SAGE: Sink-Aware Guided Emphasis for Visual Grounding in Vision-Language Decoders

Jeonghyo Song*, **YoungJoon Yoo**†
Department of Artificial Intelligence, Chung-Ang University
{thd9592s, yjyoo3312}@cau.ac.kr
*First author. †Corresponding author.

Abstract

Recent large vision-language models (VLMs) pair a visual encoder with a large language model (LLM) and perform well on diverse image–text tasks, yet their reliability is often limited by decoder attention pathologies that suppress visual evidence and exacerbate hallucinations. In this paper, we revisit visual attention sinks and uncover a structured, layer-dependent behavior: across prompts, early and late decoder layers exhibit prompt-invariant attention collapse onto the same few image regions, which we term PIS (Prompt-Invariant Sinks), whereas mid layers become prompt-conditioned and drive vision-language alignment. This split suggests that treating sinks as a uniform effect is incomplete. Building on this insight, we propose SAGE (Sink-Aware Guided Emphasis), a lightweight intervention that steers decoder attention away from PIS and toward query-dependent regions of interest (ROIs) using token-aligned ROI masks derived from standard vision backbones such as CLIP, ViT, and DINOv3. Evaluated on diverse vision-encoder + decoder-only LLM VLM families, SAGE improves visual grounding, reduces hallucinations, and yields consistent gains across public downstream vision-language benchmarks, including fine-grained visual discrimination settings where localized evidence is crucial, when instantiated with backbone-derived ROI masks.

1 Introduction

Large vision–language models (VLMs) (Liu et al., 2023, 2024b; Li et al., 2023a; Wang et al., 2024; Yang et al., 2025) have become a strong foundation for integrating visual perception with language reasoning. By connecting a visual encoder to a large language model (LLM) (Achiam et al., 2023; Touvron et al., 2023; Minaee et al., 2024) decoder via lightweight adapters (Wan et al., 2023; Van Nguyen et al., 2025), recent VLMs support a broad spectrum of applications, including image captioning, detailed image understanding, OCR, object localization, and multi-turn visual question answering within a unified interface. Despite these capabilities, reliability remains a central concern. VLMs may overlook subtle but decisive visual evidence, struggle with fine-grained recognition (Tong et al., 2024; Kim and Ji, 2024; Liao et al., 2025), and produce plausible yet incorrect statements that contradict the image, a behavior commonly referred to as hallucination (Li et al., 2023b; Huang et al., 2024; Chen et al., 2025b; Liu et al., 2024a).

Prior analyses suggest that these failures are not solely due to the visual encoder, but also to the decoder's attention dynamics over visual tokens (Kang et al., 2025). In particular, self-attention can concentrate on a small subset of tokens that consistently attracts an outsized share of the attention mass. In multimodal decoders, such visual attention sinks can dominate the attention budget and pull attention away from query-relevant image regions. This reduces precise visual grounding, especially for questions that rely on subtle, localized evidence.

In this paper, we re-examine decoder-side visual attention sinks and identify a structured, layer-dependent behavior. This behavior is easy to miss when attention is aggregated across layers rather than analyzed layer by layer. Figure 1 provides a motivating example. Across three substantially different prompts, early decoder layers and final decoder layers repeatedly assign high attention to nearly identical image locations, even though the question changes. We call these stable, prompt-agnostic regions Prompt-Invariant Sinks. In contrast, the mid layers exhibit the only consistently prompt-responsive behavior. When the query asks for the 'espresso price', mid-layer attention shifts toward the menu board. When the query asks for the current time, mid-layer attention moves toward the clock. When the prompt requests a scene description, mid-layer attention becomes broader and

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 36004–36026
October 24-29, 2026 ©2026 Association for Computational Linguistics

better aligned with semantically relevant regions. These observations reveal a dual role of the decoder stack: attention is largely static near the beginning and the end, while meaningful vision–language alignment emerges primarily in the middle layers.

This layer-wise structure has two implications. First, treating visual attention sinks as a single uniform pathology is incomplete, because the decoder simultaneously contains components that are insensitive to the query and components that must perform alignment. Second, Prompt-Invariant Sinks offer a concrete target for intervention. If attention can be discouraged from collapsing onto these static sinks and instead encouraged to concentrate on query-dependent regions of interest, VLMs may become more faithful to the image and more effective at fine-grained reasoning.

Motivated by these observations, we introduce SAGE (Sink-Aware Guided Emphasis) as a lightweight and controlled method to establish where and how decoder-side attention interventions matter in encoder–decoder VLMs such as widely-used Qwen-VL (Wang et al., 2024; Yang et al., 2025), LLaVA (Liu et al., 2023, 2024b; Li et al., 2025)-style, InternVL (Zhu et al., 2025; Chen et al., 2024) and NVILA (Liu et al., 2025) architectures. The core contribution of this work is the layer-dependent structure of visual attention sinks and alignment, and SAGE serves to make this effect measurable and reproducible. Concretely, SAGE steers decoder attention away from Prompt-Invariant Sinks and toward query-dependent regions of interest (ROIs) during generation. In our implementation, ROI masks are derived from standard pretrained vision backbones such as CLIP, ViT, and DINOv3, providing a consistent way to construct token-aligned prompt-conditioned interventions across models. With SAGE, we sweep over the intervention layers and show that simple ROI-guided edits at specific decoder layers yield reliable gains in grounding and downstream performance.

Our contributions are summarized as threefold:

- We identify Prompt-Invariant Sinks, showing that early and final decoder layers exhibit prompt-agnostic attention concentration on fixed image regions across diverse queries, as illustrated in Figure 1.
- We reveal a layer-wise duality in VLM decoders, where mid layers are primarily responsible for prompt-responsive vision-language alignment, while end layers exhibit static sink behavior.
- We propose SAGE as a controllable local ROI-guided intervention tool that enables systematic layer sweeps and provides evidence that selecting an effective decoder layer for intervention can improve grounding and benchmark performance across multiple VLMs.

2 Related Work

2.1 Vision-Language Models

Large vision-language models (VLMs) couple visual representation learning with language generation and reasoning. A dominant family follows the vision-encoder + decoder-only LLM design, in which image features are projected into visual tokens and jointly processed with text tokens by an autoregressive decoder. Representative examples include LLaVA-style models (Liu et al., 2024b; Li et al., 2025) and Qwen-VL (Bai et al., 2023; Wang et al., 2024; Yang et al., 2025), while recent work further improves efficiency and fine-grained perception by sparsifying visual tokens at inference or by predicting regions of interest to focus computation (Khaki et al., 2025; Shi et al., 2026). Other work explores alternative multimodal bridging and scaling strategies, including Flamingo (Alayrac et al., 2022), BLIP-2/InstructBLIP (Li et al., 2023a; Dai et al., 2023), explicit grounding formulations (Geigle et al., 2024), and multilingual large-scale training as in PaLI and PaLI-X (Chen et al., 2022, 2023).

2.2 Attention Sinks and Visual Attention Sinks

Transformer attention is known to exhibit attention sinks, where a few tokens absorb disproportionate attention mass independent of semantic relevance. This phenomenon has been linked to efficient long-context and streaming inference, and later work studies when sinks emerge during pretraining and how architectural or normalization choices affect their formation (Xiao et al., 2023; Gu et al., 2025; Wu et al., 2024). Similar pathologies also arise over visual tokens in multimodal models. VAR (Kang et al., 2025) extends the sink notion to LVLMs and shows repeated concentration on query-irrelevant visual tokens; related work in vision transformers and token/head-level analyses likewise studies architectural mitigation, selective steering toward salient regions, and training-time regularization for

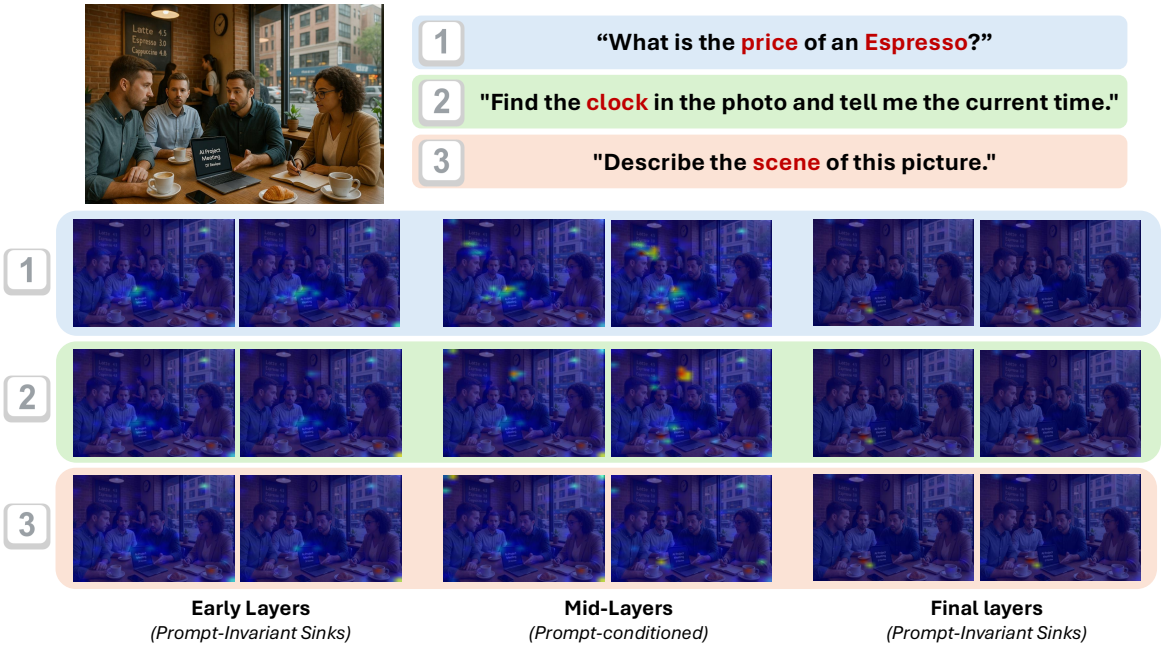


Figure 1: **Prompt-Invariant Sinks and prompt-conditioned mid-layer attention in a VLM decoder.** We visualize decoder attention over visual tokens for three different prompts (rows) at representative early, mid, and final decoder layers (columns). Early and final layers exhibit prompt-invariant attention collapse onto nearly identical image regions across prompts, forming Prompt-Invariant Sinks. In contrast, mid-layer attention is prompt-conditioned: it shifts to the menu board when asking for the espresso price (Prompt 1), to the wall clock when asking for the current time (Prompt 2), and becomes broader for holistic scene description (Prompt 3). Warmer colors indicate higher attention.

stronger grounding (Feng et al., 2026; Tang et al., 2026; Zhao et al., 2026; Esmaeilkhani and Latecki, 2026).

