HIDDEN HARM IN NON-ATOMIC DEMONSTRATIONS: SELECTIVE COUNTERFACTUAL COMPONENT INTERVENTION FOR MULTIMODAL IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal in-context learning (MICL) commonly constructs and evaluates in-context sequences at the level of complete image-text demonstrations, although prior studies show that visual and textual information can contribute differently across tasks. We ask a finer-grained question: for a query, do the image and text within the same demonstration exert the same effect? They need not. Their query-conditioned utilities can differ, and sometimes disagree in sign, such that a demonstration appearing useful at the whole-demonstration level may still contain a component that harms the prediction. We refer to this failure mode as *hidden component harm*, arising from *demonstration non-atomicity*. To mitigate it, we propose Selective Counterfactual Component Intervention (SCCI), a training-free framework that decomposes each demonstration into visual and textual components and derives query-specific counterfactual signals through leave-one-component-out inference. A reliability-aware decision rule intervenes only when the evidence is sufficiently informative, while otherwise preserving the context. Across six vision-language benchmarks and two primary multimodal language models, Qwen2.5-VL-7B and InternVL2.5-8B, our method consistently improves 4-shot MICL over the unmodified context and outperforms both random component removal and matched whole-demonstration removal. Further analyses of image-text utility disagreement, hidden component harm cases, and whole-demonstration attribution support the value of semantic component granularity, with gains remaining robust across demonstration sampling seeds and order permutations. Anonymous code is available at `https://anonymous.4open.science/r/SCCI-3F45/`.

1 INTRODUCTION

Multimodal in-context learning (MICL) enables large vision-language models (LVLMs) to adapt to tasks from a few interleaved image-text demonstrations without parameter updates (Alayrac et al., 2022). Subsequent work extends this capability through multimodal in-context and instruction tuning (Zhao et al., 2024; Li et al., 2025a; Sun et al., 2024). Yet MICL remains inconsistent across tasks and contexts (Doveh et al., 2024; Zong et al., 2025). Empirical analyses show sensitivity to demonstration content and selection, with models often relying disproportionately on textual demonstrations (Baldassini et al., 2024; Chen et al., 2025). Even well-matched demonstrations may remain underused, indicating limited exploitation of multimodal context (Li et al., 2026b). Mechanistic analyses identify unreliable task-mapping transfer to the query as a bottleneck (Wang & Li, 2026).

These limitations motivate context optimization for LVLMs. At the demonstration level, context optimization operates on complete examples, selecting informative sets for the query (Li et al., 2025b). Text ICL studies similarly estimate complete-demonstration utility to distinguish helpful and harmful examples (Xie et al., 2025). At finer granularity, visual token pruning removes redundant patches and extends to interleaved MICL settings with multiple demonstrations (Chen et al., 2024a; Alvar et al., 2025; Li et al., 2026a). However, these approaches operate either at the whole-demonstration or patch-token level and do not model distinct query-conditioned utilities for the image and text within a demonstration. This motivates an intermediate semantic granularity where visual and textual components become separate query-conditioned utility estimation and intervention units.

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.

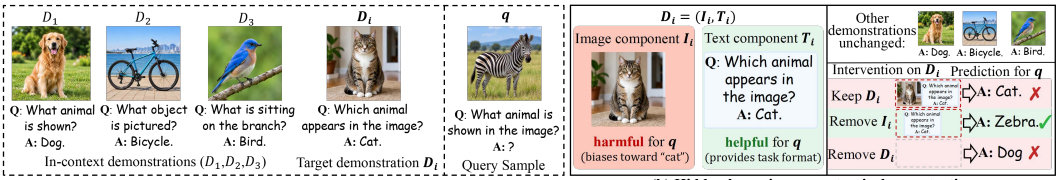


Figure 1: **Hidden component harm.** (a) Whole-demonstration attribution treats $D_i = (I_i, T_i)$ as a unit. (b) Visual and textual components may have opposing query-conditioned effects.

At this semantic granularity, prior modality-level analyses suggest that visual and textual information can contribute differently across MICL tasks (Xu et al., 2025). We therefore ask: *for a query q and demonstration $D_i = (I_i, T_i)$, can I_i and T_i have different query-conditioned utilities, even with opposite signs?* We call such a demonstration *non-atomic with respect to the query*, where whole-demonstration attribution fails to capture component-level effects. A critical case occurs when a demonstration appears beneficial at the whole-demonstration level, yet removing one of its components improves the prediction. As illustrated in Fig. 1, this hidden harmful component is concealed by the overall demonstration utility. We term this failure mode *hidden component harm*.

Motivated by this observation, we propose **Selective Counterfactual Component Intervention (SCCI)**. Given a fixed MICL sequence, SCCI probes each demonstration's visual and textual components through leave-one-component-out counterfactual inference to derive query-specific removal signals. It selectively intervenes according to counterfactual reliability while preserving the original context when evidence is insufficient. The LVLM remains frozen throughout.

We evaluate SCCI on six vision-language benchmarks with two representative LVLMs, Qwen2.5-VL-7B and InternVL2.5-8B (Bai et al., 2025; Chen et al., 2024b). SCCI consistently improves 4-shot MICL over the unmodified context and outperforms random component removal and matched whole-demonstration removal, with robustness across demonstration sampling seeds and order permutations. We further analyze component-level utilities and hidden component harm to validate query-conditioned demonstration non-atomicity. In summary, our main contributions are threefold:

- We identify **query-conditioned demonstration non-atomicity and hidden component harm** in multimodal ICL, showing that visual and textual components within a demonstration can have distinct utilities overlooked by whole-demonstration attribution.
- We propose **Selective Counterfactual Component Intervention (SCCI)**, a training-free framework that derives query-specific component signals through leave-one-component-out counterfactual probing and selectively intervenes based on evidence reliability.
- We provide extensive empirical validation across two primary LVLMs and six vision-language benchmarks, including effectiveness, hidden component harm analysis, and robustness under different demonstration sampling and ordering.

2 Related Work

Multimodal In-Context Learning and Context Optimization. Multimodal in-context learning (MICL) enables vision-language models to adapt to new tasks from a small number of image-text demonstrations (Sheng et al., 2024; Zhou et al., 2024). Early vision-language models established few-shot transfer through interleaved prompting (Alayrac et al., 2022), followed by advances in multimodal pretraining and instruction tuning (Li et al., 2022; Dai et al., 2023; Liu et al., 2023). Recent studies optimize in-context examples through demonstration retrieval and context construction (Rubin et al., 2022; Zhao et al., 2021; Li et al., 2025b). Meanwhile, recent multimodal efficiency methods reduce context redundancy through visual token pruning or informative region selection (Chen et al., 2024a; Yang et al., 2025; Li et al., 2026a). However, existing approaches operate either at the whole-demonstration or visual-token level, leaving the semantic component granularity, where visual and textual components have distinct query-conditioned utilities, underexplored.

Attribution and Counterfactual Intervention. Attribution methods quantify how inputs specifically contribute to model predictions. Local surrogate and additive methods estimate feature im-

portance (Ribeiro et al., 2016; Lundberg & Lee, 2017), while gradient-based approaches provide finer attribution (Sundararajan et al., 2017). Influence analyses measure the effects of removing or modifying inputs (Koh & Liang, 2017; Pruthi et al., 2020), and broader explainability studies extend this view to training examples (Minh et al., 2022). Counterfactual methods study how alternative inputs change predictions to support intervention (Wachter et al., 2017; Mothilal et al., 2020). Recent multimodal work analyzes visual evidence, modality contributions, and cross-modal interactions (Chefer et al., 2021; Bugliarello et al., 2021; Lai et al., 2024). However, existing methods mainly analyze tokens, modalities, or examples, without modeling the query-conditioned utility of visual and textual components within a demonstration. We instead evaluate these components separately and intervene only when counterfactual evidence is reliable.

3 Analyzing Demonstration Non-Atomicity

We examine whether multimodal demonstrations behave as atomic units in MICL. We first define query-conditioned component removal utilities and then study two signatures of non-atomicity: *component utility disagreement* and *hidden component harm*.

