 2025a) have revealed that MLLMs are prone to pay more attention to the textual context while paying inadequate attention to visual tokens, which may result in hallucinated generations. Moreover, studies (Zhou et al., 2023; Tu et al., 2025; Zhu et al., 2025b) also point out that co-occurrence bias (i.e., the tendency to predict objects that frequently appear together) and over-reliance on language priors can both contribute to hallucinations. Furthermore, prior works (Huang et al., 2024; Kang et al., 2025) have observed issues such as attention sinks and anchor tokens, which may amplify hallucinations.

2.2 Addressing Hallucinations in MLLMs

Research on hallucination mitigation has mainly focused on several directions, including fine-tuning with curated datasets (Chen et al., 2023; Wang et al., 2024b), reinforcement learning (Sun et al., 2023b; Yu et al., 2024), and incorporating external knowledge bases (Su et al., 2024). However, these approaches typically require additional data collection or retraining, making them time-consuming and computationally expensive for real-world applications. More recently, generation-time interventions have drawn growing attention. DoLa (Chuang et al., 2023), a contrastive decoding method, compares factual knowledge across shallow and deep layers to encourage faithful generation. VCD (Leng et al., 2024a) extends this by contrasting logits conditioned on visual inputs with different uncertainty levels to enhance robustness. DAMO (Wang et al., 2025) emphasizes cross-layer consistency and employs a momentum mechanism to mitigate hallucinations in deep layers. DCLA (Tang et al., 2025) aggregates hidden states to construct semantic references for consistent decoding. HalluRNN (Yu et al., 2025) applies a modified RNN block across model layers to improve consistency. Other works (Huang et al., 2024; Gong et al., 2024) identify anchor tokens and reallocate attention weights to mitigate hallucinations.

3 Method

3.1 Formulation

Following common vision-language MLLMs, we consider a model composed of a vision encoder, a projection module, and a pretrained Large Language Model (LLM). Given an input image X_v, the vision encoder $g(\cdot)$ produces visual tokens $Z_v = g(X_v) \in \mathbb{R}^{L_v \times D_v}$, where L_v denotes the number of visual tokens and D_v is the dimensionality of the visual hidden states. These tokens are then projected into the language embedding space as $H_v = P(Z_v) \in \mathbb{R}^{L_v \times D}$, where D is the hidden dimension of the LLM.

The system prompt X_s and language query X_q are tokenized and embedded as $H_s \in \mathbb{R}^{L_s \times D}$ and $H_q \in \mathbb{R}^{L_q \times D}$, respectively. The visual and textual tokens are concatenated as the LLM input:

$$H_0 = \{h_{1,0}, h_{2,0}, \ldots, h_{t,0}\}, \tag{1}$$

where t denotes the current sequence length. Initially, $t = L_s + L_v + L_q$, and it increases during autoregressive generation. This sequence is passed through the transformer layers of the LLM. Let $H_j = \{h_{1,j}, h_{2,j}, \ldots, h_{t,j}\}$ denote the hidden states at the j-th transformer layer. Standard autoregressive decoding predicts the next token from the final-layer hidden state at the last position:

$$p(x_{t+1} \mid x_{<t+1}) = \mathrm{softmax}\left(\phi(h_{t,n})\right), \tag{2}$$

where $\phi(\cdot)$ denotes the language modeling head and n is the number of transformer layers.

Instead of relying only on the final prediction token $h_{t,n}$, MAP refines next-token prediction by aggregating evidence from a layer-position semantic map. At the j-th layer, we define the semantic map as:

$$\mathcal{H}_j = \{h_{u,\ell} \mid 1 \le u \le t,\ 1 \le \ell \le j\}, \tag{3}$$

where $h_{u,\ell} \in \mathbb{R}^D$ denotes the hidden state at position u and layer ℓ. The semantic map $\mathcal{H}_j$ can be viewed as a layer-position map of size $j \times t$, where rows correspond to layers and columns correspond to token positions.

3.2 Map-Level Operation

We first introduce a general map-level operation that serves as the basis for both Layer-Wise Criss-Cross Attention and Global-Local Logit Fusion. Given a semantic map $\mathcal{H}_j$, an anchor token $h_{t,j}$, and a structured semantic neighborhood $\mathcal{M}_{t,j} \subseteq \mathcal{H}_j$, we define the map-level aggregation function as:

$$\mathcal{F}(h_{t,j}, \mathcal{M}_{t,j}) = \sum_{h_{u,\ell} \in \mathcal{M}_{t,j}} w_{u,\ell}\, h_{u,\ell}, \tag{4}$$