These attention pathologies are closely connected to hallucination and grounding failures, where outputs are not faithfully supported by the image. Benchmarks such as POPE (Li et al., 2023b) quantify object hallucination, while recent methods mitigate it through decoding-time interventions, contrastive signals, or vision-side masking of uncertain visual tokens (Kim et al., 2024; Park et al., 2025; Zhao et al., 2026; Chen et al., 2025a; Jung et al., 2025; Seo et al., 2026; Leng et al., 2024).

Our positioning and new finding. While prior work establishes the existence of visual attention sinks, our study shows that sink behavior is structured across depth rather than uniform. Early and late decoder layers tend to exhibit prompt-invariant collapse onto stable image regions, whereas mid layers remain comparatively prompt-responsive and play the main role in vision-language alignment. This layer-dependent split motivates interventions that steer attention away from prompt-invariant sink regions and query-dependent ROI.

3 Methodology

In this section, we present **SAGE** (Sink-Aware Guided Emphasis), a training-free inference-time intervention that steers decoder attention toward query-relevant visual evidence and away from attention-sink regions.

We focus on vision-encoder + decoder-only LLM VLMs, including LLaVA 1.5(Liu et al., 2024b), Qwen2-VL (Wang et al., 2024), InternVL3 (Zhu et al., 2025), and NVILA (Liu et al., 2025), where a vision encoder produces visual features that are projected or adapted into a sequence of visual tokens and processed jointly with text tokens by an autoregressive LLM via causal self-attention.

A key of SAGE is that it can be instantiated with multiple backbone-derived ROI generators as long as they can be mapped into a token-aligned ROI mask over visual tokens. In this paper, we instantiate SAGE with ROI masks derived from standard vision backbones such as CLIP (Radford et al., 2021), ViT (Dosovitskiy et al., 2020), and DINOv3 (Siméoni et al., 2025). These masks provide a simple and consistent interface for decoder-side attention steering across models.

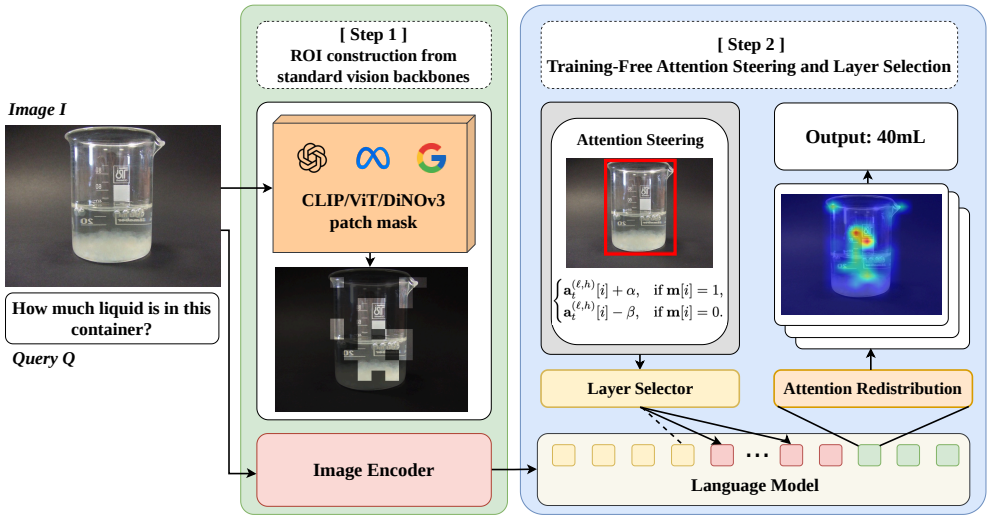


Figure 2: **Overview of SAGE.** Given an image I and a query Q, SAGE operates in two stages. **Step 1 (ROI construction from standard vision backbones):** a token-aligned ROI mask over visual tokens is derived from patch-level relevance maps obtained from CLIP, ViT, or DINOv3. **Step 2 (Training-free attention steering and layer selection):** using the ROI mask, SAGE applies an additive bias to visual-token attention logits (Eq. (1)) at selected decoder layer(s), redistributing attention away from sink regions and toward query-relevant visual evidence. Layer selection can be performed via a layer sweep and top-k selection strategy (Section 3.4). The resulting attention redistribution improves visual grounding during autoregressive generation.

3.1 Preliminaries and Notation

Given an image I and a query Q, the vision encoder produces visual features that are transformed into N visual tokens $\mathbf{V} = [v_1, \ldots, v_N]$ via a projector/adapter. The LLM based on the decoder-only processes a unified multimodal sequence consisting of visual tokens and text tokens and generates an answer autoregressively.

At decoding step t, layer $\ell \in \{0, \ldots, L-1\}$, and head h, we denote by $\mathbf{a}_t^{(\ell,h)} \in \mathbb{R}^N$ the restricted pre-softmax attention logits from the query token at step t to the N visual-token keys. SAGE represents the guided ROI in visual-token space as a binary mask $\mathbf{m} \in 0, 1^N$, where $\mathbf{m}[i] = 1$ marks the i-th token as inside the ROI.

3.2 ROI Construction from Standard Vision Backbones

SAGE requires only a token-aligned binary ROI mask $\mathbf{m} \in \{0, 1\}^N$ and makes no assumption about how it is produced. We instantiate it with three standard pretrained vision backbones, which differ in an important respect: only CLIP conditions on the query, while ViT and DINOv3 produce query-agnostic saliency.

CLIP We project and L_2-normalize the patch-token embeddings and the query text embedding into the shared CLIP space and score each patch by $\cos(\text{patch}, \text{text})$. The resulting relevance map therefore depends on the query.

ViT We use attention rollout from the CLS token at the last layer, averaged over heads. The query is not an input.

DINOv3 We use CLS-to-patch attention at the last layer, excluding register tokens. The query is not an input.

Patch-to-token alignment. For every source, the patch score map is resized to the visual-token grid of the evaluated VLM with nearest-neighbor interpolation, and the top-k tokens are kept as the binary mask.

3.3 Training-Free Attention Steering

Given a token-aligned ROI mask $\mathbf{m} \in \{0, 1\}^N$, SAGE steers attention toward query-relevant visual evidence by directly biasing attention allocation within the decoder-only LLM. For a chosen layer ℓ and head h at decoding step t, let $\mathbf{a}_t^{(\ell,h)} \in \mathbb{R}^N$ denote the (visual-token) attention logits over N visual tokens. SAGE applies a simple additive bias controlled by $\alpha > 0$ (ROI amplification) and $\beta > 0$ (non-ROI suppression):

$$\mathbf{a}_t'^{(\ell,h)}[i] = \begin{cases} \mathbf{a}_t^{(\ell,h)}[i] + \alpha, & \text{if } \mathbf{m}[i] = 1, \\ \mathbf{a}_t^{(\ell,h)}[i] - \beta, & \text{if } \mathbf{m}[i] = 0. \end{cases} \quad (1)$$

This bias reallocates attention mass toward the guided ROI, encouraging the model to consult query-relevant visual tokens while preserving the

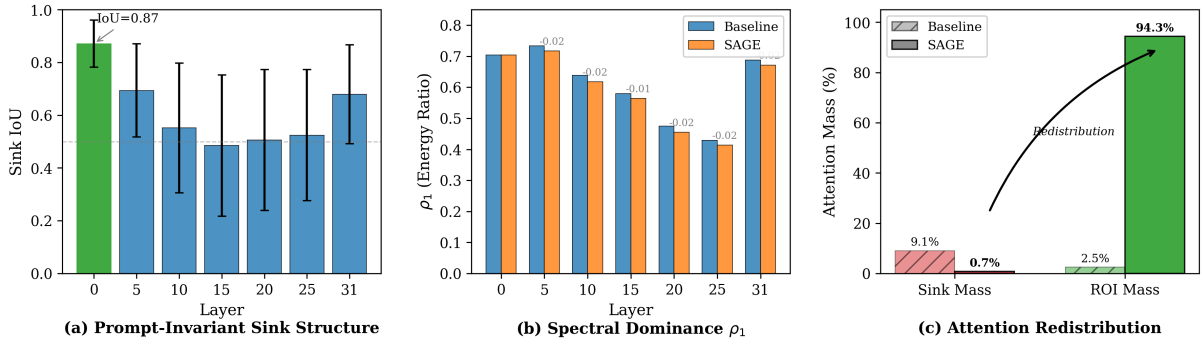


Figure 3: **Evidence for sink structure, spectral dominance, and attention redistribution.** (a) **Prompt-invariant sink structure.** For each layer, we form a sink set as the top-5% visual tokens by attention weight and report the IoU of sink sets across different prompts for the same image (error bars: standard deviation). (b) **Spectral dominance.** We stack visual-key attention logits across prompts and decoding steps to form $\mathbf{A}^{(\ell,h)}$ (Eq. (3)) and report the top-1 energy ratio ρ_1 (Eq. (5)); SAGE reduces spectral concentration, indicating weaker low-rank sink dominance. (c) **Attention redistribution.** we report attention mass assigned to sink tokens (top-5% tokens) and ROI tokens; SAGE shifts attention budget away from sink tokens and toward ROI, consistent with the odds-ratio effect in Eq. (6).

original model parameters. SAGE is therefore training-free and can be applied at inference time to arbitrary layer sets $\mathcal{L}_{\text{int}}$ (single-layer or multi-layer).

Implementation note. The ROI mask is first computed at patch level and then aligned to the visual-token grid used by each evaluated VLM. The exact patch-to-token mapping depends on the model's vision encoder and preprocessing pipeline; we keep this mapping fixed within each experiment and summarize the details in Appendix A.3.

3.4 Layer Sweep and Top-k Layer Selection

SAGE can be applied to different decoder depths, and its effect is generally layer-dependent. Rather than prescribing a single fixed layer as the only valid choice, we use a simple sweep protocol to (i) characterize where attention steering tends to be effective and (ii) form reasonable intervention sets (e.g., top-k layers) that perform well in practice.

Concretely, we run a single-layer sweep where SAGE is applied to one layer ℓ at a time (all other layers unchanged) and measure the relative change in performance against the no-intervention baseline. Let $\Delta(\ell)$ denote an aggregated relative performance change for intervening only at layer ℓ. This sweep serves as a diagnostic: it highlights layers where steering is more effective and provides a ranking that can be used to construct multi-layer intervention sets. Using the ranking by $\Delta(\ell)$, we define a top-k layers set

$$\mathcal{L}_{\text{top-}k} = \text{TopK}_{\ell}\big(\Delta(\ell)\big), \tag{2}$$

and apply SAGE to $\mathcal{L}_{\text{top-}k}$ in subsequent experiments as one practical choice. Importantly, our goal is not to claim that only a particular layer (or only one selection rule) works; instead, the sweep illustrates that attention steering itself is broadly effective, and that multiple reasonable layer selections (single-layer or multi-layer) can yield improvements.