3.1 Counterfactual Component Utilities

For single-query discussions and illustrations, we omit the query index and write q and $D_i = (I_i, T_i)$, as in Fig. 1. Formal definitions restore the index n. Let $\{(q_n, y_n^*)\}_{n=1}^N$ denote the evaluation queries and target answers. For query q_n, let $\mathcal{D}_n = (D_{n,1}, \ldots, D_{n,K_n})$ denote the ordered sequence of K_n in-context demonstrations, where $D_{n,i} = (I_{n,i}, T_{n,i})$. Define the modality set $\mathcal{M} = \{\mathrm{img}, \mathrm{txt}\}$ and indexed component occurrences $\mathcal{J}_n = [K_n] \times \mathcal{M}$, with $c_n(i, \mathrm{img}) = I_{n,i}$ and $c_n(i, \mathrm{txt}) = T_{n,i}$, where $[K_n] = \{1, \ldots, K_n\}$. We use the shared intervention space $\mathcal{A}_n = \{\varnothing\} \cup \mathcal{J}_n$, where $\varnothing$ denotes no intervention. Each action $a \in \mathcal{A}_n$ induces

$$\mathcal{P}_n(a) = \begin{cases} \mathcal{D}_n, & a = \varnothing, \\ \mathcal{D}_n \setminus c_n(i,m), & a = (i,m) \in \mathcal{J}_n, \end{cases} \tag{1}$$

where removal acts on the indexed component occurrence while preserving the order of all remaining components. For a frozen autoregressive LVLM with parameters θ, we score context $\mathcal{P}$ using the teacher-forced, length-normalized log-likelihood of the target answer:

$$\ell_n(\mathcal{P}) = \frac{1}{L_n} \sum\nolimits_{t=1}^{L_n} \log p_\theta\big(y_{n,t}^* \mid \mathcal{P}, q_n, y_{n,<t}^*\big), \tag{2}$$

where L_n is the number of tokens in y_n^*. Following the counterfactual explanation principle, we define the query-conditioned removal utility of an action $a \in \mathcal{A}_n$ as $U_n(a) = \ell_n(\mathcal{P}_n(a)) - \ell_n(\mathcal{P}_n(\varnothing))$. Thus, $U_n(\varnothing) = 0$. For $a = (i,m) \in \mathcal{J}_n$, we write $U_n(i,m) := U_n((i,m))$ for brevity. A positive utility means that removal increases the target-answer score, indicating that the component's presence is harmful for q_n; a negative utility indicates useful information. We separately define the whole-demonstration removal utility as $U_n^{\mathrm{demo}}(i) = \ell_n(\mathcal{D}_n \setminus D_{n,i}) - \ell_n(\mathcal{D}_n)$. Because components interact nonlinearly with one another and with the surrounding context, $U_n^{\mathrm{demo}}(i)$ need not equal $U_n(i, \mathrm{img}) + U_n(i, \mathrm{txt})$; we therefore measure these counterfactual effects directly.

3.2 Diagnostics of Demonstration Non-Atomicity

Component utility disagreement. We first test whether the two components of a demonstration can exert opposing query-conditioned effects. Opposite removal-utility signs provide a sufficient signature of non-atomicity: removing one component improves the target-answer score while removing the other decreases it. Equal signs do not imply atomicity, since the effects may still differ in magnitude. To treat near-zero effects as neutral, let $s_\epsilon(u) = -1$, 0, or $+1$ according to whether $u < -\epsilon$, $|u| \le \epsilon$, or $u > \epsilon$, where $\epsilon \ge 0$ is fixed before test-set analysis. With $M = \sum_{n=1}^N K_n$ query–demonstration pairs, the disagreement rate is

$$R_{\mathrm{dis}}^{(\epsilon)} = \frac{1}{M} \sum\nolimits_{n=1}^{N} \sum\nolimits_{i=1}^{K_n} \mathbb{I}[s_\epsilon(U_n(i, \mathrm{img}))\, s_\epsilon(U_n(i, \mathrm{txt})) = -1]. \tag{3}$$

Only non-neutral, mixed-sign pairs contribute to this rate. Fig. 2 visualizes the case $\epsilon = 0$ on Qwen2.5-VL-7B. We write $U_I = U_n(i, \mathrm{img})$ and $U_T = U_n(i, \mathrm{txt})$ for each query–demonstration

pair. The stacked bars summarize the sign patterns across six benchmarks, while the scatter plot shows their utility magnitudes on VQAv2. A non-negligible mass in the mixed-sign regions shows that a whole-demonstration utility cannot capture the opposing effects of its internal components.

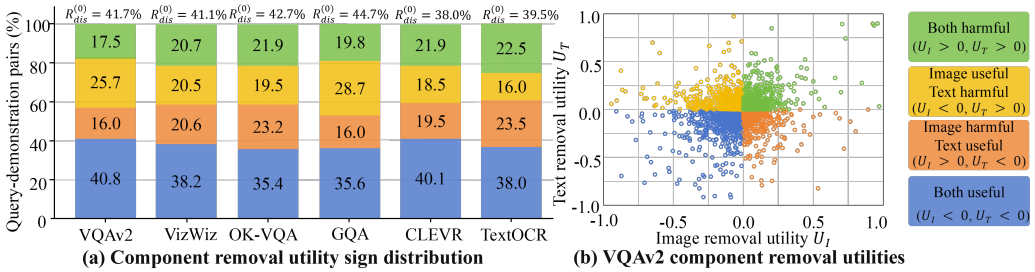


Figure 2: **Component-level removal utility disagreement at $\epsilon = 0$ on Qwen2.5-VL-7B.** (a) Strict image–text removal-utility sign patterns across six benchmarks. (b) Joint image and text removal utilities on VQAv2, highlighting substantial mixed-sign component effects.

Hidden component harm. Component disagreement reveals internal heterogeneity but does not by itself show that whole-demonstration attribution misses a harmful component. For each demonstration, define its largest component removal utility as $U^{\max}_{n,i} = \max_{m \in \mathcal{M}} U_n(i, m)$. We define *hidden component harm* as the strict cross-granularity reversal

$$H^{(\epsilon)}_{n,i} = \mathbb{I}\big[U^{\max}_{n,i} > \epsilon \ \wedge \ U^{\text{demo}}_n(i) < -\epsilon\big] \,. \tag{4}$$

The first condition identifies a harmful component, while the second indicates that the full demonstration is beneficial when retained. Thus, whole-demonstration attribution would preserve $D_{n,i}$ while leaving its harmful component in context. We measure the prevalence of this failure mode as

$$R^{(\epsilon)}_{\text{hch}} = \frac{1}{M} \sum\nolimits_{n=1}^{N} \sum\nolimits_{i=1}^{K_n} H^{(\epsilon)}_{n,i}. \tag{5}$$

Together, $R^{(\epsilon)}_{\text{dis}}$ and $R^{(\epsilon)}_{\text{hch}}$ characterize complementary aspects of demonstration non-atomicity. The former captures opposing effects within a demonstration, whereas the latter captures cross-granularity reversals in which a harmful component is concealed within a demonstration that is beneficial as a whole. All target-answer-based quantities in this section are used only for post-hoc diagnosis and are unavailable to SCCI at inference time. Appendix A provides a qualitative case.

4 Selective Counterfactual Component Intervention

Sec. 3 shows that component-level counterfactuals can reveal effects hidden by whole-demonstration attribution. Because the diagnostic utility $U_n(a)$ depends on the unknown target answer y^*_n, it is unavailable at test time. We propose **Selective Counterfactual Component Intervention (SCCI)**, which reuses $\mathcal{A}_n$ and $\mathcal{P}_n(a)$ but replaces $U_n(a)$ with target-answer-free evidence. With the LVLM frozen, SCCI evaluates the unchanged context and all single-component counterfactuals, removes at most one supported component, and otherwise retains the context. Fig. 3 gives an overview.

4.1 Target-Free Counterfactual Evidence

For every action $a \in \mathcal{A}_n$, the frozen LVLM generates $\widehat{y}_{n,a} = \text{Decode}_\theta(\mathcal{P}_n(a), q_n)$ using the same deterministic decoding rule. We measure the model's length-normalized self-log-likelihood of the generated answer for consistent branchwise confidence comparison as

$$S_n(a) = \frac{1}{\widehat{L}_{n,a}} \sum\nolimits_{t=1}^{\widehat{L}_{n,a}} \log p_\theta(\widehat{y}_{n,a,t} \mid \mathcal{P}_n(a), q_n, \widehat{y}_{n,a,<t}) \,, \tag{6}$$

where $\widehat{L}_{n,a}$ is the generated answer length. Length normalization reduces bias toward shorter outputs across actions. We define the target-free counterfactual evidence as

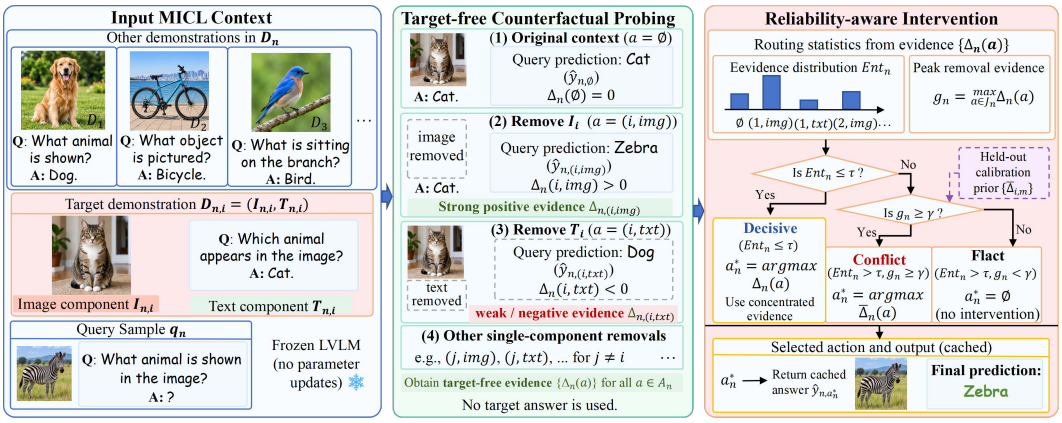


Figure 3: **Overview of SCCI.** SCCI probes the original context and single-component removals with target-free evidence. Reliability-aware routing selects decisive evidence, calibrates conflicts with a held-out prior, and abstains under weak evidence. The cached selected answer is returned.

$$\Delta_n(a) = S_n(a) - S_n(\varnothing), \qquad a \in \mathcal{A}_n, \tag{7}$$

which implies $\Delta_n(\varnothing) = 0$. For $a \in \mathcal{J}_n$, a positive $\Delta_n(a)$ indicates higher self-log-likelihood after removal and serves as SCCI's routing evidence. Unlike $U_n(a)$, which scores the same target answer across contexts, $\Delta_n(a)$ compares confidence in context-specific generated answers and is therefore target-free rather than a direct estimate of diagnostic utility. We restrict SCCI to at most one component removal, avoiding the composition of component effects that need not be additive while preserving the potentially useful remainder of a non-atomic demonstration.