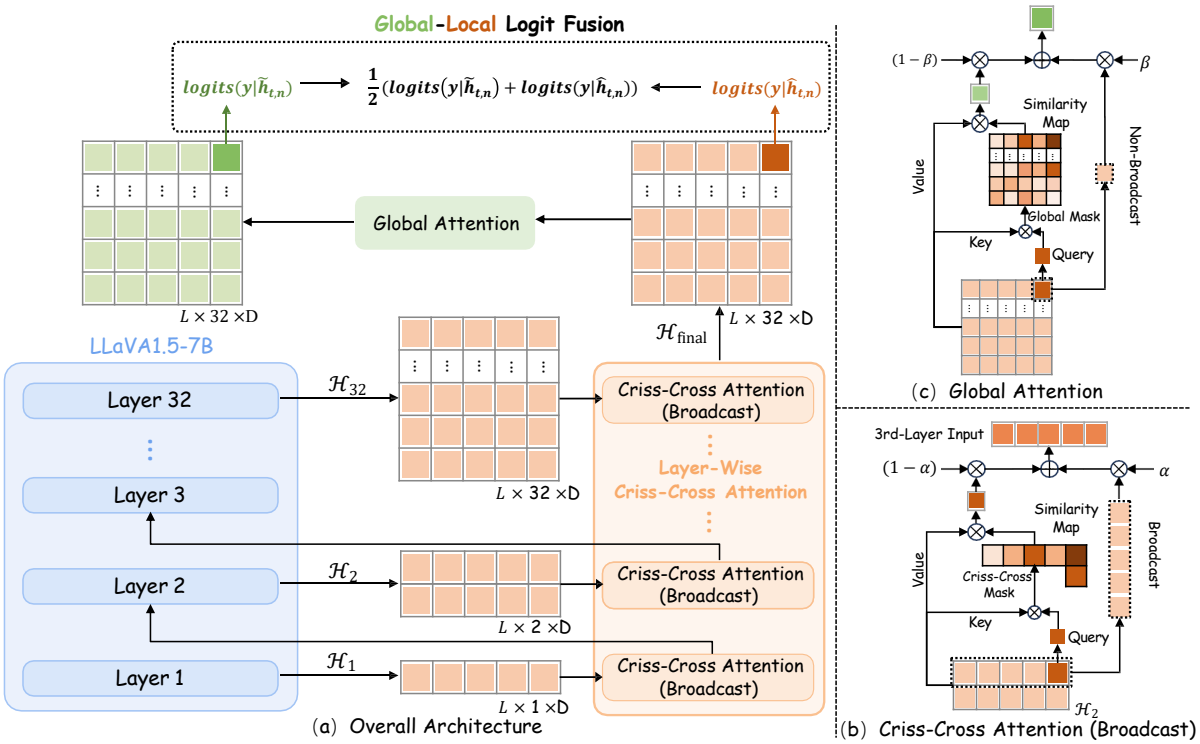

Figure 3: Overall architecture of MAP applied to LLaVA1.5-7B. Specifically, MAP introduces a Layer-Wise Criss-Cross Attention module to refine token representations by aggregating visually grounded evidence from a 2D semantic map constructed after each decoding layer. Additionally, a Global-Local Logit Fusion module is integrated to fuse hierarchical content at logit-level to produce reliable predictions.

where $w_{u,\ell}$ is a parameter-free similarity weight. Specifically, we compute the normalized similarity weight as:

$$s_{u,\ell} = \cos(h_{u,\ell}, h_{t,j}),$$
$$w_{u,\ell} = \frac{\exp(s_{u,\ell})}{\sum_{h_{a,b} \in \mathcal{M}_{t,j}} \exp(s_{a,b})}. \quad (5)$$

By changing the definition of $\mathcal{M}_{t,j}$, different map-level operations can be instantiated.

Global Semantic Neighborhood. The global semantic neighborhood aggregates information from the entire semantic map. Specifically, it contains all hidden states in $\mathcal{H}_j$ except the anchor token:

$$\mathcal{M}^g_{t,j} = \mathcal{H}_j \setminus \{h_{t,j}\}. \quad (6)$$

Criss-Cross Semantic Neighborhood. The criss-cross semantic neighborhood aggregates information along the row and column of the anchor token. Specifically, it consists of hidden states from the same layer and the same position as the anchor token:

$$\mathcal{M}^c_{t,j} = \{h_{u,j} \mid u \neq t\} \cup \{h_{t,\ell} \mid \ell \neq j\}. \quad (7)$$

This design allows the anchor token to collect evidence from both the intra-layer token dimension and the inter-layer depth dimension.

3.3 Layer-Wise Criss-Cross Attention

Using the semantic map constructed at each layer, we introduce Layer-Wise Criss-Cross Attention to aggregate dispersed visually grounded evidence during decoding, as illustrated in Figure 3. Given the semantic map $\mathcal{H}_j$ at layer j, we first compute an aggregated representation from the criss-cross semantic neighborhood:

$$r^c_{t,j} = \mathcal{F}(h_{t,j}, \mathcal{M}^c_{t,j}). \quad (8)$$

We then update the current-layer hidden states through residual mixing:

$$\hat{h}_{u,j} = (1-\alpha)\, r^c_{t,j} + \alpha\, h_{u,j}. \quad (9)$$

where $u \in \{1, \ldots, t\}$ and $\alpha \in [0, 1]$ balances the original token representation and the aggregated map-level evidence. The refined sequence $\hat{H}_j = \{\hat{h}_{1,j}, \ldots, \hat{h}_{t,j}\}$ is then passed to subsequent transformer layers, allowing the aggregated evidence to influence later decoding states.

Since next-token prediction in autoregressive MLLMs is driven by the final-position representation, we use $h_{t,j}$ as the prediction anchor rather than querying the semantic map from every position. The aggregated evidence is broadcast to the current-layer sequence through residual mixing, allowing later layers to access prediction-centered map-level evidence while retaining token-specific information. This design keeps the additional decoding overhead lightweight. In practice, `MAP` is applied only from a starting layer ℓ_{start} onward. Layers before ℓ_{start} follow the original transformer computation, while later layers progressively propagate map-level evidence through Layer-Wise Criss-Cross Attention.

3.4 Global-Local Logit Fusion

Layer-Wise Criss-Cross Attention produces a refined local prediction token $\hat{h}_{t,n}$ at the final layer. To complement this local prediction with broader map-level evidence, we further compute a globally enhanced token over the final semantic map. Let $\hat{\mathcal{H}}_n$ denote the final semantic map after layer-wise refinement, and let $\hat{\mathcal{M}}^g_{t,n} = \hat{\mathcal{H}}_n \setminus \{\hat{h}_{t,n}\}$ denote its global neighborhood. We first compute the globally aggregated evidence:

$$r^g_{t,n} = \mathcal{F}(\hat{h}_{t,n}, \hat{\mathcal{M}}^g_{t,n}). \tag{10}$$

The globally enhanced token is then computed as:

$$\tilde{h}_{t,n} = (1-\beta)\, r^g_{t,n} + \beta\, \hat{h}_{t,n}. \tag{11}$$

where $\beta \in [0,1]$ controls the fusion weight between the local prediction token and globally aggregated evidence.

