SeeMe: Mitigating Hallucinations in Large Vision-Language Models through Effective Visual Token Engineering

Kai Tang[1,4,*] **Jinhao You**[2,*] **Bohua Zhang**[1,3] **Yichen Guo**[1,4] **Yiding Sun**[1] **Dongxu Zhang**[5] **Chenxi Li**[6]
Xiande Huang[7,†] **Shanghang Zhang**[1,†]
[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2] University of Pennsylvania
[3] University of Electronic Science and Technology of China
[4] Nanyang Technological University
[5] Tsinghua University
[6] The Chinese University of Hong Kong, Shenzhen
[7] De Artificial Intelligence Lab
[*]Equal contribution. [†]Corresponding authors.

Abstract

Large Vision-Language Models (LVLMs) have achieved remarkable progress in visual understanding tasks such as image captioning and visual question answering. However, they remain susceptible to hallucinations, generating content that is inconsistent with the actual visual input. Existing methods primarily intervene at the decoding stage, while overlooking a critical source of hallucinations: irrelevant or noisy visual tokens that mislead the decoding process. To address this issue, we propose **SeeMe**, a training-free framework that introduces the concept of feature engineering from traditional machine learning into LVLMs. SeeMe restructures visual tokens through a three-stage token engineering process to suppress hallucination sources while preserving informative visual evidence. Experiments on MME, POPE, and AMBER benchmarks across four LVLMs demonstrate that SeeMe consistently reduces hallucinations and improves output consistency, providing a novel perspective for mitigating hallucinations in LVLMs.

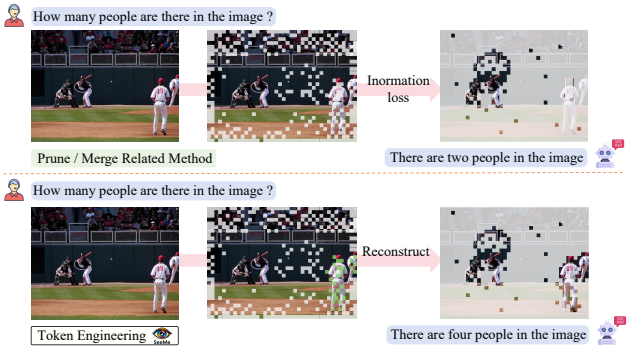

Figure 1. Comparison of visual token handling strategies. Token engineering reconstructs informative representations, leading to more accurate answers.

1. Introduction

Large Vision-Language Models (LVLMs) have achieved remarkable success in open-ended visual understanding tasks, including image captioning, visual question answering, and multimodal dialogue (Liu et al., 2023b; Hu et al., 2023; Zhu & et al., 2023; Bai et al., 2023; Ye et al., 2024). Despite these advances, LVLMs remain prone to hallucination—generating content that is inconsistent with the actual visual input (Liu et al., 2024b). This phenomenon severely undermines their reliability in practical applications, particularly in safety-critical or factual scenarios (Hartsock & Rasool, 2024; Zhou et al., 2024; Ma et al., 2024).

Recent studies have attributed hallucinations in LVLMs to several factors, including over-reliance on statistical biases in training data (Zhou et al., 2023b; Chen et al., 2024b), the dominance of language priors over weak visual grounding (Han et al., 2022; Guan et al., 2024), and the dilution of cross-modal attention in deeper layers (An et al., 2025). To mitigate these issues, recent work has explored training-free strategies that intervene during inference. These methods typically operate on the language decoder's internal dynamics: DoLa (Chuang et al., 2023) compares early and late layer outputs, VCD (Leng et al., 2024) contrasts original and distorted visual inputs, DAMO (Wang et al., 2025) accumulates past activations to stabilize hidden states, and DCLA (Tang et al., 2026) enforces inter-layer consistency

Contact: Kai Tang <kaitang030113@gmail.com>; Correspondence: Shanghang Zhang <shanghang@pku.edu.cn>.
Preprint. July 7, 2026.

to reduce semantic drift.

Although these methods have shown effectiveness, they primarily operate on the language decoder's internal dynamics, overlooking a key factor: visual tokens themselves are often a primary source of hallucinations. In many cases, hallucinations arise from irrelevant or noisy visual tokens passed from the visual encoder to the language decoder, leading the model to generate outputs inconsistent with the actual visual content (Che et al., 2025; Woo et al., 2024; An et al., 2025). Recent studies have addressed this issue by directly manipulating visual tokens: EAZY (Che et al., 2025) identifies hallucinatory tokens through attention analysis and zeros them out, while SPIN (Sarkar et al., 2025) suppresses attention heads that exhibit low attention to image tokens.

However, these suppression-based methods inherently face a trade-off between hallucination mitigation and semantic preservation: aggressive suppression effectively removes noise but risks discarding useful visual evidence. To address this challenge, we propose **SeeMe**, a training-free framework that *restructures* visual tokens through a three-stage token engineering process, rather than merely suppressing them. The first stage, **Sele**ction, prunes irrelevant visual tokens using cross-modal attention. In the **Me**rging stage, similarity-driven fusion is used to combine visual tokens from the original input with similar semantics, forming a high-quality token pool. The final **Sele**ction stage refines the merged tokens by selecting those that best align with the language context, effectively complementing the initial pruning step. Through this progressive restructuring, SeeMe improves visual grounding while reducing hallucination and redundancy, without requiring any retraining or architectural changes.

Experiments on MME, POPE, and AMBER benchmarks demonstrate that SeeMe consistently reduces hallucinations across four mainstream LVLMs: LLaVA-1.5 (Liu et al., 2023b), LLaVA-NEXT (Liu et al., 2024a), INF-MLLM (Zhou et al., 2023a), and mPLUG-Owl2 (Ye et al., 2024), without requiring additional training. The results highlight SeeMe's strong generalizability and effectiveness in hallucination mitigation.