4.2 Reliability-Aware Intervention

Counterfactual evidence can be concentrated on one action, distributed across competing actions, or uniformly weak. SCCI treats concentration and peak strength as test-time reliability proxies. We convert the evidence into

$$\pi_{n,a} = \frac{\exp(\Delta_n(a))}{\sum_{b \in \mathcal{A}_n} \exp(\Delta_n(b))}, \qquad \mathsf{Ent}_n = -\frac{\sum_{a \in \mathcal{A}_n} \pi_{n,a} \log \pi_{n,a}}{\log(2K_n + 1)}, \tag{8}$$

where $\mathsf{Ent}_n \in [0, 1]$ measures the dispersion of the evidence-induced distribution, with lower values indicating greater concentration. We also define $g_n = \max_{a \in \mathcal{J}_n} \Delta_n(a)$ as the peak component-removal evidence. Calibration statistics for routing are estimated from a held-out calibration-query set $\mathcal{Q}_{\text{cal}}$ disjoint from the reported evaluation queries. Let $\tau = Q_\rho(\{\mathsf{Ent}_r : r \in \mathcal{Q}_{\text{cal}}\})$, $\mathcal{Q}_{\text{low}}^{\text{cal}} = \{r \in \mathcal{Q}_{\text{cal}} : \mathsf{Ent}_r \leq \tau\}$, where Q_ρ denotes the empirical ρ-quantile.

For each position–modality action (i, m) with nonempty calibration support, we estimate the prior

$$\overline{\Delta}_{i,m} = \frac{1}{|\mathcal{Q}_{\text{low}}^{\text{cal},i,m}|} \sum_{r \in \mathcal{Q}_{\text{low}}^{\text{cal},i,m}} \Delta_r(i, m), \tag{9}$$

where $\mathcal{Q}_{\text{low}}^{\text{cal},i,m} = \{r \in \mathcal{Q}_{\text{low}}^{\text{cal}} : (i, m) \in \mathcal{J}_r\}$. If this set is empty, we use the neutral fallback $\overline{\Delta}_{i,m} = 0$. To define the peak-strength threshold, let $\mathcal{G}_{\text{cal}}^{+} = \{g_r : r \in \mathcal{Q}_{\text{cal}}, \mathsf{Ent}_r > \tau, g_r > 0\}$. We set $\gamma = \text{median}(\mathcal{G}_{\text{cal}}^{+})$ when nonempty and $\gamma = +\infty$ otherwise. The conflict-branch score is

$$\widetilde{\Delta}_n(a) = \begin{cases} 0, & a = \varnothing, \\ \beta \Delta_n(i, m) + (1 - \beta)\overline{\Delta}_{i,m}, & a = (i, m) \in \mathcal{J}_n, \end{cases} \tag{10}$$

where $\beta \in [0, 1]$ balances query-specific evidence and the calibration prior. SCCI selects

$$a_n^* = \begin{cases} \arg\max_{a \in \mathcal{A}_n}^{\text{TB}} \Delta_n(a), & \mathsf{Ent}_n \leq \tau, \\ \arg\max_{a \in \mathcal{A}_n}^{\text{TB}} \widetilde{\Delta}_n(a), & \mathsf{Ent}_n > \tau \ \wedge \ g_n \geq \gamma, \\ \varnothing, & \mathsf{Ent}_n > \tau \ \wedge \ g_n < \gamma. \end{cases} \tag{11}$$

Table 1: **Main results on six vision-language benchmarks under the 4-shot setting.** Accuracy (%); best and second-best results within each model are **bold** and underlined.

Model	Method	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Qwen2.5-VL-7B	Vanilla MICL	75.37	61.39	55.00	61.88	53.50	45.50
	RICES (Yang et al., 2022)	75.62	62.71	55.83	62.31	54.61	47.42
	MMICES (Chen et al., 2025)	75.81	63.18	56.21	62.70	55.46	48.31
	TACO (Li et al., 2025b)	75.95	63.51	56.47	**62.74**	56.92	48.72
	CATP (Li et al., 2026a)	75.94	62.46	55.69	62.27	55.12	46.84
	SCCI	**76.05**	**63.73**	**56.84**	62.72	**57.51**	**49.00**
InternVL2.5-8B	Vanilla MICL	61.26	44.08	50.22	30.88	19.00	58.00
	RICES (Yang et al., 2022)	64.86	47.91	51.73	35.42	20.31	59.18
	MMICES (Chen et al., 2025)	66.47	49.88	53.61	40.27	21.46	60.37
	TACO (Li et al., 2025b)	69.84	51.74	53.71	44.83	22.96	61.26
	CATP (Li et al., 2026a)	62.18	45.36	50.94	32.14	19.82	58.76
	SCCI	**72.12**	**53.03**	**54.32**	**49.50**	**24.48**	**62.01**

These cases define the *decisive*, *conflict*, and *flat* branches, respectively. The *decisive* branch uses concentrated query-specific evidence, the *conflict* branch stabilizes competing evidence with the calibration prior, and the *flat* branch abstains from intervention. The operator $\arg\max^{\mathrm{TB}}$ favors $\varnothing$ on a tie and otherwise follows a fixed order over (i, m). Because $\varnothing$ has score zero in both maximizations, either branch retains the original context whenever no component has positive evidence under its corresponding scoring rule.

Calibration and inference. The quantile ρ and mixing weight β are selected using only the held-out calibration split and are then frozen, together with τ, γ, and the priors $\{\overline{\Delta}_{i,m}\}$, before evaluation. SCCI is training-free in the sense that the LVLM parameters are never updated, although labeled held-out data are used for hyperparameter selection. No test target is used to compute Δ_n, estimate calibration statistics, route a query, or choose an intervention. For a query with K_n demonstrations, SCCI evaluates the unchanged context and all $2K_n$ single-component counterfactuals, requiring $1 + 2K_n$ generations. The outputs are cached before routing, so the selected answer requires no additional decoding; the default 4-shot setting therefore evaluates nine prompt variants per query.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), OK-VQA (Marino et al., 2019), VizWiz (Gurari et al., 2018), CLEVR (Johnson et al., 2017), and TextOCR (Singh et al., 2021), covering diverse visual reasoning and scene-text tasks. For TextOCR, we follow VL-ICL Bench (Zong et al., 2025), where the model infers from demonstrations to transcribe highlighted text. Further dataset details are provided in Appendix B.1.

Evaluation and baselines. We report accuracy (%) following the task-specific evaluation protocol for each benchmark. In addition to *Vanilla MICL*, we compare SCCI with context-optimization approaches at different granularities: *RICES* (Yang et al., 2022) and *MMICES* (Chen et al., 2025) for demonstration retrieval and selection, *TACO* (Li et al., 2025b) for sequence configuration, and *CATP* (Li et al., 2026a) for token-level pruning. We consider *random component removal* and *matched whole-demonstration removal* as controlled baselines. For these controlled comparisons, intervention opportunities are matched whenever applicable, isolating component-selection quality and intervention granularity. We further ablate entropy-based reliability gating, conflict-branch calibration, and flat-branch abstention. Baseline details are provided in Appendix B.2.

Models and inference. We evaluate Qwen2.5-VL-7B (Bai et al., 2025) and InternVL2.5-8B (Chen et al., 2024b) as primary LVLMs, with additional Idefics2-8B validation (Laurençon et al., 2024) to assess broader backbone generalization. Within each model–benchmark setting, methods share evaluation queries, support pool, shot budget, task prompt, and decoding configuration whenever compatible with their protocols. Retrieval-based methods select demonstrations from the shared support pool, while controlled comparisons use the same demonstration context. For a query with K_n demonstrations, SCCI requires $1 + 2K_n$ generations; the default 4-shot setting therefore uses nine prompt variants. Outputs are cached, so routing requires no additional decoding.

Table 2: **Router ablation study under the 4-shot setting.** Accuracy (%).