The local token $\hat{h}_{t,n}$ and the global token $\tilde{h}_{t,n}$ capture complementary information. As later shown in Table 3, the local token performs better on certain fine-grained categories such as *Count* and *Posters*, while the global token excels in categories such as *Position* and *Color*. Motivated by this complementarity, we average the logits from both representations for final prediction:

$$\begin{cases} \text{logit}_{\text{final}} = \dfrac{1}{2} \cdot \Big(\phi(\tilde{h}_{t,n}) + \phi(\hat{h}_{t,n})\Big) \\ p(x_{t+1} \mid x_{<t+1}) = \text{softmax}\left(\text{logit}_{\text{final}}\right) \end{cases} \tag{12}$$

Overall, `MAP` provides a training-free decoding framework that refines next-token prediction by recovering dispersed visually grounded evidence from the layer-position semantic map. It integrates both inter-layer and intra-layer cues without introducing additional trainable parameters.

4 Experiment

4.1 Experiment Setup

Datasets. We evaluate `MAP` on three widely used hallucination benchmarks covering both closed- and open-ended settings: POPE (Li et al., 2023), MME (Fu et al., 2024), and MMHal-Bench (Sun et al., 2023b). POPE evaluates object-level hallucinations under random, popular, and adversarial settings, with images drawn from MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019). MME provides a fine-grained perception evaluation across ten hallucination-sensitive categories. MMHal-Bench evaluates open-ended generation, where responses are assessed by GPT-4 (Hurst et al., 2024) following the benchmark protocol.

Models and Baselines. We compare `MAP` with recent training-free decoding methods, including VCD (Leng et al., 2024b), DAMO (Wang et al., 2025), DCLA (Tang et al., 2025), SPIN (Sarkar et al., 2025b), and HGAI (Jiang et al., 2025). These baselines cover contrastive decoding, inter-layer intervention, and intra-layer attention refinement. To assess generalization across MLLM architectures, we evaluate three representative vision-language MLLMs: LLaVA-1.5-7B (Liu et al., 2023b), mPLUG-Owl2-7B (Ye et al., 2024), and InstructBLIP-7B (Dai et al., 2023).

Implementation Details. All hyperparameters for `MAP` and the compared baselines are provided in the Appendix. We further analyze the sensitivity of α, β, and ℓ_{start} in Section 4.3. To ensure reproducibility, we disable sampling and set the temperature to zero for all decoding methods.

4.2 Main Experiment Results

Results on MME. Table 1 reports the results on the MME perception benchmark. `MAP` achieves the best total score across all three evaluated MLLMs, with scores of 1529.3 on LLaVA-1.5, 1466.4 on mPLUG-Owl2, and 1302.7 on InstructBLIP. Compared with vanilla decoding, `MAP` improves the total score by 37.7 and 31.3 points on LLaVA-1.5 and InstructBLIP, respectively, while bringing a smaller but consistent gain on mPLUG-Owl2. Compared with inter-layer methods such as DAMO and DCLA, `MAP` further improves the overall score on all three models, suggesting that aggregating evidence from the layer-position map provides complementary benefits beyond single-dimensional interventions.

Table 1: Results on MME. Regular , Other , Inter-Layer , Intra-Layer , **Map-Level (Ours)** methods are shaded accordingly. Best results are in **bold**.

Model	Decoding	Cele.	Color	Artw.	Count	Land.	OCR	Post.	Exis.	Posi.	Scen.	Total
LLaVA1.5	Vanilla	135.0	165.0	119.3	160.0	162.3	125.0	140.5	190.0	138.3	156.3	1491.6
	VCD 2024	**151.8**	148.3	**125.5**	148.3	165.3	130.0	131.6	190.0	118.3	150.3	1459.4
	DAMO 2024	134.1	165.0	120.0	150.0	163.8	140.0	**144.6**	**195.0**	143.3	157.8	1513.5
	DCLA 2025	132.1	**175.0**	117.3	163.3	160.5	140.0	137.4	190.0	**148.3**	156.3	1520.1
	SPIN 2025	132.9	170.0	123.0	163.3	165.3	125.0	141.5	190.0	133.3	157.5	1501.9
	HGAI 2025	135.0	165.0	118.5	160.0	160.8	125.0	140.5	190.0	138.3	156.3	1489.3
	MAP	138.2	170.0	115.8	**165.0**	**165.3**	**140.0**	142.5	190.0	143.3	**159.3**	**1529.3**
mPLUG-Owl2	Vanilla	162.9	150.0	**139.8**	165.0	160.0	102.5	163.3	185.0	78.3	152.8	1459.5
	VCD 2024	162.4	150.0	135.3	145.0	147.8	95.0	149.3	175.0	83.3	151.0	1394.0
	DAMO 2024	163.8	150.0	139.0	160.0	161.5	95.0	**163.3**	185.0	**93.3**	153.8	1464.7
	DCLA 2025	163.8	155.0	139.0	165.0	159.5	102.5	162.2	185.0	78.3	153.0	1463.4
	SPIN 2025	**166.2**	**160.0**	136.5	160.0	160.0	102.5	160.2	185.0	68.3	**158.3**	1457.0
	HGAI 2025	162.4	150.0	136.5	165.0	160.8	102.5	157.1	185.0	86.7	153.5	1459.4
	MAP	163.5	155.0	135.8	**165.0**	**162.5**	**102.5**	160.2	**185.0**	83.3	153.3	**1466.4**
InstructBLIP	Vanilla	154.7	143.3	112.5	58.3	165.8	80.0	152.7	185.0	58.3	**160.8**	1271.4
	VCD 2024	127.1	140.0	107.0	68.3	133.8	63.3	117.0	145.0	63.3	149.3	1153.2
	DAMO 2024	155.3	143.3	116.3	78.3	159.3	87.5	149.0	185.0	68.3	157.3	1299.5
	DCLA 2025	157.6	143.3	**119.5**	73.3	157.3	87.5	149.7	185.0	**68.3**	158.0	1299.6
	SPIN 2025	152.9	143.3	114.3	**86.7**	160.3	87.5	141.2	185.0	63.3	159.8	1294.2
	HGAI 2025	153.8	143.3	111.0	58.3	165.0	80.0	**152.7**	185.0	56.7	159.0	1264.9
	MAP	**158.5**	**143.3**	115.0	70.0	**165.8**	**95.0**	147.3	**185.0**	63.3	159.5	**1302.7**

