RVSD: Retrieval Vision Sparse Decoding for Mitigating Visual Hallucinations in Large Vision-Language Models

Canjie Liu[1], **Jiawen Kang**[†1], **Jinbo Wen**[2], **Zishao Zhong**[3]
[1] School of Automation, Guangdong University of Technology, Guangzhou, China
[2] Department of Computer Science, City University of Hong Kong, Hong Kong, China
[3] The Second Affiliated Hospital of Guangzhou University of Chinese Medicine, Guangzhou, China
2112404359@mail2.gdut.edu.cn, kavinkang@gdut.edu.cn
jinbowen2-c@my.cityu.edu.hk, zhongzishao@gzucm.edu.cn

Abstract

Large vision-language models have achieved remarkable success in vision-language tasks. However, they remain prone to Visual Hallucinations (VHs), undermining their reliability in real-world applications. Existing solutions typically require curated datasets, additional training, or multi-round decoding, resulting in considerable computational overhead. In this paper, we propose **RVSD** (Retrieval Vision Sparse Decoding), a training-free and plug-and-play decoding framework that, for the first time, unifies token sparsification and **Semantic-Space Visual Retrieval** (SSVR) within a single decoding pass. Within RVSD, we introduce a **semantics-directed token selection** strategy that selectively sparsifies redundant tokens while preserving critical visual information. We further propose the SSVR mechanism, which reformulates visual compensation as an on-demand cross-modal retrieval process within a shared semantic space. Extensive experiments demonstrate that RVSD achieves state-of-the-art performance in mitigating VHs while maintaining robust suppression capabilities under long-context generation settings. Our code is available here.[1]

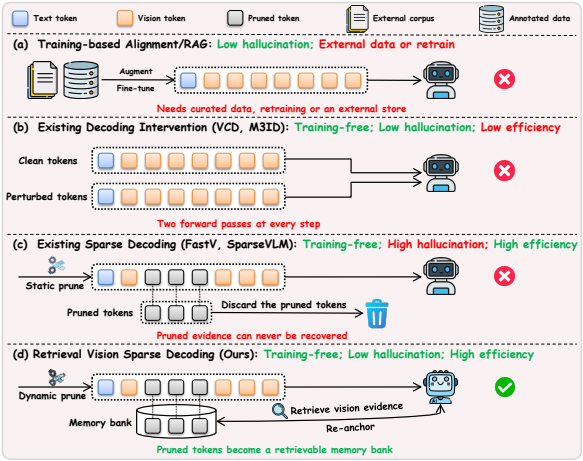

Figure 1: Comparison of representative paradigms for mitigating VHs in LVLMs: (a) training-based alignment and RAG (Sun et al., 2024), (b) decoding interventions (Leng et al., 2024; Favero et al., 2024), (c) sparse decoders (Chen et al., 2024; Zhang et al., 2025), and (d) our RVSD, which turns pruned tokens into a retrievable memory bank and restores grounding on demand in a single forward pass.

1 Introduction

Large Vision-Language Models (LVLMs) have recently demonstrated impressive capabilities across a wide range of vision-language tasks (Liu et al., 2024b; Bai et al., 2023; Liu et al., 2024c), such as image captioning, visual question answering, and open-ended visual reasoning. Despite their rapid progress, LVLMs remain highly susceptible to Visual Hallucinations (VHs) (Bai et al., 2024; Liu et al., 2024a), where generated responses contradict or lack grounding in the actual visual content. This fundamental flaw seriously compromises the reliability of LVLMs in real-world applications (Wang et al., 2023b), where misinterpretation of visual data can lead to significant consequences.

Existing solutions can be categorized into four paradigms, which we compare with RVSD in Fig. 1. Training-based alignment (Sun et al., 2024) mitigates VHs via additional supervision and preference alignment, but requires expensive retraining and annotated data. Retrieval-augmented generation (Lewis et al., 2020) grounds generation using external knowledge bases, yet introduces substantial storage and retrieval overhead. Decoding-side interventions (Liu et al., 2025) adjust token distributions during inference, but often require iterative decoding procedures that incur considerable latency. More recently, sparsification-based decoding (Dang et al., 2026) has emerged as a promising direction that simultaneously accelerates inference and suppresses distracting visual tokens, offering an attractive trade-off between efficiency and VH mitigation for LVLMs.

† Corresponding author.
[1] https://github.com/canjie-liu/RVSD

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 45789–45806
October 24-29, 2026 ©2026 Association for Computational Linguistics

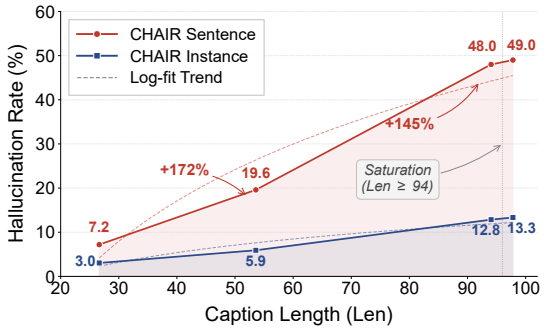

Figure 2: Sparsification-hallucination paradox, exemplified by the representative sparse decoder VASparse (Zhuang et al., 2025) on LLaVA-1.5 using the Caption Hallucination Assessment with Image Relevance (CHAIR) dataset.

However, existing sparse decoding methods remain largely modality-agnostic, often pruning visual tokens that appear statistically redundant yet contain semantically critical information (Zhuang et al., 2025). Once such fine-grained visual tokens are discarded, the resulting information loss becomes irreversible. Our empirical analysis further reveals that the hallucination rates of State-of-the-Art (SOTA) sparse decoders (Zhuang et al., 2025) increase substantially as generation length grows, as shown in Fig. 2. We refer to this phenomenon as the *sparsification-hallucination paradox*, highlighting the fundamental tension between inference efficiency and reliable visual grounding in LVLMs.

In this paper, we propose **RVSD** (Retrieval Vision Sparse Decoding), a training-free and plug-and-play decoding framework for mitigating VHs in LVLMs. RVSD addresses the sparsification-hallucination paradox by combining semantics-directed token sparsification with on-demand visual compensation in a unified single-pass inference process. Our contributions are as follows:

- We propose RVSD, a novel training-free and plug-and-play decoding framework that addresses the sparsification-hallucination paradox. RVSD jointly integrates token sparsification and dynamic visual compensation within a single inference pass, achieving an effective balance between inference efficiency and visual faithfulness.

- We develop a semantics-directed token selection strategy that selectively distinguishes redundant tokens from visually critical ones. Unlike prior modality-agnostic pruning methods, the proposed strategy performs targeted sparsification while preserving essential visual grounding information.

- We propose a Semantic-Space Visual Retrieval (SSVR) mechanism that performs on-demand cross-modal retrieval in a shared semantic space. When visual grounding becomes insufficient during generation, SSVR dynamically retrieves relevant visual evidence from a deferred visual memory bank to recalibrate hallucinated output distributions.

- Extensive experiments across multiple prominent benchmarks demonstrate that RVSD achieves SOTA performance, simultaneously improving both generation quality and decoding efficiency while exhibiting strong robustness under long-context generation settings.

2 Related Work

Large Vision-Language Models. Building upon the rapid advances in large language models (Zhao et al., 2026), recent research has extended them to the multimodal domain, giving rise to LVLMs (Liu et al., 2023, 2024b,c). LVLMs are typically trained via large-scale vision-language pretraining followed by instruction-based fine-tuning, while subsequent studies further align generation behavior with human preferences via Reinforcement Learning from Human Feedback (RLHF) (Sun et al., 2024) and preference optimization techniques. Despite their strong multimodal capabilities, LVLMs remain prone to VHs (Bai et al., 2024; Yin et al., 2024; Wu et al., 2023), severely undermining the reliability of LVLMs in real-world applications.

Visual Hallucination Mitigation. Existing efforts mitigate VHs from several perspectives. Training-centric approaches (Sun et al., 2024) improve multimodal alignment through additional supervision, RLHF, or preference optimization techniques. Decoding-time methods intervene during inference, where contrastive strategies, such as VCD (Leng et al., 2024) and M3ID (Favero et al., 2024), suppress linguistic priors via perturbed visual inputs, while attention-guided methods, including VTI (Liu et al., 2025), AvisC (Woo et al., 2025), and SHARP (Wu et al., 2025), enhance attention toward visually relevant regions. Adaptive visual reinforcement methods (Zhu et al., 2026) further strengthen visual evidence on demand during generation. Closest to our setting, RVCD (Lee and Song, 2025) performs retrieval-based contrastive

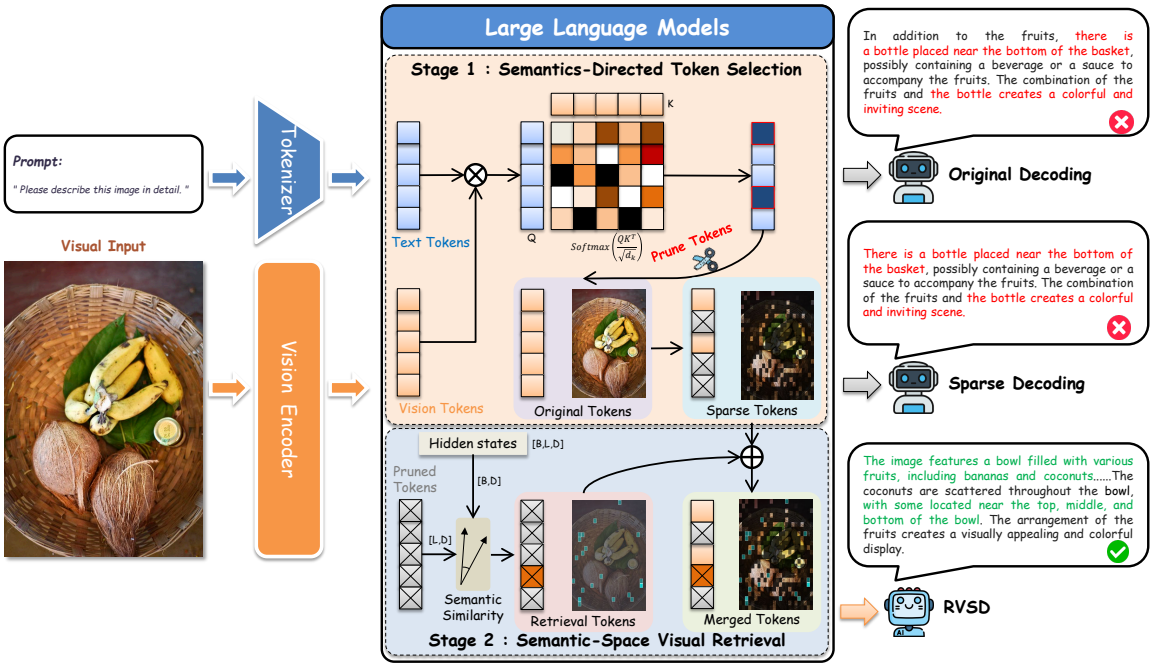

Figure 3: Overview of the proposed RVSD framework, which consists of two stages. Stage 1 performs semantics-directed token selection, partitioning visual tokens into an active sparse set and a deferred memory bank through cross-modal attention. Stage 2 applies the SSVR mechanism to retrieve on-demand visual evidence from the deferred bank and re-anchor decoding within a single forward pass, transforming hallucinated content from the original and sparse decoders into faithful generation.

decoding by referencing an offline database of generated single-concept images at the logit level. In contrast, RVSD requires no external image store, detector, or draft decoding pass, and retrieves the model's own deferred visual tokens directly in the semantic space of a single forward pass. Despite their different formulations, these methods still rely on dense visual processing throughout decoding, leaving substantial room for improving efficiency-aware visual grounding.