3.5 Mechanistic Rationale

We provide a mechanistic view of why SAGE can improve visual grounding in terms of Sink Subspace and Attention Redistribution. Our analysis in Figure 3 suggests that visual attention sinks in VLM decoders are not arbitrary noise but exhibit structured, prompt-invariant patterns across depth. We formalize this intuition using a simple spectral perspective and show how the SAGE bias induces an attention-redistribution effect. Figure 3 summarizes the key empirical signatures discussed below.

Sink subspace perspective. Fix a decoder layer ℓ and head h. For each decoding step t and prompt instance p, let $\mathbf{a}_{p,t}^{(\ell,h)} \in \mathbb{R}^N$ denote the pre-softmax attention logits over the N visual-token keys (as in Section 3.1). To quantify prompt-invariant sink, we define a sink set as the top-r% visual tokens by attention weight and measure its overlap across prompts for the same image using set IoU (Fig. 3a).

We stack these logit vectors across prompts and decoding steps into a matrix

$$\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{M\times N}, \qquad \mathbf{A}^{(\ell,h)}[r,:] = \mathbf{a}_{p,t}^{(\ell,h)}, \tag{3}$$

where r indexes the pair (p,t) and M is the total number of stacked instances. We then compute an SVD,

$$\mathbf{A}^{(\ell,h)} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}. \tag{4}$$

If attention repeatedly collapses onto a small set of prompt-invariant sink tokens, the spectrum becomes more concentrated, and the top right-singular vectors (columns of $\mathbf{V}$) capture stable token-level directions corresponding to sink-dominated patterns. A convenient summary is the energy captured by the top-k components:

$$\rho_k(\mathbf{A}) = \frac{\sum_{j=1}^{k} \sigma_j^2}{\sum_{j=1}^{\min(M,N)} \sigma_j^2}, \tag{5}$$

where $\{\sigma_j\}$ are singular values. Our hypothesis is that prompt-invariant sinks induce high ρ_k for small k, reflecting structured low-rank dominance (Fig. 3b).

Attention redistribution induced by SAGE. SAGE applies an additive logit bias to visual-token attention as defined in Eq. (1). Let $p_i = \mathrm{Softmax}(\mathbf{a})_i$ and $p'_i = \mathrm{Softmax}(\mathbf{a}')_i$ be the attention weights over visual keys before and after applying Eq. (1), respectively. For any ROI index i and non-ROI index j, the odds ratio satisfies

$$\frac{p'_i}{p'_j} = \exp(\alpha + \beta) \cdot \frac{p_i}{p_j}. \tag{6}$$

Consequently, even when sink locations remain spatially stable, suppressing non-ROI tokens (which may include sink tokens) and amplifying ROI tokens can substantially redistribute attention budget toward query-relevant visual evidence in our analysis setting (Fig. 3c).

Testable implications. This view suggests two measurable signatures of effective attention steering: (i) reduced sink dominance reflected by lower spectral concentration (e.g., smaller ρ_k for small k), and (ii) increased attention mass on ROI tokens accompanied by decreased mass on sink tokens. We examine these signatures with layer-wise spectral and mass-based analyses, with details provided in Appendix E.

Cross-model evidence. The two diagnostics are not independent descriptions of the same plot. Across LLaVA-1.5 and Qwen2-VL on TextVQA and GQA, layer-wise prompt-invariant Sink IoU correlates strongly with top-1 spectral concentration (Pearson $r = 0.886$–0.942). This supports reading prompt-invariant sinks as structured low-rank routing directions rather than image-specific heatmap artifacts, and it is why the two can be combined into a single label-free depth criterion (Section 3.4).

4 Experiments

4.1 Experimental Setup

Models. We evaluate SAGE on vision-encoder + decoder-only LLM VLMs from the LLaVA 1.5 (Liu et al., 2024b), Qwen2-VL (Wang et al., 2024), InternVL3 (Zhu et al., 2025) and NVILA (Liu et al., 2025). Unless stated otherwise, we use the instruct-tuned checkpoints listed in Table 1.

ROI sources. Our main results use ROI masks derived from standard vision backbones, specifically CLIP, ViT, and DINOv3. Given an image and query, each source produces patch-level relevance scores, which are thresholded to obtain a binary ROI mask aligned to the visual-token grid of the evaluated VLM. Detailed source-specific settings are summarized in Appendix C.

Intervention settings. Unless stated otherwise, we use a fixed bias strength across models. In all experiments, we set $\alpha = 5.0$ (ROI amplification) and $\beta = 2.0$ (non-ROI bias). For layer-wise ablations, we apply a single-layer intervention by sweeping the intervention layer across the decoder stack of each model, with all other layers unchanged.

Normalized decoder depth. Because different VLM families can have different numbers of decoder layers, we use normalized decoder depth when comparing layer-wise behavior across models. For a model with L decoder layers and zero-indexed layer ℓ, we define the normalized depth as $d(\ell) = \ell/(L-1)$. We use raw layer indices only in model-specific analyses such as LLaVA-1.5 case studies and appendix tables.

4.2 Main Evaluations

We evaluate SAGE on representative vision-encoder + decoder-only LLM VLM families and summarize the main results in Table 1. For each model, benchmark, and ROI source, the reported SAGE score corresponds to the best single-layer result obtained from the layer sweep. We use these results to quantify the peak effect of SAGE at effective decoder depths, rather than to claim a universal fixed layer rule. Because the layer is selected per benchmark, these values indicate an upper bound on the achievable effect. Section 4.5 reports the result obtained with a single layer fixed in advance, without benchmark labels. For transparency, the exact selected layer indices are listed in Appendix Table 6; when discussing layer behavior across models, we refer to normalized decoder depth rather

Table 1: **Main results of SAGE across VLM families.** For each model, benchmark, and ROI source, the reported SAGE score corresponds to the best single-layer result obtained from the layer sweep. Exact selected layer indices are listed in Appendix Table 6.

Model	TextVQA	POPE		GQA	ScienceQA	MME	MMVP		MMVP Original		MMVet
		Acc	F1				Image	Pair	Image	Pair	
LLaVA 1.5-13B	55.50	82.75	81.55	61.65	63.98	1783.11	61.48	11.85	33.58	10.65	31.20
+ SAGE (CLIP)	**61.94** (+6.44)	**86.84** (+4.09)	**86.58** (+5.03)	62.65 (+1.00)	68.63 (+4.65)	1842.58 (+59.47)	62.96 (+1.48)	37.78 (+25.93)	50.00 (+16.42)	**22.00** (+11.35)	**42.20** (+11.00)
+ SAGE (ViT)	61.85 (+6.35)	85.81 (+3.06)	84.82 (+3.27)	**65.50** (+3.85)	69.15 (+5.17)	**1895.55** (+112.44)	**68.89** (+7.41)	**38.52** (+26.67)	**58.58** (+25.00)	21.27 (+10.62)	34.50 (+3.30)
+ SAGE (DINOv3)	61.54 (+6.04)	83.00 (+0.25)	81.72 (+0.17)	62.55 (+0.90)	**69.17** (+5.19)	1791.44 (+8.33)	63.33 (+1.85)	14.07 (+2.22)	34.00 (+0.42)	12.00 (+1.35)	36.47 (+5.27)
Qwen2-VL-7B	81.32	75.94	75.42	62.04	68.34	1792.60	78.15	57.78	69.33	40.67	59.60
+ SAGE (CLIP)	81.43 (+0.11)	76.59 (+0.65)	75.52 (+0.10)	**62.38** (+0.34)	**74.16** (+5.82)	1895.10 (+102.50)	81.15 (+3.00)	58.28 (+0.50)	**71.21** (+1.88)	41.33 (+0.66)	61.10 (+1.50)
+ SAGE (ViT)	**81.52** (+0.20)	77.15 (+1.21)	76.81 (+1.39)	62.28 (+0.24)	71.28 (+2.94)	**1920.00** (+127.40)	78.88 (+0.73)	57.94 (+0.16)	70.24 (+0.91)	**42.53** (+1.86)	**61.80** (+2.20)
+ SAGE (DINOv3)	81.41 (+0.09)	**78.35** (+2.41)	**77.56** (+2.14)	62.15 (+0.11)	68.47 (+0.13)	1795.23 (+2.63)	**91.85** (+13.70)	**85.19** (+27.41)	70.67 (+1.34)	**42.53** (+1.86)	59.90 (+0.30)
InternVL3-2B	47.60	87.97	87.41	56.85	61.95	2087.90	73.44	49.37	72.67	47.33	61.40
+ SAGE (CLIP)	60.20 (+12.60)	88.27 (+0.30)	88.36 (+0.95)	57.15 (+0.30)	62.25 (+0.30)	2092.58 (+4.68)	76.30 (+2.86)	**58.52** (+9.15)	73.33 (+0.66)	49.33 (+2.00)	**62.20** (+0.80)
+ SAGE (ViT)	47.80 (+0.20)	88.48 (+0.51)	87.89 (+0.48)	**59.85** (+3.00)	**62.45** (+0.50)	**2097.29** (+9.39)	**77.78** (+4.34)	57.78 (+8.41)	72.33 (-0.34)	46.00 (-1.33)	62.10 (+0.70)
+ SAGE (DINOv3)	**73.50** (+25.90)	**88.89** (+0.92)	**88.60** (+1.19)	57.55 (+0.70)	62.02 (+0.07)	2072.23 (-15.67)	73.70 (+0.26)	49.63 (+0.26)	**75.00** (+2.33)	**50.70** (+3.37)	61.50 (+0.10)
NVILA-8B	74.44	85.37	84.81	62.20	87.57	2107.29	78.52	58.52	70.67	43.33	61.79
+ SAGE (CLIP)	**78.13** (+3.69)	**92.20** (+6.83)	**91.99** (+7.18)	62.80 (+0.60)	89.37 (+1.80)	2125.59 (+18.30)	78.88 (+0.36)	59.25 (+0.73)	**72.67** (+2.00)	**47.33** (+4.00)	62.16 (+0.37)
+ SAGE (ViT)	74.84 (+0.40)	88.95 (+3.58)	88.26 (+3.45)	64.90 (+2.70)	**92.24** (+4.67)	**2133.59** (+26.30)	79.12 (+0.60)	59.82 (+1.30)	71.48 (+0.81)	45.13 (+1.80)	61.94 (+0.15)
+ SAGE (DINOv3)	74.75 (+0.31)	89.00 (+3.63)	88.18 (+3.37)	**64.95** (+2.75)	91.61 (+4.04)	2118.85 (+11.56)	**79.94** (+1.42)	**60.02** (+1.50)	71.86 (+1.19)	44.94 (+1.61)	**62.36** (+0.57)

than raw layer number.

Across the evaluated VLMs, SAGE improves over the no-intervention baseline on many benchmarks and across multiple ROI sources. These gains are consistent with our motivating hypothesis that decoder-side attention pathologies can suppress query-relevant visual evidence, and that a lightweight logit bias toward an ROI helps the model focus on more relevant visual tokens. Across CLIP, ViT, and DINOv3, the qualitative trend is similar: the strongest improvements appear at selected decoder depths, while the effective depth can vary across model families and benchmarks.