Contributions. Our main contributions are as follows:

- We introduce the concept of *feature engineering* from traditional machine learning into LVLM hallucination mitigation for the first time, proposing to actively restructure visual tokens rather than merely suppressing them.
- We present SeeMe, a training-free, three-stage framework that first filters irrelevant tokens, then enriches the token pool through semantic fusion, and finally refines the representation by selecting high-quality candidates aligned with textual context.
- Extensive experiments on MME, POPE, and AMBER benchmarks across four LVLMs (LLaVA-1.5, LLaVA-NEXT, INF-MLLM, mPLUG-Owl2) demonstrate that SeeMe consistently improves hallucination mitigation, highlighting its effectiveness and generalizability.

2. Related Work

2.1. Large Vision-Language Models

Large Vision-Language Models (LVLMs) have evolved from BERT-style multimodal encoders (Lu et al., 2019; Devlin et al., 2019; Liu et al., 2019) to decoder-based architectures powered by Large Language Models (LLMs) (Touvron et al., 2023; Chiang et al., 2023). In this paradigm, visual inputs are encoded into tokens and passed to a pretrained LLM, enabling unified decoding across modalities. End-to-end training methods (Jia et al., 2021; Radford et al., 2021) have improved cross-modal alignment, while instruction-tuned models such as LLaVA (Liu et al., 2023b) and InstructBLIP (Hu et al., 2023) have further enhanced performance on open-ended vision-language tasks. Recent work has also explored efficient and explicit reasoning for multimodal systems, including chain-of-thought compression, adaptive coarse-to-fine refinement, late-stage fragility analysis, and 3D geometric reasoning benchmarks (Zhang et al., 2026c;a; 2025; 2026b).

2.2. Hallucination in LVLMs

Hallucination—generating content that is inconsistent with the source input—has long been a concern in natural language generation (Yao et al., 2023; Zhu et al., 2025; Xu et al., 2024). With the rise of LVLMs, this problem becomes more prominent, as the model must align high-dimensional visual inputs with text generation in a multimodal setting (Liu et al., 2024b; Ye et al., 2024; Liu et al., 2023b). Hallucinations in LVLMs typically manifest as descriptions of non-existent objects, incorrect attributes, or globally plausible but visually unfaithful outputs (Liu et al., 2023a; Li et al., 2023).

Prior approaches to mitigating hallucinations typically rely on supervised fine-tuning or data augmentation (Wang et al., 2024; Xiao et al., 2025), or train correction modules to detect and revise hallucinated content (Liu et al., 2023a; Gunjal et al., 2024). However, these solutions require additional annotation and computation, limiting their scalability. Recent works have explored training-free mitigation strategies at inference time, such as DCLA (Tang et al., 2026), which enforces inter-layer consistency, DAMO (Wang et al., 2025), which accumulates internal activations to stabilize hidden states, and FADE (Guo et al., 2026), which reduces language-prior dominance. Yet these methods operate entirely on the language decoder's internal dynamics.

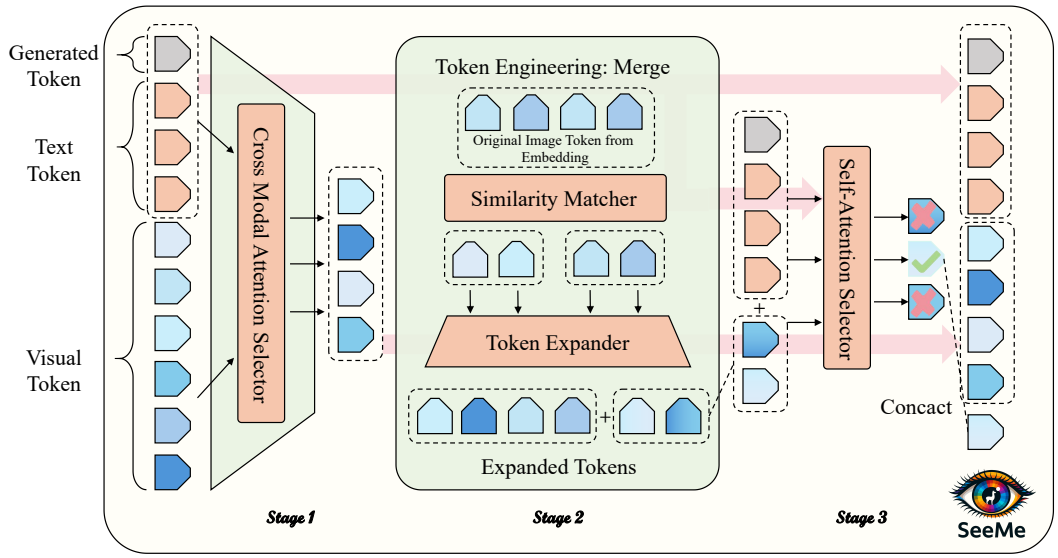

Figure 2. Architecture of the proposed SeeMe framework, which restructures visual tokens through three stages: cross-modal pruning, semantic merging, and self-attention-based refinement.

2.3. Visual Token Manipulation

In typical LVLM architectures, an image is processed by a vision encoder into a sequence of visual tokens, which are then passed to the language decoder (Radford et al., 2021; Hu et al., 2023). A 336×336 image typically yields 576 visual tokens, and higher-resolution images can produce several thousand (Bolya et al., 2022; Shang et al., 2024; Kim et al., 2024). Prior studies have found that many of these tokens receive minimal attention from text queries and contribute little to the final output (Zhang et al., 2024; Guo et al., 2025; Chen et al., 2024a).

To address this inefficiency, token reduction methods have been proposed. ToMe (Bolya et al., 2022) merges similar tokens to reduce computation, while STAR (Guo et al., 2025) prunes visual tokens based on cross-modal attention for efficient inference. More recently, methods have directly targeted hallucination through token manipulation: EAZY (Che et al., 2025) identifies hallucinatory tokens via attention analysis and zeros them out, while SPIN (Sarkar et al., 2025) suppresses attention heads with low image attention. However, these suppression-based approaches face a fundamental trade-off between hallucination mitigation and semantic preservation—aggressive suppression risks discarding useful visual evidence.