Sparse Decoding for Efficient Inference. Token sparsification has been utilized to improve both inference efficiency and visual faithfulness (Wu et al., 2026). General-purpose accelerators, such as SparseVLM (Zhang et al., 2025) and FastV (Chen et al., 2024), prune visual tokens using modality-agnostic criteria primarily designed for acceleration. Recent methods, such as VASparse (Zhuang et al., 2025), LTS-FS (Dang et al., 2026), PruneHal (Sun et al., 2025), MINT (Wang et al., 2025), and EAZY (Che et al., 2025), further explore hallucination-aware sparsification by enabling LVLMs to focus on task-relevant visual tokens. Nevertheless, existing methods still perform irreversible pruning, which may remove fine-grained visual evidence. Unlike prior methods, RVSD reorganizes discarded tokens into a deferred visual memory bank, enabling on-demand restoration of visual grounding within the same decoding pass without additional training or iterative decoding.

3 Problem Formulation

Given an LVLM $\mathcal{M}_\theta^{\text{MLLM}}$ parameterized by θ, the model consists of a text embedding layer, a vision encoder, a vision-language interface module, a decoder with L transformer layers, and an affine output layer $\varsigma(\cdot)$ for next-token prediction. For an image-grounded generation task with textual query x and input image v, the vision encoder first extracts visual features from v, which are then transformed into aligned visual tokens z_v via a Multi-Layer Perceptron (MLP) or Q-Former interface. The visual tokens z_v are concatenated with the query x and fed into the decoder to autoregressively generate the response y, given by

$$y_t \sim p_\theta(\cdot|v, x, y_{<t}) \propto \text{softmax}(f_\theta(\cdot|v, x, y_{<t})), \quad (1)$$

where y_t denotes the token generated at step t, $y_{<t}$ denotes the previously generated prefix, and f_θ represents the unnormalized logits from $\mathcal{M}_\theta^{\text{MLLM}}$.

When the generated response y contradicts or is unsupported by the input image v, the LVLM is considered to exhibit VHs. Accordingly, the objective of VH mitigation is to suppress unfaithful generations while preserving overall response quality. In the following, we present RVSD, which

improves visual faithfulness during inference with minimal computational overhead.

4 Methodology

As illustrated in Fig. 3, we present the architecture of the proposed RVSD. In the following, we describe its single-pass inference process and two key components: semantics-directed token selection and the SSVR mechanism.

4.1 Single Forward Pass

Let the visual token bank be denoted by $Z_v = [z_1, \ldots, z_{N_v}] \in \mathbb{R}^{N_v \times D}$, where N_v is the total number of visual tokens and D is the hidden dimension. The decoder hidden states at decoding step t in layer l are denoted by $H_t^{(l)} = [h_1^{(l)}, \ldots, h_t^{(l)}]$, in which each $h_i^{(l)} \in \mathbb{R}^D$ denotes the hidden state at textual position i. Thus, the text-to-vision attention weight $A_{i,j}^{(l)}$ from the i-th text token to the j-th visual token at layer l is defined as

$$A_{i,j}^{(l)} = \frac{\exp\left((W_Q^{(l)} h_i^{(l)})^\top (W_K^{(l)} z_j)/\sqrt{D}\right)}{\sum_{r=1}^{N_v} \exp\left((W_Q^{(l)} h_i^{(l)})^\top (W_K^{(l)} z_r)/\sqrt{D}\right)}, \quad (2)$$

where $W_Q^{(l)}, W_K^{(l)}$ denote the query and key projection matrices at layer l, respectively.

At each decoding step, RVSD maintains two complementary visual token sets: an active sparse set $\mathcal{S}_t$ for standard decoding and a deferred set $\mathcal{P}_t$ that forms a deferred visual memory bank, where $\mathcal{S}_t \cup \mathcal{P}_t = 1, \ldots, N_v$ and $\mathcal{S}_t \cap \mathcal{P}_t = \varnothing$. Decoding is performed over $\mathcal{S}_t$, while $\mathcal{P}_t$ remains semantically retrievable and provides relevant visual evidence on demand. Based on Eq. (1), RVSD autoregressively predicts the token at generation step t as

$$p_\theta^{\text{RVSD}} = \text{softmax}\left(f_\theta(\cdot | Z_v[\mathcal{S}_t], x, y_{<t}; \Delta_t)\right), \quad (3)$$

where Δ_t denotes an optional SSVR branch activated only when visual grounding becomes insufficient. This design preserves decoding efficiency under normal generation while enabling query-driven visual retrieval during uncertain steps, without requiring multi-round decoding. Algorithm 1 presents the overall single forward pass of RVSD. The dynamic sparsity module for creating $\mathcal{S}_t$ and $\mathcal{P}_t$ and the SSVR module realizing Δ_t are detailed in Sections 4.2 and 4.3, respectively.

4.2 Dynamic Sparsity

The objective of dynamic sparsity is to retain visual tokens most relevant to the salient textual semantics in the current generation trajectory. Rather than performing modality-agnostic pruning, RVSD explicitly scores visual tokens by aggregating cross-modal attention from salient textual positions. At decoding step t, we first compute textual saliency over generated tokens $\{1, \ldots, t-1\}$ as

$$\beta_i^t = \frac{\exp(\|h_i^{(L)}\|_2/\tau)}{\sum_{r=1}^{t-1} \exp(\|h_r^{(L)}\|_2/\tau)}, \quad (4)$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm, $h_i^{(L)}$ is the last-layer hidden state of the i-th token, and τ is a temperature parameter. We then select the major textual token set $\mathcal{T}_t = \text{TopK}(\beta^t, m)$, where m is the number of major textual tokens retained. Given a layer subset $\mathcal{L}_s$ for stable attention aggregation, we define the visual relevance score s_j^t as

$$s_j^t = \sum_{i \in \mathcal{T}_t} \beta_i^t \Big(\frac{1}{|\mathcal{L}_s|} \sum_{l \in \mathcal{L}_s} A_{i,j}^{(l)}\Big). \quad (5)$$

To suppress step-wise noise, we apply temporal smoothing, which is expressed as

$$\tilde{s}^t = \eta \tilde{s}^{t-1} + (1-\eta) s^t, \quad (6)$$

where $\eta \in [0, 1)$ is the smoothing coefficient, and $\tilde{s}^t$ represents the smoothed relevance vector at step t. We then construct the active sparse set as

$$\mathcal{S}_t = \text{TopK}(\tilde{s}^t, k), \;\; \mathcal{P}_t = \{1, \ldots, N_v\} \setminus \mathcal{S}_t, \quad (7)$$

where k is the budget that controls the size of the retained sparse set. Therefore, RVSD retains the visual tokens most relevant to the current decoding

Algorithm 1 Single Forward Pass of RVSD

Require: An LVLM $\mathcal{M}_\theta$, an image v, and a query x
Encode visual tokens: $Z_v \leftarrow \mathcal{M}_\theta^{\text{enc}}(v)$; initialize $\tilde{s}^0 \leftarrow \mathbf{0}$
for $t = 1, 2, \ldots$ **do**
  Select token subset $\mathcal{T}_t$ via β^t using Eq. (4)
  Update smoothed saliency $\tilde{s}^t$ from s^t using Eq. (6)
  Partition visual tokens into $\mathcal{S}_t$ and $\mathcal{P}_t$ using Eq. (7)
  trigger $\leftarrow$ TRUE
  for $l = l_s$ **to** l_e **do**
    Compute uncertainty score $u_t^{(l)}$ using Eq. (8)
    if *trigger* **and** $u_t^{(l)} > \gamma$ **then**
      Retrieve deferred tokens from $Z_v[\mathcal{P}_t]$ via query $q_t^{(l)}$ using Eq. (9)
      Aggregate retrieved features $\tilde{Z}_v$ from top-k_c candidates using Eq. (10)
      Construct adaptive adapter weights $W_{1,2}^{\text{adpt}}$ using Eq. (11)
      Fuse retrieved visual evidence using Eq. (13)
      trigger $\leftarrow$ FALSE
    end if
  end for
  Predict token y_t over $Z_v[\mathcal{S}_t]$ using Eq. (3)
end for
return Response sequence y

context while deferring the remaining tokens for potential later retrieval. With $k \ll N_v$, the per-layer attention complexity is reduced from $\mathcal{O}((t+N_v)^2)$ to $\mathcal{O}((t+k)^2)$.

4.3 Semantic-Space Visual Retrieval

Dynamic sparsity alone may still discard fine-grained visual evidence required at later decoding stages. To address this issue, RVSD reformulates visual compensation as on-demand cross-modal retrieval in a shared semantic space. Instead of directly restoring pruned features, RVSD treats the deferred set $\mathcal{P}_t$ as a deferred visual memory bank and uses the current decoder state as a semantic query to retrieve relevant visual evidence whenever grounding becomes insufficient.

4.3.1 Uncertainty-triggered retrieval gate

At decoding step t, we scan intermediate layers and compute the normalized predictive uncertainty as

$$u_t^{(l)} = -\frac{1}{\log N} \sum_{n=1}^{N} p_{\theta,n}^{(l)} \log p_{\theta,n}^{(l)}, \tag{8}$$

where N is the vocabulary size, and $p_{\theta,n}^{(l)}$ represents the predicted probability of the n-th vocabulary token derived from the layer-l hidden state. When $u_t^{(l)} > \gamma$ for $l \in [l_\mathrm{s}, l_\mathrm{e}]$, where γ is a predefined entropy threshold and $[l_\mathrm{s}, l_\mathrm{e}]$ denotes the intermediate-layer scanning window, the model triggers a single retrieval call at this step, indicating that the current generation trajectory has insufficient visual grounding and requires re-anchoring.

