SAPS-CoT: Step-Aware Visual Grounding with Dynamic Residual Injection for Interpretable Multimodal Reasoning

Anonymous authors
Paper under double-blind review

Abstract

While Large Vision-Language Models (LVLMs) enhanced by Chain-of-Thought (CoT) prompting have significantly advanced multimodal reasoning, they predominantly suffer from a "visual black box" dilemma and a "perceptual lock-in effect." Existing paradigms typically extract visual features once and freeze them, forcing the model to rely heavily on language priors during intermediate reasoning steps. This often leads to visual hallucinations and cascading errors. To address this, we propose SAPS-CoT, a two-stage reasoning framework that seamlessly integrates Step-Aware Patch Selection (SAPS) with a Dynamic Residual Injection mechanism. In the discovery phase, SAPS establishes a fine-grained, explicit alignment between textual reasoning steps and visual regions by leveraging multi-dimensional metrics, including semantic consistency, gradient saliency, and location bias. In the refinement phase, we introduce an in-place latent residual injection mechanism as an alternative to token concatenation. It dynamically updates the model's visual memory while preserving the original visual-token sequence length. Controlled comparisons show modest but consistent accuracy gains over concatenation, with a small implementation-dependent latency premium. Extensive experiments on complex real-world scenarios demonstrate that SAPS-CoT mitigates visual hallucinations and outperforms state-of-the-art multimodal CoT methods in both accuracy and interpretability.

1 Introduction

In recent years, the intersection of Computer Vision (CV) and Natural Language Processing (NLP) has witnessed remarkable progress. Visual Question Answering (VQA), a core task for evaluating machine intelligence in multimodal understanding, has achieved a qualitative leap in performance, driven by the explosive growth of Large Vision-Language Models (LVLMs) Liu et al. (2023; 2024a); Bai et al. (2023); Li et al. (2023a). To address complex spatial, logical, and compositional reasoning tasks that cannot be solved by a single forward pass, researchers have successfully introduced the Chain-of-Thought (CoT) paradigm into the multimodal domain Wei et al. (2022); Zhang et al. (2024c); Lu et al. (2022). By guiding models to decompose intricate problems into sequential, intermediate reasoning steps, multimodal CoT has significantly enhanced the logical coherence of model outputs. Yet these methods suffer from a fundamental disconnect between visual perception and textual reasoning, manifesting as a coupled **visual black box dilemma** and **perceptual lock-in effect**. Stemming from the conventional "one-pass encoding" strategy, visual tokens are statically extracted and frozen initially, causing attention to them to decay as the reasoning chain lengthens. Lacking a mechanism to dynamically re-focus on the image based on newly derived textual clues, the model becomes "locked" into its initial perception, forcing it to rely on language priors rather than verifiable visual evidence – frequently triggering object hallucinations and rendering reasoning uninterpretable.

Recent interventions to alleviate these limitations generally fall into three categories, each with critical trade-offs. Prompting-based methods (e.g., DDCoT Zheng et al. (2023), CCoT Mitra et al. (2023)) attempt to decouple reasoning duties but remain "text-bottlenecked," losing fine-grained visual cues. Visual prompting (e.g., SCAFFOLD Lei et al. (2025)) introduces unnatural distribution shifts via artificial coordinate overlays. Meanwhile, retrieval or supervised grounding methods

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

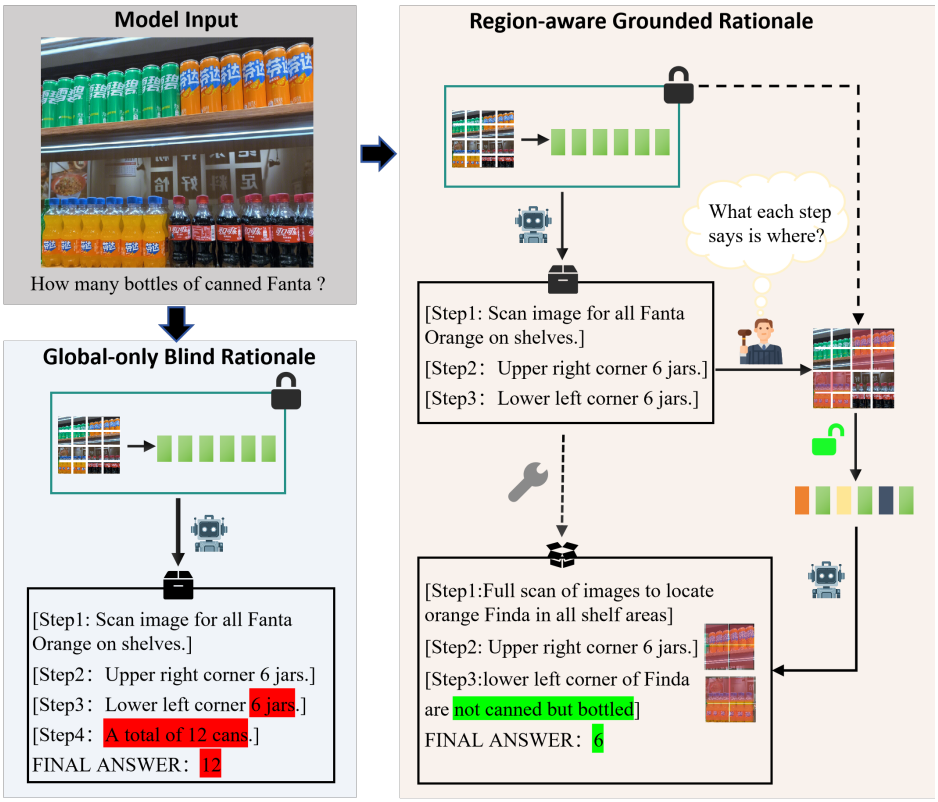


Figure 1: Comparison of implicit multimodal CoT (Left) and our grounded visual CoT (Right). Unlike baseline "black-box" reasoning, our paradigm explicitly aligns each reasoning step with its corresponding visual patches, effectively mitigating hallucinations and ensuring faithful grounding.

(e.g., ICoT Gao et al. (2025), Visual CoT Shao et al. (2024)) either concatenate selected patches—increasing the visual-token sequence and its theoretical self-attention cost—or require expensive manual annotations and full-model fine-tuning, severely restricting their plug-and-play usability.