3. Motivation

To investigate the relationship between visual token manipulation and hallucination, we conducted a preliminary experiment in which we pruned visual tokens based on cross-modal attention at different decoder layers. The results in Figure 3 show that pruning visual tokens in middle layers significantly reduces hallucinations, but excessive pruning causes information loss and degrades performance. To avoid loss of visual information, we design a visual token engineering process to construct an informative visual token pool, from which task-relevant tokens are further selected to supplement the visual token sets.

Based on this observation, we design **SeeMe** to address hallucination at its root by restructuring visual tokens rather than merely suppressing them.

4. Method

Inspired by the classical philosophy of feature engineering (Zhang et al., 2023), SeeMe treats visual tokens as editable features and applies a three-stage editing process within the frozen decoder of an LVLM. As shown in Figure 2, SeeMe consists of the following steps: (1) an initial **Selection** stage prunes semantically irrelevant visual tokens using cross-modal attention; (2) a **Merging** stage fuses locally similar tokens through similarity-weighted aggregation to recover fine-grained evidence; and (3) a final attention-guided **Selection** retains only the most linguistically aligned fused tokens. This progressive restructuring not only reduces redundancy, but also generates high quality, semantically faithful visual representations—ultimately lowering the risk of hallucination without additional training or architectural changes.

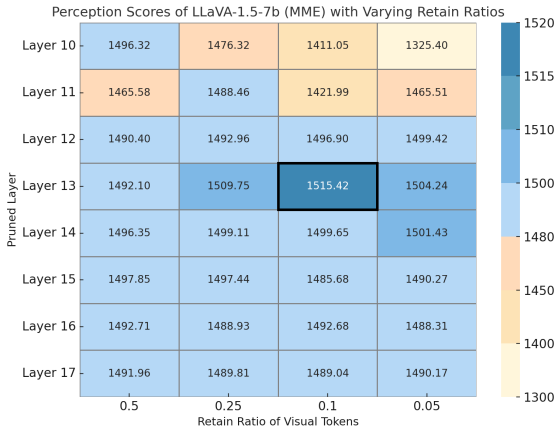

Figure 3. Perception performance of LLaVA-1.5-7B on the MME dataset with varying retain ratios of visual tokens at different intermediate layers.

4.1. Stage 1: Cross-Modal Attention Selector

To reduce unnecessary computation and suppress hallucination sources early in the pipeline, we introduce a cross-modal attention-based filtering mechanism to eliminate semantically irrelevant visual tokens before further processing. We adopt a pruning strategy inspired by STAR (Guo et al., 2025). Specifically, we perform token filtering after a decoder layer where visual-textual attention has matured. This allows the model to first attend to relevant visual details and establish cross-modal grounding. By pruning at this point, we remove semantically irrelevant tokens while preserving those essential for grounding, thereby reducing the risk of hallucination in the language generation stage.

Let $H_v \in \mathbb{R}^{L_v \times d}$ denote the visual token embeddings, and let $H_q \in \mathbb{R}^{L_q \times d}$ and $H_{\text{resp}} \in \mathbb{R}^{L_o \times d}$ represent the embeddings of the input query and the generated response, respectively. We concatenate the textual components to form:

$$\hat{H}_q = [H_q; H_{\text{resp}}] \in \mathbb{R}^{(L_q+L_o)\times d}.$$

At decoder layer K, we extract the cross-modal attention weights produced by the model at layer $K-1$:

$$C_{K-1} = \text{Softmax}\left(\frac{\hat{H}_q H_v^\top}{\sqrt{d}}\right) \in \mathbb{R}^{(L_q+L_o)\times L_v}.$$

We then compute an importance score r_i for each visual token i by averaging its attention weights over all textual tokens:

$$r_i = \frac{1}{L_q + L_o} \sum_{j=1}^{L_q+L_o} C_{K-1}[j, i], \quad \text{for } i = 1, \ldots, L_v.$$

This results in an importance vector $\vec{r} = [r_1, r_2, \ldots, r_{L_v}] \in \mathbb{R}^{L_v}$. We select the top-$k$ visual tokens according to their scores, where $k = \lfloor P \cdot L_v \rfloor$ and $P \in (0, 1)$ is a predefined retention ratio:

$$\text{RetainedIndices} = \text{TopK}(\vec{r}, k).$$

We apply this hard pruning by directly removing the unselected visual tokens from the sequence, while maintaining the relative order of the remaining ones. This design improves efficiency and helps reduce hallucination by filtering out visually irrelevant context early in the decoding process.

Algorithm 1 SeeMe

Input: Visual tokens $\mathbf{Z}_v$, system prompt $\mathbf{X}_s$, textual query $\mathbf{X}_q$
Output: Restructured token sequence $\tilde{\mathbf{Z}}$
Stage 1: Cross-modal Selection
Obtain attention map $\mathbf{A}$ at decoder layer $K-1$
Compute cross-modal attention scores from $[\mathbf{X}_q; \mathbf{X}_{\text{resp}}]$ to $\mathbf{Z}_v$
Select top-$k_1 = r \cdot |\mathbf{Z}_v|$ visual tokens as $\mathbf{Z}_v^{\text{sel}}$
Stage 2: Similarity-guided Token Fusion
Normalize original $\mathbf{Z}_v$ from embedding layer, compute cosine similarity
For each token, retrieve k-nearest neighbors
Merge token with neighbors using similarity-weighted fusion $\rightarrow \mathbf{Z}_v^{\text{fused}}$
Concatenate: $\mathbf{Z}_v^{\text{enh}} = [\mathbf{Z}_v; \mathbf{Z}_v^{\text{fused}}]$
Stage 3: Final Attention-based Selection
Compute cross-modal attention from text to $\mathbf{Z}_v^{\text{enh}}$
Select top-n tokens $\mathbf{Z}_v^{\text{final}}$ by attention score
Output: Concatenate final sequence:
$\tilde{\mathbf{Z}} = [\mathbf{X}_s; \mathbf{Z}_v^{\text{final}}; \mathbf{Z}_v^{\text{sel}}; \mathbf{X}_q]$

4.2. Stage 2: High Quality Token Expander

How to Generate High-Quality Tokens? Although large-scale pruning of visual tokens in Stage 1 helps reduce hallucination by removing semantically irrelevant inputs, it also introduces a potential risk: some tokens that are globally low in attention may still carry locally important visual details and could be mistakenly discarded. These tokens may contain fine-grained information such as edge structures, small objects, or background evidence that contribute to overall scene understanding. Relying solely on global attention for hard filtering can thus lead to semantic degradation.