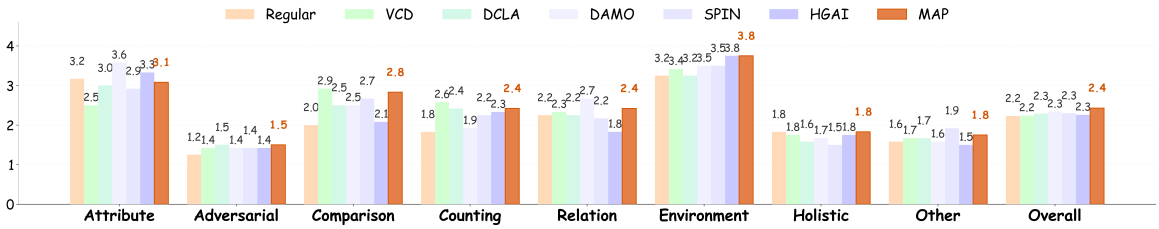

Figure 4: Results on MMHal-Bench using LLaVA-1.5 across different categories and decoding strategies.

Results on POPE. POPE provides a targeted evaluation of object-level hallucinations. As shown in Table 2, MAP generally improves performance across random, popular, and adversarial settings, with particularly clear gains on LLaVA-1.5 and InstructBLIP. On the challenging GQA adversarial split, MAP improves accuracy over VCD by 4.47% on LLaVA-1.5 and 2.87% on mPLUG-Owl2. We also observe that different decoding paradigms show varying stability across datasets and models. For example, HGAI obtains strong results on MSCOCO with LLaVA-1.5, but its gains are less consistent on A-OKVQA and GQA. These results suggest that MAP provides robust object-level hallucination mitigation across diverse POPE settings.

Results on MMHal-Bench. MMHal-Bench evaluates hallucination mitigation in open-ended generation across eight categories. As shown in Figure 4, MAP obtains the highest overall score among the compared decoding strategies on LLaVA-1.5. MAP also shows balanced improvements across several categories, reaching 3.8 on *Environment*, which improves vanilla decoding by 0.5. Other decoding methods improve certain categories but can be less stable across tasks; for example, DAMO improves *Relation* but drops on *Counting*, while VCD performs well on *Comparison* but poorly on *Counting*. These results indicate that MAP can also benefit open-ended hallucination mitigation.

4.3 Ablation Study

Evaluation on Map-Level Operations. We first disable Global-Local Logit Fusion and evaluate different map-level operations on MME, as shown in Table 4. In addition to global and criss-cross neighborhoods, we also consider a local semantic neighborhood: $\mathcal{M}^l_{t,j} = \{h_{u,\ell} \mid |u-t| \le r,\ |\ell - j| \le r,\ (u,\ell) \ne (t,j)\}$, where r denotes the window radius. For non-layer-wise variants, the operation is applied only to the final semantic map $\mathcal{H}_{32}$. The

Table 2: Results on POPE. Rows are shaded by decoding paradigm accordingly, and best results are in **bold**.