Model	Variant	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Qwen2.5-VL-7B	PCP	75.85	63.43	55.67	62.60	54.50	46.00
	Gated PCP	75.87	63.67	56.11	62.64	57.02	46.48
	SCCI w/o Prior	75.93	63.69	56.38	62.67	57.24	47.53
	SCCI w/o Flat	75.97	63.66	56.55	62.69	57.35	48.08
	SCCI	**76.05**	**63.73**	**56.84**	**62.72**	**57.51**	**49.00**
InternVL2.5-8B	PCP	65.91	52.06	53.28	48.60	21.51	60.00
	Gated PCP	71.99	52.90	53.47	49.34	22.03	60.96
	SCCI w/o Prior	72.03	52.96	53.81	49.40	23.04	61.42
	SCCI w/o Flat	72.07	53.00	54.02	49.44	23.51	61.57
	SCCI	**72.12**	**53.03**	**54.32**	**49.50**	**24.48**	**62.01**

Table 3: **Controlled removal comparisons under the 4-shot setting.** Accuracy (%).

Model	Method	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Qwen2.5-VL-7B	Random component	75.36	59.42	54.58	61.30	51.83	44.83
	Whole-demo removal	75.64	59.65	55.00	61.82	56.50	46.50
	SCCI	**76.05**	**63.73**	**56.84**	**62.72**	**57.51**	**49.00**
InternVL2.5-8B	Random component	63.24	43.00	50.48	34.77	18.32	57.53
	Whole-demo removal	66.68	45.64	52.29	43.58	18.50	58.12
	SCCI	**72.12**	**53.03**	**54.32**	**49.50**	**24.48**	**62.01**

Calibration. SCCI uses a held-out calibration split disjoint from the evaluation queries. The entropy quantile ρ and mixing weight β are selected using calibration performance, while τ, γ, and $\overline{\Delta}_{i,m}$ are derived from target-free calibration evidence following Sec. 4.2. All are fixed before evaluation. Calibration labels are used only for hyperparameter selection; no evaluation target answer is used for evidence computation, routing, or intervention. Details are provided in Appendix B.3.

5.2 Main Results

Table 1 compares SCCI with MICL context-optimization methods at different granularities. SCCI achieves the best performance in 11 of 12 model–benchmark settings and ranks second in the remaining one. On Qwen2.5-VL-7B, it performs best on five benchmarks and second on GQA; on InternVL2.5-8B, it performs best on all six benchmarks. Its advantage over demonstration retrieval and selection, sequence configuration, and token-pruning baselines indicates that component-level intervention provides complementary information beyond these context-optimization granularities. This is consistent with SCCI's component-level design, which removes at most one component while preserving the useful remainder. In the only setting where SCCI ranks second, TACO performs best, suggesting that sequence-level configuration may complement component intervention.

5.3 Ablation and Controlled Comparisons

We examine where SCCI's gains come from through router ablations and controlled removal comparisons. The former isolates contributions of reliability-aware routing, while the latter tests whether the gains depend on query-conditioned component selection and component-level intervention.

Router ablation. Table 2 evaluates SCCI's reliability-aware router. *PCP* routes using the raw target-free evidence Δ, while *Gated PCP* restricts intervention to low-entropy queries. *SCCI w/o Prior* and *SCCI w/o Flat* isolate the calibration prior and flat-branch abstention, respectively. Full SCCI performs best in all 12 model–benchmark settings. The consistent improvement of Gated PCP over PCP shows the benefit of reliability-aware routing, while both targeted ablations remaining below full SCCI indicate complementary roles for conflict calibration and flat-branch abstention.

Controlled removal comparisons. Table 3 tests whether SCCI's gains depend on component selection and granularity. *Random component removal* removes a random component under matched opportunities, while *matched whole-demonstration removal* deletes the selected component's parent demonstration. SCCI outperforms both controls across all 12 model–benchmark settings, showing that gains depend on query-conditioned selection rather than deletion alone. Its advantage over whole-demonstration removal is consistent with demonstration non-atomicity: component-level intervention can preserve useful information that whole-demonstration removal may discard.

5.4 Counterfactual Evidence and Routing Reliability

We analyze whether SCCI's target-free evidence reflects decision-relevant component information and whether its routing branches distinguish different levels of intervention reliability. The former relates $\Delta_n(a)$ to the diagnostic utility $U_n(a)$, while the latter compares *decisive*, *conflict*, and *flat* queries. All analyses involving $U_n(a)$ are post-hoc and unavailable to SCCI during inference.

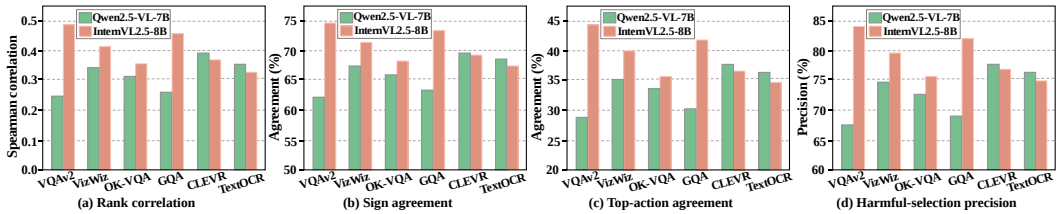


Figure 4: **Alignment between target-free evidence and diagnostic utility.** Across six benchmarks and two models, $\Delta_n(a)$ captures decision-level information about effects measured by $U_n(a)$.

Alignment between target-free evidence and diagnostic utility. Fig. 4 evaluates the agreement between $\Delta_n(a)$ and $U_n(a)$. All 12 model–benchmark settings show positive rank correlation, with sign agreement above 60% and top-action agreement of 28.6%–44.6%. Harmful-selection precision reaches 67.4%–84.1%, indicating that components favored for removal by the target-free signal are often harmful under the post-hoc utility. These results support $\Delta_n(a)$ as an informative routing signal without treating it as a direct estimate of $U_n(a)$. Metric definitions are provided in Appendix C.

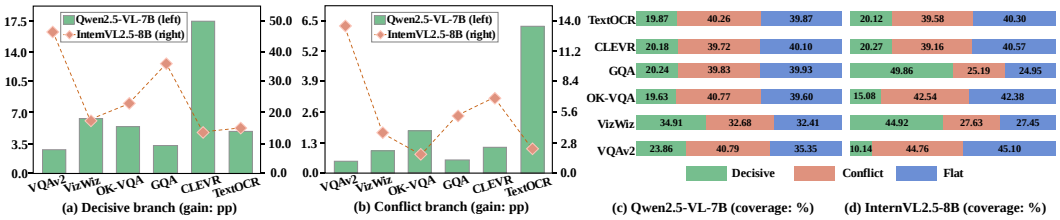


Figure 5: **Routing reliability and coverage.** (a)–(b) Accuracy gains (percentage points) of SCCI over Vanilla MICL for *decisive* and *conflict* queries. (c)–(d) Fractions (%) of queries assigned to the *decisive*, *conflict*, and *flat* branches for Qwen2.5-VL-7B and InternVL2.5-8B.

Routing reliability and coverage. Fig. 5(a)–(b) reports branch-wise accuracy gains in percentage points relative to Vanilla MICL. *Decisive* queries generally show larger gains than *conflict* queries, indicating that concentrated counterfactual evidence provides a more reliable basis for intervention. *Conflict* queries still improve in all 12 model–benchmark settings, supporting calibration under competing evidence. *Flat* queries are omitted because SCCI abstains and retains the original context, yielding zero gain by construction. Fig. 5(c)–(d) shows the fraction of queries assigned to each branch. All three regimes appear across models and benchmarks, while non-trivial *flat* coverage confirms that abstention is used under weak evidence. Overall, *decisive* evidence enables the largest gains, *conflict* evidence remains actionable after calibration, and weak evidence triggers abstention.

5.5 Robustness and Complementarity

We further examine whether SCCI generalizes beyond the default 4-shot configuration and whether component-level intervention remains beneficial when combined with demonstration retrieval.

Varying the number of demonstrations. Figs. 6(a)–(b) evaluate SCCI by setting $K_n \in \{2, 4, 6, 8\}$ uniformly across queries for each evaluated shot budget. On Qwen2.5-VL-7B, SCCI improves Vanilla MICL in 22 of 24 shot–benchmark settings; the only negative case is CLEVR at 2-shot, while 8-shot CLEVR is unchanged. On InternVL2.5-8B, SCCI yields positive gains across all shot budgets. These results show that component-level intervention remains effective beyond the default 4-shot setting as the amount of in-context evidence changes.

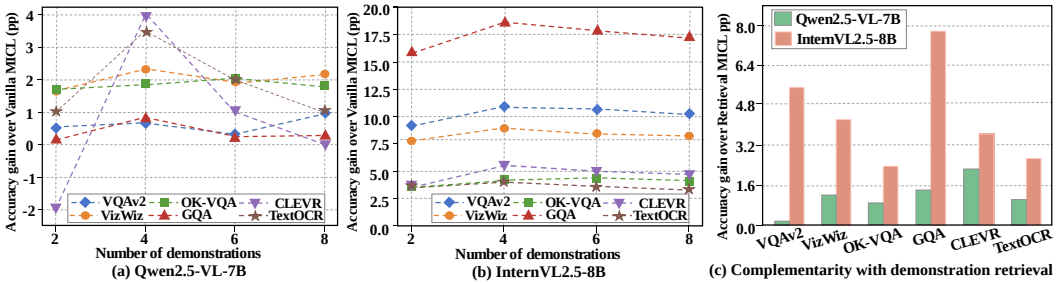


Figure 6: **Robustness and complementarity of SCCI.** SCCI remains effective across demonstration counts and complementary to retrieval-based context optimization.