To address this issue, we introduce a similarity-guided visual token enhancement mechanism in Stage 2. Our method is inspired by classical feature engineering pipelines, particularly those based on the expand-and-compress paradigm such as OpenFE (Zhang et al., 2023). In these frameworks, new features are first expanded from base attributes through composition and then selectively compressed via performance-driven ranking or pruning (in Stage 3). This process allows

models to explore a richer set of representations while maintaining computational tractability. Motivated by this, we propose to treat tokens as features, and design a multi-stage editing framework that performs explicit token-level expansion and compression within large vision-language models. In this view, visual tokens serve as raw features extracted from images; we then enhance and refine their semantic structure by applying multiple rounds of selective retention and similarity-guided fusion. This process, which we term token engineering, reshapes the token space into a more structured and expressive representation.

By combining expansion with targeted selection, our method not only removes redundancy but also generates high-quality, semantically enriched tokens that better support downstream inference and alignment. This module leverages local semantic similarities among the original visual tokens to generate fused representations that recover potentially useful information lost during aggressive pruning. These enriched tokens serve as structural complements to reinforce visual grounding in the decoding process.

Why Use Original Tokens for Fusion? A natural question arises: if Stage 1 causes information loss, why does Stage 2 fuse tokens from the *original* embedding layer rather than directly recovering the pruned tokens? We argue that the pruned tokens are discarded precisely because they have low cross-modal relevance—re-introducing them directly would re-inject noise. Instead, Stage 2 takes a different approach: it generates *new* high-quality tokens by fusing semantically similar original tokens. These fused tokens serve as a "semantic buffer" that may capture fine-grained visual evidence missed by the attention-based pruning. Stage 3 then selects from this enriched pool based on textual alignment, ensuring that only genuinely useful information is retained. Our ablation study in Table 4 validates this design: Stage 2's similarity-guided fusion achieves comparable or better recovery than naively appending original tokens, while being more efficient.

Token Expansion Let $X \in \mathbb{R}^{B\times k\times d}$ denote the original visual tokens set in the embedding layer, where B is the batch size, $k = L_v$ is the number of visual tokens, and d is the dimension of features. We first apply ℓ_2 normalization to each token along the feature dimension and compute the cosine similarity matrix between all token pairs:

$$S = \left(\frac{X}{\|X\|_2}\right) \cdot \left(\frac{X}{\|X\|_2}\right)^{\top} \in \mathbb{R}^{B\times k\times k},$$

To avoid self-similarity, we mask the diagonal entries of S with $-\infty$. Then, for each token, we retrieve its top-t most similar neighbors. Let x_i be the i-th token, and $N_i = \{x_{j_1}, \ldots, x_{j_t}\}$ be its top-t neighbors. We compute the fusion result by:

$$\hat{x}_i^{(j)} = \alpha_{ij} \cdot x_i + (1 - \alpha_{ij}) \cdot x_j, \quad \text{where } \alpha_{ij} = \frac{s_{ij}}{s_{ij} + 1}$$

where s_{ij} is the similarity between token x_i and its neighbor x_j. All $\hat{x}_i^{(j)}$ are concatenated and reshaped into enhanced $k \cdot t$ tokens. These enhanced tokens form an enriched visual representation and serve as input to the next stage, where global attention is used to select the most semantically aligned tokens. In this way, the enriched representation acts as an intermediate buffer that supports a more robust cross-modal grounding in Stage 3.

4.3. Stage 3: Cross-Modal Token Refiner

4.3.1. Stage 3 is Essential for Hallucination Suppression

Figure 4 illustrates the impact of Stage 3 on perception performance across different refinement settings. It is evident that simply retaining all tokens (Without Stage 3) does not effectively suppress hallucination. However, overly aggressive pruning leads to information loss, while insufficient pruning retains excessive noise and redundancy. Only the reasonable setting—corresponding to our proposed SeeMe method—achieves a favorable trade-off, significantly enhancing performance. These results highlight the necessity of stage 3 as a crucial step for balancing semantic retention and hallucination suppression.

4.3.2. Cross-Modal Token Refiner

While the fused tokens generated in Stage 2 help recover fine-grained local information lost during pruning, not all of them are equally relevant to the textual context. To further enhance semantic alignment, we introduce a cross-modal refinement mechanism that selects the most text-aligned fused tokens using global self-attention.

We feed the entire sequence into the decoder layer K and extract its self-attention map $A \in \mathbb{R}^{B\times H\times L\times L}$, where B is the batch size, H is the number of attention heads, and L is the total length of the sequence. For each fused token i, we compute its average attention received from all textual tokens:

$$\text{CrossScore}(i) = \frac{1}{L_q + L_o} \sum_{j=1}^{L_q+L_o} A[:, :, j, i] \in \mathbb{R}^{B\times H}.$$

We average across all heads and the batch dimension to obtain a single scalar score per fused token, and select the top n tokens with the highest scores. These selected fused tokens are then concatenated with the retained visual tokens from Stage 1. This global selection step refines the visual representation by retaining only those fused tokens that are highly

Table 1. Experimental results of various decoding strategies on MME dataset across four models: LLaVA-1.5, LLaVA-NEXT, INF-MLLM and mPLUG-Owl2. The best values are highlighted in **bold**.