Drawing inspiration from the saccadic eye movement mechanism in human cognition—where attention iteratively refocuses on local, high-frequency regions as logical deduction unfolds Yao et al. (2023)—we propose **SAPS-CoT**, a novel framework featuring explicit visual grounding and dynamic feature rewriting. Our approach effectively mitigates the lock-in effect by ensuring that each step of the reasoning chain is not merely a linguistic derivation, but is tightly coupled with physical visual evidence.

In summary, the main contributions of this paper are threefold. First, to resolve the visual black box issue, we introduce Step-Aware Visual Grounding (SAPS), a multi-dimensional scoring strategy combining semantic consistency, gradient saliency, and geometric location to achieve fine-grained alignment between textual reasoning and visual patches. Second, distinct from existing retrieval-based methods that rely on token concatenation, we propose Dynamic Residual Injection. This in-place enhancement mechanism utilizes forward hooks to inject visual evidence directly into the feature space via attention-based residual fusion, avoiding growth of the visual-token sequence and KV cache. Finally, we construct a Refined Real-World Benchmark tailored for complex interactions. Comprehensive experiments demonstrate that SAPS-CoT consistently outperforms state-of-the-art baselines, offering robust hallucination suppression and pixel-level interpretability.

2 Related Work

2.1 Large Vision-Language Models (LVLMs)

The advent of LVLMs, such as LLaVA Liu et al. (2023; 2024a), InstructBLIP Dai et al. (2023), and Qwen-VL Bai et al. (2023), has transformed multimodal understanding. Built predominantly upon the Perceiver-LLM paradigm Alayrac et al. (2022), these architectures bridge a pre-trained visual encoder with an LLM via an alignment adapter Li et al. (2023a). While excelling in generic visual perception, they inherently rely on a "One-pass Encoding" strategy. This static global feature representation struggles to dynamically adapt during complex, multi-step reasoning, rendering models susceptible to visual hallucinations Li et al. (2023b); Fang et al. (2026); Zhao et al. (2025) and underscoring the need for dynamic visual grounding.

2.2 Multimodal Chain-of-Thought and Visual Grounding

To bridge the logical reasoning gap in LVLMs, the Multimodal Chain-of-Thought paradigm has emerged Wei et al. (2022). Early methods like DDCoT Zheng et al. (2023) and CCoT Mitra et al. (2023) decompose reasoning duties or build scene graphs, while Cognitive CoT (CoCoT) Park et al. (2026) structures reasoning across perception and situation stages. Despite these advances, they remain text-bottlenecked, discarding fine-grained visual cues by compressing signals into discrete text. To overcome this, recent studies pursue explicit visual alignment. SCAFFOLD Lei et al. (2025) overlays coordinate matrices to bind text with spatial anchors, yet alters the image distribution. RSVP Lu et al. (2025) instead employs visual prompting to achieve reasoning-driven region localization, but still relies on external visual markers. These prompting-based methods introduce unnatural visual artifacts.

A more direct line of work retrieves and re-injects visual evidence. ICOT Gao et al. (2025) pioneers interleaved text-image reasoning via Attention-driven Selection (ADS), yet relies on coarse, noisy single-token attention. VoCoT Li et al. (2025) grounds multi-step reasoning through object-centric visual features, but requires explicit object detection modules. On the data-driven front, Visual CoT Shao et al. (2024) constructs bounding-box-annotated datasets for region cropping and re-encoding, and Zebra-CoT Li et al. (2026) further scales this paradigm to 182K interleaved reasoning trajectories across 18 domains. However, these approaches rely on costly manual labeling and full model fine-tuning, precluding plug-and-play deployment.

In contrast, our proposed **Step-Aware Patch Selection (SAPS)** significantly advances this line of research. Without requiring any manual annotation or fine-tuning, SAPS achieves adaptive, step-aware visual grounding by mining gradient signals and semantic alignment within the frozen model. Instead of relying on noisy single-token attention or costly labeled data, SAPS constructs a rigorous multi-dimensional scoring matrix. By integrating explicit cross-modal semantic consistency, Gradient Saliency (to trace the exact causal regions minimizing the generation loss), geometric location bias, and multi-scale window compensation, SAPS mines the model's internal causal attention flow, enabling a much more robust, accurate, and highly interpretable visual grounding process.

2.3 Token Concatenation and Sequence-Length Growth

Existing retrieval-augmented visual models, including ICoT, predominantly adopt token concatenation, in which retrieved visual patches are appended to the input sequence. If K patches are added to an original sequence of length N, the sequence grows to $N + K$; since standard Transformer self-attention scales quadratically with sequence length Vaswani et al. (2017), this increases the theoretical attention and KV-cache costs. Excessive visual tokens may also dilute attention density Li et al. (2024); Huang et al. (2024); Wang et al. (2024). Nevertheless, the practical wall-clock effect depends on K, the original context length, implementation, and hardware. We therefore evaluate both predictive accuracy and measured latency rather than inferring practical efficiency solely from asymptotic complexity. Prior work on hidden-state modulation Zhang et al. (2024b; 2023) motivates our constant-length residual injection design.