Fine-grained visual discrimination. The MMVP-family benchmarks are particularly relevant to our setting because they emphasize fine-grained pairwise visual discrimination, where the decisive evidence is often local and easy to miss. Table 1 shows strong improvements on MMVP and MMVP-Original, especially on pairwise metrics where both related samples must be answered correctly. To further examine whether these gains are tied only to a single selected layer, Figure 4 reports layer-wise averaged normalized improvement on the four MMVP-family sub-benchmarks for LLaVA-1.5. For CLIP-, ViT-, and DINOv3-derived ROI masks, the averaged improvement remains positive across a broad range of decoder layers, indicating that the benefit on fine-grained visual discrimination is not restricted to one isolated intervention depth.

4.3 Layer-wise Effects

This section analyzes how SAGE's effectiveness varies with normalized decoder depth on LLaVA-1.5 and Qwen2-VL. Although SAGE uses the same intervention rule (Eq. (1)) and the same ROI construction procedure within each experiment setting, its effect can vary substantially depending on where in the decoder stack the intervention is applied. To isolate this factor, we conduct a single-layer sweep over the decoder stack for each model family by setting $\mathcal{L}_{\text{int}} = \{\ell\}$ and intervening at one layer at a time. Figure 5 summarizes the layer-wise improvements over the no-intervention baseline as a function of normalized decoder depth, aggregated across benchmarks (median), and highlights the top-performing depths.

Layer-wise variability. Figure 5 shows that SAGE exhibits clear layer-wise variability. Across both model families, intervening at a subset of layers yields consistent gains, whereas intervening at other layers leads to smaller changes and can oc-

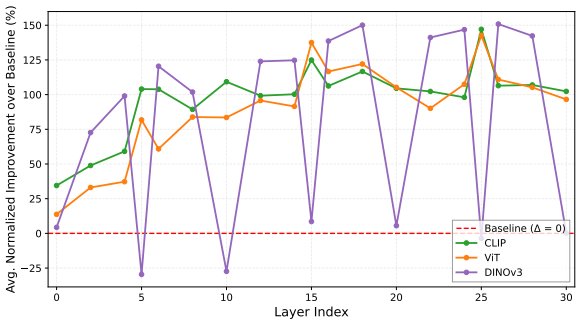


Figure 4: **Layer-wise averaged normalized improvement on MMVP-family benchmarks.** For each ROI source (CLIP, ViT, DINOv3), we apply SAGE to one decoder layer and evaluate on four fine-grained sub-benchmarks: MMVP image, MMVP pair, MMVP-Original image, and MMVP-Original pair. For each sub-benchmark, we compute the normalized improvement over the no-intervention baseline as $(s-b)/b \times 100$, where s is the score with SAGE and b is the baseline score. The plotted value is the average of these four normalized improvements at each intervention layer. The x-axis shows the LLaVA-1.5 decoder layer index, and values above zero indicate improvement over the baseline.

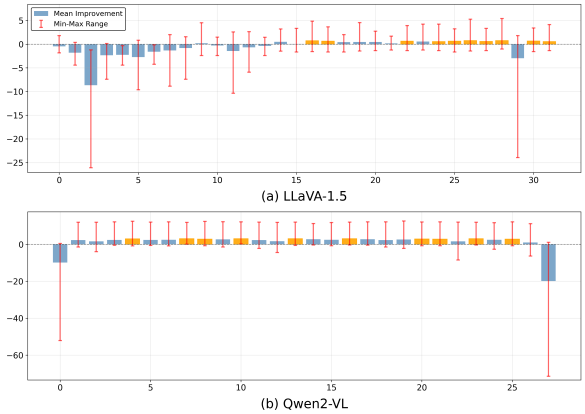


Figure 5: **Layer-wise single-layer sweep of SAGE interventions.** We report relative performance improvement over the no-intervention baseline when applying SAGE to a single decoder layer at a time. The x-axis is normalized decoder depth $d(\ell) = \ell/(L-1)$, allowing comparison across models with different decoder depths. The horizontal reference at 0% indicates the no-intervention baseline. Bars aggregate results across all benchmarks evaluated in our experiments, and error bars show the min–max range. Orange bars indicate the top-performing normalized depth positions.

casionally be neutral. This indicates that attention steering is not uniformly beneficial throughout the decoder and that layer selection is a key factor in determining the effectiveness of SAGE.

Connection to visual attention sinks. Prior work reports that VLMs can suffer from visual attention sinks, where attention repeatedly collapses onto a small set of query-irrelevant visual tokens and limits the use of task-relevant visual evidence. SAGE is designed to counteract this pathology by reallocating attention toward a query-dependent ROI. The concentration of gains in selected decoder layers suggests that correcting attention allocation at those depths can improve grounding and downstream performance in the evaluated setting.

4.4 Depth-Grouped Multi-layer Comparisons

We further examine multi-layer SAGE by grouping the tested sets according to coarse normalized-depth composition. Table 2 reports mean TextVQA accuracy within each group, while selected exact LLaVA-1.5 layer combinations are deferred to Appendix Tables 7, 8, and 9.

The summary shows a clear depth-dependent pattern. Very-early groups are substantially weaker, whereas mid–final and final–final groups are stronger and more stable. This indicates that multi-layer gains are not tied to a single exact layer set, but appear more broadly among non-early depth compositions. Using normalized-depth groups reduces reliance on any single selected layer. The gap between very-early and non-early groups suggests that the earliest decoder depths are more fragile intervention sites. Additional exact combinations and a layer-skip ablation are provided in Appendix B.3 and Appendix B.4.

Table 2: **Depth-grouped summaries of multi-layer SAGE sets on TextVQA.** Each value reports the mean accuracy within a coarse depth-composition group.

Depth composition	Mean Acc.
2-layer sets	
Very-early only	19.19
Early-mixed	49.85
Mid–mid	61.96
Mid–final	**62.60**
Final–final	62.42
3-layer sets	
Very-early $\times 3$	2.35
Early $\times 3$	55.39
Early–early–mid	59.10
Early–mid–mid	60.73
Mid $\times 3$	62.11
Mid–mid–final	62.10
Mid–final–final	62.40
Final $\times 3$	62.40

Table 3: **Fixed-layer result and ROI content control**. **Best** selects the layer on the benchmark; **L24** uses one fixed layer chosen by label-free diagnostics. **Rand.** replaces the guided ROI with a random token set of the same size, at the same layer and bias strength.

		SAGE		
Benchmark	Base	Best	L24	Rand.
TextVQA	44.0	55.4	54.6	43.1
GQA	57.1	62.8	62.3	57.0
POPE_{acc}	72.7	73.6	73.1	72.1
POPE_{F1}	69.3	73.0	69.9	69.1
ScienceQA	68.1	68.6	68.6	68.0
MME	1681	1707	1688	1677
MMVP_{img}	67.0	68.1	67.2	66.7
$\text{MMVP}_{\text{pair}}$	35.6	37.8	35.7	35.2

4.5 Fixed-Layer Result and ROI Content Control

Table 1 selects the intervention layer per benchmark. We therefore also evaluate a fixed-layer setting in which the intervention depth is chosen in advance, without benchmark labels, using the two diagnostics already introduced in Section 3.5: prompt-invariant Sink IoU (Fig. 3a) and top-1 spectral concentration ρ_1 (Eq. (5), Fig. 3b). Both are computed from attention statistics alone and require no labels. In the LLaVA-1.5 diagnostic sweep, layer L24 attains the lowest value under both, indicating the weakest prompt-invariant sink structure among the examined depths. We select it before running any downstream evaluation and keep it fixed across all benchmarks, never re-selecting per benchmark or per metric.

Table 3 reports this setting against the re-measured baseline under one unified protocol (LLaVA-1.5-7B). With one fixed layer, SAGE raises TextVQA from 44.0 to 54.6 and GQA from 57.1 to 62.3, while the changes on POPE, ScienceQA, MME, and MMVP are small. This is consistent with the premise that the intervention helps when the decisive evidence is localized.

ROI content control. Within this same fixed-layer setting, we further ask whether the gain depends on what the ROI contains. A logit bias of this strength reallocates attention regardless of where the mask points, so we hold the intervention layer, the bias strength, and the mask size fixed and replace the guided ROI with a randomly selected set of visual tokens. This control leaves every benchmark at or below baseline. The effect therefore requires the content of the target region and is not a generic consequence of perturbing the attention distribution.

5 Conclusion

In this paper, we uncover a consistent decoder attention pathology in modern VLMs: early and late layers repeatedly collapse onto prompt-invariant visual regions, while mid layers remain the main locus of prompt-conditioned grounding. Building on this diagnosis, we propose SAGE, a simple training-free inference-time method that reweights attention with a token-aligned ROI mask to prioritize relevant evidence. SAGE can be instantiated with backbone-derived ROI masks from standard vision backbones such as CLIP, ViT, and DINOv3. More broadly, our results suggest that decoder attention allocation is not only a symptom of grounding failure, but also a useful intervention point for improving visually grounded generation. We further show strong layer dependence and that intervening at appropriately chosen layers yields the most reliable improvements. Across multiple VLM families and benchmarks, SAGE yields broad improvements in many evaluated settings, suggesting that correcting decoder-side attention allocation is an effective handle for reducing visually unfaithful outputs.

Limitations

SAGE relies on token-aligned ROI masks derived from standard vision backbones, and its effectiveness depends on the quality of these estimates. The three sources differ in what signal they encode: CLIP scores each patch by its similarity to the query embedding, while ViT and DINOv3 derive relevance from CLS-token attention and therefore capture image saliency. SAGE benefits from both kinds of signal, which suggests that steering toward task-relevant regions does not require the mask itself to be query-conditioned; extending saliency-based sources with query conditioning is a natural direction for future work. As illustrated in Figure 8, fragmented or misaligned ROI estimates can steer attention toward irrelevant visual tokens and suppress the true evidence. The random-ROI control in Section 4.5 bounds the other end of this dependence, since an uninformative mask of the same size and bias strength removes the gain entirely. A deployed system would therefore benefit from a confidence signal that decides when to intervene, for instance a margin on the patch relevance scores; we leave this, together with improved patch-to-token alignment, to future work.

The intervention itself is deliberately simple. We apply one global (α, β) across all models and

benchmarks to avoid per-task tuning, and the mask is held fixed across decoding steps and broadcast to all attention heads. This keeps SAGE easy to deploy and reproduce, but it also emphasizes the same region for every generated token, which suits questions with a single decisive region better than multi-region descriptions or multi-step reasoning. Per-step masks, head-selective steering, and per-model tuning of the bias strength are all compatible with the formulation in Eq. (1), as is extending the single-image mask to multi-image and video.