4.3.2 Cross-modal retrieval

Once triggered, we project the current decoder hidden state into the shared visual-text semantic space and use it as a query $q_t^{(l)} = h_t^{(l)}[-1,:]$, i.e., the hidden state at the last token position of layer l, to query the deferred visual memory bank:

$$r_j = \frac{q_t^{(l)} z_j^\top}{\sqrt{D}},\ j \in \mathcal{P}_t,\ \ w = \mathrm{softmax}(r), \tag{9}$$

where r_j measures the semantic relevance between the textual query and each deferred visual token, and w defines a normalized retrieval distribution over $\mathcal{P}_t$. With $k_c = \min(K, |\mathcal{P}_t|)$, where K is a predefined upper bound, we select the top-k_c tokens $(\tilde{w}, \mathcal{I}_t) = \mathrm{TopK}(w, k_c)$, where $\tilde{w}$ denotes the corresponding weights, and $\mathcal{I}_t \subseteq \mathcal{P}_t$ denotes their indices in the deferred memory bank. The weights are renormalized and aggregated into a query-conditioned visual representation, which is expressed as

$$\tilde{Z}_v = \mathrm{diag}(\tilde{w}) Z_v[\mathcal{I}_t], \tag{10}$$

where $\mathrm{diag}(\tilde{w})$ denotes a diagonal matrix whose main diagonal is formed by $\tilde{w}$.

4.3.3 Lightweight injection

To inject the retrieved evidence into the decoding stream without disturbing the model's intrinsic distribution, we introduce a transient adapter whose weights are scale-matched to the host Feed-Forward Network (FFN), expressed as

$$W_1^{\mathrm{adpt}} = \frac{s_{\mathrm{up}}}{\tilde{s}_v} \tilde{Z}_v,\ \ W_2^{\mathrm{adpt}} = \frac{s_{\mathrm{down}}}{\tilde{s}_v} \tilde{Z}_v^\top. \tag{11}$$

Here, W_1^{adpt} and W_2^{adpt} denote the up- and down-projection weights of the transient adapter, respectively. W_{up} and W_{down} denote the corresponding up- and down-projection matrices of the host FFN, respectively. $s_{\mathrm{up}} = \mathrm{mean}(|W_{\mathrm{up}}|)$, $s_{\mathrm{down}} = \mathrm{mean}(|W_{\mathrm{down}}|)$, and $\tilde{s}_v = \mathrm{mean}(|\tilde{Z}_v|)$ denote the mean absolute magnitudes used for amplitude alignment. The adapter performs a low-rank projection that re-grounds the textual representation using the retrieved visual evidence, expressed as

$$\mathcal{G}(x) = x(W_1^{\mathrm{adpt}})^\top (W_2^{\mathrm{adpt}})^\top = \frac{s_{\mathrm{up}} s_{\mathrm{down}} x \tilde{Z}_v^\top \tilde{Z}_v}{\tilde{s}_v^2}, \tag{12}$$

where x denotes the input residual-stream hidden state fed into the adapter. After amplitude normalization, the retrieved visual signal $\hat{a}$ is fused with the block-output residual stream $h_t^{(l+1)}$, i.e., the hidden state after FFN residual addition in layer l that feeds into the attention sublayer of layer $l+1$:

$$o = (1-\alpha) \cdot h_t^{(l+1)} + \alpha \cdot \hat{a}, \tag{13}$$

where o denotes the fused hidden state passed to layer $l+1$, and $\alpha \in [0,1]$ denotes the injection ratio that balances the original residual stream and the retrieved visual evidence. After retrieval, the trigger is reset, and the transient adapter is released. This enables parameter-free and on-demand visual retrieval in the semantic space, i.e., visual evidence is retrieved only when grounding fails, and the overhead is removed after each step, ensuring efficiency while mitigating VHs.

5 Experiments

5.1 Experimental Settings

Benchmarks. We evaluate RVSD on five complementary benchmarks. For discriminative hallucination, we use POPE (Li et al., 2023) with random,

Models	Methods	Random		Popular		Adversarial		Average	
		Acc (%)↑	F1 (%)↑	Acc (%)↑	F1 (%)↑	Acc (%)↑	F1 (%)↑	Acc (%)↑	F1 (%)↑
LLaVA-1.5	Vanilla	83.8 ↑0.0	84.2 ↑0.0	82.0 ↑0.0	82.6 ↑0.0	75.8 ↑0.0	78.1 ↑0.0	80.5 ↑0.0	81.6 ↑0.0
	+VCD (Leng et al., 2024)	85.0 ↑1.2	84.2 ↑0.0	82.1 ↑0.1	83.2 ↑0.6	76.3 ↑0.5	78.7 ↑0.6	81.1 ↑0.6	82.0 ↑0.4
	+M3ID (Favero et al., 2024)	86.1 ↑2.3	85.0 ↑0.8	82.8 ↑0.8	84.1 ↑1.5	77.1 ↑1.3	78.9 ↑0.8	82.0 ↑1.5	82.7 ↑1.1
	+VTI (Liu et al., 2025)	83.0 ↓0.8	83.6 ↓0.6	80.4 ↓1.6	81.8 ↓0.8	76.0 ↑0.2	78.8 ↑0.7	79.8 ↓0.7	81.4 ↓0.2
	+AvisC (Woo et al., 2025)	82.3 ↓1.5	83.5 ↓0.7	78.2 ↓3.8	80.5 ↓2.1	74.2 ↓1.6	77.7 ↓0.4	78.2 ↓2.3	80.6 ↓1.0
	+RVSD (Ours)	**87.3** ↑3.5	**86.3** ↑2.1	**86.6** ↑4.6	**85.3** ↑2.7	**84.7** ↑8.9	**83.5** ↑5.4	**86.2** ↑5.7	**85.0** ↑3.4
LLaVA-NEXT	Vanilla	84.4 ↑0.0	82.3 ↑0.0	83.2 ↑0.0	81.5 ↑0.0	79.5 ↑0.0	78.1 ↑0.0	82.4 ↑0.0	80.6 ↑0.0
	+VCD (Leng et al., 2024)	86.0 ↑1.6	84.3 ↑2.0	84.5 ↑1.3	82.9 ↑1.4	80.9 ↑1.4	79.7 ↑1.6	83.8 ↑1.4	82.3 ↑1.7
	+M3ID (Favero et al., 2024)	85.5 ↑1.1	83.6 ↑1.3	84.2 ↑1.0	82.4 ↑0.9	80.6 ↑1.1	79.2 ↑1.1	83.4 ↑1.0	81.7 ↑1.1
	+VTI (Liu et al., 2025)	84.8 ↑0.4	83.1 ↑0.8	81.8 ↓1.4	81.6 ↑0.1	79.0 ↓0.5	78.1 ↑0.0	81.9 ↓0.5	80.9 ↑0.3
	+AvisC (Woo et al., 2025)	85.2 ↑0.8	82.8 ↑0.5	83.9 ↑0.7	81.6 ↑0.1	81.8 ↑2.3	80.2 ↑2.1	83.6 ↑1.2	81.5 ↑0.9
	+RVSD (Ours)	**88.5** ↑4.1	**87.6** ↑5.3	**87.9** ↑4.7	**86.7** ↑5.2	**86.5** ↑7.0	**85.4** ↑7.3	**87.6** ↑5.2	**86.6** ↑6.0
Qwen-VL	Vanilla	84.9 ↑0.0	82.9 ↑0.0	84.0 ↑0.0	81.9 ↑0.0	82.1 ↑0.0	80.2 ↑0.0	83.7 ↑0.0	81.7 ↑0.0
	+VCD (Leng et al., 2024)	85.5 ↑0.6	83.6 ↑0.7	84.9 ↑0.9	**83.6** ↑1.7	84.0 ↑1.9	82.0 ↑1.8	84.8 ↑1.1	83.1 ↑1.4
	+M3ID (Favero et al., 2024)	85.3 ↑0.4	83.4 ↑0.5	84.2 ↑0.2	82.7 ↑0.8	83.2 ↑1.1	80.8 ↑0.6	84.2 ↑0.5	82.3 ↑0.6
	+VTI (Liu et al., 2025)	85.3 ↑0.4	83.5 ↑0.6	83.0 ↓1.0	82.3 ↑0.4	83.2 ↑1.1	81.8 ↑1.6	83.8 ↑0.1	82.5 ↑0.8
	+AvisC (Woo et al., 2025)	82.9 ↓2.0	80.0 ↓2.9	82.8 ↓1.2	80.1 ↓1.8	81.2 ↓0.9	78.5 ↓1.7	82.3 ↓1.4	79.5 ↓2.2
	+RVSD (Ours)	**86.2** ↑1.3	**84.7** ↑1.8	**85.3** ↑1.3	83.3 ↑1.4	**84.5** ↑2.4	**82.6** ↑2.4	**85.3** ↑1.6	**83.5** ↑1.8

Table 1: Performance comparison of RVSD with the baselines on the POPE-MSCOCO discrimination benchmark across the random, popular, and adversarial splits. The average column reports the mean accuracy and F1 scores over all splits. Green (●) and red (●) indicate improvements and degradations relative to the base model, respectively, while (↑/↓) denotes the absolute change. Best results are highlighted in **bold**.

popular, and adversarial splits, the MME hallucination subset (Fu et al., 2026) covering object-level and attribute-level probes, and the discrimination subset of AMBER (Wang et al., 2023a), reporting accuracy and F1 scores. For generative hallucination, we evaluate on CHAIR (Rohrbach et al., 2018) over 500 MS-COCO images and the AMBER generative subset, and also report AMBER discrimination accuracy and F1 scores for a unified view of caption-level performance. To assess whether VH mitigation affects general capability, we further evaluate on MM-Vet (Yu et al., 2024), which covers six vision-language abilities and uses GPT-4 as the evaluator.

Models. We evaluate RVSD on the 7B variants of LLaVA-1.5 (Liu et al., 2024b), LLaVA-NEXT (Liu et al., 2024c), and Qwen-VL (Bai et al., 2023). LLaVA-1.5 adopts a two-layer MLP for vision-language alignment, and LLaVA-NEXT further introduces the AnyRes mechanism to support higher-resolution visual understanding, while Qwen-VL instead couples a cross-attention-based vision-language adapter with a non-LLaMA language backbone. Together, they cover both standard- and high-resolution settings as well as heterogeneous vision-language interfaces, enabling a comprehensive evaluation of RVSD. Moreover, RVSD is model-agnostic and can be seamlessly integrated into different LVLM architectures.

Baselines. We compare RVSD with representative training-free baselines from three categories: (i) contrastive decoders, including VCD (Leng et al., 2024) and M3ID (Favero et al., 2024); (ii) the attention-intervention method AvisC (Woo et al., 2025); and (iii) the latent-space steering method VTI (Liu et al., 2025). All baselines leverage the hyperparameters recommended in their original papers under the same decoding budget, while the vanilla LVLM with sampling decoding serves as the base model.