Decoding	LLaVA-1.5			LLaVA-NEXT			INF-MLLM			mPLUG-Owl2		
	Perc.	Cog.	Total	Perc.	Cog.	Total	Perc.	Cog.	Total	Perc.	Cog.	Total
Regular	1491.56	294.29	1785.85	1519.30	330.00	1849.30	1491.96	266.07	1758.03	1459.54	345.71	1805.25
VCD	1484.96	287.50	1772.46	1418.27	**351.07**	1769.34	1444.36	270.71	1715.07	1311.52	329.29	1640.81
DoLa	1495.02	**318.21**	1813.23	1515.41	262.14	1777.55	1491.15	265.00	1756.15	1462.33	265.00	1722.33
DCLA	**1520.14**	280.00	1800.14	1525.73	330.00	1855.73	1509.05	**273.21**	**1782.26**	1463.40	334.29	1797.69
SPIN	1491.91	295.71	1787.62	1506.37	326.07	1832.44	1493.21	268.33	1761.54	1465.38	332.56	1797.94
SeeMe	1519.49	310.00	**1829.49**	**1527.34**	346.43	**1873.77**	**1511.17**	269.29	1780.46	**1472.04**	**345.71**	**1817.75**

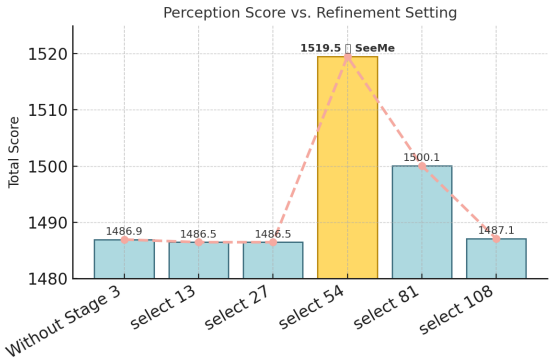

Figure 4. This figure shows that both omitting the final refinement stage (Stage 3) and selecting suboptimal token counts (too few or too many) lead to lower overall scores. Selecting 54 tokens in Stage 3 yields the highest performance, validating its necessity.

aligned with the language stream. It enhances grounding accuracy while suppressing residual hallucinations introduced by low-confidence visual content.

5. Experiment

5.1. Setup

We apply SeeMe with model-specific configurations across different LVLMs. For all models, attention-based cross-modal pruning is applied at a single decoder layer (without additional training), with a certain ratio of visual tokens retained. In the merging stage, each token is fused with its top k most similar tokens using cosine similarity. In the final selection stage, a fixed number of visual tokens is retained based on alignment with the linguistic context. The specific parameter design is reported in Section 5.5 and Table 6.

For other decoding methods, we uniformly set the temperature to 0 for the fairness of the experiment. The specific parameters are presented in Section 5.5.

Datasets To thoroughly evaluate the effectiveness of our SeeMe method in addressing hallucination issues in Large Vision-Language Models(LVLMs), we employed the MME benchmark (Fu et al., 2023), which includes 14 tasks categorized into perception and cognition. Additionally, we focused on assessing SeeMe's performance in mitigating object hallucinations using the POPE benchmark (Polling-based Object Probing Evaluation) (Li et al., 2023), which utilizes SEEM-annotated datasets such as MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022) and GQA (Hudson & Manning, 2019). Furthermore, we used AMBER (Wang et al., 2023) to evaluate the effectiveness of the SeeMe method in suppressing model hallucinations in multi-modal discrimination tasks. AMBER is a multi-dimensional benchmark dataset designed for LVLMs to evaluate the impact of different types of hallucinations (such as presence hallucinations, attribute hallucinations, and relational hallucinations) on model performance.

Models and Baselines We conducted experiments on four recent LVLMs with 7B parameters: LLaVA-1.5 (Liu et al., 2023b), LLaVA-NEXT (Liu et al., 2024a), INF-MLLM (Zhou et al., 2023a), and mPLUG-Owl2 (Ye et al., 2024). These models are commonly used in LVLM benchmarks and vary significantly in their vision-language fusion approaches and pre-training strategies, offering a robust framework for testing the applicability of our method.

For baseline comparisons, we evaluated SeeMe against several prominent decoding techniques, including standard decoding, contrastive decoding methods such as VCD (Leng et al., 2024) and DoLa (Chuang et al., 2023), an inter-layer mechanism: DCLA (Tang et al., 2026), and an image-guided inference approach: SPIN (Sarkar et al., 2025). These methods were selected due to their representation of different conceptual approaches in the current landscape of decoding optimization: VCD and DoLa focus on enhancing decoding strategies, DCLA prioritizes maintaining semantic consistency across layers, and SPIN leverages visual input to guide inference, thus addressing a range of mainstream optimization directions. To ensure fairness and reproducibility, all decoding strategies were evaluated under the same conditions, with the decoding temperature consistently set to zero across all experiments.

Table 2. Experimental results of various decoding strategies on POPE dataset across four models: LLaVA-1.5, LLaVA-NEXT, INF-MLLM and mPLUG-Owl2. We used the average accuracy and F1 score of the Random, Popular and Adversarial splits. The best values are highlighted in **bold**.

Model		*MSCOCO*		*A-OKVQA*		*GQA*	
	Decoding	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
LLaVA-1.5	Regular	85.19	86.10	78.84	82.51	76.57	80.98
	VCD	84.66	85.35	77.86	81.41	75.94	80.36
	DoLa	85.08	85.99	78.70	82.28	76.56	80.84
	DCLA	85.82	86.45	**79.76**	**83.12**	77.41	81.51
	SPIN	85.34	86.12	79.41	82.86	77.22	81.40
	SeeMe	**86.07**	**86.73**	79.66	82.99	**77.50**	**81.52**
LLaVA-NEXT	Regular	87.62	86.42	87.41	87.46	86.58	86.41
	VCD	79.65	74.82	79.24	75.80	78.85	75.27
	DoLa	84.91	82.55	86.46	85.63	84.64	83.20
	DCLA	**87.71**	**86.49**	87.55	87.58	86.61	86.41
	SPIN	86.89	85.35	87.11	87.06	85.99	85.67
	SeeMe	87.62	86.44	**87.61**	**87.77**	**86.62**	**86.47**
INF-MLLM	Regular	88.29	88.36	83.66	84.37	82.77	84.73
	VCD	85.56	85.73	81.70	83.70	79.79	82.05
	DoLa	88.28	88.36	84.12	85.89	82.88	84.81
	DCLA	88.43	88.46	84.14	85.91	83.04	84.94
	SPIN	88.31	88.41	83.46	84.33	83.47	84.98
	SeeMe	**88.47**	**88.56**	**84.70**	**86.01**	**83.67**	**85.01**
mPLUG-Owl2	Regular	86.39	85.91	83.03	83.69	81.18	81.61
	VCD	79.03	78.56	79.03	78.56	79.03	78.00
	DoLa	85.98	**86.34**	83.42	83.42	**83.49**	80.38
	DCLA	86.51	86.19	83.20	84.31	81.61	81.94
	SPIN	86.41	86.03	83.43	84.25	82.11	81.78
	SeeMe	**86.77**	86.32	**83.76**	**84.42**	82.33	**82.27**