Model	Decoding	MSCOCO						A-OKVQA						GQA					
		Random		Popular		Adversarial		Random		Popular		Adversarial		Random		Popular		Adversarial	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
LLaVA1.5	Vanilla	89.57	89.66	86.10	86.70	79.50	81.48	87.10	88.30	80.10	83.08	69.03	75.90	86.50	87.83	74.73	79.45	68.77	75.73
	VCD 2024	89.07	89.00	85.63	86.03	79.27	81.01	86.97	87.23	78.90	81.82	67.90	75.01	86.00	87.33	74.50	79.29	68.10	75.15
	DAMO 2024	90.03	90.02	86.73	87.13	80.47	82.15	87.87	88.91	81.30	83.87	70.53	76.75	87.50	88.58	76.17	80.28	70.07	76.42
	DCLA 2025	90.03	89.99	86.73	87.10	80.70	82.28	87.93	88.98	81.13	83.78	70.23	76.60	87.90	88.94	75.20	79.68	69.13	75.91
	SPIN 2025	89.63	89.67	86.43	86.90	79.90	81.74	87.70	88.79	80.63	83.41	69.97	76.43	87.53	88.66	75.27	79.76	68.87	76.43
	HGAI 2025	**90.70**	**90.77**	**87.67**	**88.12**	81.00	82.80	87.37	88.50	80.37	83.20	70.00	76.42	87.07	88.24	74.37	79.10	69.20	75.90
	MAP	90.10	89.86	87.40	87.44	**81.80**	**82.82**	**89.03**	**89.70**	**82.70**	**84.70**	**72.77**	**77.84**	**88.57**	**89.32**	**77.90**	**81.22**	**72.57**	**77.70**
mPLUG2	Vanilla	88.27	87.56	86.37	85.82	84.07	83.82	88.27	88.23	84.73	85.21	77.57	79.67	86.56	**87.20**	**80.23**	**80.67**	78.43	**79.27**
	VCD 2024	86.77	86.11	83.97	83.59	82.10	82.07	86.83	86.93	82.07	82.96	75.50	78.13	84.87	84.08	77.60	78.00	75.70	76.38
	DAMO 2024	88.40	87.75	86.40	85.92	83.90	83.75	88.13	88.10	84.43	84.95	77.23	79.42	87.03	86.38	79.67	80.18	77.97	78.88
	DCLA 2025	88.53	87.99	86.03	86.05	82.67	83.25	87.87	88.23	83.77	84.82	75.33	78.63	**87.07**	86.98	78.90	80.37	76.77	78.81
	SPIN 2025	88.47	87.71	86.63	86.03	84.37	84.04	88.13	87.96	84.43	84.77	77.60	79.46	86.10	85.40	79.80	80.11	78.13	78.81
	HGAI 2025	88.43	87.67	86.47	85.87	84.30	83.97	88.33	88.26	84.57	85.04	77.57	79.64	86.90	86.15	79.63	80.00	78.03	78.76
	MAP	**88.80**	**88.16**	**86.70**	**86.25**	**84.37**	**84.21**	**88.33**	**88.32**	**84.77**	**85.27**	**77.70**	**79.82**	86.93	86.31	80.13	80.33	**78.57**	79.10
InstBLIP	Vanilla	87.13	85.72	83.50	84.09	80.77	81.93	88.60	88.34	79.47	81.85	71.37	76.38	86.93	87.34	76.47	**79.30**	71.60	76.04
	VCD 2024	84.10	83.07	80.17	79.27	78.17	78.13	82.03	82.36	76.90	78.40	71.67	74.93	81.43	81.60	73.73	75.81	71.93	74.80
	DAMO 2024	88.23	87.46	84.93	84.49	82.50	**82.31**	87.33	87.77	79.07	81.29	71.40	76.07	87.00	87.47	75.03	78.42	72.13	76.50
	DCLA 2025	**88.40**	**87.57**	83.97	82.44	82.77	81.37	88.13	88.28	79.83	81.59	72.37	76.39	87.47	87.65	75.93	78.70	73.00	76.71
	SPIN 2025	86.33	84.61	84.10	82.53	82.83	81.40	88.57	88.11	79.60	**81.97**	71.47	76.47	86.70	86.04	76.20	79.15	71.37	75.93
	HGAI 2025	87.30	85.87	84.73	83.49	82.80	81.79	88.57	88.35	79.30	81.77	71.33	76.40	87.30	86.99	75.57	77.66	**74.03**	76.68
	MAP	88.30	87.43	**85.43**	**84.82**	**82.97**	82.14	**88.60**	**88.41**	**80.20**	81.85	**73.03**	**76.80**	**87.57**	**87.69**	**76.53**	79.06	73.83	**77.20**

Table 3: Complementarity between local and global prediction tokens on MME using LLaVA-1.5.

	Celebrity	Color	Artwork	Count	Landmark	Existence	Posters	OCR	Position	Scene	Total
Local Token	137.4	165.0	115.0	**165.0**	**165.3**	190.0	**144.6**	140.0	138.3	**159.3**	1520.3
Global Token	**139.1**	**170.0**	115.0	160.0	164.5	190.0	138.4	140.0	**143.3**	158.5	1519.4

Table 4: Comparison of different map operations on MME without Global-Local Logit Fusion.

Model	Map-Level Operation	Total MME Score
LLaVA1.5	Vanilla	1491.6
	Global Attention	1509.7
	Local Attention (7×7)	1502.4
	Local Attention (5×5)	1502.5
	Criss-Cross Attention	1507.9
	Layer-Wise Criss-Cross Attention	**1520.3**

Table 5: The efficiency results of MAP compared with other intra- or inter-layer approaches.

	Prefill Latency (ms/token)	Decode Latency (ms/token)	GPU Usage (GB)	TFLOPs
Vanilla	73.16	19.37	15.28	8.49
DAMO	90.85	38.69	15.36	8.51
DCLA	90.72	28.60	15.29	8.49
SPIN	74.88	22.19	15.27	8.49
MAP	77.33	26.69	15.29	8.49

results show that non-layer-wise local and global operations provide limited gains, reaching MME scores of 1502.5 and 1509.7, respectively. In contrast, applying criss-cross attention in a layer-wise manner increases the score from 1507.9 to 1520.3, validating the benefit of progressive map-level evidence aggregation.

Complementarity between Local and Global Tokens. We further analyze the local prediction token $\hat{h}_{t,n}$ and the globally enhanced token $\tilde{h}_{t,n}$ used in Global-Local Logit Fusion. As shown in Table 3, the two tokens achieve comparable overall MME scores, but exhibit complementary strengths across fine-grained categories. The local token performs better on such as *Count* and *Scene*, while the global token performs better on *Celebrity* and *Position*. This category-level complementarity motivates our logit-level fusion strategy.

Evaluation on Computational Efficiency. We compare the efficiency of MAP with other decoding methods in terms of prefill latency, decoding latency, GPU memory usage, and TFLOPs. As shown in Table 5, MAP achieves a decoding latency of 26.69 ms per token, which is lower than DAMO

Table 6: Generalization of MAP to stronger MLLMs.

	Qwen2.5-VL-7B	InternVL2.5-8B	InternVL3-14B
Vanilla	1704.2	1697.2	1729.6
DAMO	1699.9	1709.9	1725.8
DCLA	1704.0	1694.6	1726.4
SPIN	1686.3	1685.2	1722.8
MAP	1712.0	1704.3	1737.0

Table 7: Accuracy sensitivity to α and ℓ_{start} on MSCOCO POPE.