5.2 Experimental Results

Results on POPE. In Table 1, we see that RVSD achieves the best performance across all POPE splits on the three backbones, obtaining the highest accuracy on every split and the highest F1 scores in almost all cases. On LLaVA-1.5-7B, RVSD improves the average Accuracy/F1 from 80.5/81.6 to 86.2/85.0. On LLaVA-NEXT, the scores increase from 82.4/80.6 to 87.6/86.6, and on Qwen-VL, from 83.7/81.7 to 85.3/83.5. The largest gains are observed on the adversarial split, with improvements of +8.9/+5.4, +7.0/+7.3, and +2.4/+2.4 on the three backbones, respectively. In contrast, VCD

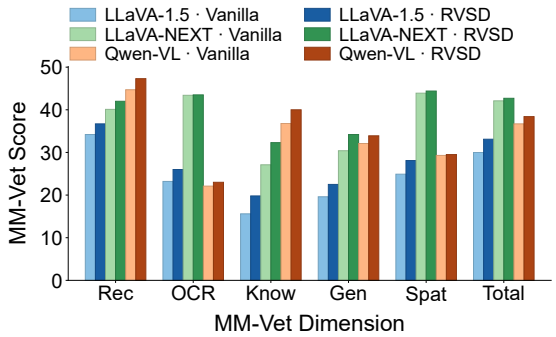

Figure 4: Per-dimension MM-Vet scores of Vanilla decoding and RVSD on LLaVA-1.5, LLaVA-NEXT, and Qwen-VL, covering the Rec, OCR, Know, Gen, and Spat capability dimensions as well as the overall total score. RVSD improves every reported dimension on all three backbones.

and M3ID yield only limited improvements, while VTI and AvisC occasionally degrade performance. These results demonstrate the effectiveness of combining semantics-directed sparsification with on-demand retrieval for reliable visual grounding.

Results on MM-Vet. To verify that VH mitigation does not compromise general multimodal capability, we evaluate RVSD on MM-Vet and report the Rec, OCR, Know, Gen, and Spat dimensions together with the overall total score in Fig. 4. We observe that RVSD improves the total score over the vanilla model on all three backbones, from 30.0 to 33.1 on LLaVA-1.5, from 42.1 to 42.7 on LLaVA-NEXT, and from 36.7 to 38.4 on Qwen-VL, with gains observed across every reported dimension. The improvements are most pronounced on Know, suggesting that restoring pruned visual evidence also benefits knowledge-intensive reasoning. The results demonstrate that RVSD mitigates VHs while preserving general multimodal competence.

Results on MME. As shown in Fig. 5, RVSD obtains the highest total score on the MME hallucination subset across all three backbones, reaching 651.7 on LLaVA-1.5, 668.3 on LLaVA-NEXT, and 593.4 on Qwen-VL, and consistently surpassing the vanilla decoder on the total axis of every panel. On the two LLaVA backbones, the improvements are concentrated on the attribute-level probes, where RVSD achieves the highest scores on both Color and Position, whereas the scores of most baselines decline markedly on Position for LLaVA-NEXT. On Qwen-VL, the gain is instead driven by object Existence, while Count and Position remain below the strongest baselines. These results further demonstrate the effectiveness of RVSD in mitigating fine-grained VHs.

Results on CHAIR and AMBER. As shown in Table 2, we further evaluate caption-level behavior on CHAIR and AMBER, jointly assessing discrimination (Accuracy/F1) and generative hallucination (CHAIR_S, CHAIR_I, AMBER CHAIR, HalRate, Score). We observe that RVSD achieves the best discrimination performance and the highest AMBER score across all three backbones, and obtains the lowest CHAIR_I on the two LLaVA backbones and the lowest CHAIR_S on Qwen-VL, indicating effective suppression of object-level hallucinations during long-form generation. In contrast, existing baselines provide only limited or inconsistent improvements. Overall, RVSD consistently improves both faithfulness and caption quality across different LVLM architectures, demonstrating the effectiveness of combining semantics-directed sparsification with on-demand visual retrieval.

5.3 Ablation Studies

Mechanism Contribution. To evaluate the contribution of each component, we conduct an ablation study on the full MME benchmark using LLaVA-1.5 by separately removing the semantics-directed sparse selection (w/o Sparse) and the SSVR module (w/o SSVR). As shown in Table 3, removing either component degrades the overall total score, demonstrating that both modules are essential to RVSD. In particular, removing SSVR causes the largest performance drop, reducing Cognition from 371.79 to 348.21 and yielding the greatest decline in the total score, highlighting the importance of on-demand visual retrieval for restoring visual grounding during reasoning-intensive decoding. In contrast, removing sparse selection slightly improves the perception score by retaining all visual tokens, but noticeably reduces the cognition score by 14.29, indicating that dense decoding introduces redundant visual information that weakens effective grounding. Overall, RVSD achieves the best total and cognition scores, demonstrating that semantics-directed sparsification and SSVR complement each other to improve visual grounding.

5.4 Inference Latency

As shown in Table 4, we further benchmark inference efficiency under the same decoding budget. We observe that RVSD is the only method that maintains the same latency as the Vanilla baseline (1.00×) while simultaneously reducing memory usage by 2.2%, Tera Floating-Point Operations Per Second (TFLOPs) by 15.6%, and Time-to-First-

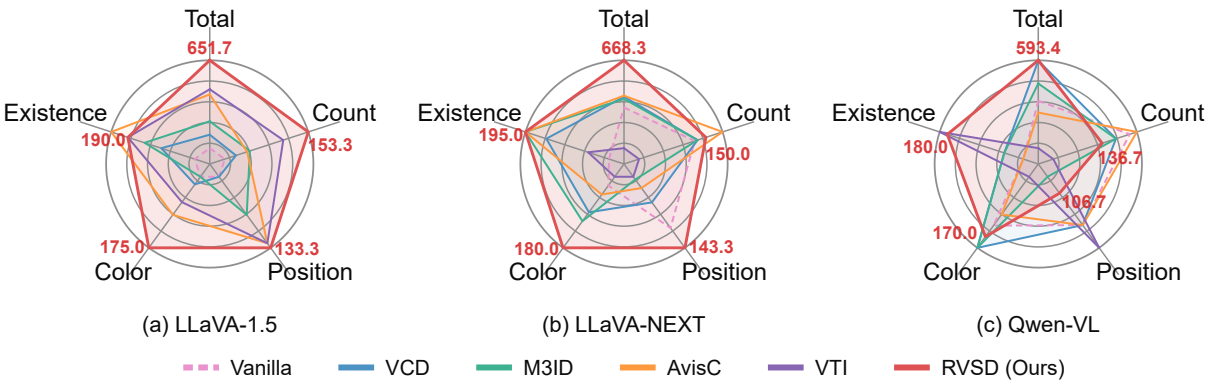

Figure 5: Performance comparison of RVSD with the baselines on the MME hallucination subset for (a) LLaVA-1.5, (b) LLaVA-NEXT, and (c) Qwen-VL, covering object-level (Existence, Count) and attribute-level (Position, Color) probes, with Total denoting their sum. Values annotated in red are those of RVSD, and each axis is scaled independently.

Models	Methods	CHAIR				AMBER				
		CHAIR_S↓	CHAIR_I↓	Recall↑	Length	Acc (%)↑	F1 (%)↑	CHAIR↓	HalRate↓	Score↑
LLaVA-1.5	Vanilla	52.8 ↓0.0	15.9 ↓0.0	77.3 ↑0.0	93.4	67.0 ↑0.0	71.0 ↑0.0	12.0 ↓0.0	51.0 ↓0.0	79.5 ↑0.0
	+VCD (Leng et al., 2024)	51.0 ↓1.8	14.9 ↓1.0	77.2 ↓0.1	101.9	67.3 ↑0.3	71.1 ↑0.1	10.0 ↓2.0	43.6 ↓7.4	80.6 ↑1.1
	+M3ID (Favero et al., 2024)	56.2 ↑3.4	17.0 ↑1.1	**79.3** ↑2.0	97.1	67.3 ↑0.3	70.9 ↓0.1	9.1 ↓2.9	42.8 ↓8.2	80.9 ↑1.4
	+AvisC (Woo et al., 2025)	44.0 ↓8.8	13.7 ↓2.2	72.9 ↓4.4	89.8	70.7 ↑3.7	74.2 ↑3.2	11.8 ↓0.2	51.2 ↑0.2	81.2 ↑1.7
	+VTI (Liu et al., 2025)	**36.8** ↓16.	13.6 ↓2.3	66.6 ↓10.	66.7	66.5 ↓0.5	70.6 ↓0.4	**6.9** ↓5.1	**27.0** ↓24.	81.9 ↑2.4
	+RVSD (Ours)	38.4 ↓14.	**10.3** ↓5.6	74.4 ↓2.9	90.6	**73.6** ↑6.6	**76.9** ↑5.9	7.3 ↓4.7	31.2 ↓19.	**84.8** ↑5.3
LLaVA-NEXT	Vanilla	35.8 ↓0.0	12.0 ↓0.0	59.5 ↑0.0	179.0	72.9 ↑0.0	78.6 ↑0.0	12.0 ↓0.0	59.6 ↓0.0	83.3 ↑0.0
	+VCD (Leng et al., 2024)	40.2 ↑4.4	10.7 ↓1.3	62.1 ↑2.6	171.2	74.3 ↑1.4	79.6 ↑1.0	11.8 ↓0.2	58.6 ↓1.0	83.9 ↑0.6
	+M3ID (Favero et al., 2024)	35.2 ↓0.6	10.3 ↓1.7	61.4 ↑1.9	152.7	75.0 ↑2.1	80.0 ↑1.4	10.2 ↓1.8	52.8 ↓6.8	84.9 ↑1.6
	+AvisC (Woo et al., 2025)	40.4 ↑4.6	12.4 ↑0.4	60.0 ↑0.5	183.8	77.7 ↑4.8	82.1 ↑3.5	14.2 ↑2.2	65.2 ↑5.6	84.0 ↑0.7
	+VTI (Liu et al., 2025)	26.4 ↓9.4	7.7 ↓4.3	58.5 ↓1.0	187.6	75.2 ↑2.3	81.9 ↑3.3	10.3 ↓1.7	58.3 ↓1.3	87.1 ↑3.8
	+RVSD (Ours)	**26.2** ↓9.6	**6.7** ↓5.3	**63.7** ↑4.2	185.0	**80.0** ↑7.1	**83.8** ↑5.2	**8.9** ↓3.1	**47.9** ↓11.	**87.5** ↑4.2
Qwen-VL	Vanilla	2.8 ↓0.0	3.0 ↓0.0	31.0 ↑0.0	5.3	82.9 ↑0.0	86.9 ↑0.0	4.8 ↓0.0	9.2 ↓0.0	91.1 ↑0.0
	+VCD (Leng et al., 2024)	1.4 ↓1.4	**1.2** ↓1.8	30.8 ↓0.2	4.0	84.1 ↑1.2	87.9 ↑1.0	3.5 ↓1.3	7.7 ↓1.5	92.2 ↑1.1
	+M3ID (Favero et al., 2024)	1.7 ↓1.1	1.3 ↓1.7	31.8 ↑0.8	3.4	84.0 ↑1.1	87.1 ↑0.2	3.7 ↓1.1	8.2 ↓1.0	91.7 ↑0.6
	+AvisC (Woo et al., 2025)	1.6 ↓1.2	1.6 ↓1.4	**32.0** ↑1.0	4.4	84.1 ↑1.2	87.9 ↑1.0	4.3 ↓0.5	8.6 ↓0.6	91.8 ↑0.7
	+VTI (Liu et al., 2025)	1.9 ↓0.9	1.9 ↓1.1	30.2 ↓0.8	4.5	83.5 ↑0.6	87.3 ↑0.4	**3.1** ↓1.7	**6.8** ↓2.4	92.4 ↑1.3
	+RVSD (Ours)	**1.2** ↓1.6	1.6 ↓1.4	31.9 ↑0.9	5.9	**84.3** ↑1.4	**88.6** ↑1.7	3.4 ↓1.4	7.3 ↓1.9	**92.6** ↑1.5

Table 2: Performance comparison on the CHAIR benchmark (CHAIR_S, CHAIR_I, Recall, Length) and the AMBER benchmark for open-ended captioning, where the AMBER block jointly reports discrimination metrics (Accuracy, F1) and generative metrics (CHAIR, HalRate, Score). Best results are highlighted in **bold**.