5.2. Results

Result on MME We evaluated the performance of SeeMe and several decoding strategies across four models: LLaVA-1.5, LLaVA-NEXT, INF-MLLM, and mPLUG-Owl2, with the experimental results summarized in Table 1. The table presents the Perception, Cognition and Total scores for each model under different decoding strategies. In the LLaVA-1.5, LLaVA-NEXT and mPLUG-Owl2 models, SeeMe outperformed regular decoding, VCD, DoLa, DCLA, and SPIN methods, achieving total scores of 1829.49, 1873.77, and 1817.75, respectively. In the INF-MLLM model, SeeMe also demonstrated strong performance, attaining a total score of 1780.46, significantly surpassing most other decoding strategies. These results clearly highlight the advantages of SeeMe in improving decoding performance and mitigating hallucinations.

Result on POPE To evaluate the effectiveness of our proposed method in mitigating object-level hallucinations, we conducted experiments on the SEEM-annotated versions of the *MSCOCO*, *A-OKVQA*, and *GQA* datasets, which are part of the POPE benchmark. The POPE benchmark is widely used for assessing hallucinations in visual question answering (VQA) tasks, focusing on the accuracy and reliability of generated object mentions in image captions. The benchmark includes three distinct splits: Random Split (any object from the dataset), Popular Split (the most frequent objects in the dataset), and Adversarial Split (objects that are closely related but misleading). These splits represent different levels of challenge, from general object recognition to dealing with tricky or misleading objects that may cause hallucinations. In this study, we compare the performance of SeeMe with several representative decoding strategies, including regular decoding, VCD, DoLa, DCLA and SPIN. For each of the three splits, we calculate the average score across all categories and use it as the final evaluation score to assess the overall effectiveness of the decoding strategies in reducing hallucinations.

As shown in Table 2, SeeMe achieved high accuracy and F1 scores on the *MSCOCO*, *A-OKVQA*, and *GQA* datasets across the LLaVA-1.5, LLaVA-NEXT, INF-MLLM, and mPLUG-Owl2, surpassing other decoding strategies and demonstrating substantial improvements. These results highlight SeeMe's ability to consistently enhance model performance, particularly in terms of inference accuracy and effectively mitigating hallucinations in LVLMs.

Results on AMBER To thoroughly evaluate the effectiveness of our SeeMe method in addressing hallucination issues in Large Vision-Language Models (LVLMs), we em-

Table 3. Experimental results on discriminative tasks from AMBER dataset across four models: LLaVA-1.5, LLaVA-NEXT, INF-MLLM and mPLUG-Owl2. We used the overall accuracy and F1 score of discriminative tasks. The best values are highlighted in **bold**.

Model	Decoding	Accuracy	F1 Score
LLaVA-1.5	Regular	71.5	74.1
	VCD	72.0	74.9
	DoLa	71.5	74.2
	DCLA	72.6	75.7
	SPIN	72.3	75.3
	SeeMe	**72.7**	**75.9**
LLaVA-NEXT	Regular	83.6	87.7
	VCD	82.6	87.1
	DoLa	83.3	87.7
	DCLA	83.5	87.7
	SPIN	83.1	87.5
	SeeMe	**83.7**	**87.8**
INF-MLLM	Regular	71.4	74.3
	VCD	71.5	74.4
	DoLa	72.2	74.8
	DCLA	72.7	75.1
	SPIN	70.9	74.1
	SeeMe	**73.4**	**75.1**
mPLUG-Owl2	Regular	76.3	78.8
	VCD	75.9	78.4
	DoLa	76.1	79.1
	DCLA	76.5	79.3
	SPIN	76.4	78.8
	SeeMe	**76.8**	**79.7**

ployed the AMBER benchmark, which offers a comprehensive framework for hallucination evaluation. AMBER provides a multi-dimensional evaluation that focuses on three primary types of hallucinations: existence, attribute, and relation hallucinations, making it a valuable tool for assessing the impact of hallucinations on model performance. We use discriminative tasks in AMBER to examine SeeMe's ability to mitigate hallucinations by improving the model's judgment of object existence, attributes and relationships, ensuring the alignment of model responses with actual visual content. The experimental data in Table 3 indicates that, compared to other decoding strategies, SeeMe achieved the best results across LLaVA-1.5, LLaVA-NEXT, INF-MLLM and mPLUG-Owl2. Overall, SeeMe exhibits superior anti-hallucination capability and discrimination accuracy, thereby validating its effectiveness in mitigating hallucinations in multimodal discrimination tasks.

5.3. Ablation Study

We conducted a comprehensive ablation study to analyze the contributions of different components in SeeMe's three-stage pipeline. As shown in Table 4, we examine the effect of each stage under different retain ratios.

The results reveal several key findings: (1) Pure pruning (Stage 1 only) causes significant information loss, especially at low retain ratios (↓166 at r=0.05); (2) Appending original embedding tokens recovers most of the lost performance, proving that pruned tokens contain valuable visual semantics; (3) Stage 2's similarity-guided fusion achieves comparable or better recovery than naive appending, while being more efficient; (4) Stage 3 consistently provides additional refinement (+2∼4 points) through cross-modal alignment. A more detailed layer/ratio analysis is provided in Section 5.6.