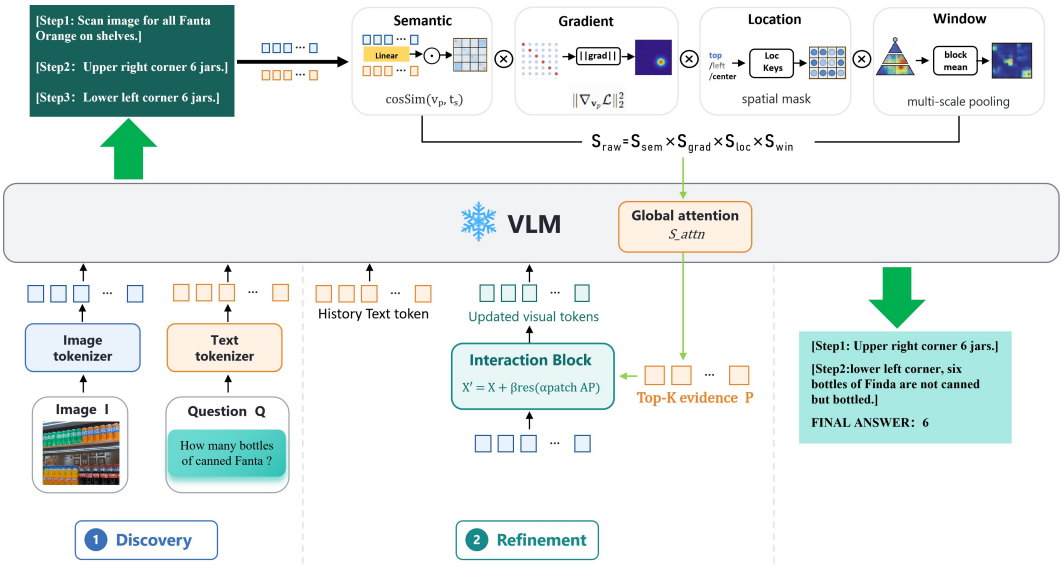


Figure 2: Overview of SAPS-CoT. During discovery, SAPS selects step-specific visual evidence using semantic, gradient, spatial, and multi-scale scores modulated by global attention. During refinement, Dynamic Residual Injection updates visual features for second-pass reasoning without increasing sequence length. Model parameters remain frozen; scoring details are provided in Section 3.2.

3 METHODOLOGY

To mitigate the "visual black box" phenomenon and the "Lock-in Effect" prevalent in current Multimodal CoT methods, we propose a two-stage reasoning framework based on Step-Aware Patch Selection (SAPS), as illustrated in Figure 2. The process initiates with **Stage 1 (Discovery Phase)**, where the model generates an initial reasoning chain. Leveraging the SAPS mechanism, we precisely localize the key visual patches corresponding to each reasoning step via a multi-dimensional scoring strategy, thereby establishing an explicit mapping between textual reasoning and visual evidence. Subsequently, in **Stage 2 (Refinement Phase)**, we abandon traditional token concatenation in favor of a Dynamic Residual Injection mechanism. By implementing a forward hook within the visual encoder, the identified visual evidence is injected into the original feature space in a residual manner, guiding the model to perform a second-pass reasoning with enhanced focus on critical regions.

3.1 PRELIMINARIES

Standard Vision-Language Models (VLMs) typically consist of a visual encoder E and a generative LLM Liu et al. (2023); Li et al. (2023a). Given an image I and a question Q, the model generates the final answer A: $A = \text{VLM}(I, Q)$. Multimodal CoT further introduces intermediate reasoning steps $R = \{r_1, r_2, \dots\}$ to guide generation Zhang et al. (2024c), i.e., $P(A|I, Q, R)$. However, in existing pipelines, visual features are input only in the initial stage. Subsequent reasoning r_t often relies solely on r_{t-1} and language priors, causing visual information to decay with reasoning depth. Our goal is to establish a dynamic association $P(r_t|I_{patch}, r_{t-1})$ between r_t and local image regions I_{patch}.

3.2 STAGE 1: STEP-AWARE PATCH SELECTION (SAPS)

Previous visual grounding attempts often rely on single global features or coarse attention maps, which fail to capture the semantic focal shifts during complex logical deduction. To bridge this gap, we design SAPS, which constructs a multi-dimensional criteria matrix to localize the optimal visual patches for reasoning step t. For a candidate patch p, we compute a Raw Relevance Score:

$$Score_{raw}(p, t) = S_{sem} \cdot N \cdot S_{grad} \cdot S_{loc} \cdot S_{win} \tag{1}$$

where N is a scaling normalization factor. The specific scoring dimensions are formulated as follows:

Semantic Consistency (S_{sem}): We compute the cosine similarity between the patch visual embedding $\mathbf{v}_p$ and the step-specific text embedding $\mathbf{t}_s$. To ensure distribution stability across diverse samples, we apply Z-score normalization followed by a Sigmoid activation:

$$S_{sem} = \sigma\left(\frac{(\mathbf{v}_p \cdot \mathbf{t}_s) - \mu_{cos}}{\sigma_{cos}}\right) \tag{2}$$

Gradient Saliency (S_{grad}): While semantic matching is intuitive, it often ignores the underlying causal relationships in the model's decision-making. We backpropagate the text generation loss $\mathcal{L}$ of step t to the frozen visual features and calculate the L_2 norm of the gradients. This isolates regions that strictly minimize the generation uncertainty:

$$S_{grad} = \sigma\left(\frac{\|\nabla_{\mathbf{v}_p}\mathcal{L}\|_2^2 - \mu_{grad}}{\sigma_{grad}}\right) \tag{3}$$

Location Bias (S_{loc}): Spatial geometry plays a pivotal role in visual reasoning Zhang et al. (2024a). We parse directional keywords (e.g., "top-left", "center") in r_t. If matched, regions are weighted via a predefined spatial mask $\text{Map}_{spatial}$; otherwise, a general edge-based structural bias is applied, scaled by α_{loc}:

$$S_{loc} = \text{Map}_{spatial}(p, \text{keywords}) \cdot \alpha_{loc} \tag{4}$$

Multi-scale Window Compensation (S_{win}): To overcome the restricted receptive field of isolated patches and capture complete object-level semantics, we aggregate the scores of adjacent patches using sliding windows across multiple scales ws:

$$S_{win} = 1.0 + \alpha \cdot \left(\sum_{w \in ws} \frac{\text{Pool}(\mathbf{v}_p \cdot \mathbf{t}_s, w)}{w^2} - \mu_{win}\right) \tag{5}$$

Finally, this robust local score is aligned with the global intent of the LLM decoder via a global cross-modal attention score (S_{attn}). The refined metric is used to select the Top-K key visual tokens $P \in \mathbb{R}^{K \times D}$ serving as the verified visual evidence.

$$Score_{final}(p,t) = Score_{raw}(p,t) \cdot (1 + \gamma \cdot S_{attn}) \tag{6}$$

3.3 Stage 2: Dynamic Residual Injection

Identifying critical evidence solves only half the problem; integrating it without extending the visual-token sequence is equally important. Traditional retrieval methods concatenate selected patches, increasing the sequence length processed by self-attention. As a constant-length alternative, we propose a **Dynamic Residual Injection** mechanism leveraging PyTorch-style forward hooks.

Specifically, we utilize a hook function to dynamically intercept the hidden states during the forward pass. Let $X \in \mathbb{R}^{N \times D}$ be the original visual tokens and $P \in \mathbb{R}^{K \times D}$ be the selected SAPS patches ($K \ll N$). Rather than appending P to the sequence, we execute an attention-based residual fusion logic:

Feature Normalization: Both token sets are normalized to a hypersphere to stabilize the variance during similarity computation: $\hat{X} = \text{LN}(X), \hat{P} = \text{LN}(P)$.