Sensitivity to α						
α	0.80	0.82	0.84	0.86	0.88	0.90
Random	90.10	90.07	90.10	90.00	90.13	**90.16**
Popular	**87.40**	87.33	87.37	87.33	87.30	87.27
Adversarial	**81.80**	81.67	81.50	81.53	81.50	81.47
Sensitivity to ℓ_{start}						
ℓ_{start}	26th	27th	28th	29th	30th	31st
Random	88.87	89.17	90.00	90.10	**90.23**	90.13
Popular	84.90	85.70	87.00	87.40	**87.47**	87.37
Adversarial	79.03	80.17	81.47	**81.80**	81.70	81.60

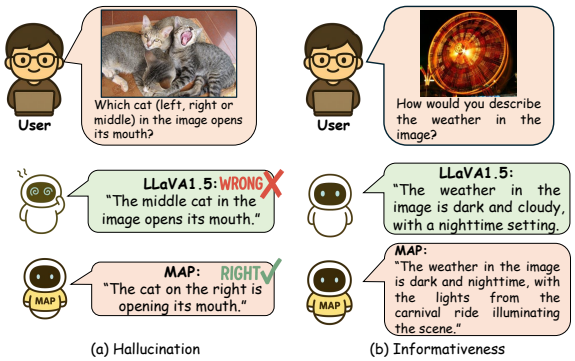

Figure 5: Qualitative comparison between vanilla LLaVA-1.5 and MAP on open-ended generation.

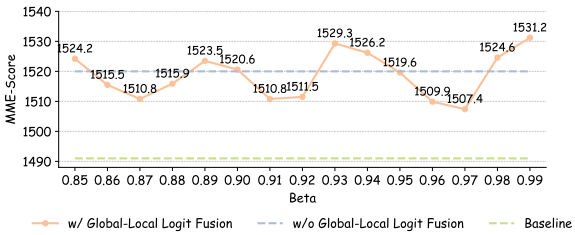

Figure 6: MME performance across different β.

and DCLA while remaining close to lightweight intra-layer methods. The GPU memory usage and TFLOPs are also comparable across methods. This efficiency comes from our prediction-centered map aggregation design, which queries the semantic map only from the final-position token rather than from every token position.

Evaluation on Stronger MLLMs. To examine whether MAP generalizes to more recent MLLMs, we further evaluate Qwen2.5-VL-7B (Bai et al., 2025), InternVL2.5-8B (Chen et al., 2024), and InternVL3-14B (Zhu et al., 2025a). As shown in Table 6, MAP improves the MME total score from 1704.2 to 1712.0 on Qwen2.5-VL-7B, from 1697.2 to 1704.3 on InternVL2.5-8B, and from 1729.6 to 1737.0 on InternVL3-14B. Although the gains are more moderate on stronger models, the consistent improvements suggest that MAP remains complementary to stronger multimodal alignment.

Sensitivity to α and ℓ_{start}. We evaluate the sensitivity of α and ℓ_{start} on the MSCOCO split of POPE using LLaVA-1.5. As shown in Table 7, MAP remains stable across α values from 0.80 to 0.90, with variations within 0.40%. Table 7 shows that starting refinement from later layers generally performs better, with the best average performance achieved around $\ell_{\text{start}} = 29$.

Sensitivity to β. We further evaluate the impact of β in Global-Local Logit Fusion on MME. As shown in Figure 6, all tested β values from 0.85 to 0.99 outperform vanilla decoding, indicating that the fusion module is generally beneficial. The best score is obtained at $\beta = 0.99$, suggesting that a small amount of globally aggregated evidence can complement the local prediction token.

Case Study. We qualitatively evaluate MAP under an open-ended generation setting, the results of which are presented in Figure 5. As illustrated, MAP effectively corrected the hallucinated response given by the vanilla model. Additionally, it can also be observed that MAP could further enhance the overall informativeness of the generated response, which added more details of scene description.

5 Conclusion

In this work, we study hallucination in MLLMs from a map-level perspective. Through a logit-lens analysis, we show that visually grounded factual evidence is not confined to final-layer or token-local representations, but is dispersed across layer-position hidden states. Motivated by this observation, we propose MAP, a training-free decoding method that recovers such dispersed evidence from a semantic map. MAP combines Layer-Wise Criss-Cross Attention for progressive map-level evidence aggregation with Global-Local Logit Fusion for final prediction. Experiments on closed- and open-ended benchmarks across multiple MLLM architectures show that MAP consistently improves factuality with modest decoding overhead.

Limitations

MAP has several limitations. First, as a training-free decoding method, it requires access to intermediate hidden states and is therefore mainly applicable to open-source or white-box MLLMs rather than API-only models. Second, its gains become more moderate on stronger recent MLLMs, suggesting that map-level evidence aggregation is complementary to stronger multimodal pretraining and alignment. Third, `MAP` relies on parameter-free cosine weighting and manually defined neighborhoods; while this keeps it architecture-agnostic, learnable map-level operators may further improve performance when calibration data is available. Finally, our evaluation focuses on image-text hallucination benchmarks and does not cover video, multi-image, or domain-specific settings. MMHal-Bench also relies on automatic LLM-based judging, which may introduce evaluation bias.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

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

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 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.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ngoc Dung Huynh, Mohamed Reda Bouadjenek, Sunil Aryal, Imran Razzak, and Hakim Hacid. 2025. Visual question answering: from early developments to recent advances–a survey. *arXiv preprint arXiv:2501.03939*.

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 Computer Vision and Pattern Recognition Conference*, pages 25004–25014.

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

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*.

Minchan Kim, Minyeong Kim, Junik Bae, Suhwan Choi, Sungkyung Kim, and Buru Chang. 2024. Exploiting semantic reconstruction to mitigate hallucinations in vision-language models. In *European Conference on Computer Vision*, pages 236–252. Springer.

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*, pages 13872–13882.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

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*.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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*.