Table 4. Ablation study on MME at decoder layer 10. Pure pruning causes information loss; appending original tokens recovers it. Stage 2 achieves similar recovery, and Stage 3 further refines alignment.

Method	r=0.5	r=0.25	r=0.1	r=0.05
A: S1 only	1496.3	1476.3	1411.1	1325.4
B: A + Orig. tokens	1496.4	1483.4	1488.3	1473.7
C: S1 + S2	1497.7	1486.7	1490.5	1473.6
D: SeeMe (Full)	**1499.8**	**1489.6**	**1492.4**	**1478.1**
Key comparisons:				
Δ(B−A): Orig. tokens	+0.1	+7.1	+77.3	+148.3
Δ(C−A): S2 recovery	+1.4	+10.4	+79.4	+148.2
Δ(C−B): S2 vs append	+1.3	+3.3	+2.2	−0.1
Δ(D−C): S3 refine	+2.1	+2.1	+2.0	+4.5

Table 5. Efficiency comparison of different decoding strategies on LLaVA-1.5-7B. ↓ indicates lower is better; ↑ indicates higher is better. The best values are highlighted in **bold**.

	Prefill Latency↓ (ms/token)	Decode Latency↓ (ms/token)	GPU Usage↓ (GB)	Throughput↑ (Images/s)
Regular	50.46	**21.65**	14.72	9.14
VCD	174.18	174.65	**14.04**	2.85
DoLa	135.11	130.52	14.92	1.39
DCLA	67.46	30.08	18.77	7.45
SPIN	74.88	22.19	15.27	7.80
SeeMe	**45.56**	22.73	15.88	**9.23**

5.4. Efficiency Analysis

We analyze the computational overhead of different decoding strategies on LLaVA-1.5-7B. As shown in Table 5, SeeMe achieves the lowest prefill latency (45.56 ms/token), while maintaining comparable decode latency (22.73 ms/token). More importantly, SeeMe attains the highest throughput (9.23 images/s) among all methods, while not incurring significant memory overhead compared to regular decoding. These results demonstrate that SeeMe's three-stage pipeline introduces negligible computational cost while delivering superior hallucination mitigation, making it highly practical for real-world deployment.

5.5. Hyperparameter Settings

For the MME and POPE benchmarks, SeeMe is activated at the 14th decoder layer of LLaVA-1.5 and mPLUG-Owl2, at the 16th layer of LLaVA-NEXT, and at the 13th layer

Table 6. Hyperparameter settings of SeeMe for each model.

	LLaVA-1.5	LLaVA-NEXT	INF-MLLM	mPLUG-Owl2
K	14	16	13	14
ratio	0.05	0.05	0.05	0.05
top-k	2	1	2	2
select-n	54	27	27	54

of INF-MLLM, each with a pruning retain ratio of 0.05. During merging, each token is fused with its most similar neighbors: top-2 by cosine similarity for LLaVA-1.5, INF-MLLM, and mPLUG-Owl2, and top-1 for LLaVA-NEXT. The final number of selected tokens (n) is fixed at 54 for LLaVA-1.5 and mPLUG-Owl2, and 27 for INF-MLLM and LLaVA-NEXT. When evaluating on AMBER, we adjust the retain ratio to 0.01.

For all other methods, we use the official open-source parameters, while AMBER retains the same parameters as the other datasets. As a special case, we explicitly set the temperature of VCD to 0 to ensure experimental fairness. SPIN uses its POPE parameters when evaluated on AMBER.

Hyperparameter Selection Guidelines. To facilitate the application of SeeMe to new models, we provide practical guidelines for hyperparameter selection. We recommend selecting K around 40–50% of the total decoder layers. For a 32-layer model, $K \in [13, 16]$ typically works well. The key insight is that cross-modal attention should have matured, meaning that visual-textual alignment has been established, but should not yet be diluted by overly dispersed attention. Pruning too early loses relevant tokens, while pruning too late misses the opportunity to filter noise before it propagates.

A retain ratio of 0.05 works robustly across all tested models. Lower ratios, such as 0.01, may improve hallucination suppression but risk information loss, which Stage 2 and Stage 3 must compensate for. Higher ratios, such as 0.25, are safer but provide less noise reduction. For the merging and final selection stages, top-$k \in \{1, 2\}$ and select-$n \approx 0.1 \times |\mathbf{Z}_v|$ provide a good balance. For example, selecting 54 tokens works well when the input contains 576 visual tokens. Larger k generates more fused tokens but increases redundancy, while smaller n is more aggressive and may discard useful fused tokens.

For a new model, we suggest starting with $K = \lfloor 0.45 \times \text{num\_layers} \rfloor$, ratio $= 0.05$, top-$k = 2$, and select-$n = 54$. A small validation set, such as 100 samples from MME, can then be used to adjust K by ± 2 layers and tune the ratio or select-n according to the trade-off between hallucination suppression and information retention.

5.6. Extended Ablation Study

As shown in Figure 3, the layer/retain-ratio sweep measures how aggressively to apply cross-modal pruning on the MME perception task. We sweep the retain ratio of visual tokens (50%, 25%, 10%, 5%) at decoder layers 10–17 in LLaVA-1.5-7B, using the regular decoding score of 1491.56 as the baseline. The results show that pruning in intermediate layers, especially layers 12–14, consistently exceeds the baseline, whereas pruning too early or too late is less stable.

6. Conclusion

In this paper, we identify that redundant and noisy visual tokens can mislead LVLMs, leading to output that is inconsistent with the actual visual information. To address this issue, we proposed **SeeMe**, the first method to integrate the concept of feature engineering from traditional machine learning into LVLMs, offering an innovative solution to the hallucination problem. Through a three-stage visual tokens reconstruction process, SeeMe effectively reduces redundancy and noise in visual tokens without requiring additional training, minimizing the inconsistency between the generated content and the actual visual input, and enhancing the overall reliability and effectiveness of the model. Extensive experiments across multiple benchmark datasets demonstrate the effectiveness of SeeMe in various multimodal tasks, particularly in visual question answering, where it significantly improves the accuracy and stability of the generated results.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, specifically in improving the reliability of Large Vision-Language Models by mitigating hallucinations. Our method, SeeMe, is a training-free approach that enhances the accuracy of visual understanding without requiring additional data or computational resources for retraining. The potential societal benefits include more reliable AI systems for applications such as medical image analysis, autonomous driving, and assistive technologies. We do not foresee any specific negative societal consequences that must be highlighted here, though we encourage responsible deployment and continued research into AI safety and reliability.