Cross-Modal Attention Alignment: We compute a spatial affinity matrix $A \in \mathbb{R}^{N \times K}$ by measuring the dot product between the original sequence and the selected patches, scaled by the feature dimension D:

$$A_{i,j} = \frac{\exp(\langle \hat{X}_i, \hat{P}_j \rangle / \sqrt{D})}{\sum_{k=1}^{K} \exp(\langle \hat{X}_i, \hat{P}_k \rangle / \sqrt{D})} \tag{7}$$

This softmax-normalized distribution ensures that each original token X_i dynamically routes the most relevant information from the SAPS pool back to its proper geometric location.

Gated Latent Fusion: The fused residual $\mathcal{R} = A \cdot P$ is subsequently injected back into the original representations. To balance the preservation of the original global visual semantics and the injection

of step-aware local evidence, we decouple the fusion process into two gating factors: α_{patch} (patch gate) scaling the retrieved evidence, and β_{res} (residual gate) controlling the injection intensity:

$$X_i' = X_i + \beta_{res} \cdot (\alpha_{patch} \cdot \mathcal{R}_i) \tag{8}$$

In our implementation, these factors can be either learned or empirically set as hyperparameters (detailed sensitivity analysis is provided in Appendix A.3). By utilizing the forward hook to perform this "in-place" enhancement, our mechanism explicitly rewrites the visual memory to focus on step-specific evidence while preserving the original visual-token sequence length ($\Delta L = 0$). The mechanism facilitates an "attention refocusing" process during second-stage reasoning: selected SAPS evidence is routed back to the original visual representations, helping the model rectify perceptual biases or descriptive inaccuracies incurred in the discovery phase. This feature-space fusion itself introduces additional computation, whose measured latency is evaluated in Section 4.3.

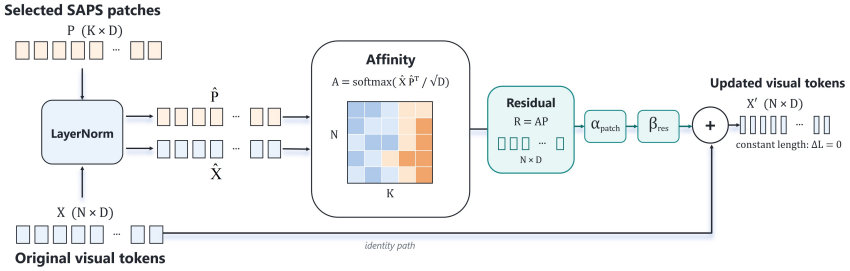


Figure 3: Dynamic Residual Injection. Original visual tokens $X \in \mathbb{R}^{N \times D}$ and selected SAPS patches $P \in \mathbb{R}^{K \times D}$ are separately layer-normalized to compute the affinity matrix $A = \mathrm{softmax}(\hat{X}\hat{P}^\top/\sqrt{D})$, with softmax applied over the K selected patches. The residual $\mathcal{R} = AP$ aggregates the original selected patch features and is scaled by the patch gate α_{patch} and residual gate β_{res} before being added to X. The resulting tokens X' retain the original shape $N \times D$, with no increase in sequence length ($\Delta L = 0$).

4 Experiments

We evaluate SAPS-CoT on four multimodal benchmarks using two frozen backbones, with controlled ablations and qualitative analyses of visual grounding and reasoning errors.

4.1 Experimental Setup

Datasets and Metrics. Existing benchmarks often feature iconic images with simple backgrounds, which are insufficient to evaluate complex reasoning in open-world scenarios. To comprehensively assess our method, we select four diverse datasets: **MathVista** Lu et al. (2024), **M^{3}CoT** Chen et al. (2024), **GQA** Hudson & Manning (2019), and **ScienceQA** Saikh et al. (2022). To better evaluate practical visual grounding and spatial reasoning capabilities, we construct a **Refined Real-World Subset** from M^{3}CoT. By filtering out abstract mathematical problems and chart-based data, we retain exactly 1,500 samples featuring complex physical scenarios, multi-object interactions, and severe occlusions. We report **Accuracy** using the same dataset-specific answer-matching rule for all compared methods and fusion variants.

Baselines. We implement SAPS-CoT on two cutting-edge backbones: **LLaVA-NeXT-8B** Liu et al. (2024b) and **Qwen2-VL-7B** Bai et al. (2023). We compare against two categories of methods: (1) **Base LVLMs (No-CoT):** Direct zero-shot inference. (2) **Plug-and-Play CoT Paradigms:** Multimodal-CoT Zhang et al. (2024c), CCoT Mitra et al. (2023), DDCoT Zheng et al. (2023), and SCAFFOLD Lei et al. (2025). Note that the recent ICoT Gao et al. (2025) framework is excluded from direct empirical comparison as it requires extensive architectural modifications and domain-specific fine-tuning, violating the training-free premise of our evaluation; instead, we conceptually contrast our injection mechanism with their token concatenation approach in the ablation studies.

Implementation Details. All baselines are re-implemented on our frozen unified backbones for fair comparison. We use CLIP-ViT-L/14 as the visual encoder and keep all original parameters frozen. Accuracy in the controlled fusion comparison is evaluated on the complete M^3CoT and GQA evaluation sets, containing 1,500 and 2,020 samples, respectively. Second-stage latency is measured on a fixed subset of 200 samples for each model–dataset pair and reported as mean $\pm$ standard deviation. All latency measurements use a single NVIDIA L20 GPU under identical generation settings; the remaining experiments use up to two NVIDIA L20 GPUs.

4.2 Main Results

Table 1: Main results on standard multimodal reasoning benchmarks. We compare SAPS-CoT with state-of-the-art multimodal CoT methods under two strong backbones. Bold numbers indicate the best performance. SAPS-CoT consistently outperforms all baselines across most benchmarks, especially on tasks requiring fine-grained visual reasoning and spatial understanding.