Methods	Perception↑	Cognition↑	Total↑
RVSD	1511.40	**371.79**	**1883.19**
w/o Sparse	1514.60	357.50	1872.10
w/o SSVR	**1520.36**	348.21	1868.57

Table 3: Component-level ablation study on the full MME benchmark using LLaVA-1.5, evaluating the contributions of the semantics-directed sparse selection module and the SSVR module. Best results are highlighted in **bold**.

Token (TTFT) by 20.2%. In contrast, contrastive decoding methods (VCD and M3ID) introduce over 2× latency overhead due to auxiliary forward passes, while AvisC incurs the highest cost from dual forward inference and additional attention extraction. Although VTI is relatively lightweight, it still increases latency by 1.09×. These results demonstrate that RVSD effectively improves visual grounding without sacrificing inference efficiency, benefiting from semantics-directed sparsification and the uncertainty-triggered retrieval mechanism that is activated only when necessary.

6 Conclusion

In this paper, we have proposed RVSD, a training-free and plug-and-play decoding framework that, for the first time, resolves the previously overlooked sparsification-hallucination paradox in existing sparse decoders. By integrating semantics-

Methods	Latency ↓	Memory ↓	TFLOPs ↓	TTFT ↓
Vanilla	**61.9**	14334	4.36	243.8
+VCD	131.6	14979	8.72	445.4
+M3ID	132.5	14977	8.72	477.8
+AvisC	173.6	15501	8.45	725.2
+VTI	67.3	14363	4.36	266.3
+RVSD	62.0	**14015**	**3.68**	**194.6**

Table 4: Inference efficiency comparison between RVSD and representative decoding-side VH mitigation methods, reporting per-token decoding latency (ms/token), peak GPU memory usage (MB), TFLOPs (T), and TTFT (ms). Best results are highlighted in **bold**.

directed token selection with the SSVR mechanism within a single decoding pass, RVSD transforms irreversible visual token pruning into reversible and on-demand cross-modal retrieval, enabling visual grounding to be restored only when necessary. Extensive experiments on five representative benchmarks across three LVLM backbones demonstrate that RVSD achieves SOTA performance in VH mitigation while simultaneously improving decoding efficiency and maintaining robustness under long-context generation, where existing sparse and decoding-based methods often deteriorate.

For future work, we will extend RVSD to more challenging multimodal settings, such as long-form video understanding, where extended temporal dependencies and richer modality interactions further exacerbate the sparsification-hallucination paradox. We hope this work provides a promising direction toward trustworthy and efficient LVLM inference.

Limitations

While RVSD shows promising results, several limitations remain. Firstly, the uncertainty-triggered retrieval gate in SSVR relies on a predefined entropy threshold to determine when cross-modal retrieval should be activated. Although RVSD remains robust across a reasonable range of threshold values, a more adaptive gating strategy that dynamically adjusts the trigger according to input complexity can further reduce unnecessary retrieval calls and improve efficiency on simpler queries.

Additionally, the current implementation uses a simple dot-product similarity to retrieve evidence from the deferred visual memory bank. More advanced retrieval mechanisms, such as learned semantic projections or multi-head cross-modal attention, may further improve performance on fine-grained tasks. Although our evaluation already spans two distinct architectural families, i.e., the LLaVA series and Qwen-VL, we also plan to extend RVSD to larger-scale and more heterogeneous LVLMs to further evaluate its generality.

Ethical Considerations

This work aims to improve the trustworthiness of LVLMs by mitigating VHs, thereby contributing to the safer deployment of multimodal artificial intelligence systems. RVSD is a training-free and inference-only framework that requires no additional data collection or human annotation. All datasets (i.e., POPE, MME, AMBER, CHAIR, MM-Vet, MSCOCO, A-OKVQA, and GQA) and backbone models (LLaVA-1.5, LLaVA-NEXT, and Qwen-VL) used in this study are publicly available for academic research and are used strictly in accordance with their original licenses. While RVSD substantially reduces VHs, it does not eliminate them. Therefore, we recommend pairing it with downstream verification or human oversight before deployment in safety-critical applications such as medical diagnosis or assistive technologies.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 62572132, and the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2025A1515010137.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. 2025. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21635–21644. IEEE.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. 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*, pages 19–35. Springer.

Tiantian Dang, Chao Bi, Shufan Shen, Jinzhe Liu, Qingming Huang, and Shuhui Wang. 2026. Locate-then-sparsify: Attribution guided sparse strategy for visual hallucination mitigation. *arXiv preprint arXiv:2603.16284*.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2026. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38.

Jihoon Lee and Min Song. 2025. Retrieval visual contrastive decoding to mitigate object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8200–8219.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llavanext: Improved reasoning, ocr, and world knowledge.

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

Sheng Liu, Haotian Ye, and James Y Zou. 2025. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, volume 2025, pages 72402–72419.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4035–4045.

Fengyuan Sun, Hui Chen, Xinhao Xu, Dandan Zheng, Jingdong Chen, Jun Zhou, Jungong Han, and Guiguang Ding. 2025. Prunehal: Reducing hallucinations in multi-modal large language models through adaptive kv cache pruning. *arXiv preprint arXiv:2510.19183*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pages 13088–13110.

Chao Wang, Jianming Yang, and Yang Zhou. 2025. Mint: Mitigating hallucinations in large vision-language models via token reduction. *arXiv preprint arXiv:2502.00717*.

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

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023b. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2025. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1927–1951.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE international conference on big data (bigdata)*, pages 2247–2256. IEEE.

Junfei Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2025. Sharp: Steering hallucination in lvlms via representation engineering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14357–14372.

Qiong Wu, Wenhao Lin, Yiyi Zhou, Weihao Ye, Zhanpeng Zeng, Xiaoshuai Sun, and Rongrong Ji. 2026. Accelerating multimodal large language models via dynamic visual-token exit and the empirical findings. *Advances in Neural Information Processing Systems*, 38:168378–168403.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. MM-vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57730–57754.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2025. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 74840–74857.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, and 1 others. 2026. A survey of large language models. *Frontiers of Computer Science*, 20(12):2012627.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. 2026. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. In *International Conference on Learning Representations*, volume 2026, pages 81235–81256.

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. 2025. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4189–4199. IEEE.

Appendix

A More Experiments and Results

Results on POPE A-OKVQA and GQA. To further evaluate the cross-dataset generalization of RVSD, we conduct additional experiments on the POPE A-OKVQA and GQA subsets, as shown in Tables 5 and 6. Compared with POPE-MSCOCO (refer to Table 1), these datasets contain denser scenes and stronger object co-occurrence biases, making the popular and adversarial splits more challenging. RVSD achieves the best accuracy across all splits on all three backbones and the best F1 scores on the two LLaVA backbones, with the largest improvements observed on the adversarial split. In particular, RVSD improves Accuracy/F1 by 11.6/5.4 on LLaVA-1.5 A-OKVQA and by 10.9/3.8 on LLaVA-1.5 GQA, demonstrating strong robustness against misleading linguistic priors. On Qwen-VL, RVSD still obtains the highest accuracy on every split, improving the adversarial accuracy by 2.4 on A-OKVQA and by 4.4 on GQA, while its F1 scores remain comparable to the strongest contrastive baseline. In contrast, existing baselines provide only limited or inconsistent improvements, while AvisC notably degrades the average performance on both A-OKVQA and GQA. These results demonstrate that the effectiveness of RVSD generalizes consistently across different POPE datasets and is not dependent on a specific data distribution.

B Detailed Experimental Setup

B.1 Datasets and Benchmarks

We provide additional details for the five benchmarks used in the main paper, all of which are publicly released and standard in the VH literature.

POPE (Li et al., 2023) evaluates object-level hallucination through binary "Yes/No" existence questions and reports accuracy and F1 scores under three sampling strategies, i.e., random, popular, and adversarial splits. The popular split samples frequently co-occurring object categories. In contrast, the adversarial split selects objects with strong co-occurrence priors relative to the image, making it particularly challenging for VH mitigation. We evaluate POPE on three data sources, namely MSCOCO (Rohrbach et al., 2018), A-OKVQA, and GQA, each containing 3,000 yes/no queries.

MME Hallucination Subset (Fu et al., 2026) evaluates hallucination through four probing tasks, including object-level probes (Existence, Count) and attribute-level probes (Position, Color). Each task contains 30 binary questions and is evaluated using Acc/Acc+, where Acc+ requires both yes/no questions associated with the same image to be answered correctly. The total score is computed as the sum of the four subtask scores.

AMBER (Wang et al., 2023a) contains both a discrimination subset and a generative subset. The discrimination subset uses binary yes/no questions and is evaluated with accuracy and F1 scores, while the generative subset evaluates open-ended captioning using CHAIR, HalRate, and a holistic Score that jointly measures caption coverage and hallucination penalty. AMBER covers existence-, attribute-, and relation-level hallucinations, with a particular emphasis on long-form caption generation.

CHAIR (Rohrbach et al., 2018) evaluates object hallucination in generated captions by extracting mentioned objects and matching them against MSCOCO ground-truth annotations. It reports two metrics, which are given by

$$\mathrm{CHAIR}_I = \frac{|\{\mathrm{hallucinatedobjects}\}|}{|\{\mathrm{allmentionedobjects}\}|}, \tag{14}$$

$$\mathrm{CHAIR}_S = \frac{|\{\mathrm{captionsw/hallucination}\}|}{|\{\mathrm{allcaptions}\}|}. \tag{15}$$

Following (Liu et al., 2025), we randomly sample 500 MSCOCO Val2014 images and prompt the model with "*Please describe this image in detail*" to elicit long-form responses.

MM-Vet (Yu et al., 2024) consists of 218 evaluation samples designed to probe six vision-language capabilities: Recognition (Rec), OCR, Knowledge (Know), Generation (Gen), Spatial reasoning (Spat), and Math. We follow the official evaluation protocol and use GPT-4 as the judge to assign a continuous score for each response, with the total score computed as the mean across all dimensions.

B.2 Baselines

We compare RVSD against four representative training-free decoders spanning three paradigms. All baselines follow the recommended hyperparameters and are evaluated under the same decoding budget as RVSD for a fair comparison.