Complementarity with demonstration retrieval. Fig. 6(c) evaluates SCCI on similarity-retrieved demonstrations. SCCI yields additional gains across six benchmarks and both models, showing that intervention remains useful after demonstration selection. This supports their complementarity: retrieval selects which demonstrations enter the context, while SCCI intervenes on individual image or text components within them. Sampling and ordering robustness are reported in Appendix D.

5.6 Computational Cost and Selected-Context Savings

We examine SCCI's computational trade-off in context savings and probing cost. Fig. 7 reports token and prefill reductions and leave-one-component-out (LOO-CF) overhead across six benchmarks.

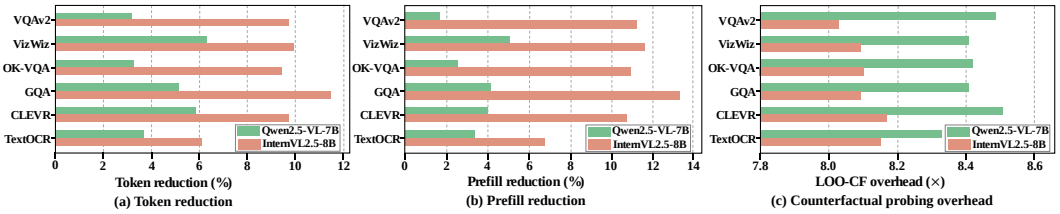


Figure 7: **Computational trade-off of SCCI.** Token and prefill reductions are measured against the full context, and LOO-CF overhead against a single full-context inference.

Selected-context savings. Figs. 7(a)–(b) show that SCCI consistently selects shorter contexts across all six benchmarks. Averaged across benchmarks, token count is reduced by 4.57% on Qwen2.5-VL-7B and 9.40% on InternVL2.5-8B, with corresponding prefill reductions of 3.47% and 10.78%. The larger reductions on InternVL2.5-8B further indicate that the computational effect of component removal depends on the model-specific multimodal context representation.

Counterfactual probing cost. Fig. 7(c) shows that LOO-CF probing dominates the overhead. Under 4-shot, SCCI evaluates the original context and eight single-component counterfactuals, with average overheads of $8.43\times$ on Qwen2.5-VL-7B and $8.11\times$ on InternVL2.5-8B. Routing reuses cached branch outputs without additional decoding. Thus, SCCI trades extra probing cost for selective context simplification rather than end-to-end acceleration. Detailed results are in Appendix E.

6 Conclusion

We show that multimodal demonstrations need not act as atomic units: their image and text components can exert different query-conditioned effects, including *hidden component harm*. Motivated by this observation, SCCI performs selective component-level intervention using target-free counterfactual evidence and reliability-aware routing. Experiments across six benchmarks and two primary LVLMs consistently support the effectiveness, robustness, and complementarity of this design. More broadly, our findings suggest that semantic component granularity is an important but underexplored dimension of MICL context optimization. A remaining challenge is to reduce the cost of counterfactual probing while retaining reliable intervention decisions.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing and improve the readability and presentation of the manuscript. All AI-assisted revisions were reviewed by the authors for technical accuracy and consistency with the intended scientific meaning. Generative AI was not used to produce or modify experimental results or to make scientific conclusions. We take responsibility for the final content of this work, including all text, claims, results, and interpretations.

ETHICS STATEMENT

We are not aware of any significant ethical concerns specific to this work. Our experiments use established vision-language benchmarks and pretrained LVLMs following their standard research protocols, and we do not collect new human-subject data. The proposed method operates only at inference time and does not modify the underlying model parameters. As with existing LVLMs, model predictions may still reflect biases or limitations inherited from the underlying models and datasets.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1539–1550, 2024.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994, 2021.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 782–791, 2021.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision – ECCV 2024*, pp. 19–35, 2024a.

Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. Can multimodal large language models truly perform multimodal in-context learning? In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 6000–6010, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu,

Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267, 2023.

Sivan Doveh, Shaked Perek, M. Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. Towards multimodal in-context learning for vision and language models. In *Computer Vision – ECCV 2024 Workshops*, pp. 250–267, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3608–3617, 2018.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2901–2910, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894, 2017.

Songning Lai, Lijie Hu, Junxiao Wang, Laure Berti-Equille, and Di Wang. Faithful vision-language interpretation via concept bottleneck models. In *International Conference on Learning Representations*, 2024.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 87874–87907, 2024.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7543–7557, 2025a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900, 2022.

Yanshu Li, Jianjiang Yang, Tian Yun, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. TACO: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 736–763, 2025b.

Yanshu Li, Jianjiang Yang, Zhennan Shen, Ligong Han, Haoyan Xu, and Ruixiang Tang. CATP: Contextually adaptive token pruning for efficient and enhanced multimodal in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6619–6627, 2026a.

Yanshu Li, Jianjiang Yang, Ziteng Yang, Bozheng Li, Ligong Han, Hongyang He, Zhengtao Yao, Yingjie Victor Chen, Songlin Fei, Dongfang Liu, and Ruixiang Tang. Make LVLMs focus: Context-aware attention modulation for better multimodal in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6610–6618, 2026b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3190–3199, 2019.

Dang Minh, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen. Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, 55(5):3503–3568, 2022.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671, 2022.

Dianmo Sheng, Dongdong Chen, Zhentao Tan, Qiankun Liu, Qi Chu, Jianmin Bao, Tao Gong, Bin Liu, Shengwei Xu, and Nenghai Yu. Towards more unified in-context visual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13362–13372, 2024.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8802–8812, 2021.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14398–14409, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328, 2017.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31: 841–887, 2017.

Yu Wang and Sharon Li. Why multimodal in-context learning lags behind? unveiling the inner mechanisms and bottlenecks. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13670–13685, 2026.

Shan Xie, Man Luo, Chadly Daniel Stern, Mengnan Du, and Cheng Lu. DemoShapley: Valuation of demonstrations for in-context learning. In *Proceedings of the 2025 IEEE International Conference on Big Data (BigData)*, pp. 4081–4090, 2025.

Nan Xu, Fei Wang, Sheng Zhang, Hoifung Poon, and Muhao Chen. From introspection to best practices: Principled analysis of demonstrations in multimodal in-context learning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3299–3324, 2025.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802, 2025.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3081–3089, 2022.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. MMICL: Empowering vision-language model with multi-modal in-context learning. In *International Conference on Learning Representations*, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706, 2021.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15890–15902, 2024.

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. VL-ICL bench: The devil in the details of multimodal in-context learning. In *International Conference on Learning Representations*, 2025.

APPENDIX

This appendix provides additional analyses, implementation details, and experimental results that complement the main paper:

- **A: Hidden Component Harm Case Study.** We provide an intuitive example illustrating how a demonstration can be beneficial as a whole while containing a harmful visual or textual component.
- **B: Additional Experimental Details.** We provide dataset, baseline, implementation, evaluation, and calibration details for the main experiments.
- **C: Evidence–Utility Alignment Metrics.** We define the metrics used to quantify the alignment between target-free counterfactual evidence and diagnostic component utility.
- **D: Additional Robustness.** We report additional robustness results under different demonstration sampling seeds and demonstration orderings.
- **E: Detailed Computational Cost and Context Savings.** We provide per-benchmark measurements of selected-context token, prefill, and memory savings, together with counterfactual probing overhead.
- **F: Hyperparameter and Calibration Sensitivity.** We analyze sensitivity to the routing hyperparameters and to the amount of held-out calibration data.
- **G: Sensitivity of Non-Atomicity Diagnostics.** We examine the robustness of component utility disagreement and hidden component harm under different diagnostic thresholds.
- **H: Additional Intervention Analysis.** We analyze when intervention helps or harms and compare physical deletion with superficial masking under controlled selection.
- **I: Additional Validation on Idefics2-8B.** We provide additional backbone validation of SCCI on Idefics2-8B, including overall performance, routing, controlled removal, and sampling robustness.

A HIDDEN COMPONENT HARM CASE STUDY

Fig. 8 provides an intuitive example of the hidden component harm condition in Eq. 4. For clarity, we omit the query index and use simplified notation. The target demonstration $D_i = (I_i, T_i)$ contains a harmful image component I_i and a useful text component T_i for the zebra query.

We assign illustrative target-answer probabilities for the fixed answer $y^* =$ Zebra under this simplified counterfactual scoring setup. Keeping the whole demonstration yields an incorrect prediction of *Cat*. Removing the image component increases the target-answer probability and corrects the prediction, corresponding to $U(i, \mathrm{img}) > 0$, whereas removing the text component decreases the probability, giving $U(i, \mathrm{txt}) < 0$. Removing the whole demonstration also decreases the probability, so $U^{\mathrm{demo}}(i) < 0$. Thus, the demonstration is beneficial as a whole while containing a harmful image component: whole-demonstration attribution would retain D_i, whereas component-level intervention can remove only I_i and preserve the useful text. The probabilities and predictions are illustrative; dataset-level prevalence is quantified by $R^{(\epsilon)}_{\mathrm{hch}}$ in Eq. 5.