Backbone	Methods	MathVista ↑	M^3CoT ↑	GQA ↑	ScienceQA ↑	Avg ↑
LLaVA-NeXT-8B	No-CoT	42.1	50.5	60.9	67.0	55.12
	Multimodal CoT Zhang et al. (2024c)	50.4	59.0	66.2	**70.4**	61.50
	CCoT Mitra et al. (2023)	44.4	56.2	62.8	65.7	57.27
	DDCoT Zheng et al. (2023)	44.7	59.4	62.4	65.4	57.97
	SCAFFOLD Lei et al. (2025)	24.1	52.8	60.4	64.9	50.55
	SAPS-CoT (Ours)	**51.3**	**62.2**	**67.8**	69.7	**62.75**
Qwen2-VL-7B	No-CoT	51.7	73.7	65.1	64.8	63.82
	Multimodal CoT Zhang et al. (2024c)	55.2	78.7	68.3	**68.8**	67.75
	CCoT Mitra et al. (2023)	48.4	74.7	66.9	63.9	63.47
	DDCoT Zheng et al. (2023)	51.6	77.4	67.9	65.0	65.47
	SCAFFOLD Lei et al. (2025)	28.9	74.9	62.4	64.4	57.65
	SAPS-CoT (Ours)	**55.8**	**79.9**	**69.8**	67.5	**68.25**

Table 1 summarizes the performance across multiple benchmarks. Our SAPS-CoT consistently outperforms state-of-the-art baselines like LLaVA-NeXT and Qwen2-VL. Notably, in the GQA dataset where multi-hop spatial reasoning is required, SAPS-CoT achieves a significant performance gain, proving the effectiveness of our gradient-guided visual grounding in resolving complex visual queries.

Superiority over Base LVLMs. SAPS-CoT consistently outperforms the strong No-CoT baselines across both LLaVA-NeXT and Qwen2-VL backbones. This significant improvement confirms that relying on static global visual encoding is insufficient for complex physical scenarios, whereas our dynamic look-back mechanism effectively recovers missing details.

Superiority over Existing CoT Paradigms. Compared to text-bottlenecked methods (DDCoT, CCoT) that compress images into discrete language or scene graphs, SAPS-CoT achieves superior accuracy by preserving continuous visual representations. Furthermore, our method significantly outperforms SCAFFOLD. While SCAFFOLD relies on artificial coordinate overlays that introduce unnatural distribution shifts to the image, SAPS inherently mines the model's internal causal attention, leading to more faithful and robust reasoning.

4.3 Ablation Studies

To verify the effectiveness of our core SAPS module and its multi-dimensional scoring mechanism, we conduct comprehensive ablation studies on the LLaVA-NeXT-8B backbone.

1) Effectiveness of the SAPS Module. We validate the core design of SAPS using LLaVA-NeXT-8B by comparing it with a vanilla two-stage Multi-CoT baseline (which performs a second-pass reasoning without visual grounding enhancement). As shown in Table 2, the lack of step-aware visual evidence means the model can hardly achieve further clear accuracy gains, demonstrating that our grounding-refinement cycle is critical for effective error correction.

2) Contribution of SAPS Scoring Factors. We further ablate each scoring component in the SAPS module: Location Bias (Loc.), Gradient Saliency (Grad.), and Semantic Consistency (Sem.). Results

Table 2: Ablation of the SAPS pipeline on LLaVA-NeXT-8B. Comparison between the full SAPS-CoT and a variant without SAPS demonstrates that step-aware visual enhancement significantly boosts performance across all benchmarks, validating its effectiveness in multimodal reasoning.

Methods	MathVista ↑	M^3CoT ↑	GQA ↑	Avg ↑
Full Model	**51.3**	**62.2**	**67.8**	**60.43**
w/o SAPS	49.5	60.0	66.6	58.70

in Table 3 show that *all three factors are indispensable*. Removing Gradient Saliency specifically causes a severe drop, highlighting the importance of capturing the internal causal sensitivity of the model.

Table 3: Ablation study on SAPS scoring components based on LLaVA-NeXT-8B. We ablate each scoring term in SAPS: Semantic Consistency, Gradient Saliency, Location Bias, and Multi-scale Window Compensation. All factors are indispensable.

Methods	MathVista ↑	M^3CoT ↑	GQA ↑	Avg ↑
Full Model	**51.3**	**62.2**	**67.8**	**60.43**
w/o Semantic	47.7	59.9	66.1	57.90
w/o Gradient	46.2	59.1	64.8	56.70
w/o Location	45.0	53.7	62.1	53.60
w/o Multi-scale	43.0	56.5	63.1	54.20

3) Score Aggregation Strategy. To validate the effectiveness of our product-based score aggregation, we compare it with three widely used alternatives: sum, max, and mean pooling. As shown in Table 4, multiplicative fusion consistently outperforms other strategies across both backbones. This is because product-based aggregation imposes a strict joint constraint—only patches scoring highly across all criteria (semantic, gradient, location, and multi-scale) survive, effectively suppressing noisy patches that may dominate a single dimension. In contrast, sum and mean pooling dilute discriminative signals, while max pooling overly relies on a single criterion, leading to unstable grounding.

Table 4: Comparison of different score aggregation methods on M^3CoT and GQA.

	LLaVA-NeXT-8B		Qwen2-VL-7B	
Aggregation Method	M^3CoT	GQA	M^3CoT	GQA
Sum	55.57	59.95	78.14	68.13
Max	56.50	60.85	77.21	68.78
Mean	55.36	61.74	79.03	69.25
Product (Ours)	**62.21**	**67.81**	**79.93**	**69.85**

4) Concept Comparison: Injection vs. Concatenation. We isolate the second-stage fusion operator while keeping the backbone, SAPS selection procedure, per-backbone Top-k, prompt, and decoding configuration unchanged. Table 5 shows that residual injection improves accuracy in all four settings by 0.74–1.19 percentage points (0.94 points on average) while preserving the original visual-token sequence ($\Delta L = 0$). Token concatenation is slightly faster in the tested sparse-patch regime. Residual injection incurs a second-stage latency premium of 1.1–6.0% (2.8% on average), owing to its attention-based feature-fusion computation. Thus, its empirical advantage is a modest but consistent accuracy improvement and bounded sequence length, rather than lower wall-clock latency.

5) Sensitivity to Top-k Patches. We analyze the impact of the number of selected patches k and observe an information-noise trade-off, where performance generally peaks at $k \in [10, 20]$ across different backbones; comprehensive results and visualizations are provided in Appendix A.2.