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

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

Xiaoye Qu, Qiyuan Chen, Wei Wei, Jiashuo Sun, Daizong Liu, and Jianfeng Dong. 2024. Alleviating hallucination in large vision-language models with active retrieval augmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter A Beerel, and Souvik Kundu. 2025a. Mitigating hallucinations in vision-language models through image-guided head suppression. *arXiv preprint arXiv:2505.16411*.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter A Beerel, and Souvik Kundu. 2025b. Mitigating hallucinations in vision-language models through image-guided head suppression. *arXiv preprint arXiv:2505.16411*.

SM Sarwar. 2025. Filterrag: zero-shot informed retrieval-augmented generation to mitigate hallucinations in vqa. *arXiv preprint arXiv:2502.18536*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large vision-language model alignment and misalignment: A survey through the lens of explainability. *arXiv preprint arXiv:2501.01346*.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216.

Weihang Su, Yichen Tang, Qingyao Ai, Changyue Wang, Zhijing Wu, and Yiqun Liu. 2024. Mitigating entity-level hallucination in large language models. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 23–31.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023a. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

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. 2023b. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Kai Tang, Jinhao You, Xiuqi Ge, Hanze Li, Yichen Guo, and Xiande Huang. 2025. Mitigating hallucinations via inter-layer consistency aggregation in large vision-language models. *arXiv preprint arXiv:2505.12343*.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *arXiv preprint arXiv:2503.08342*.

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

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, and 1 others. 2023. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Kaishen Wang, Hengrui Gu, Meijun Gao, and Kaixiong Zhou. 2025. Damo: Decoding by accumulating activations momentum for mitigating hallucinations in vision-language models. In *The Thirteenth International Conference on Learning Representations*.

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

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*, pages 37–53. Springer.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051.

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

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. 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.

Jianfei Zhao, Feng Zhang, Xin Sun, and Chong Feng. 2025. Aligning attention distribution to information flow for hallucination mitigation in large vision-language models. *arXiv preprint arXiv:2505.14257*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

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*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025b. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1624–1633.

A Additional Results for Preliminary Experiments

Here, we provide a more detailed analysis of our preliminary experiments described in the introduction. Specifically, we construct an analysis dataset consisting of 3,000 images sampled from a random split of the MSCOCO subset in the POPE benchmark. Each image is paired with the prompt: "Please describe this image in detail." We use LLaVA-1.5 to process these image-text pairs and project all hidden states across different layers and time steps into the vocabulary distribution using the language head. Using the object annotations provided in the MSCOCO 2014 validation set, we analyze the confidence scores assigned to each object token to investigate whether the model inherently attends to factual visual entities in its internal representations. An interesting pattern is illustrated in Figure 7.

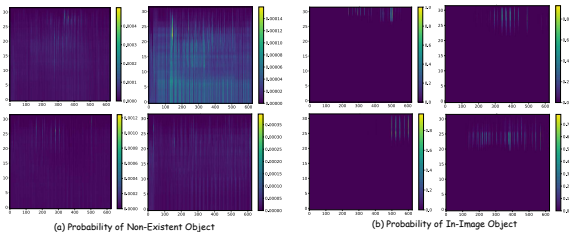

Figure 7: Examples of confidence distribution for non-existent vs. in-image objects across all hidden states.

As shown in Figure 7, we observe that at least one region of the hidden states exhibits high con-

fidence for in-image objects, while most hidden states assign very low confidence to non-existent objects. Here, tokens with high confidence for in-image objects are considered to contain real semantics related to input image, thereby possessing a certain factual information. In contrast, tokens with low confidence assigning to hallucinated objects represent non-existing semantics and provide supportive information. These results further suggest that factual information is not localized, but distributed across the model's hidden states in different layers and time steps. Notably, we observe that factual signals tend to converge in deeper layers (i.e., layers beyond 24), which supports our decision to set a deeper refinement start layer when applying Layer-Wise Criss-Cross Attention.

B More Implementation Details

B.1 Dataset Description

We provide descriptions of the datasets used in our experiments. POPE is designed to evaluate object-level hallucinations by checking whether a model generates non-existent visual objects. It consists of three subsets (MSCOCO, A-OKVQA, and GQA), each containing 3,000 VQA samples. MME is a comprehensive benchmark covering 14 multimodal subtasks, offering a broad evaluation of model capabilities. MMHal-Bench is a AI-assisted benchmark that evaluates open-ended responses in terms of hallucination and informativeness. For evaluation, we use the GPT-4.1 API to automatically rate model outputs.

B.2 Implementation Details

For LVLMs, we use three models in our experiments: LLaVA-v1.5-7B (based on Vicuna-1.5), mPLUG-Owl2 (based on LLaMA2-7B), and InstructBLIP (based on Vicuna-1.1-7B). When experimenting on the mPLUG-Owl2 and InstructBLIP, we append the prompt "Answer the question using a single word or phrase." after each question on both MME and POPE. Here, the detailed hyperparameters used in `MAP` and other methods are listed below:

VCD. For VCD, we adopt the hyperparameter settings reported in the original implementation and fix $\alpha = 1.0$, $\beta = 0.1$, and $\gamma = 0.1$ for all backbones. Following the prior work, the noise step T is set to 999 on POPE and 500 on MME.

HGAI. HGAI reports hyperparameters only for LLaVA, using ls $= 3$, le $= 13$, and $\alpha = 0.5$. Since prior work does not provide configurations for other backbones, we directly reuse the LLaVA settings for mPLUG-Owl2 and InstructBLIP without additional tuning.

SPIN. SPIN requires specifying a start layer, a routing ratio r, and a blending weight α for each backbone: LLaVA (start-layer $= 0$, $r = 0.2$, $\alpha = 0.1$), mPLUG-Owl2 (start-layer $= 16$, $r = 0.1$, $\alpha = 0.05$), and InstructBLIP (start-layer $= 16$, $r = 0.1$, $\alpha = 0.08$). We observe that the performance is relatively stable under small perturbations around these values, and therefore we fix them for all datasets for each backbone.