B ADDITIONAL EXPERIMENTAL DETAILS

B.1 DATASETS

For VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), OK-VQA (Marino et al., 2019), VizWiz (Gurari et al., 2018), and CLEVR (Johnson et al., 2017), demonstrations are drawn from support/training data disjoint from the evaluation queries. We follow the TextOCR task construction of VL-ICL Bench (Zong et al., 2025), which formulates TextOCR (Singh et al., 2021) as an image-to-text in-context learning task. A target text region is highlighted by a red bounding box, and the model is required to transcribe the text inside the box from the preceding demonstrations. Following this protocol, 800 examples from the original TextOCR training split form the support pool and 200 examples from the validation split form the query set. Invalid text annotations and rotated instances

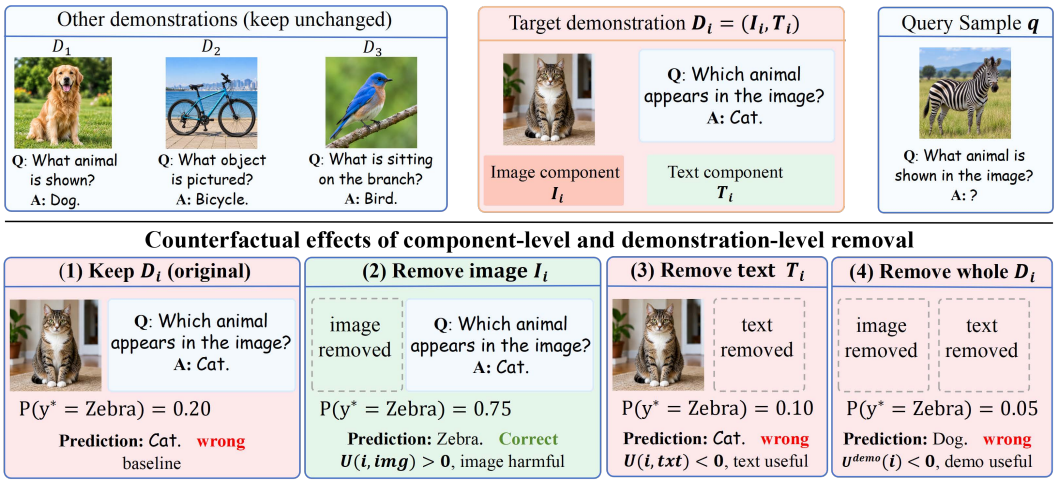


Figure 8: **Illustrative hidden component harm.** Image and text components can have opposing query-conditioned effects. Removing the harmful image corrects the prediction, whereas removing the whole demonstration discards useful context. Probabilities and predictions are illustrative.

are filtered, and when multiple valid text regions occur in an image, the largest region is selected as the target. For all TextOCR experiments, the query set and support pool are fixed.

B.2 Baselines

We compare SCCI with representative approaches operating at different granularities of multimodal in-context optimization. Unless otherwise noted, all methods use the same LVLM, evaluation queries, support pool, 4-shot budget, task prompt, and decoding configuration.

Vanilla MICL uses the unmodified multimodal in-context sequence without context optimization or component intervention. Demonstrations are sampled according to the default sampling protocol used in our main experiments. *RICES* (Yang et al., 2022) constructs a query-dependent context by retrieving four semantically similar image–text demonstrations from the support pool. *MMICES* (Chen et al., 2025) performs multimodal demonstration selection using both visual and textual information. We follow its selection procedure to construct a 4-shot context. *TACO* (Li et al., 2025b) represents sequence-level context configuration, using task-mapping signals and a lightweight task-aware transformer to dynamically configure the MICL sequence. We follow its original configuration procedure whenever applicable. *CATP* (Li et al., 2026a) represents token-level context optimization for MICL. We apply it to the 4-shot context following its original pruning procedure.

Controlled removal baselines. We construct random component removal and matched whole-demonstration removal to isolate the source of SCCI's gains. Random component removal replaces SCCI's counterfactual component selection with random selection under the same intervention opportunities. Matched whole-demonstration removal deletes the complete parent demonstration of the component selected for intervention. These controls therefore isolate component-selection quality and intervention granularity rather than merely the frequency of context modification.

B.3 Implementation and Calibration

SCCI prompt construction and component intervention. For a K-shot prompt, the MICL context consists of an ordered sequence $\mathcal{D}_n = (D_{n,1}, \ldots, D_{n,K})$ followed by query q_n, where each demonstration is $D_{n,i} = (I_{n,i}, T_{n,i})$. The unchanged branch preserves the complete prompt. For an image-component intervention (i, img), only $I_{n,i}$ is physically removed while $T_{n,i}$ and all other demonstrations remain unchanged. For a text-component intervention (i, txt), only $T_{n,i}$ is removed while the associated image is retained. The order of all remaining components is preserved. Whole-demonstration removal is used only for diagnostic analysis and the matched whole-demonstration

removal baseline, and is not part of SCCI's test-time action space. All counterfactual branches use the same task instruction, query, and deterministic decoding configuration.

Evaluation protocol. We report accuracy (%) using the corresponding task-specific answer-matching protocol. Within each model–benchmark setting, all methods use the same evaluation queries, support pool, shot budget, task prompt, and decoding configuration whenever compatible with their original protocols. Controlled intervention comparisons additionally use identical demonstrations, whereas retrieval-based baselines construct their contexts from the shared support pool according to their respective selection strategies. The target-answer score in Sec. 3 is used only for post-hoc diagnosis. Specifically, the diagnostic utility U_n evaluates the ground-truth answer under counterfactual contexts and is unavailable to SCCI during test-time routing. SCCI instead uses the target-free evidence $\Delta_n(a)$ defined in Eq. 7.

SCCI calibration and hyperparameter selection. Calibration is performed on a labeled held-out split disjoint from the reported evaluation queries. For each calibration query, we compute the target-free evidence $\Delta_n(a)$, normalized entropy Ent_n, and peak evidence g_n following Sec. 4.2. The entropy quantile ρ and conflict mixing weight β are the two explicit routing hyperparameters and are selected using calibration accuracy. For each candidate ρ, the entropy threshold is set to $\tau = Q_\rho(\{\mathsf{Ent}_r : r \in \mathcal{Q}_{\text{cal}}\})$; low-entropy calibration queries are then used to estimate the position–modality prior $\overline{\Delta}_{i,m}$, while γ is set to the median positive peak evidence among high-entropy calibration queries. Thus, τ, γ, and $\{\overline{\Delta}_{i,m}\}$ are calibration-derived quantities rather than independently tuned hyperparameters. After selecting ρ and β, all calibration quantities are fixed throughout evaluation. Calibration target answers are used only for hyperparameter selection, not for computing $\Delta_n(a)$ or routing statistics, and no evaluation target answer is used for hyperparameter selection, routing, or intervention. Sensitivity to ρ and β is reported in Appendix F.1.

C Evidence–Utility Alignment Metrics

We provide the formal definitions of the four metrics used in Fig. 4 to quantify the alignment between the target-free evidence $\Delta_n(a)$ and the diagnostic removal utility $U_n(a)$. All metrics are computed post-hoc on the evaluation queries and separately for each model–benchmark setting. The target-answer-dependent utility $U_n(a)$ is used only for analysis and is never available to SCCI during inference. For query q_n, recall that $\mathcal{J}_n = [K_n] \times \mathcal{M}$ denotes the set of single-component removal actions. We compute all alignment metrics over $\mathcal{J}_n$ and exclude the no-intervention action $\varnothing$, since $\Delta_n(\varnothing) = U_n(\varnothing) = 0$ by definition and would otherwise introduce a trivial agreement.

Rank correlation. For each query, we compute the Spearman rank correlation between the target-free evidence and diagnostic utilities over all component-removal actions, which measures the consistency of their induced action rankings: $r_n = \rho_{\text{S}}\left(\{\Delta_n(a)\}_{a\in\mathcal{J}_n}, \{U_n(a)\}_{a\in\mathcal{J}_n}\right)$. The reported rank correlation is the mean over queries,

$$R_{\text{rank}} = \frac{1}{N_{\text{valid}}} \sum\nolimits_{n\in\mathcal{N}_{\text{valid}}} r_n, \tag{12}$$

where $\mathcal{N}_{\text{valid}}$ contains queries for which the Spearman correlation is defined. A larger value indicates better agreement between the action rankings induced by Δ_n and U_n.

Sign agreement. We measure whether the target-free evidence and diagnostic utility assign the same sign to each component-removal action, capturing whether they agree on the direction of the removal effect:

$$R_{\text{sign}} = \frac{1}{N} \sum\nolimits_{n=1}^{N} \frac{1}{|\mathcal{J}_n|} \sum\nolimits_{a\in\mathcal{J}_n} \mathbb{I}\left[\text{sgn}(\Delta_n(a)) = \text{sgn}(U_n(a))\right]. \tag{13}$$

Thus, sign agreement evaluates whether $\Delta_n(a)$ consistently distinguishes removals with positive versus negative diagnostic effects. Exact zeros are assigned sign zero.