4.4 Qualitative Analysis

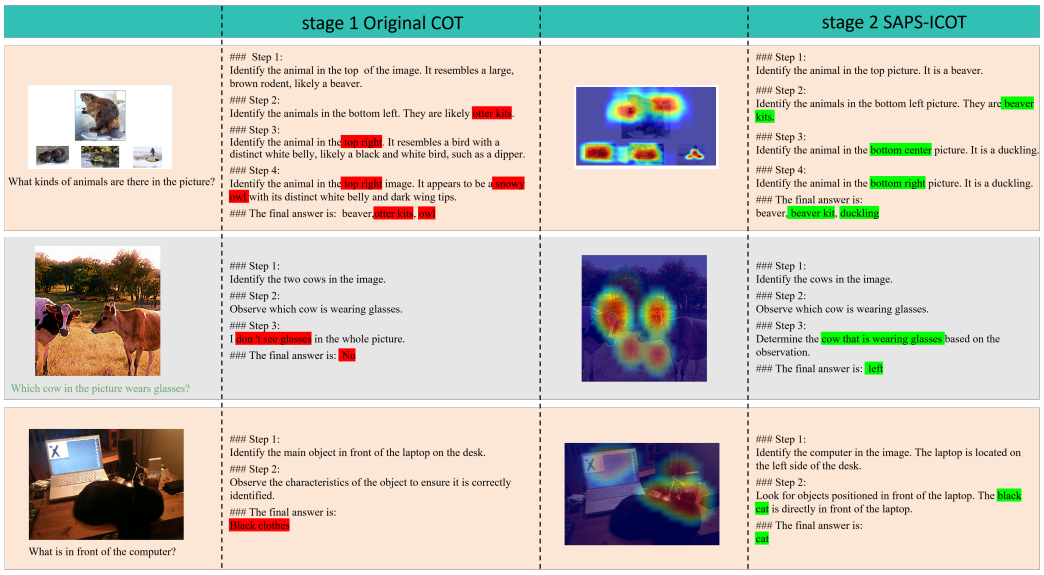


Figure 4: Qualitative visualization of SAPS-CoT. Heatmaps (red) highlight the critical visual patches automatically identified by SAPS for each reasoning step. By leveraging these fine-grained visual cues, our model successfully corrects hallucinations and inaccurate reasoning from the baseline, producing more faithful and reliable answers.

We provide qualitative examples in Figure 4. The visualizations cover three representative challenging scenarios: scenes with multiple interacting objects, targets that occupy only a small fraction of the image, and objects with subtle or ambiguous visual features. In each case, the base model's initial reasoning overlooks these fine-grained visual cues, producing incomplete or inaccurate descriptions. SAPS successfully highlights the neglected regions (shown in red on the generated heatmaps) by integrating semantic, gradient, and spatial signals, and the subsequent residual injection guides the model to self-correct its perceptual inertia and generate the faithful answer. These examples demonstrate that our step-aware grounding mechanism is particularly effective at recovering visually subtle but semantically critical details that static one-pass encoding tends to miss.

5 Discussion and Conclusion

We presented SAPS-CoT, a framework that combines Step-Aware Patch Selection with Dynamic Residual Injection to ground intermediate reasoning in visual evidence without increasing the visual-token sequence length. Experiments across MathVista, M^3CoT, GQA, and ScienceQA demonstrate consistent improvements over multimodal CoT baselines without requiring manual annotation, external modules, or model fine-tuning. Beyond predictive performance, SAPS-CoT provides an explicit visual trace that helps identify whether reasoning errors originate from perception or logical inference. Although residual injection preserves the original sequence and KV-cache lengths, its feature-fusion operation introduces a small latency overhead; its primary benefit is therefore improved accuracy with bounded sequence length rather than faster inference. The method also relies on the initial CoT remaining reasonably relevant and is less effective for severely misaligned reasoning or knowledge-dominated tasks. Future work will investigate bidirectional feedback and more efficient feature fusion to improve robustness and runtime efficiency.

AI use statement

Generative AI tools were used to assist with language editing, experimental-planning discussions, code debugging, and migration of the manuscript to the ICLR LaTeX format. The authors manually

reviewed all AI-assisted text and code, executed and verified all reported experiments, and take full responsibility for the methodology, claims, results, and final content of this work.

Reproducibility statement

The SAPS scoring procedure and Dynamic Residual Injection mechanism are specified in Section 3 through their equations and implementation description. Dataset, backbone, metric, and inference settings are reported in the experimental setup, while component ablations, hyperparameter sensitivity, and qualitative failure cases are provided in the main paper and appendix. Together, these details document the configurations used to obtain the reported results and support independent replication.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Barr, et al. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8199–8221, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.446. URL https://aclanthology.org/2024.acl-long.446/.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in LVLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=W5HnIf2jla.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-Modal Chain-of-Thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19520–19529, 2025. doi: 10.1109/CVPR52734.2025.01818.

Minbin Huang, Runhui Huang, Han Shi, Yimeng Chen, Chuanyang Zheng, Xiangguo Sun, Xin Jiang, Zhenguo Li, and Hong Cheng. Efficient multi-modal large language models via visual token grouping, 2024. URL https://arxiv.org/abs/2411.17773.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering, 2019. URL https://arxiv.org/abs/1902.09506.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2886–2903, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.195/.

Ang Li, Charles L. Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. Zebra-cot: A dataset for interleaved vision-language reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=c6XIVI3TiQ.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023a.

Junyan Li, Delin Chen, Tianle Cai, Peihao Chen, Yining Hong, Zhenfang Chen, Yikang Shen, and Chuang Gan. Flexattention for efficient high-resolution vision-language models, 2024. URL https://arxiv.org/abs/2407.20228.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuanjing Huang, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models, 2025. URL https://arxiv.org/abs/2405.16919.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2024a. doi: 10.1109/CVPR52733.2024.02484.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: multimodal reasoning via thought chains for science question answering. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Yi Lu, Jiawang Cao, Yongliang Wu, Bozheng Li, Licheng Tang, Yangguang Ji, Chong Wu, Jay Wu, and Wenbo Zhu. Rsvp: Reasoning segmentation via visual prompting and multi-modal chain-of-thought, 2025. URL https://arxiv.org/abs/2506.04277.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14420–14431, 2023. URL https://api.semanticscholar.org/CorpusID:265498786.