DCLA. DCLA provides a single set of hyperparameters (α, τ) for each backbone, and these settings are stable across benchmarks. For LLaVA, we adopt the values reported in the original paper ($\alpha = 0.82$, $\tau = 0.74$). For mPLUG-Owl2, we use the corresponding configuration ($\alpha = 0.90$, $\tau = 0.95$) provided by the original implementation. Since prior work does not report DCLA settings for InstructBLIP, we directly reuse the LLaVA configuration.

DAMO. The original DAMO paper reports hyperparameter settings (β_1, β_2, τ) for LLaVA and mPLUG. We adopt these configurations directly without additional tuning. For InstructBLIP, where no hyperparameters are available in prior work, we simply reuse the LLaVA configuration. The values used in our experiments are summarized in Table 10.

MAP. Our method involves selecting a start layer ℓ_{start} and two coefficients (α, β) for each backbone–dataset pair. Across different models, we observe consistent behavior: the optimal intervention point always lies in the late layers of the vision encoder, and the coefficients fall within stable ranges ($\alpha \in [0.80, 0.99]$, $\beta \in [0.90, 0.99]$), with one exception on LLaVA–MME where a smaller β yields slightly better validation performance. The hyperparameters used in all experiments are summarized in Table 11.

Table 8: Ablation results of different map-level operations on MME containing all subtask scores.

Model	Map Operation	Celebrity	Color	Artwork	Count	Landmark	Existence	Posters	Ccr	Position	Scene	Total
LLaVA1.5	Vanilla	135.00	165.00	**119.25**	160.00	162.25	190.00	140.48	125.00	138.33	156.25	1491.56
	Global Attention	135.29	170.00	118.00	156.67	161.25	190.00	137.41	140.00	143.33	157.75	1509.71
	Local Attention (7×7)	133.53	165.00	118.25	156.67	160.50	190.00	137.41	140.00	143.33	157.75	1502.44
	Local Attention (5×5)	135.29	165.00	118.00	156.67	159.75	190.00	137.41	140.00	143.33	157.00	1502.46
	Criss-Cross Attention	133.53	170.00	117.25	161.67	160.50	190.00	136.39	132.50	**148.33**	157.75	1507.92
	Layer-Wise Criss-Cross Attention	**137.94**	165.00	115.00	**165.00**	**165.25**	**190.00**	**144.56**	**140.00**	138.33	**159.25**	**1520.33**

Table 9: Per-category MME scores on LLaVA1.5 for different values of β. All runs share the same ℓ_{start} and α as in the β-ablation, and "Total" denotes the overall MME score.

β	Celebrity	Color	Artwork	Count	Landmark	Existence	Posters	OCR	Position	Scene	Total
vanilla	135.00	165.00	**119.25**	160.00	162.25	190.00	140.48	125.00	138.33	156.25	1491.56
0.90	138.24	165.00	115.00	165.00	165.25	190.00	144.56	140.00	138.33	159.25	1520.63
0.91	137.94	165.00	114.25	160.00	164.50	190.00	141.50	140.00	138.33	159.25	1510.77
0.92	137.94	165.00	115.00	160.00	164.50	190.00	141.50	140.00	138.33	159.25	1511.52
0.93	138.24	**170.00**	115.75	**165.00**	165.25	**190.00**	142.52	**140.00**	**143.33**	**159.25**	**1529.34**
0.94	**139.12**	170.00	115.75	160.00	165.25	190.00	143.54	140.00	143.33	159.25	1526.24
0.95	137.94	165.00	114.25	165.00	**165.25**	190.00	**144.56**	140.00	138.33	159.25	1519.58

Table 10: Hyperparameter settings of DAMO on LLaVA and mPLUG. "M" denotes MME and "P" denotes POPE.

	LLaVA M	LLaVA P	mPLUG M	mPLUG P
β_1	0.05	0.20	0.60	0.40
β_2	0.20	0.40	0.80	0.60
τ	-0.30	-0.30	-0.60	-0.60

Table 11: Hyperparameters of MAP on different models and datasets. "M" denotes MME and "P" denotes POPE.

	LLaVA1.5		mPLUG-Owl2		InstructBLIP	
	M	P	M	P	M	P
ℓ_{start}	29	25	28	28	28	19
α	0.80	0.84	0.90	0.94	0.90	0.98
β	0.10	0.93	0.95	0.96	0.99	0.98

C Additional Results for Ablation Studies

C.1 Detailed Results across Map-Level Operation

Table 8 presents the full MME subtask results corresponding to the ablation study in the main paper. Overall, Layer-Wise Criss-Cross Attention achieves the strongest performance across most categories, confirming the effectiveness of progressively aggregating map-level information along the depth of the vision encoder. The variant with broadcast further improves the overall MME score.

Across subtasks, we observe several consistent patterns. All map-level operations substantially enhance OCR-related performance compared to the vanilla baseline, indicating improved fine-grained factual grounding. Spatially oriented subtasks such as *Position* and *Scene* also benefit from stronger structural fusion mechanisms: global attention and criss-cross attention perform well, while the layer-wise variant yields the highest scores in both cases. On the other hand, the *Artwork* subtask shows a slight drop across all variants, suggesting that this category may be less sensitive to enhanced token-level structural aggregation.

C.2 Detailed Results for β Ablation

Table 9 presents the per-category MME scores on LLaVA1.5 for different values of β in the range $[0.90, 0.95]$, corresponding to the ablation study in the main paper. Across all settings, MAP consistently outperforms the vanilla baseline (1491.56 total), indicating that the method is stable under moderate variations of β. It can be observed that most categories, such as *OCR*, *Landmark* and *Posters*, exhibit minimal sensitivity to β, with all values yielding strong and nearly identical scores. Overall, $\beta = 0.93$ provides the best trade-off across categories, achieving the highest MME total value of 1529.34.