VCD (Visual Contrastive Decoding) (Leng et al., 2024) introduces a perturbed visual input by adding diffusion-style Gaussian noise and performs contrastive decoding between the clean and corrupted

Models	Methods	Random		Popular		Adversarial		Average	
		Acc (%)↑	F1 (%)↑	Acc (%)↑	F1 (%)↑	Acc (%)↑	F1 (%)↑	Acc (%)↑	F1 (%)↑
LLaVA-1.5	Vanilla	81.8 ↑0.0	83.5 ↑0.0	75.3 ↑0.0	78.7 ↑0.0	67.4 ↑0.0	73.7 ↑0.0	74.8 ↑0.0	78.6 ↑0.0
	+VCD (Leng et al., 2024)	81.2 ↓0.6	83.2 ↓0.3	74.7 ↓0.6	78.5 ↓0.2	68.1 ↑0.7	74.6 ↑0.9	74.7 ↓0.1	78.8 ↑0.2
	+M3ID (Favero et al., 2024)	82.9 ↑1.1	84.6 ↑1.1	75.8 ↑0.5	79.4 ↑0.7	68.3 ↑0.9	74.6 ↑0.9	75.7 ↑0.9	79.5 ↑0.9
	+AvisC (Woo et al., 2025)	79.1 ↓2.7	82.4 ↓1.1	71.8 ↓3.5	77.2 ↓1.5	64.4 ↓3.0	73.0 ↓0.7	71.8 ↓3.0	77.5 ↓1.1
	+VTI (Liu et al., 2025)	83.0 ↑1.2	83.6 ↑0.1	76.1 ↑0.8	79.5 ↑0.8	68.2 ↑0.8	74.6 ↑0.9	75.8 ↑1.0	79.2 ↑0.6
	+RVSD (Ours)	**87.5** ↑5.7	**86.4** ↑2.9	**85.4** ↑10.	**84.5** ↑5.8	**79.0** ↑11.	**79.1** ↑5.4	**84.0** ↑9.2	**83.3** ↑4.7
LLaVA-NEXT	Vanilla	83.8 ↑0.0	83.0 ↑0.0	81.4 ↑0.0	80.8 ↑0.0	73.2 ↑0.0	74.5 ↑0.0	79.5 ↑0.0	79.4 ↑0.0
	+VCD (Leng et al., 2024)	84.8 ↑1.0	83.9 ↑0.9	81.5 ↑0.1	81.2 ↑0.4	74.7 ↑1.5	76.0 ↑1.5	80.3 ↑0.8	80.4 ↑1.0
	+M3ID (Favero et al., 2024)	85.3 ↑1.5	84.3 ↑1.3	82.2 ↑0.8	81.6 ↑0.8	74.6 ↑1.4	75.7 ↑1.2	80.7 ↑1.2	80.5 ↑1.1
	+AvisC (Woo et al., 2025)	85.8 ↑2.0	84.4 ↑1.4	83.9 ↑2.5	82.8 ↑2.0	75.6 ↑2.4	76.6 ↑2.1	81.8 ↑2.3	81.3 ↑1.9
	+VTI (Liu et al., 2025)	84.8 ↑1.0	84.2 ↑1.2	80.4 ↓1.0	79.7 ↓1.1	72.8 ↓0.4	74.3 ↓0.2	79.3 ↓0.2	79.4 ↑0.0
	+RVSD (Ours)	**90.1** ↑6.3	**89.6** ↑6.6	**88.6** ↑7.2	**88.2** ↑7.4	**81.8** ↑8.6	**82.4** ↑7.9	**86.8** ↑7.3	**86.7** ↑7.3
Qwen-VL	Vanilla	86.8 ↑0.0	85.8 ↑0.0	85.6 ↑0.0	84.7 ↑0.0	80.4 ↑0.0	80.5 ↑0.0	84.3 ↑0.0	83.7 ↑0.0
	+VCD (Leng et al., 2024)	87.4 ↑0.6	**86.6** ↑0.8	86.3 ↑0.7	**85.1** ↑0.4	80.7 ↑0.3	80.8 ↑0.3	84.8 ↑0.5	**84.2** ↑0.5
	+M3ID (Favero et al., 2024)	87.1 ↑0.3	85.9 ↑0.1	85.9 ↑0.3	84.6 ↓0.1	80.5 ↑0.1	80.4 ↓0.1	84.5 ↑0.2	83.6 ↓0.1
	+AvisC (Woo et al., 2025)	84.7 ↓2.1	83.0 ↓2.8	83.9 ↓1.7	82.4 ↓2.3	78.1 ↓2.3	77.3 ↓3.2	82.2 ↓2.1	80.9 ↓2.8
	+VTI (Liu et al., 2025)	85.0 ↓1.8	85.3 ↓0.5	82.5 ↓3.1	83.3 ↓1.4	74.5 ↓5.9	75.0 ↓5.5	80.7 ↓3.6	81.2 ↓2.5
	+RVSD (Ours)	**87.4** ↑0.6	86.2 ↑0.4	**87.7** ↑2.1	84.5 ↓0.2	**82.8** ↑2.4	**81.1** ↑0.6	**86.0** ↑1.7	83.9 ↑0.2

Table 5: Performance comparison on the POPE discrimination benchmark over the A-OKVQA dataset. The average column reports the mean accuracy and F1 scores across the random, popular, and adversarial splits. Green (●) and red (●) indicate improvements and degradations relative to the base model (sampling baseline), respectively, while (↑/↓) denotes the absolute change. Best results within each model block are highlighted in **bold**.

visual conditions, thereby suppressing unimodal language priors.

M3ID (Multi-Modal Mutual-Information Decoding) (Favero et al., 2024) reweights output logits using a mutual-information-based score between visual and textual modalities, aiming to enhance visual grounding during generation.

AvisC (Attentional Vision Calibration) (Woo et al., 2025) identifies high-attention "blind tokens" that dominate cross-modal attention and recalibrates predictions via dual forward passes with and without these tokens.

VTI (Visual Token Intervention) (Liu et al., 2025) steers latent representations in the large language model hidden space along a pre-computed direction associated with improved visual grounding.

B.3 Models

We instantiate RVSD on three LVLMs drawn from two distinct architectural families. **LLaVA-1.5-7B** (Liu et al., 2024b) adopts CLIP-ViT-L/14-336px as the visual encoder, a two-layer MLP as the cross-modal projector, and Vicuna-7B-v1.5 as the language backbone, producing 576 visual tokens per image. **LLaVA-NEXT-7B** (Liu et al., 2024c) extends LLaVA-1.5 with the AnyRes mechanism, which partitions the input image into up to four high-resolution sub-images, increasing the number of visual tokens to as many as 2,880 and substantially amplifying the burden on visual grounding. **Qwen-VL** (Bai et al., 2023) instead couples OpenCLIP-ViT-bigG with a single-layer position-aware cross-attention adapter that compresses the visual feature map into a fixed set of 256 visual tokens, and adopts Qwen-7B as the language backbone, differing from the LLaVA series in both the vision-language interface and the language backbone. Overall, the backbones cover standard and high-resolution settings as well as both MLP-projection and cross-attention resampler interfaces, and demonstrate that RVSD is model-agnostic.

C Implementation Details

Hyperparameter Settings of RVSD. RVSD is implemented in PyTorch on top of the official LLaVA codebase and introduces no additional trainable parameters. Table 7 summarizes the default hyperparameter configuration used across all experiments unless otherwise specified. We adopt a unified pruning layer set $\mathcal{L}_s = 2, 6, 15$ for all three backbones, where the visual-token bank is progressively pruned at these layers. For LLaVA-1.5-7B, the sparse-set budget is set to $k = 192$, approximately one third of $N_v = 576$, resulting

Models	Methods	Random		Popular		Adversarial		Average	
		Acc (%)↑	F1 (%)↑	Acc (%)↑	F1 (%)↑	Acc (%)↑	F1 (%)↑	Acc (%)↑	F1 (%)↑
LLaVA-1.5	Vanilla	81.6 ↑0.0	83.5 ↑0.0	73.1 ↑0.0	77.5 ↑0.0	68.0 ↑0.0	74.5 ↑0.0	74.2 ↑0.0	78.5 ↑0.0
	+VCD (Leng et al., 2024)	82.2 ↑0.6	84.1 ↑0.6	71.5 ↓1.6	76.8 ↓0.7	67.6 ↓0.4	74.5 ↑0.0	73.8 ↓0.4	78.5 ↑0.0
	+M3ID (Favero et al., 2024)	83.3 ↑1.7	**84.5** ↑1.0	72.3 ↓0.8	77.1 ↓0.4	67.2 ↓0.8	74.0 ↓0.5	74.3 ↑0.1	78.5 ↑0.0
	+AvisC (Woo et al., 2025)	79.0 ↓2.6	82.2 ↓1.3	67.4 ↓5.7	74.8 ↓2.7	64.1 ↓3.9	72.9 ↓1.6	70.2 ↓4.0	76.6 ↓1.9
	+VTI (Liu et al., 2025)	79.8 ↓1.8	82.3 ↓1.2	73.5 ↑0.4	78.0 ↑0.5	68.0 ↑0.0	74.6 ↑0.1	73.8 ↓0.4	78.3 ↓0.2
	+RVSD (Ours)	**85.3** ↑3.7	83.8 ↑0.3	**81.3** ↑8.2	**80.3** ↑2.8	**78.9** ↑11.	**78.3** ↑3.8	**81.8** ↑7.6	**80.8** ↑2.3
LLaVA-NEXT	Vanilla	83.1 ↑0.0	82.5 ↑0.0	78.5 ↑0.0	78.5 ↑0.0	73.3 ↑0.0	74.5 ↑0.0	78.3 ↑0.0	78.5 ↑0.0
	+VCD (Leng et al., 2024)	83.4 ↑0.3	83.4 ↑0.9	78.2 ↓0.3	78.6 ↑0.1	74.2 ↑0.9	75.6 ↑1.1	78.6 ↑0.3	79.2 ↑0.7
	+M3ID (Favero et al., 2024)	84.3 ↑1.2	84.3 ↑1.8	78.8 ↑0.3	78.9 ↑0.4	74.2 ↑0.9	75.4 ↑0.9	79.1 ↑0.8	79.5 ↑1.0
	+AvisC (Woo et al., 2025)	84.9 ↑1.8	83.5 ↑1.0	80.5 ↑2.0	79.5 ↑1.0	73.2 ↓0.1	74.3 ↓0.2	79.5 ↑1.2	79.1 ↑0.6
	+VTI (Liu et al., 2025)	83.3 ↑0.2	82.6 ↑0.1	78.6 ↑0.1	78.2 ↓0.3	73.2 ↓0.1	74.3 ↓0.2	78.4 ↑0.1	78.4 ↓0.1
	+RVSD (Ours)	**88.2** ↑5.1	**87.2** ↑4.7	**84.9** ↑6.4	**84.3** ↑5.8	**81.6** ↑8.3	**81.4** ↑6.9	**84.9** ↑6.6	**84.3** ↑5.8
Qwen-VL	Vanilla	81.3 ↑0.0	79.2 ↑0.0	75.9 ↑0.0	74.9 ↑0.0	75.5 ↑0.0	74.4 ↑0.0	77.6 ↑0.0	76.2 ↑0.0
	+VCD (Leng et al., 2024)	82.0 ↑0.7	80.5 ↑1.3	75.9 ↑0.0	75.6 ↑0.7	76.7 ↑1.2	76.2 ↑1.8	78.2 ↑0.6	77.4 ↑1.2
	+M3ID (Favero et al., 2024)	82.4 ↑1.1	79.7 ↑0.5	76.8 ↑0.9	77.0 ↑2.1	77.1 ↑1.6	76.7 ↑2.3	78.8 ↑1.2	77.8 ↑1.6
	+AvisC (Woo et al., 2025)	80.5 ↓0.8	77.8 ↓1.4	74.2 ↓1.7	72.3 ↓2.6	75.5 ↑0.0	73.4 ↓1.0	76.7 ↓0.9	74.5 ↓1.7
	+VTI (Liu et al., 2025)	82.0 ↑0.7	**82.4** ↑3.2	76.5 ↑0.6	77.9 ↑3.0	71.0 ↓4.5	74.5 ↑0.1	76.5 ↓1.1	78.3 ↑2.1
	+RVSD (Ours)	**84.3** ↑3.0	81.7 ↑2.5	**82.7** ↑6.8	**80.3** ↑5.4	**79.9** ↑4.4	**77.8** ↑3.4	**82.3** ↑4.7	**79.9** ↑3.7