Eunkyu Park, Wesley Hanwen Deng, Gunhee Kim, Motahhare Eslami, and Maarten Sap. Cognitive chain-of-thought (cocot): Structured multimodal reasoning about social situations, 2026. URL https://arxiv.org/abs/2507.20409.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: a novel resource for question answering on scholarly articles: Scienceqa: a novel resource for question answering on scholarly articles. *Int. J. Digit. Libr.*, 23(3):289–301, September 2022. ISSN 1432-5012. doi: 10.1007/s00799-022-00329-y. URL https://doi.org/10.1007/s00799-022-00329-y.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Proceedings of the 38th International*

Conference on Neural Information Processing Systems, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Zekun Wang, Jingchang Chen, Wangchunshu Zhou, Haichao Zhu, Jiafeng Liang, Liping Shan, Ming Liu, Dongliang Xu, Qing Yang, and Bing Qin. SmartTrim: Adaptive tokens and attention pruning for efficient vision-language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14937–14953, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1300/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. CounterCurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15481–15495, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.915. URL https://aclanthology.org/2024.findings-acl.915/.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=d4UiXAHN2W.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024c. URL https://arxiv.org/abs/2302.00923.

Jianfei Zhao, Feng Zhang, Xin Sun, and Chong Feng. Mitigating hallucination in large vision-language models through aligning attention distribution to information flow, 2025. URL https://arxiv.org/abs/2505.14257.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

A Additional Ablation Studies

A.1 Comparison with Token Concatenation

Table 5 reports the detailed controlled comparison between token concatenation and dynamic residual injection discussed in Section 4.3.

Table 5: Controlled comparison of token concatenation and dynamic residual injection. Accuracy is evaluated on the complete M^{3}CoT (1,500 samples) and GQA (2,020 samples) evaluation sets. Second-stage latency (seconds, mean $\pm$ standard deviation) is measured on a fixed 200-sample subset using one NVIDIA L20 GPU. Bold values indicate the better result within each backbone and dataset.

		M^{3}CoT		GQA	
Backbone	Fusion	Acc. (%) ↑	Latency ↓	Acc. (%) ↑	Latency ↓
LLaVA-NeXT-8B	Token Concat.	61.14	**6.81±1.44**	67.06	**4.57±1.32**
LLaVA-NeXT-8B	Residual (Ours)	**62.20**	6.91±1.40	**67.80**	4.68±1.44
Qwen2-VL-7B	Token Concat.	78.71	**6.92±1.69**	69.05	**4.64±0.69**
Qwen2-VL-7B	Residual (Ours)	**79.90**	7.00±1.76	**69.80**	4.92±0.70

A.2 Sensitivity to Top-k Patches

To investigate the effect of the selected visual evidence scale, we vary the number of top-k patches injected during the SAPS-CoT refinement stage. As illustrated in Figure 5, we observe a consistent trend across different LLM backbones:

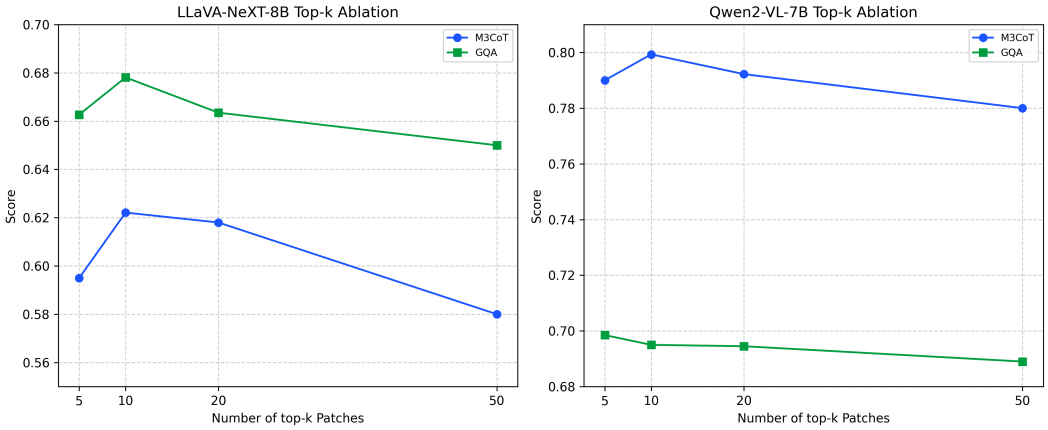


Figure 5: Impact of the selected patch number (k) on SAPS-CoT performance. We evaluate LLaVA-NeXT-8B (left) and Qwen2-VL-7B (right) across M^{3}CoT and GQA datasets. Performance generally peaks at $k \in [10, 20]$ and declines as k further increases, illustrating the crucial trade-off between supplementing sufficient visual evidence and introducing redundant noise.

The Information-Noise Trade-off: Initially, as k increases from 5 to 10 (or 20), the accuracy improves. This indicates that a minimal number of patches is insufficient to ground complex reasoning, and providing more top-scoring patches successfully injects critical visual clues. However, when k exceeds the optimal threshold (e.g., $k = 50$), performance steadily drops. We attribute this degradation to the introduction of irrelevant background noise, which dilutes the attention mechanism and distracts the reasoning process.

Dataset Complexity Variations: Interestingly, the optimal k varies slightly depending on the dataset's reasoning complexity. For instance, on the LLaVA-NeXT-8B backbone, M^{3}CoT (which requires complex, multi-step reasoning) peaks at $k = 20$, whereas GQA (which involves more direct

visual queries) peaks earlier at $k = 10$. This suggests that more intricate queries demand a broader visual context.

Backbone Robustness: Compared to LLaVA-NeXT, Qwen2-VL-7B exhibits a higher overall baseline and achieves optimal performance at a smaller k (peaking at $k = 10$ for M^{3}CoT and $k = 5$ for GQA). This implies that a stronger inherent vision-language alignment requires less residual injection to correct its reasoning trajectory.

A.3 Hyperparameter Sensitivity in Residual Injection

Table 6: **Ablation on Dynamic Residual Injection hyperparameters.** We evaluate the impact of the patch gate (α_{patch}) and residual gate (β_{res}) on LLaVA-NeXT-8B. The bold text highlights the optimal configuration used in our main experiments.