Two aspects of the diagnosis remain open. SAGE is sink-aware at the level of depth rather than of individual tokens: the PIS diagnostics identify where in the decoder steering is useful, and the additive bias then reallocates the attention budget through Eq. (6). This design needs no per-token sink classifier, but whether explicitly suppressing sink tokens would add to the depth-level effect is untested, as is whether the same normalized depth transfers across model families without re-running the diagnostic. Our characterization of prompt-invariant sinks is likewise empirical: the structure is consistent across models and across two independent diagnostics, but we do not explain why early and late layers develop it while mid layers remain prompt-responsive, and we regard this formation mechanism as an interesting open problem.

Ethical Considerations

This work studies decoder-side attention steering in vision-language models and aims to improve faithfulness to visual evidence. Although improved grounding may reduce visually unsupported outputs, the proposed method does not guarantee correctness and may still fail when ROI masks are inaccurate or unstable. Accordingly, SAGE should not be treated as a safety mechanism in high-stakes settings, and human oversight and task-specific validation remain necessary before deployment. Because the method operates on image inputs that may contain sensitive content, practitioners should also follow appropriate privacy and data-governance practices in downstream use.

Acknowledgment

This research was supported by the AI Seoul Tech Research Support Program of the Seoul Future Foundation. This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

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

Boxu Chen, Ziwei Zheng, Le Yang, Zeyu Geng, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025a. Seeing it or not? interpretable vision-aware latent steering to mitigate object hallucinations. *arXiv preprint arXiv:2505.17812*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, and 1 others. 2023. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Zhiyuan Chen, Yuecong Min, Jie Zhang, Bei Yan, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. 2025b. A survey of multimodal hallucination evaluation and detection. *arXiv preprint arXiv:2507.19024*.

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

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

Parsa Esmaeilkhani and Longin Jan Latecki. 2026. Direct visual grounding by directing attention of visual tokens. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5787–5797. IEEE.

Wenfeng Feng, Hongxiang Wang, Jianlong Wang, Xin Zhang, Jingjing Zhao, Yueyue Liang, Xiang Chen, and Duokui Han. 2026. Edit: Enhancing vision transformers by mitigating attention sink through an encoder-decoder architecture. In *International Conference on Optoelectronics, Computer Science, and Algorithms (OCSA 2025)*, volume 14008, pages 246–259. SPIE.

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.

Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024. Does object grounding really reduce hallucination of large vision-language models? *arXiv preprint arXiv:2406.14492*.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, volume 2025, pages 97114–97144.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. 2024. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9614–9631.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702. IEEE.

Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. 2025. Visual attention never fades: Selective progressive attention recalibration for detailed image captioning in multimodal large language models. *arXiv preprint arXiv:2502.01419*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pages 87676–87703.

Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N Plataniotis, Yao Lu, Song Han, and Zhijian Liu. 2025. Sparsevila: Decoupling visual sparsity for efficient vlm inference. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23784–23794. IEEE.

Jeonghwan Kim and Heng Ji. 2024. Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. *arXiv preprint arXiv:2402.16315*.

Junho Kim, Hyun J Kim, Yeon J Kim, and Yong M Ro. 2024. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 37:133571–133599.

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.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2025. Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pages 81182–81199.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PmLR.

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

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. 2025. Can large vision-language models correct semantic grounding errors by themselves? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14667–14678.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, and 1 others. 2025. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4122–4134.

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

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. 2025. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6434–6442.

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

Hoigi Seo, Dong Un Kang, Hyunjin Cho, Joohoon Lee, and Se Young Chun. 2026. On epistemic uncertainty of visual tokens for object hallucinations in large vision-language models. *Advances in Neural Information Processing Systems*, 38:130605–130660.

Yuheng Shi, Xiaohuan Pei, Minjing Dong, and Chang Xu. 2026. Catching the details: Self-distilled roi predictors for fine-grained mllm perception. In *International Conference on Learning Representations*, volume 2026, pages 136133–136155.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, and 1 others. 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Lexiang Tang, Xianwei Zhuang, Bang Yang, Zhiyuan Hu, Hongxiang Li, Lu Ma, Jinghan Ru, and Yuexian Zou. 2026. Not all tokens and heads are equally important: Dual-level attention intervention for hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9439–9447.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

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

Chien Van Nguyen, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, and 1 others. 2025. A survey on small language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 807–821.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, and 1 others. 2023. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. 2024. On the role of attention masks and layernorm in transformers. *Advances in Neural Information Processing Systems*, 37:14774–14809.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Jianfei Zhao, Feng Zhang, Xin Sun, Chong Feng, and Zhixing Tan. 2026. Tell model where to look: Mitigating hallucinations in mllms by vision-guided attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 32582–32591.

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

A Additional Implementation Details

A.1 Hyperparameters and Training Hardware Settings

Table 4 summarizes the key configuration choices used in our SAGE implementation. Unless otherwise noted, these settings are shared across the experiments reported in the main paper. We use $(\alpha, \beta) = (5, 2)$ as the default bias strengths unless otherwise stated. For all models we use deterministic greedy decoding (`do_sample = False`) for every arm, including baselines and comparison methods, and otherwise follow the standard decoding settings recommended by each model family. All experiments are run on NVIDIA GPUs (A6000) using FP16 mixed precision.

A.2 Intervention Scope and Implementation

SAGE is implemented as a training-free, inference-time intervention that modifies attention allocation inside selected decoder layers without updating model parameters. In our implementation, we instrument the self-attention modules of target layers $\ell \in \mathcal{L}_{\text{int}}$ so that the additive bias in Eq. (1) is applied during attention computation.

Scope. The intervention uses a token-aligned ROI mask $\mathbf{m} \in \{0, 1\}^N$ and applies the additive bias only to attention logits corresponding to visual-token key positions; text-token keys remain unchanged. Within each selected layer, the same ROI mask is broadcast to all heads and all query positions in the layer's attention operation. All settings are kept fixed within each experiment configuration.

Layer configuration. The intervention layer set $\mathcal{L}_{\text{int}}$ supports both single-layer and multi-layer modes and is specified by a layer index list or range in the experiment scripts (e.g., "0", "0,5,10", "0–5").

Algorithm 1 Implementation sketch of SAGE instrumentation (inference-time)

Require: Model M, ROI mask $\mathbf{m}$, layer set $\mathcal{L}_{\text{int}}$, strengths α, β

Configure M to apply Eq. (1) in all $\ell \in \mathcal{L}_{\text{int}}$
$Y \leftarrow M.\text{generate}(I, Q)$
return Y

A.3 ROI Mask Construction Details

We summarize the local ROI construction pipeline used in our main experiments and the alignment from patch-level relevance maps to token-aligned ROI masks.

Local backbone-derived ROI masks. We instantiate SAGE with ROI masks derived from standard vision backbones such as CLIP, ViT, and DINOv3. For each image–query pair, a source-specific patch-level relevance score is computed and thresholded to obtain a binary patch mask.

Patch-to-token alignment. The binary patch mask is then aligned to the visual-token grid used by each evaluated VLM. Although the effective grid size can differ across model families depending on the vision encoder and preprocessing pipeline, the mapping from patch-level relevance to token-aligned mask is fixed within each experiment.

Model-specific patch grids. Different VLM families may use different effective patch grids depending on the vision encoder and image preprocessing. We align the patch-level ROI mask to the corresponding model's visual-token grid using the model's configured grid size and release the exact settings in the experiment scripts.

A.4 Benchmark Protocols and Evaluation Prompts

We follow the standard evaluation protocols released by each benchmark and keep decoding settings fixed within each benchmark across all compared methods. Below we briefly summarize what each benchmark probes in VLMs and the primary metric we report.

TextVQA. (Singh et al., 2019) OCR-centric VQA that tests whether a model can read scene text and ground it to answer questions. We report VQA-style exact-match accuracy.

POPE. (Li et al., 2023b) A hallucination/faithfulness diagnostic using yes/no questions about object existence. We report accuracy and F1 following the official protocol.

GQA. (Hudson and Manning, 2019) A compositional visual reasoning benchmark that probes object/attribute/relation reasoning with grounding. We report standard accuracy on the evaluated split.

ScienceQA. (Lu et al., 2022) A multimodal multiple-choice benchmark for scientific reasoning requiring integration of perception and reasoning. We report multiple-choice accuracy.

MME. (Fu et al., 2026) A broad diagnostic suite covering perception (e.g., OCR, count, color) and cognition (e.g., commonsense, numerical reasoning). We report the aggregated score under the

Table 4: **Key hyperparameters and settings for SAGE.** We use $(\alpha, \beta) = (5, 2)$ as the default bias strengths unless otherwise stated.

Component	Value / Description
Local VLM	Vision-encoder + decoder-only LLM VLMs (e.g., LLaVA-1.5 (Liu et al., 2024b)/Qwen2-VL (Wang et al., 2024)/InternVL3 (Zhu et al., 2025)/NVILA (Liu et al., 2025)) with standard preprocessing and decoding settings.
ROI sources	Patch-level ROI masks derived from standard pretrained vision backbones (CLIP/ViT/DINOv3).
Patch relevance scoring	Backbone-specific patch relevance scores are thresholded to form a binary ROI mask; the same scoring rule is used consistently within each source.
Patch-to-token mapping	The patch-level mask is aligned to the visual-token grid used by the evaluated VLM; the mapping is fixed per model and preprocessing configuration.
Patch grid	Model-specific patch grid size determined by the vision encoder and preprocessing; grid configuration is fixed per model and recorded in the experiment scripts.
Intervention layers	Layer set $\mathcal{L}_{\text{int}}$ specified as a list or range (single-layer or multi-layer), fixed per experiment.
Bias strengths	Default $(\alpha, \beta) = (5, 2)$ for ROI amplification and non-ROI suppression in Eq. (1).
Decoding	Greedy `do_sample = False`; maximum new tokens set per model family; decoding settings are held fixed within each benchmark.

official protocol.

MMVP. (Tong et al., 2024) A paired-image visual pattern benchmark that stresses sensitivity to subtle visual differences and discourages language priors. We report strict pair accuracy (a pair is correct only if both questions are answered correctly).

MMVet. (Yu et al., 2023) An integrated capability benchmark with open-ended questions scored by an LLM-based evaluator. We report the resulting score following the benchmark evaluator protocol.

Evaluation prompts. We use each benchmark's recommended input formatting and evaluation scripts; any additional prompt templates used in our runs are included in the released supplementary materials.

Artifact use. We use publicly released research artifacts, including model checkpoints, vision backbones, benchmark datasets, and official evaluation protocols, under their respective licenses and terms. We use these artifacts only for research evaluation and do not redistribute the underlying datasets or model weights.

A.5 Detailed Algorithm

Algorithm 2 summarizes the full SAGE pipeline. Given an input image I, a query Q, and a local VLM M (vision encoder + decoder-only LLM), SAGE performs a training-free intervention during generation by selectively biasing the logits of visual tokens at a chosen set of decoder layers $\mathcal{L}_{\text{int}}$.