Top-action agreement. For each query, let $a_n^{\Delta} = \arg\max_{a\in\mathcal{J}_n} \Delta_n(a)$ and $a_n^{U} = \arg\max_{a\in\mathcal{J}_n} U_n(a)$, with deterministic tie-breaking. Top-action agreement measures whether the two signals select the same highest-ranked component-removal action:

$$R_{\text{top}} = \frac{1}{N} \sum\nolimits_{n=1}^{N} \mathbb{I}\left[a_n^{\Delta} = a_n^{U}\right]. \tag{14}$$

This metric evaluates exact decision agreement, namely whether the component most strongly favored for removal by the target-free evidence coincides with the component whose removal yields the largest diagnostic utility.

Harmful-selection precision. We finally measure how often a component positively favored for removal by the target-free evidence is genuinely harmful according to the post-hoc diagnostic utility, thereby assessing the reliability of positive removal proposals. Define $\mathcal{S}_\Delta^+ = \{n : \Delta_n(a_n^\Delta) > 0\}$. The harmful-selection precision is

$$P_{\text{harm}} = \frac{\sum_{n \in \mathcal{S}_\Delta^+} \mathbb{I}\left[U_n(a_n^\Delta) > 0\right]}{|\mathcal{S}_\Delta^+|}. \tag{15}$$

Because $U_n(a) > 0$ means that removing component a increases the target-answer score, the selected component is diagnostically harmful when this condition holds. The denominator therefore includes only queries for which the target-free evidence positively favors at least one component removal, while the numerator counts those selections confirmed as harmful by the diagnostic utility.

D Additional Robustness

We further evaluate SCCI under changes in demonstration sampling and ordering, complementing the shot-budget and retrieval analyses in Sec. 5.5. The main results report mean accuracy over three demonstration-sampling seeds. We therefore use the same means in the sampling analysis below and additionally report their standard deviations. The ordering analysis instead averages over different demonstration permutations and thus has a different averaging protocol.

Table 4: **Robustness to demonstration sampling across three sampling seeds.** Accuracy (%) is reported as mean$_{\pm\text{std}}$.

Model	Method	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Qwen2.5-VL-7B	Vanilla MICL	$75.37_{\pm 0.11}$	$61.39_{\pm 0.32}$	$55.00_{\pm 0.27}$	$61.88_{\pm 0.33}$	$53.50_{\pm 2.08}$	$45.50_{\pm 2.18}$
	SCCI	$\mathbf{76.05}_{\pm 0.17}$	$\mathbf{63.73}_{\pm 0.22}$	$\mathbf{56.84}_{\pm 0.35}$	$\mathbf{62.72}_{\pm 0.15}$	$\mathbf{57.51}_{\pm 3.51}$	$\mathbf{49.00}_{\pm 1.44}$
InternVL2.5-8B	Vanilla MICL	$61.26_{\pm 0.36}$	$44.08_{\pm 0.87}$	$50.22_{\pm 0.41}$	$30.88_{\pm 0.33}$	$19.00_{\pm 1.15}$	$58.00_{\pm 2.47}$
	SCCI	$\mathbf{72.12}_{\pm 0.22}$	$\mathbf{53.03}_{\pm 0.95}$	$\mathbf{54.32}_{\pm 0.40}$	$\mathbf{49.50}_{\pm 0.33}$	$\mathbf{24.48}_{\pm 2.63}$	$\mathbf{62.01}_{\pm 1.00}$

Robustness to Demonstration Sampling. Table 4 reports the mean and standard deviation across the same three demonstration-sampling seeds used for the main results, while keeping the evaluation queries fixed. Consequently, the reported means match those in the main results, while the standard deviations quantify variation across sampling seeds. SCCI achieves higher mean accuracy on all six benchmarks for both models, showing that its gains are robust to demonstration sampling.

Table 5: **Robustness to demonstration ordering over the original order and two random permutations.** Accuracy (%) is reported as mean$_{\pm\text{std}}$.

Model	Method	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Qwen2.5-VL-7B	Vanilla MICL	$74.97_{\pm 0.35}$	$61.67_{\pm 0.34}$	$54.64_{\pm 0.33}$	$61.75_{\pm 0.12}$	$50.50_{\pm 2.76}$	$46.67_{\pm 1.04}$
	SCCI	$\mathbf{75.76}_{\pm 0.28}$	$\mathbf{63.79}_{\pm 0.25}$	$\mathbf{56.70}_{\pm 0.15}$	$\mathbf{62.51}_{\pm 0.19}$	$\mathbf{55.00}_{\pm 2.36}$	$\mathbf{49.00}_{\pm 0.00}$
InternVL2.5-8B	Vanilla MICL	$61.27_{\pm 0.19}$	$44.51_{\pm 0.40}$	$49.88_{\pm 0.34}$	$30.83_{\pm 0.14}$	$16.83_{\pm 2.93}$	$58.67_{\pm 0.76}$
	SCCI	$\mathbf{72.13}_{\pm 0.16}$	$\mathbf{52.90}_{\pm 0.46}$	$\mathbf{54.07}_{\pm 0.29}$	$\mathbf{49.44}_{\pm 0.30}$	$\mathbf{21.00}_{\pm 2.65}$	$\mathbf{62.33}_{\pm 0.76}$

Robustness to Demonstration Ordering. We further test sensitivity to demonstration ordering by keeping the same four demonstrations and evaluating the original order together with two random permutations. Table 5 reports the mean and standard deviation across these three orderings under this controlled permutation-based evaluation protocol. Unlike the sampling analysis, these means need not match the main results because they are averaged over demonstration permutations rather than demonstration-sampling seeds. SCCI achieves higher mean accuracy on every benchmark for both models, indicating that its gains are not tied to a particular demonstration ordering.

E Detailed Computational Cost and Context Savings

We provide the per-benchmark measurements underlying the computational trade-off summarized in Fig. 7. All results use the default 4-shot setting and the same inference configuration as the main experiments. We distinguish the savings associated with the context selected by SCCI from the additional computational cost required to evaluate counterfactual branches. For each query, the *full context* denotes the unchanged MICL prompt, whereas the *selected context* denotes the branch returned by SCCI after routing. If SCCI abstains, the selected context is identical to the full context. Because the two LVLMs use different native multimodal input representations, their absolute input-token counts and memory footprints are model specific and should not be compared directly across architectures. We therefore report savings relative to each model's own unchanged context.

Let $\overline{T}_{\mathrm{full}}$ and $\overline{T}_{\mathrm{sel}}$ denote the average input-token counts of the full and selected contexts, respectively. Token saving is defined as $100(1 - \overline{T}_{\mathrm{sel}}/\overline{T}_{\mathrm{full}})$. Similarly, given average prefill latencies $\bar{t}_{\mathrm{full}}$ and $\bar{t}_{\mathrm{sel}}$, prefill saving is defined as $100(1 - \bar{t}_{\mathrm{sel}}/\bar{t}_{\mathrm{full}})$. For the memory measurements, we analogously report the relative reduction between the full and selected contexts: $\mathrm{Save}_{\mathrm{mem}} = 100\left(1 - \overline{M}_{\mathrm{sel}}/\overline{M}_{\mathrm{full}}\right)$, where $\overline{M}_{\mathrm{full}}$ and $\overline{M}_{\mathrm{sel}}$ are measured under the same model-specific inference configuration. These quantities characterize the context selected by SCCI and should not be interpreted as end-to-end inference savings. Under 4-shot MICL, SCCI evaluates the unchanged context together with eight single-component counterfactuals, yielding nine prompt variants per query. We therefore report the LOO-CF overhead relative to one full-context inference separately from selected-context savings. Since all branch outputs are cached before routing, selecting and returning the final branch requires no additional decoding. The intervention rate is the fraction of evaluation queries for which the selected action satisfies $a_n^* \neq \varnothing$.

Table 6: **Selected-context token and prefill measurements on Qwen2.5-VL-7B.** Savings are measured relative to the model's own unchanged full context.

Benchmark	Full Input	Selected Input	Token Saving	Full Prefill (ms)	Selected Prefill (ms)	Prefill Saving
VQAv2	370.3	358.4	3.20%	76.76	75.53	1.61%
VizWiz	377.8	353.9	6.32%	79.09	75.08	5.07%
OK-VQA	382.3	369.8	3.26%	79.08	77.08	2.54%
GQA	379.3	359.8	5.15%	78.63	75.35	4.16%
CLEVR	346.0	325.9	5.82%	75.57	72.52	4.03%
TextOCR	392.1	377.8	3.66%	81.46	78.71	3.38%

Table 7: **Selected-context token and prefill measurements on InternVL2.5-8B.** Savings are measured relative to the model's own unchanged full context.