Ablation Target	Value	M^{3}CoT ↑	GQA ↑
Patch Gate (α_{patch}) (*with* $\beta_{res} = 0.4$)	0.2	56.71	65.60
	0.4	54.21	66.78
	0.8	**62.21**	**67.81**
Residual Gate (β_{res}) (*with* $\alpha_{patch} = 0.8$)	0.2	55.67	63.53
	0.4	**62.21**	**67.81**
	0.8	54.17	66.48

To comprehensively evaluate the robustness of our Dynamic Residual Injection mechanism, we conduct a sensitivity analysis on the two core gating hyperparameters introduced in Section 3.3: the patch gate (α_{patch}) and the residual gate (β_{res}). We use the LLaVA-NeXT-8B backbone and evaluate on the M^{3}CoT and GQA datasets.

As presented in Table 6, the framework demonstrates varying degrees of sensitivity to these scaling factors:

- **Impact of Patch Gate (α_{patch}):** When fixing the residual injection intensity ($\beta_{res} = 0.4$), increasing α_{patch} from 0.2 to 0.8 consistently improves performance, culminating at 62.21% on M^{3}CoT and 67.81% on GQA. A lower α_{patch} suppresses the fine-grained visual cues retrieved by SAPS, causing the model to revert to its baseline "blind" reasoning state.
- **Impact of Residual Gate (β_{res}):** Conversely, when fixing the patch evidence scale ($\alpha_{patch} = 0.8$), the optimal residual injection strength is found at $\beta_{res} = 0.4$. If the injection weight is too low (0.2), the correction signal is too weak to alter the model's locked-in perceptual bias. However, if it is too high (0.8), we observe a catastrophic drop in performance (e.g., M^{3}CoT drops from 62.21% to 54.17%). This degradation occurs because an overly aggressive residual addition corrupts the original structural integrity and global semantics of the visual tokens X_i, overwhelming the language model with localized noise.

These findings confirm that an optimal balance between preserving original global semantics and injecting step-aware local evidence ($\alpha_{patch} = 0.8$, $\beta_{res} = 0.4$) is crucial for the success of SAPS-CoT.

B More Visualization and Error Analysis

To provide a comprehensive evaluation of our SAPS-CoT framework, we present additional qualitative examples. This section includes both a successful correction of baseline hallucinations and an explicit analysis of challenging failure cases where our method struggles due to inherent visual limitations.

B.1 Successful Hallucination Correction

We first demonstrate how SAPS-CoT effectively eliminates hallucinations by grounding the reasoning process in specific visual evidence.

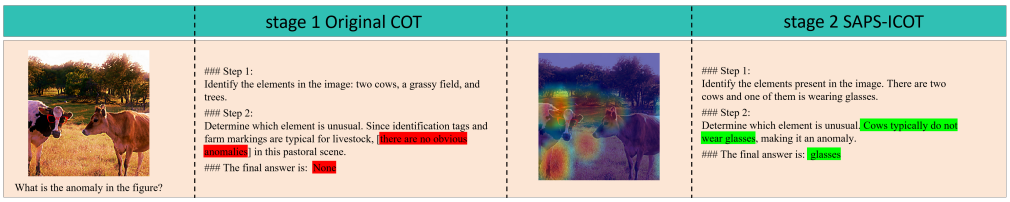


Figure 6: **SAPS-CoT corrects baseline hallucination.** While the baseline model generates hallucinatory reasoning by failing to identify the abnormal visual clue, SAPS-CoT explicitly grounds the critical visual regions. This dynamic residual injection ensures the reasoning is anchored to the actual image content, leading to a reliable and correct prediction.

B.2 Failure Case Analysis

To better understand the boundaries of our framework, we present two distinct failure cases. These examples illustrate that while SAPS-CoT excels at visual grounding, it is inherently constrained by the quality of the raw visual input.

Failure Case 1: Severe Physical Occlusion.

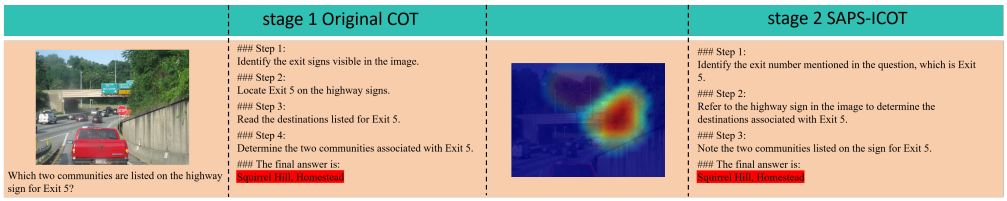


Figure 7: **Failure Case 1: Partially occluded visual information.** The community names on the highway exit sign are physically occluded. Because the required text is fundamentally unreadable, SAPS-CoT cannot extract valid visual evidence. Consequently, the model exhibits a confirmation bias, defaulting to the hallucinatory textual priors generated during the initial coarse-grained reasoning stage.

Failure Case 2: Extreme Blur and Low Resolution.

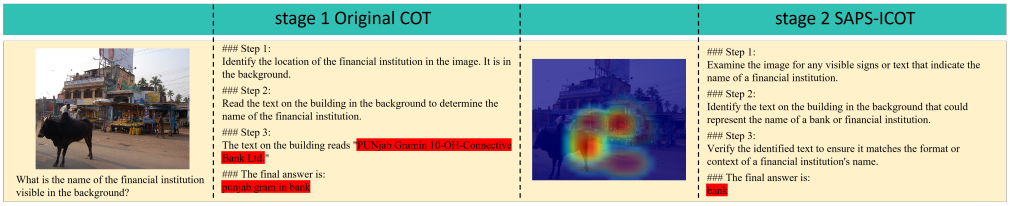


Figure 8: **Failure Case 2: Unrecoverable low-resolution text.** The financial institution's name in the background is extremely blurry and indistinct. SAPS-CoT operates as a visual grounding mechanism rather than an image restoration module; therefore, it cannot recover fine-grained information that is absent from the original pixel space.

Summary. These failure cases highlight a fundamental boundary of current visual grounding systems. SAPS-CoT significantly mitigates hallucinations under normal perceptual conditions. However, when critical information is lost due to *physical occlusion* or *irretrievably low resolution*, the model may fall back on language priors. Future work could explore incorporating confidence-based abstention mechanisms (e.g., rejecting to answer when visual evidence scores are excessively low) to handle such severe visual degradation gracefully.