Step 1: ROI mask construction. SAGE first constructs a binary ROI mask $\mathbf{m}$ over the visual tokens produced by the vision encoder:

$$\mathbf{m} \leftarrow \textsc{GetROIMask}(I, Q, \mathcal{S}).$$

Here, the ROI source $\mathcal{S}$ is a local vision backbone such as CLIP, ViT, or DINOv3. For the selected source, we compute a patch-level relevance map, threshold it to obtain a binary patch mask, and align the result to the visual-token grid of the evaluated VLM. The output mask $\mathbf{m} \in \{0, 1\}^{N_v}$ marks which visual tokens belong to the guided ROI.

Step 2: Enable logit-level steering at selected layers. Next, SAGE activates a lightweight forward hook that applies Eq. (1) to the visual-token logits at each layer $\ell \in \mathcal{L}_{\text{int}}$:

$$\textsc{EnableSteering}(M, \mathcal{L}_{\text{int}}, \mathbf{m}, \alpha, \beta).$$

Intuitively, α controls the strength of encouraging ROI tokens, while β controls the strength of suppressing non-ROI tokens (as defined in Eq. (1)). Importantly, this intervention does not require gradient updates or any modification to model parameters; it only alters the logits used to form attention

Algorithm 2 SAGE: Local ROI-Guided, Training-Free Attention Steering

Require: Image I, Query Q, local VLM M (vision encoder + decoder-only LLM)
Require: ROI source $\mathcal{S}$ (e.g., CLIP/ViT/DINOv3 patch mask)
Require: Intervention layer set $\mathcal{L}_{\text{int}}$, hyperparameters α, β
Ensure: Answer Y

$\mathbf{m} \leftarrow \textsc{GetROIMask}(I, Q, \mathcal{S})$
$\textsc{EnableSteering}(M, \mathcal{L}_{\text{int}}, \mathbf{m}, \alpha, \beta)$ ▷ apply Eq. (1) to visual-token logits at selected layers
$Y \leftarrow M.\text{generate}(I, Q)$
$\textsc{DisableSteering}(M)$
return Y

over visual tokens during decoding. By biasing visual-token logits, this intervention implicitly reshapes the attention distribution over visual tokens during decoding, without any parameter updates.

Step 3: Generate with SAGE enabled. With steering enabled, the model generates the answer as usual:

$$Y \leftarrow M.\texttt{generate}(I, Q).$$

All other components of decoding (e.g., sampling strategy, max tokens, temperature) remain unchanged unless specified elsewhere.

Step 4: Disable steering and return. Finally, SAGE removes the hook to restore the original model behavior:

$$\textsc{DisableSteering}(M),$$

and returns the generated output Y. This explicit enable/disable design ensures SAGE is modular and can be applied on a per-query basis without persistent side effects.

Practical notes. In practice, $\mathcal{L}_{\text{int}}$ can be chosen as a single layer or a small set of layers to trade off effectiveness and overhead. The only additional per-query cost comes from computing $\mathbf{m}$ and applying a constant-time logit bias at the selected layers; no extra training or finetuning is required.

A.6 Hyperparameters and Sensitivity

SAGE is controlled by two scalar hyperparameters in Eq. (1): α, the ROI-side amplification, and β, the non-ROI-side suppression. Table 5 reports a seven-point sensitivity study over (α, β) on MMVP with the ROI source and intervention layer held fixed. Performance is not flat across settings: one-sided variants (pull-only, push-only) and overly aggressive bias both fall well below the default, spanning roughly nine points. We therefore fix $(\alpha, \beta) = (5, 2)$ globally across all models and benchmarks rather than tuning per task, and treat the bias strength as a parameter that must be kept moderate rather than one that can be pushed for larger gains.

Table 5: **Sensitivity of SAGE to logit-level intervention strength (α, β).** We report MMVP (%).

Variant	(α, β)	MMVP (%)
No re-weighting	$(0, 0)$	61.48
Pull-only	$(0, 2)$	54.07
Push-only	$(5, 0)$	57.49
Weak	$(3, 1)$	61.91
Default	$(5, 2)$	**62.96**
Strong	$(7, 3)$	59.39
Aggressive	$(9, 4)$	55.93

When reporting sensitivity, we recommend fixing all other settings (ROI source, layer set, decoding parameters) and varying only one parameter at a time.

B Additional Layer-wise Analyses

B.1 Per-benchmark Single-layer Sweep Curves

Figures 9–11 provide per-benchmark single-layer sweep curves for LLaVA-1.5 using local ROI masks derived from CLIP, ViT, and DINOv3, respectively. For each decoder layer index ℓ, we apply SAGE to a single layer $\mathcal{L}_{\text{int}} = \{\ell\}$ and report the resulting benchmark score using the official evaluation scripts. Across sources and tasks, the response to intervention is clearly layer-dependent: some layers yield consistent gains over the no-intervention baseline, while others are neutral or occasionally degrade performance. These curves complement the aggregated summaries in the main paper and show that the effectiveness of steering is not tied to a single local ROI source.

B.2 Exact Best-Layer Indices for Main Results

For transparency, we list in Table 6 the exact single-layer indices used to produce the **Best** results in Table 1. These indices are provided as supplementary detail; in the main text, cross-model discus-

sion uses normalized decoder depth rather than raw layer number.

B.3 Additional Multi-layer Set Comparisons

To complement the group-level summaries in Table 2, we provide a small number of illustrative exact layer combinations for the multi-layer setting. Rather than exhaustively listing every tested set, we report selected examples that reflect the stronger non-early compositions and a few failure-prone early-layer cases.

Tables 7–9 summarize illustrative 2-layer and 3-layer configurations, together with selected degradation cases. Across these examples, sets concentrated on mid and/or final depths tend to produce stronger performance than sets involving the earliest layers. These examples are consistent with the depth sensitivity observed in the main paper and help ground the group-level averages with concrete layer indices.

B.4 Layer Skip Ablation

To further probe depth-wise roles of the decoder, we perform a layer-skip ablation on LLaVA-1.5 that removes computation from selected decoder layers at inference time. For a chosen skip set $\mathcal{L}_{\text{skip}}$, we bypass each layer $\ell \in \mathcal{L}_{\text{skip}}$ by forwarding its input hidden states to the next layer without applying the layer's transformation. This ablation adds no new parameters and isolates how sensitive the model's final behavior is to perturbations at different decoder depths.

Figure 6 reports the impact of skipping different layers on downstream performance. Skipping is not uniformly tolerated across depth: certain layers lead to more noticeable degradation than others. This provides complementary evidence to the single-layer intervention sweep in the main paper and supports the interpretation that some early decoder layers can be more sensitive to perturbations.

C ROI Sources from Standard Vision Backbones

SAGE only requires a token-aligned binary mask $\mathbf{m} \in \{0,1\}^N$ over visual tokens. In this section, we summarize the three ROI sources used in our study: CLIP, ViT, and DINOv3. These sources produce different patch-level relevance patterns, but the intervention in Eq. (1) is identical across sources; only the procedure that produces $\mathbf{m}$ changes.

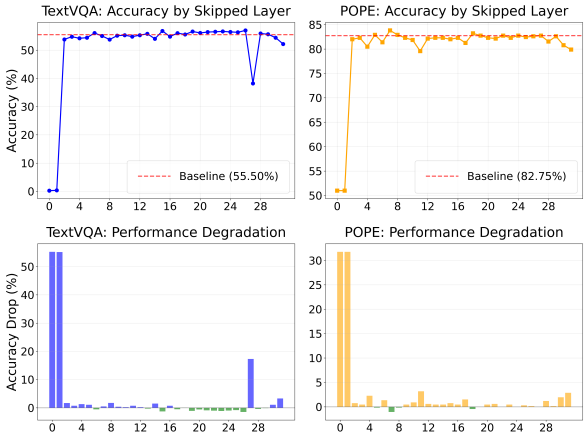


Figure 6: **Layer-skip ablation on LLaVA-1.5.** For each decoder layer, we skip only that layer at inference and evaluate performance. The top row reports TextVQA and POPE accuracy versus the skipped depth, and the bottom row shows the change relative to the no-skip baseline.

D Additional Controlled Comparisons

This appendix collects two comparisons that share the fixed-protocol setting of Section 4.5 but are not needed to follow the main argument. Both use no benchmark-specific selection of the intervention layer.

D.1 Comparison with an Inference-Time Decoding Baseline

Section 4.5 compares the fixed-layer setting against a re-measured baseline and a matched random-ROI control, but not against another published intervention. To place SAGE relative to existing inference-time methods, we run visual contrastive decoding (VCD (Leng et al., 2024)) under exactly the same protocol with its official hyperparameters ($\alpha = 1.0$, $\beta = 0.1$, noise step 500). VCD reaches 50.8 on TextVQA and 59.5 on GQA, and is neutral-to-negative on POPE (-0.2) and ScienceQA (-0.3). Fixed-layer SAGE reaches 54.6 and 62.3 on the same two benchmarks (Table 3). The two methods therefore differ not only in magnitude on the localized-evidence benchmarks but also in whether they are neutral elsewhere. This comparison covers a single method on a single model family, and we do not read it as a general ranking of inference-time interventions.

D.2 Fixed-Protocol Control Across Model Families

The controls in Section 4.5 are established on LLaVA-1.5. To check that the fixed-protocol set-

Table 6: **Exact model-specific best-layer indices used for the Best results in Table 1.** Indices are raw layer numbers within each model and should not be compared directly across architectures; cross-model discussion in the main text uses normalized decoder depth.

Model	TextVQA	POPE	GQA	SQA	MME	MMVet	MMVP Img	MMVP Pair	MMVP-O Img	MMVP-O Pair
LLaVA 1.5-13B										
+ SAGE (CLIP)	L17	L3	L20	L18	L24	L20	L26	L26	L18	L6
+ SAGE (ViT)	L29	L10	L19	L20	L22	L18	L18	L18	L2	L18
+ SAGE (DINOv3)	L30	L10	L30	L25	L16	L20	L26	L26	L18	L18
Qwen2-VL-7B										
+ SAGE (CLIP)	L11	L18	L21	L15	L25	L14	L18	L18	L15	L15
+ SAGE (ViT)	L13	L18	L15	L15	L20	L17	L27	L27	L15	L15
+ SAGE (DINOv3)	L16	L15	L18	L18	L27	L14	L18	L18	L12	L12
InternVL3-2B										
+ SAGE (CLIP)	L11	L18	L21	L15	L25	L25	L18	L18	L8	L8
+ SAGE (ViT)	L15	L18	L14	L19	L24	L17	L27	L27	L16	L16
+ SAGE (DINOv3)	L9	L18	L27	L19	L27	L15	L21	L21	L13	L13
NVILA-8B										
+ SAGE (CLIP)	L19	L20	L12	L15	L15	L10	L10	L10	L20	L20
+ SAGE (ViT)	L13	L20	L18	L17	L16	L12	L15	L15	L13	L13
+ SAGE (DINOv3)	L10	L10	L15	L18	L5	L21	L15	L15	L20	L20

Table 7: **Illustrative 2-layer SAGE sets on TextVQA (LLaVA-1.5).** These examples are selected to represent stronger non-early depth compositions.