Table 6: Performance comparison on the POPE discrimination benchmark over the GQA dataset. The average column reports the mean accuracy and F1 scores across the random, popular, and adversarial splits. Green (●) and red (●) indicate improvements and degradations relative to the base model (sampling baseline), respectively, while (↑/↓) denotes the absolute change. Best results within each model block are highlighted in **bold**.

Symbols	Descriptions	LLaVA-1.5	LLaVA-NEXT	Qwen-VL
k	Sparse-set budget	192	Scaled by `image_shape`	192 (tier) → [133, 89, 52]
m	Major textual tokens	$\lfloor \lvert y_{<t} \rvert / 2 \rfloor$	$\lfloor \lvert y_{<t} \rvert / 2 \rfloor$	$\lfloor \lvert y_{<t} \rvert / 2 \rfloor$
τ	Saliency softmax temperature	1.0	1.0	1.0
η	Temporal smoothing coefficient	0	0	0
$\mathcal{L}_s$	Pruning layer set	{2, 6, 15}	{2, 6, 15}	{2, 6, 15}
γ	Entropy trigger threshold	0.75	0.5	0.75
$[l_s, l_e]$	SSVR layer scanning window	[6,27]	[5,12]	[5,16]
K	Retrieval top-K bound	128	7	57
α	Injection ratio	0.2	0.09	0.2

Table 7: Default hyperparameter settings of RVSD used across all experiments unless otherwise specified.

in a 67% reduction in visual tokens. For LLaVA-NEXT-7B, the budget is rescaled according to the image shape to maintain a consistent retention ratio under the AnyRes high-resolution setting. For Qwen-VL, whose cross-attention adapter emits only $N_v = 256$ visual tokens, the budget starts from $k = 192$ and is progressively tightened to [133, 89, 52] at the three pruning layers, yielding a final retention ratio comparable to that of LLaVA-1.5. The number of major textual tokens is defined as $m = \lfloor |y< t|/2 \rfloor$, which selects the most recent half of the generated prefix for visual-saliency aggregation and ensures a length-adaptive allocation consistent with the Top-K formulation in the main paper. Temporal smoothing in Eq. (6) is disabled by setting $\eta = 0$, i.e., $\tilde{s}^t = s^t$. For SSVR, the entropy threshold γ, the intermediate-layer window $[l_s, l_e]$, the retrieval bound K, and the injection ratio α are tuned per backbone to accommodate differences in visual-token granularity across LLaVA-1.5 (single 24×24 grid), LLaVA-NEXT (multi-granularity AnyRes grids), and Qwen-VL (256 resampled tokens without an explicit spatial grid).

Decoding Protocol. Following (Leng et al., 2024; Liu et al., 2025), all methods, including RVSD and baselines, are evaluated under a unified decoding setup using nucleus sampling with temperature $T = 1.0$, top-$p = 0.9$, no repetition penalty, and

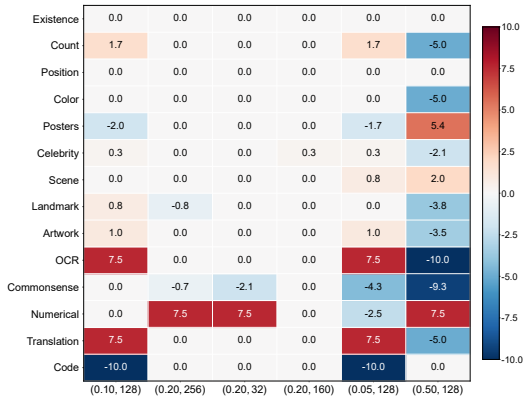

Figure 6: Subtask-level Δ heatmap on MME under non-baseline (α, k_c) configurations, where $\Delta = \text{score(cfg)} - \text{score(base)}$ is computed relative to the baseline setting $(\alpha = 0.20, k_c = 128)$.

batch size 1 for a fair comparison. The generation budget, i.e., the maximum number of new tokens, is benchmark-dependent: It is set to 2 for binary yes/no tasks in POPE and the MME hallucination subset, and to 1,024 for open-ended generation benchmarks including AMBER (generative), MM-Vet, and CHAIR. For the inference latency evaluation in Table 4, we follow the protocol specified in the table caption, setting the maximum number of new tokens to 32 with batch size 1.

Implementation Details. We conduct all experiments on a single NVIDIA RTX A6000 GPU (48 GB) with NVIDIA driver 535.183.01 and CUDA 12.2 on Ubuntu 22.04. The software stack includes PyTorch 2.0 and Hugging Face Transformers 4.40.

D Hyperparameter Sensitivity Analysis

Subtask-Level Sensitivity on MME. We first conduct a one-factor-at-a-time analysis around the baseline $(\alpha = 0.20, k_c = 128)$ on MME with LLaVA-1.5, where α denotes the retrieval injection ratio and k_c the top-k retrieval size. As shown in Fig. 6, α is the primary driver of subtask-level variations: Small values of α (0.05/0.10) consistently improve OCR and translation by up to +7.5, whereas a larger value ($\alpha = 0.50$) leads to broad degradations across OCR, commonsense, counting, and color reasoning. In contrast, varying k_c within [32, 256] yields negligible changes across all subtasks. This analysis identifies $(\alpha = 0.10, k_c = 128)$ as a stable operating point and further reveals that the aggregate MME score can obscure task-specific regressions, particularly in reasoning-intensive categories.

Hyperparameter Sensitivity. To complement the subtask-level analysis above, we further investigate the sensitivity of RVSD to its four key hyperparameters on the POPE-MSCOCO adversarial split with LLaVA-1.5, which is the most discriminative setting for hallucination evaluation. We perform a one-factor-at-a-time analysis, varying each hyperparameter while keeping the others fixed at their default values in Table 7, and the results are reported in Table 8.

Hyperparameters	Setting	Acc (%) ↑	F1 (%) ↑
Injection ratio α	0.05	84.7	83.4
	0.10	84.7	83.4
	0.20	**84.7**	**83.5**
	0.50	84.4	83.2
Retrieval bound K	32	84.6	83.4
	64	84.6	83.4
	128	**84.7**	**83.5**
	256	84.6	83.4
Entropy threshold γ	0.3	**84.7**	83.4
	0.5	**84.7**	83.4
	0.7	84.6	83.4
Sparse budget k	96	83.0	81.1
	192	84.7	83.5
	288	**85.0**	**84.1**
	576 ($=N_v$)	**85.0**	**84.1**

Table 8: Hyperparameter sensitivity analysis on the POPE-MSCOCO Adversarial split with LLaVA-1.5. The best result in each block is highlighted in **bold**.

Effect of Injection Ratio α. RVSD is robust to the injection ratio α in the practical range [0.05, 0.20], where Accuracy/F1 remains stable at 84.7/83.4-84.7/83.5. Increasing α to 0.50 over-injects retrieved evidence relative to the residual-stream magnitude, leading to a mild degradation to 84.4/83.2. This behavior is consistent with the amplitude-alignment design in Eq. (11), where the retrieval branch is intended to re-anchor rather than overwrite the host representation.

Effect of Retrieval Bound K. Accuracy/F1 scores vary by less than 0.1 across the retrieval bound $K \in \{32, 64, 128, 256\}$, indicating strong robustness of the softmax-weighted aggregation in Eq. (9). Irrelevant entries in the deferred memory bank receive near-zero weights and are effectively suppressed, such that increasing K neither introduces noise nor yields additional benefit.

Effect of Entropy Threshold γ. RVSD is stable across $\gamma \in 0.3, 0.5, 0.7$, with the accuracy score remaining at 84.7 for the two lower thresholds and

decreasing by only 0.1 at $\gamma = 0.7$. This suggests that, on the POPE adversarial split, the SSVR gate activates on a well-separated subset of uncertain steps, and moderate variations in γ do not materially affect which steps are re-anchored.

Effect of Sparse Budget k. The sparse budget k is the most influential among the four hyperparameters. A small budget ($k = 96$, approximately one-sixth of N_v) removes too much fine-grained visual evidence and reduces Accuracy/F1 to 83.0/81.1. Increasing k to 288 or 576 slightly improves performance to 85.0/84.1. The default setting ($k = 192$) achieves competitive Accuracy/F1 (84.7/83.5) while maintaining a substantially lower attention cost than $k = 576$, and is therefore adopted as the efficiency–faithfulness trade-off point reported in Table 4. This trend empirically supports the claim in Section 4.2 that visual grounding in RVSD is primarily driven by semantics-directed selection rather than the sheer number of retained tokens.

E Algorithmic Complexity

We analyze the asymptotic overhead introduced by RVSD over vanilla decoding. Let D denote the hidden dimension, N_v the number of visual tokens, t the current decoding step, and k the sparse-set budget. The dominant cost of vanilla decoding at step t arises from per-layer self-attention, with complexity $\mathcal{O}((t + N_v)^2 D)$. In RVSD, this is reduced to $\mathcal{O}((t + k)^2 D)$ through the sparse set $\mathcal{S}_t$, leading to computational savings when $k \ll N_v$.