References

An, W., Tian, F., Leng, S., Nie, J., Lin, H., Wang, Q., Chen, P., Zhang, X., and Lu, S. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29915–

29926, 2025.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Che, L., Liu, T. Q., Jia, J., Qin, W., Tang, R., and Pavlovic, V. Eazy: Eliminating hallucinations in lvlms by zeroing out hallucinatory image tokens. *arXiv preprint arXiv:2503.07772*, 2025.

Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., and Chang, B. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024a.

Chen, X., Ma, Z., Zhang, X., Xu, S., Qian, S., Yang, J., Fouhey, D., and Chai, J. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418, 2024b.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., and He, P. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., and Ji, R. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. URL https://arxiv.org/abs/2306.13394.

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

Gunjal, A., Yin, J., and Bas, E. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024.

Guo, Y., Li, H., Zhang, Z., You, J., Tang, K., and Huang, X. Star: Stage-wise attention-guided token reduction for efficient large vision-language models inference. *arXiv preprint arXiv:2505.12359*, 2025.

Guo, Y., Tang, K., Lin, F., Sun, Y., Zhang, D., Wang, W., Cong, L. W., and Zhang, S. Fade: Mitigating hallucinations by reducing language-prior dominance in large vision-language models, 2026. URL https://arxiv.org/abs/2606.29431.

Han, Y., Nie, L., Yin, J., Wu, J., and Yan, Y. Visual perturbation-aware collaborative learning for overcoming the language prior problem. *arXiv preprint arXiv:2207.11850*, 2022.

Hartsock, I. and Rasool, G. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in Artificial Intelligence*, 7:1430984, 2024.

Hu, X., Gao, J., Li, C., and et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

Hudson, D. A. and Manning, C. D. 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*, pp. 6700–6709, 2019.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Kim, M., Gao, S., Hsu, Y.-C., Shen, Y., and Jin, H. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1383–1392, 2024.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. 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*, pp. 13872–13882, 2024.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

Liu, H., Zhang, P., Yang, Z., Yang, J., Yuan, L., and Zhang, L. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llavanext: Improved reasoning, ocr, and world knowledge, 2024a.

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., and Peng, W. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024b.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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

Ma, Y., Song, Z., Zhuang, Y., Hao, J., and King, I. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sarkar, S., Che, Y., Gavin, A., Beerel, P. A., and Kundu, S. Mitigating hallucinations in vision-language models through image-guided head suppression. *arXiv preprint arXiv:2505.16411*, 2025.

Schwenk, D., Khandelwal, A., Clark, C., Marino, K., and Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Shang, Y., Cai, M., Xu, B., Lee, Y. J., and Yan, Y. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

Tang, K., You, J., Guo, Y., Sun, Y., Zhang, D., Wang, W., Li, H., Luo, T., Li, R., Huang, X., and Zhang, S. Mitigating hallucinations via inter-layer consistency aggregation in large vision-language models, 2026. URL https://arxiv.org/abs/2505.12343.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

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

Wang, L., He, J., Li, S., Liu, N., and Lim, E.-P. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pp. 32–45. Springer, 2024.

Woo, S., Kim, D., Jang, J., Choi, Y., and Kim, C. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *arXiv preprint arXiv:2405.17820*, 2024.

Xiao, W., Huang, Z., Gan, L., He, W., Li, H., Yu, Z., Shu, F., Jiang, H., and Zhu, L. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25543–25551, 2025.

Xu, Z., Jain, S., and Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

Yao, J.-Y., Ning, K.-P., Liu, Z.-H., Ning, M.-N., Liu, Y.-Y., and Yuan, L. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023.

Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., and Huang, F. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 13040–13051, 2024.

Zhang, D., Wu, Y., Sun, Y., Zhu, J., Yang, J., Xin, M., and Tian, B. Not all errors are created equal: Ascot addresses late-stage fragility in efficient llm reasoning. *arXiv Prepr. arXiv:2508.05282*, 2025.

Zhang, D., Lin, H., Sun, Y., Wang, P., Wang, Q., Yang, N., and Zhu, J. Not all queries need deep thought: Coficot for adaptive coarse-to-fine stateful refinement. In *Ann. Conf. Uncertain. Artif. Intell.*, 2026a.

Zhang, D., Sun, Y., Li, P., Liu, Y., Lin, H., Xu, H., Mu, X., Lin, L., Yan, W., Yang, N., et al. Pointcot: A multi-modal benchmark for explicit 3d geometric reasoning. *arXiv Prepr. arXiv:2602.23945*, 2026b.

Zhang, D., Sun, Y., Tan, C., Yan, W., Yang, N., Zhu, J., and Zhang, H. Chain-of-thought compression should not be blind: V-skip for efficient multimodal reasoning via dual-path anchoring. In *Ann. Meet. Assoc. Comput. Linguist.*, 2026c.

Zhang, T., Zhang, Z. A., Fan, Z., Luo, H., Liu, F., Liu, Q., Cao, W., and Jian, L. Openfe: Automated feature generation with expert-level performance. In *International Conference on Machine Learning*, pp. 41880–41901. PMLR, 2023.

Zhang, Y., Fan, C.-K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Zhou, Q., Wang, Z., Chu, W., Xu, Y., Li, H., and Qi, Y. Infmllm: A unified framework for visual-language tasks. *arXiv preprint arXiv:2311.06791*, 2023a.

Zhou, X., Liu, M., Yurtsever, E., Zagar, B. L., Zimmer, W., Cao, H., and Knoll, A. C. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, 2024.

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023b.

Zhu, D. and et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Zhu, L., Ji, D., Chen, T., Xu, P., Ye, J., and Liu, J. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1624–1633, 2025.