Layer set	Depth composition	Accuracy (%)
(14,16)	Mid–mid	62.43
(17,21)	Mid–final	62.72
(28,31)	Final–final	62.58

Table 8: **Illustrative 3-layer SAGE sets on TextVQA (LLaVA-1.5).** These examples are selected to represent stronger multi-layer compositions away from the earliest decoder layers.

Layer set	Depth composition	Accuracy (%)
(14,18,21)	Mid $\times 3$	62.33
(17,24,30)	Mid–final–final	62.63
(24,27,30)	Final $\times 3$	62.53

Table 9: **Illustrative failure-prone early-layer sets on TextVQA (LLaVA-1.5).** These examples show that sets concentrated near the earliest decoder layers can still lead to substantial degradation.

Layer set	Depth composition	Accuracy (%)
(0,2)	Very-early only	12.07
(0,2,4)	Very-early $\times 3$	7.12
(1,2,3)	Very-early $\times 3$	1.12

ting is not specific to that model, we repeat it across four families (LLaVA-1.5, Qwen2-VL, InternVL3, NVILA) using one fixed target layer per family, fixed $(\alpha, \beta) = (5, 2)$, a fixed top-48 mask, and no benchmark-specific selection at any point. Under this setting the same push–pull intervention yields moderate gains: GQA $+2.55$, TextVQA $+1.58$, and MMVP-pair $+6.67$. We report them as evidence that the intervention is not confined to one architecture, not as a like-for-like comparison with the main table.

D.3 ROI Sources

Backbone-derived ROI (CLIP/ViT/DINOv3). We derive ROI masks from standard pretrained vision backbones without explicit detector-based grounding. For each source, we compute patch-level relevance scores and threshold them to obtain a binary patch mask, which is then aligned to the visual-token grid used by the evaluated VLM. Although the resulting masks can differ in sharpness and spatial coverage across sources, all three provide usable local guidance signals for SAGE.

D.4 Single-layer Sweep with Backbone-Derived ROI Masks

Figures 9–11 report additional single-layer sweep results using ROI masks derived from CLIP, ViT, and DINOv3. These results further support that SAGE remains effective under multiple local ROI construction mechanisms.

D.5 Qualitative Comparison of ROI Masks Across Sources

Figure 8 visualizes ROIs produced by different local ROI sources for the same image–query pairs. CLIP, ViT, and DINOv3 can produce different mask shapes and degrees of spatial concentration, reflecting differences in their patch-level relevance patterns. Despite these differences, all three sources can be converted into token-aligned masks

and used by SAGE without changing the intervention rule.

E Mechanistic Analysis Details

This appendix specifies the exact definitions and computation procedures used in our mechanistic analyses (sink structure, spectral dominance, and attention-mass redistribution).

E.1 Constructing the Logit Matrix for Spectral Analysis

Fix a decoder layer ℓ and head h. For each evaluation instance, we collect pre-softmax attention logits from text-query positions to visual-token keys. Let $\mathbf{a}_{p,t}^{(\ell,h)} \in \mathbb{R}^N$ denote the restricted logit vector over the N visual-token keys at decoding step t for prompt (instance) p. We stack these vectors across instances and decoding steps into a matrix

$$\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{M\times N}, \qquad \mathbf{A}^{(\ell,h)}[r,:] = \mathbf{a}_{p,t}^{(\ell,h)}, \quad (7)$$

where r indexes the pair (p,t) and M is the total number of stacked vectors. We then compute $\mathbf{A}^{(\ell,h)} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and report spectral concentration using

$$\rho_k(\mathbf{A}) = \frac{\sum_{j=1}^{k}\sigma_j^2}{\sum_{j=1}^{\min(M,N)}\sigma_j^2}, \quad (8)$$

where $\{\sigma_j\}$ are singular values.

E.2 Prompt-Invariant Sink Definition and Sink IoU

For each instance p and decoding step t, we convert the visual-token attention weights into a set of sink tokens by selecting the top-$r\%$ tokens by attention mass. We use $r = 5$ throughout unless otherwise specified. Given two sink sets S_p and $S_{p'}$ computed from different prompts (or different queries) on the same image, we measure prompt-invariance using the intersection-over-union (IoU):

$$\mathrm{IoU}(S_p, S_{p'}) = \frac{|S_p \cap S_{p'}|}{|S_p \cup S_{p'}|}. \quad (9)$$

We report mean IoU with standard deviation across prompt pairs.

E.3 Attention-Mass Metrics: Sink Mass and ROI Mass

Let $\mathbf{p} \in \mathbb{R}^N$ be attention weights over visual keys (after softmax) and let $\mathbf{m} \in \{0,1\}^N$ be the ROI mask. We define ROI mass as the total attention assigned to ROI tokens:

$$\mathrm{Mass}_{\mathrm{ROI}}(\mathbf{p}) = \sum_{i=1}^{N}\mathbf{p}[i]\cdot\mathbf{m}[i]. \quad (10)$$

We define sink mass as the total attention assigned to the sink-token set S (top-$r\%$ tokens by attention mass):

$$\mathrm{Mass}_{\mathrm{sink}}(\mathbf{p}) = \sum_{i\in S}\mathbf{p}[i]. \quad (11)$$

Unless stated otherwise, we use the same $r = 5$ as in Appendix E.2.

E.4 Implementation Notes for Mechanistic Measurements

All mechanistic measurements are computed under matched decoding settings between baseline and SAGE runs. When reporting layer-wise results, we either fix a representative head or average the reported quantities across heads within the layer; we explicitly state the choice in each figure caption.

E.5 Evidence Retention to the Final Layer

To test whether the steering effect survives to late decoding rather than being a single-layer artifact, we measure a retention ratio: final-layer ROI attention mass relative to its mid-layer peak. On LLaVA-1.5, averaged over four benchmarks, guided steering raises this ratio from 0.77 to 0.96, indicating reduced mid-to-final evidence loss and a change in routing rather than a local change in one layer's heatmap.

F Qualitative Examples

We provide qualitative examples to illustrate how SAGE changes the model's visual evidence usage during generation. Figure 7 compares baseline predictions with SAGE under the same input image and query, together with attention visualizations over the image. Across examples, baseline models can allocate attention to visually salient but query-irrelevant regions, which can lead to weak grounding in fine-grained questions. In contrast, with the same ROI guidance, SAGE shifts attention toward query-relevant regions and encourages the model to rely on localized visual evidence when forming its answer.

In Figure 7, SAGE is particularly helpful on queries that require fine-grained visual evidence. For instance, when asked to read the text on a distant sign, the baseline answer can be driven by a

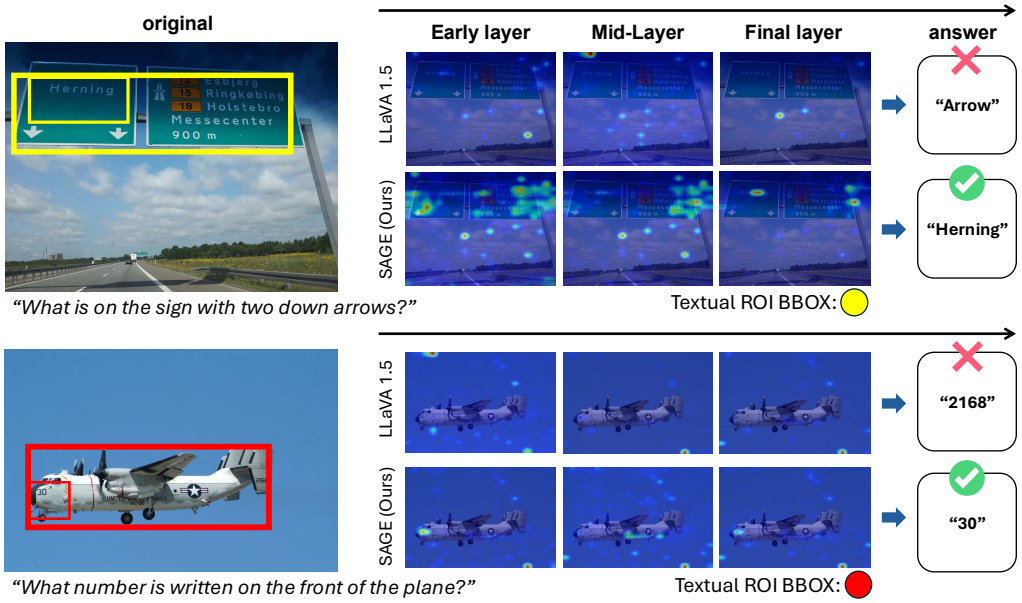


Figure 7: **Qualitative comparison of baseline vs. SAGE.** For each example, we show the input image and query, baseline and SAGE predictions, and attention visualizations. SAGE encourages attention to concentrate on query-relevant visual evidence, improving grounding in fine-grained cases.

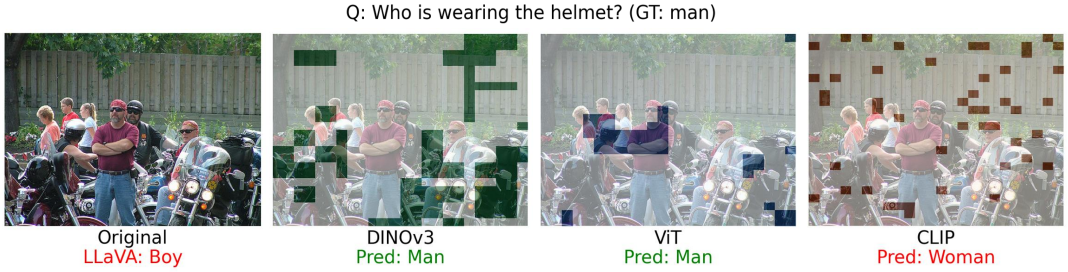
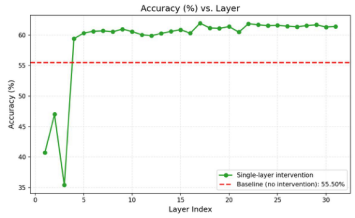


Figure 8: **Failure case under unstable local ROI estimates.** Different local ROI sources can produce fragmented or partially misaligned masks, which may steer attention away from the true evidence and lead SAGE to fail.

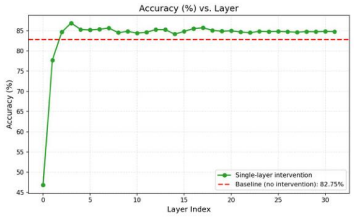
nearby salient cue and miss the actual characters, while SAGE increases attention on the sign region and enables the correct reading. Similarly, for the aircraft example, focusing on the front fuselage is necessary to identify the printed number; SAGE concentrates attention on this localized area and corrects the prediction. These cases suggest that improving ROI-focused attention can be beneficial for tasks such as reading small text or verifying localized details.