The additional overhead of RVSD consists of three components. First, the textual saliency computation in Eq. (4) and the relevance scoring in Eq. (5) incur costs of $\mathcal{O}(t)$ and $\mathcal{O}(mN_v|\mathcal{L}_s|)$, respectively. Both are linear and negligible compared with the quadratic attention term. Second, SSVR retrieval in Eq. (9) requires $\mathcal{O}(|\mathcal{P}_t|D)$ for similarity computation and $\mathcal{O}(|\mathcal{P}_t| \log K)$ for top-K selection, which remains dominated by self-attention. Third, the residual-stream adapter in Eq. (11) and the fusion step in Eq. (13) incur $\mathcal{O}(KD)$ per triggered step. However, the entropy-based gate in Eq. (8) activates retrieval only on a small subset of steps. Overall, the additional per-step cost is bounded by $\mathcal{O}(N_v D)$ and is negligible relative to attention. Empirically, Table 4 further shows that RVSD reduces per-token latency, maximum memory, TFLOPs, and TTFT compared with vanilla decoding, a profile not achieved by previous decoding-side methods for VH mitigation.

F License of Artifacts

We use all datasets and models strictly for academic research in accordance with their original licenses and terms. MSCOCO is released under the Creative Commons Attribution 4.0 License, and POPE is released under the MIT License. AMBER, MME, and MM-Vet are used under their respective official non-commercial academic research licenses. LLaVA-1.5 and LLaVA-NEXT are released under the Apache 2.0 License, while their LLaMA-2/Vicuna backbones are subject to the Meta LLaMA-2 community license. Qwen-VL is released under the Tongyi Qianwen License, which allows use for academic research. We do not redistribute any of these resources beyond the permissions granted by their licenses, and we will release our RVSD implementation under a permissive open-source license upon publication.

G Case Study

To provide an intuitive understanding of how RVSD mitigates VHs during generation, we present qualitative case studies in Figs. 7 and 8. Compared with the vanilla decoder and representative sparse decoding baselines, RVSD produces captions that are more faithfully grounded in the visual content, successfully suppressing hallucinated objects, attributes, and spatial relations that are absent from the input image. As shown in Fig. 7, RVSD consistently mitigates object- and attribute-level hallucinations across diverse scenes, while Fig. 8 further demonstrates its effectiveness under more challenging long-form generation settings with denser visual content and finer-grained semantic dependencies. These cases collectively demonstrate that the semantics-directed token selection effectively preserves visually critical evidence, while the SSVR mechanism re-anchors decoding through on-demand cross-modal retrieval whenever visual grounding becomes insufficient.

H Usage of AI Assistant

We used GPT-4o exclusively for the language polishing of this manuscript, including grammar correction and sentence-level rephrasing. All generated outputs were manually reviewed and edited to ensure technical accuracy. No part of the research ideation, method design, experimental analysis, or interpretation of the results involved AI assistance.

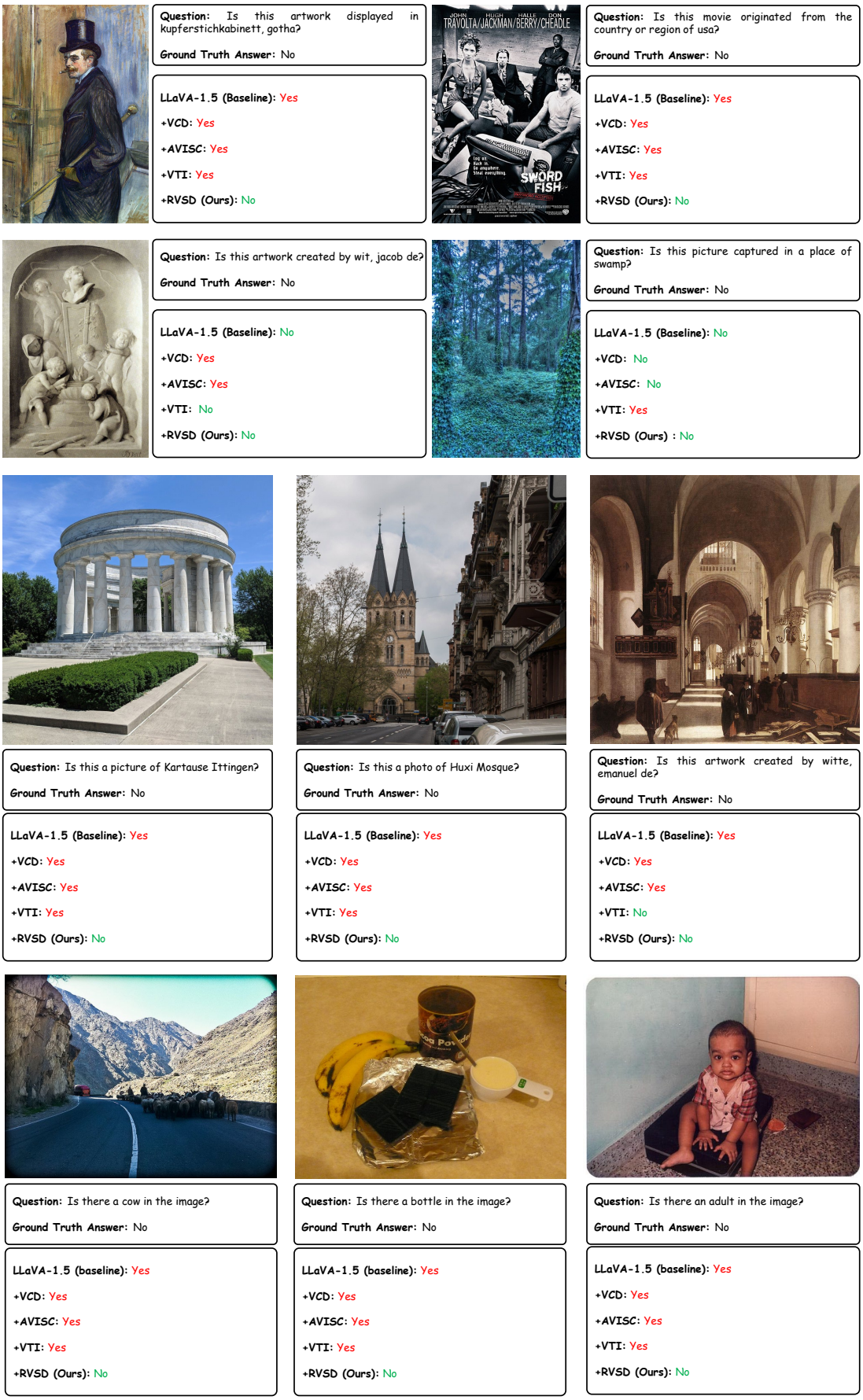

Figure 7: Qualitative case studies of RVSD compared with the vanilla decoder and baseline methods. Hallucinated content is highlighted in red, while faithful content grounded in the input image is highlighted in green. RVSD consistently produces visually faithful descriptions across diverse scenes.

Textual input: Please describe this image in detail.

LLaVA-1.5: The image is a gathering of several children sitting on benches and waiting areas, likely at an airport. There are a bunch of young kids and their luggage, including backpacks and suitcases, scattered throughout the scene. Some suitcases are placed on the floor close to the children, while others are further away. Multiple children and their belongings are visible in the central area of the image, with their luggage close by. **There are also several chairs within the area, along with a couple of benches that the kids are seated on.** The scene gives a sense of anticipation as the children wait for their transportation, likely for their next flight.

VCD: The image portrays a group of people, including children, sitting on multiple benches inside a building, likely a mall or airport. They appear to be waiting or relaxing in the public space. There are several benches scattered throughout the scene, with some individuals occupying them. **In addition to the people, there are numerous personal belongings like backpacks, handbags, and a suitcase. Two backpacks are placed on the ground near some of the benches, while another backpack is placed further away. A handbag is situated near a bench, and a suitcase is placed close to one of the people. A cup can also be seen on a surface in the scene.Overall, the atmosphere is lively, with people engaged in conversation, waiting, or simply enjoying their time in the public space.**

VTI: The image features a very busy and crowded airport terminal with numerous people present. There is a small family consisting of a man, a woman, and two little children seated together on a bench close to a large mirror. The woman is possibly accompanying the group, comforting them as they wait. **Various luggage is scattered throughout the scene, including a suitcase close to the center bench, another suitcase towards the top right, and a handbag lying closer to the woman on the large bench.** The airport's entrance can be seen in the upper part of the image, and it is filled with people trying to find their flights or looking for their gates.

AVISC: The image features a group of people sitting on a bench at a travel place. The bench is occupied by several individuals, including children and adults. A chair is visible towards the right side of the scene. The setting is likely a travel spot, as several backpacks can be seen placed around the people. **There are also two handbags in the area, one close to the center of the scene and the other towards the left side.** The people seem to be enjoying their time together, waiting for their journey to continue.

RVSD: The image features a busy airport scene with several people scattered throughout the location. **A group of children can be seen walking around, while other individuals are sitting on multiple benches or chairs. One girl, in particular, is sitting at the edge of a bench, seemingly posing for the photograph. There are multiple backpacks placed in various locations around the children, with some being close to the benches where they're sitting and others further away.** The children appear to be enjoying themselves as they navigate through the busy transport area.

Textual input: Please describe this image in detail.

LLaVA-1.5: The scene features a basket filled with tropical fruit and vegetables, including a group of bananas, coconuts, a guava, and other vegetables. There are bananas scattered throughout the basket with some placed at the top, center, and sides. **In addition to the fruits, there is a bottle placed near the bottom of the basket, possibly containing a beverage or a sauce to accompany the fruits. The combination of the fruits and the bottle creates a colorful and inviting scene.** This presentation of fruit and vegetables makes for a vibrant and visually appealing arrangement.

VCD: **The image features a bowl placed on a table, filled with a variety of fruits and vegetables. There are several bananas in the bowl, some of which are partially eaten or rotting. A few apples are also present in the bowl, adding to the assortment of fruit. In addition to the bananas and apples, the bowl also contains a handful of coconuts, adding a tropical touch to the display.** A leaf is situated close to the bananas, further enhancing the natural and organic appearance of the scene.

VTI: The image depicts a basket containing ripe, yellow fruits, possibly bananas or cantaloupes. These fruits are spread across the basket and partially overlapping each other. **In addition to the fruit, there is a small jar of honey sitting on top of the basket.** The combination of fruits and honey creates a natural and appealing arrangement for consumption.

AVISC: **The image shows a table with a bunch of bananas and two bowls.** One bowl contains a variety of bananas and coconuts, while the other bowl appears to have some fruit in it as well. **A bottle can be seen placed near the table, and the bananas are placed in a variety of spots on the table.**

RVSD: The image depicts a basket filled with a variety of fruits including bananas and coconuts. **There are six bananas situated in different areas of the basket, ranging from the top left to the bottom right. Close to the bananas, the basket also contains eight coconuts, spread out across the top, middle, and bottom parts of the container.** This colorful and diverse fruits and coconuts arrangement creates an appetizing display.

Figure 8: Additional qualitative case studies of RVSD on more challenging long-form generation samples. Hallucinated content is highlighted in red, while faithful content grounded in the input image is highlighted in green. RVSD continues to deliver visually faithful generations even under denser scenes and finer-grained semantic dependencies, where baseline decoders frequently exhibit object-, attribute-, and relation-level hallucinations.