F.1 Failure Case: Sensitivity to Unstable Local ROI Estimates

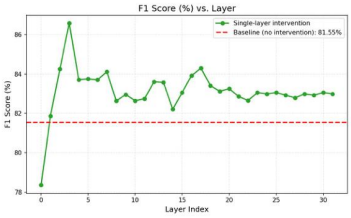
Figure 8 shows a representative failure case where SAGE does not correct the answer when the local ROI estimate is unstable. Since SAGE intervenes only on the region specified by the ROI source, inaccurate or fragmented local masks may steer attention toward irrelevant visual tokens and suppress the true evidence. As a result, the model can still produce an incorrect prediction even though logit-level steering is applied.



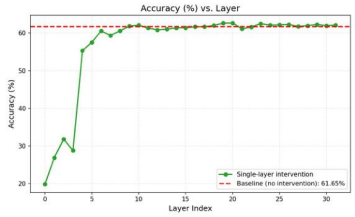
TextVQA



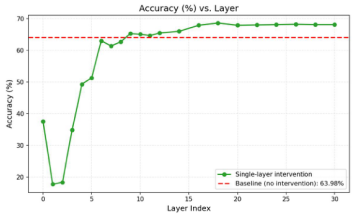
POPE



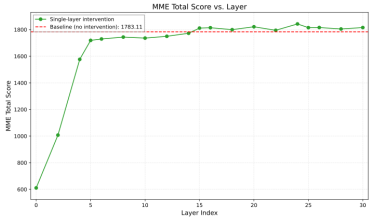
POPE



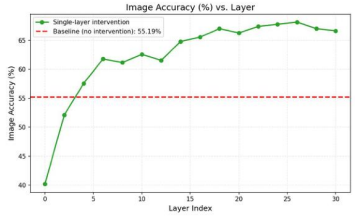
GQA



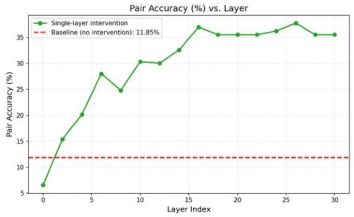
ScienceQA



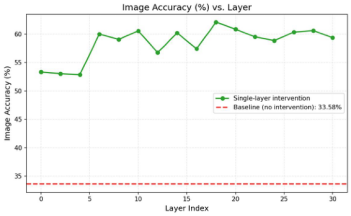
MME



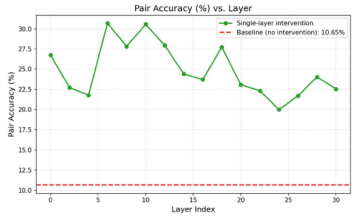
MMVP (image)



MMVP (pair)



MMVP Original (image)



MMVP Original (pair)

Figure 9: **Single-layer sweep with a CLIP-derived ROI mask.** We report per-benchmark performance as a function of the intervened layer using a CLIP-based patch-level ROI mask.

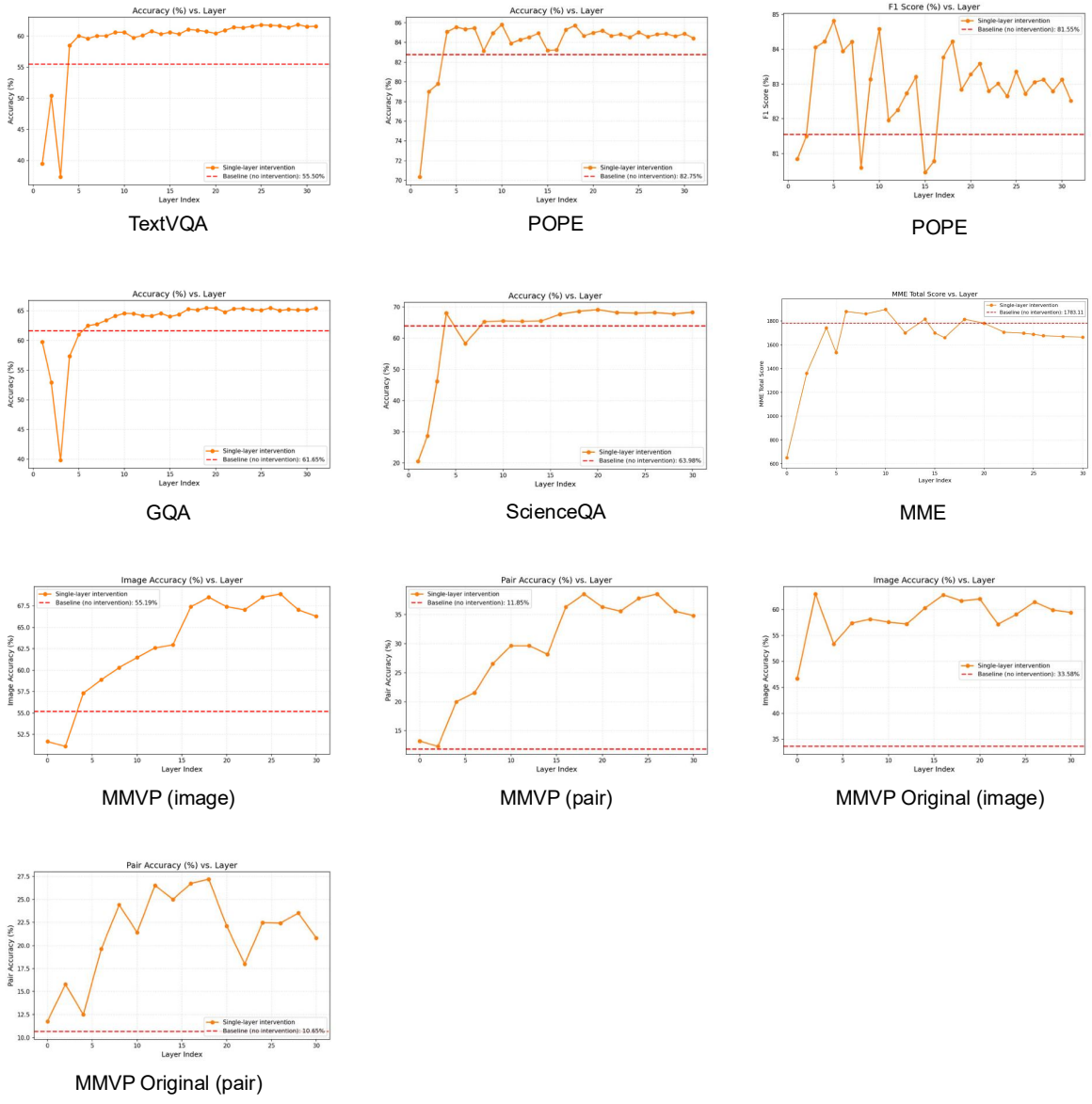


Figure 10: **Single-layer sweep with a ViT-derived ROI mask.** We report per-benchmark performance as a function of the intervened layer using a ViT-based patch-level ROI mask.

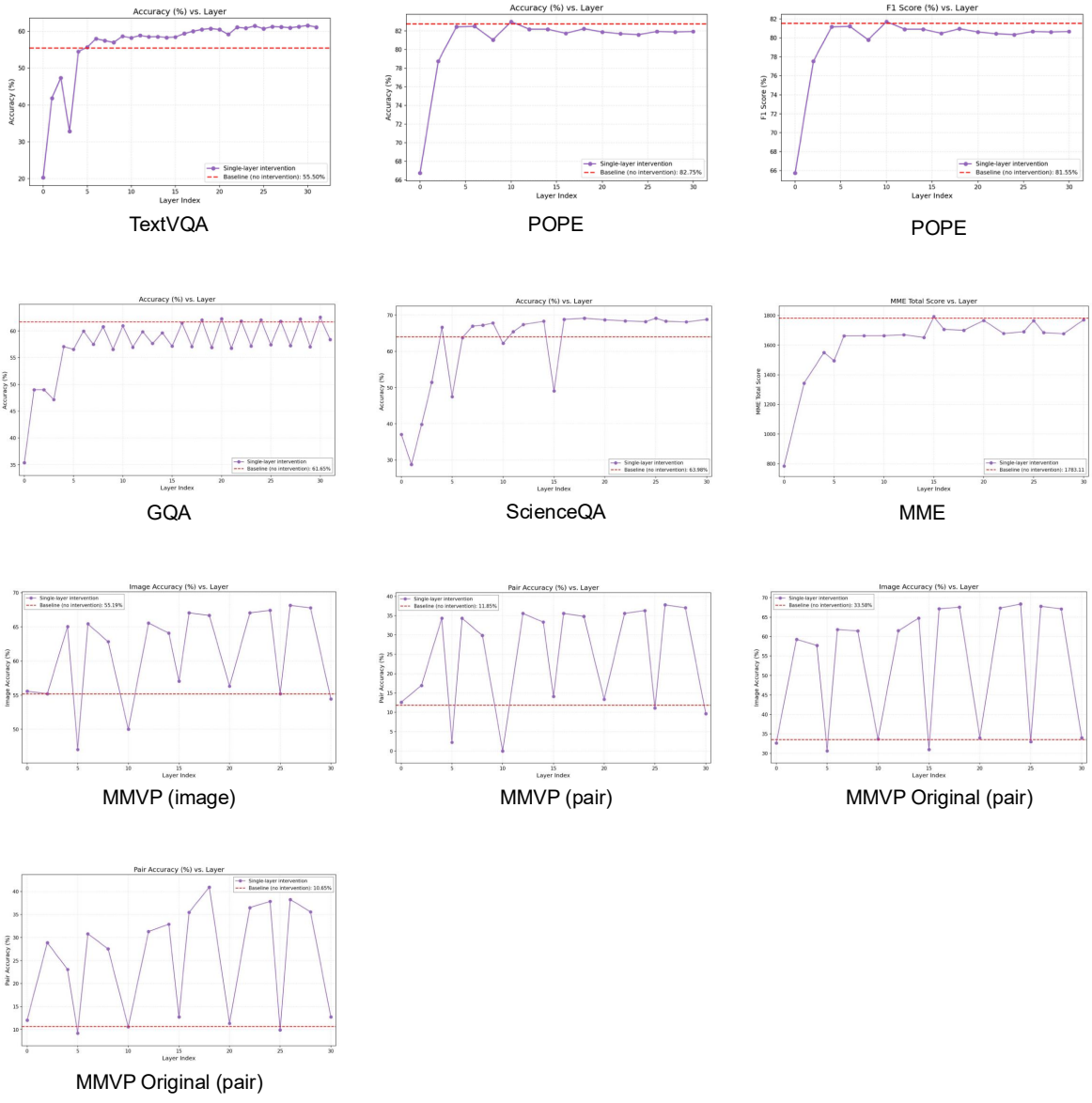


Figure 11: **Single-layer sweep with a DINOv3-derived ROI mask.** We report per-benchmark performance as a function of the intervened layer using a DINOv3-based patch-level ROI mask.