Benchmark	Full Input	Selected Input	Token Saving	Full Prefill (ms)	Selected Prefill (ms)	Prefill Saving
VQAv2	1415.0	1277.2	9.74%	306.01	271.36	11.32%
VizWiz	1425.5	1283.8	9.94%	303.53	268.25	11.62%
OK-VQA	1427.6	1293.1	9.42%	307.44	274.02	10.87%
GQA	1425.0	1261.5	11.47%	306.84	265.97	13.32%
CLEVR	1394.0	1258.7	9.71%	293.09	261.67	10.72%
TextOCR	1426.9	1339.9	6.10%	304.22	283.46	6.82%

Tables 6 and 7 provide the detailed measurements corresponding to Fig. 7. Macro-averaged across the six benchmarks, selected-context token savings are 4.57% and 9.40%, with corresponding prefill savings of 3.47% and 10.78% for Qwen2.5-VL-7B and InternVL2.5-8B, respectively. The different absolute input-token counts reflect model-specific multimodal representations; the relevant comparison is therefore the reduction relative to each model's own full context.

Table 8: **Selected-context memory savings.** Memory reductions are reported relative to each LVLM's own unchanged context rather than as cross-architecture absolute footprints.

Model	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Qwen2.5-VL-7B	3.14%	6.17%	3.34%	5.09%	5.91%	3.76%
InternVL2.5-8B	10.50%	10.62%	10.19%	12.35%	10.30%	6.36%

Selected-context memory savings. Table 8 reports normalized memory savings. Averaged across the six benchmarks, the selected context reduces the measured memory footprint by 4.57% on Qwen2.5-VL-7B and 10.05% on InternVL2.5-8B. The benchmark-wise pattern broadly follows the token and prefill results, with larger reductions on InternVL2.5-8B and smaller savings on TextOCR. As with token and prefill savings, these values describe only the branch ultimately selected by SCCI and not the total memory cost of evaluating all counterfactual branches. As summarized in Fig. 7, the LOO-CF overhead averages $8.43\times$ on Qwen2.5-VL-7B and $8.11\times$ on InternVL2.5-8B, while intervention rates average 61.5% and 63.2%, respectively. Together with the token, prefill, and memory results above, these measurements show that SCCI simplifies the selected context for a substantial fraction of queries while incurring the separate cost of counterfactual probing.

F Hyperparameter and Calibration Sensitivity

We examine SCCI's sensitivity to its routing hyperparameters and the amount of held-out calibration data. SCCI has two routing hyperparameters, the entropy quantile ρ and conflict mixing weight β, while the entropy threshold τ, peak-strength threshold γ, and position–modality prior $\overline{\Delta}_{i,m}$ are derived from the calibration split. We first vary ρ and β individually and then study how performance changes as the available calibration data are reduced. Hyperparameter selection and all calibration-derived quantities are determined using calibration data only and fixed before evaluation.

F.1 Routing Hyperparameter Sensitivity

We vary one hyperparameter at a time around its calibration-selected setting, while keeping the other fixed to its selected value. This one-at-a-time analysis isolates the sensitivity to each routing hyperparameter. The sweeps are used only to assess robustness on the evaluation queries and do not influence hyperparameter selection.

Entropy quantile ρ. The quantile ρ controls the calibration-derived entropy threshold $\tau = Q_\rho(\{\mathsf{Ent}_r : r \in \mathcal{Q}_{\text{cal}}\})$ and thus the partition between decisive and higher-entropy queries. A larger ρ raises τ and admits more queries into the decisive branch, whereas a smaller ρ routes a larger fraction of queries through the conflict or flat branches. We evaluate $\rho \in \{0.05, 0.15, 0.25, 0.35, 0.45, 0.55\}$ for all model–benchmark settings while fixing β to its calibration-selected value. For every candidate ρ, we recompute τ, γ, and $\{\overline{\Delta}_{i,m}\}$ from the held-out calibration split. As shown in Fig. 9, performance is generally stable over a broad range of entropy quantiles on both LVLMs, although several benchmarks exhibit larger task-specific variation. Overall, the results indicate that SCCI does not rely on a narrowly chosen entropy partition.

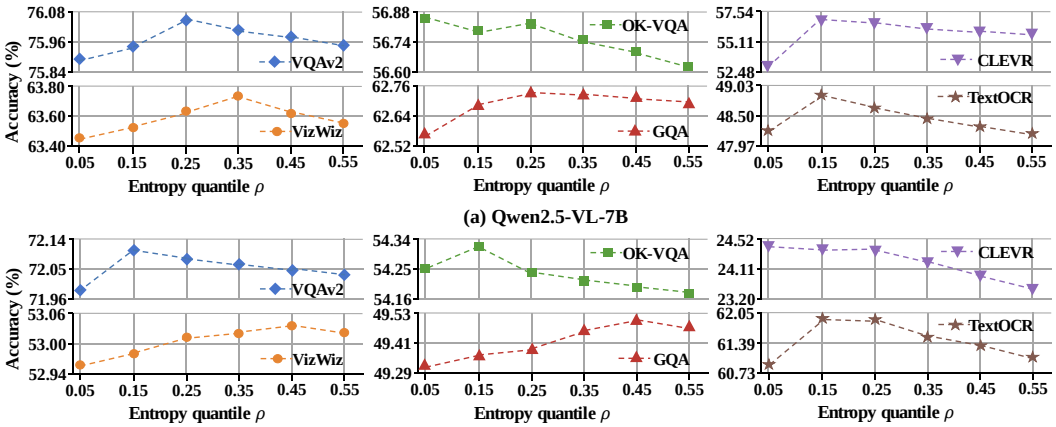


Figure 9: **Sensitivity to the entropy quantile ρ.** SCCI remains stable across a broad range of entropy quantiles on both LVLMs, with larger variation confined to a few model–benchmark settings.

Conflict mixing weight β. The mixing weight β controls the balance between query-specific counterfactual evidence and the calibration-derived position–modality prior in the conflict branch:

$\widetilde{\Delta}_n(i,m) = \beta\Delta_n(i,m) + (1-\beta)\overline{\Delta}_{i,m}$. The two endpoints have direct interpretations: $\beta = 0$ relies entirely on the calibration prior, whereas $\beta = 1$ relies entirely on query-specific evidence. We evaluate $\beta \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ for all model–benchmark settings while fixing ρ to its calibration-selected value. Fig. 10 shows that SCCI maintains competitive performance across a wide range of mixing weights in most settings. The variation on several benchmarks also suggests that the relative contribution of query-specific evidence and the calibration prior is task dependent, motivating calibration-based selection rather than a single manually chosen mixing weight.

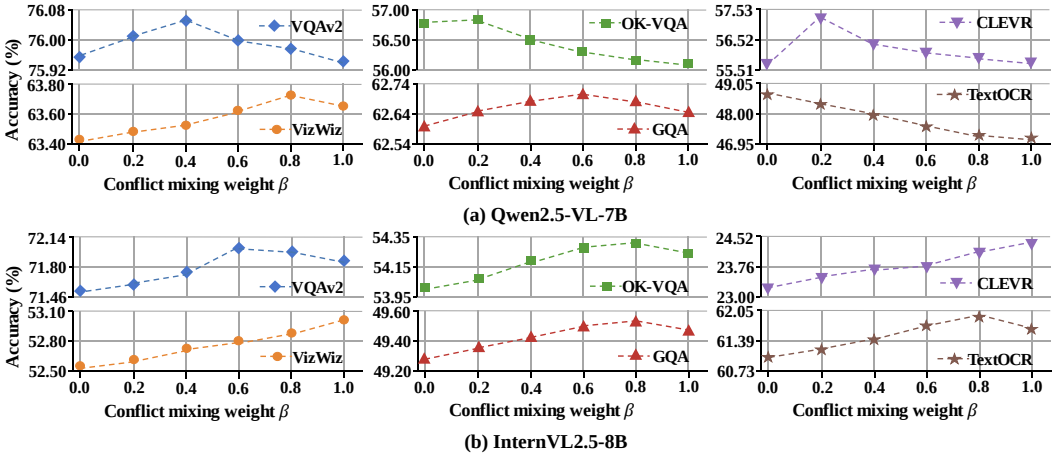


Figure 10: **Sensitivity to the conflict mixing weight β.** Performance is broadly robust to the balance between query-specific evidence and the calibration prior across the two LVLMs.

F.2 Sensitivity to Calibration Size

We next examine SCCI's dependence on the amount of held-out calibration data. We consider calibration fractions of 25%, 50%, 75%, and 100% while keeping the evaluation queries fixed. For each fraction, we repeat the full calibration procedure: ρ and β are selected using the reduced split, while τ, γ, and $\{\overline{\Delta}_{i,m}\}$ are recomputed from the same data. The resulting configuration is then fixed throughout evaluation. Calibration target answers are used only for hyperparameter selection; neither $\Delta_n(a)$ nor evaluation-time routing uses evaluation targets. We repeat random subsampling over three seeds and report mean accuracy.

As shown in Fig. 11, accuracy generally improves with more calibration data, although the magnitude varies across benchmarks and LVLMs. VQAv2 and VizWiz change relatively little, whereas CLEVR is more sensitive for both models, with the remaining benchmarks showing intermediate and model-dependent behavior. In most settings, using 50% or 75% of the calibration split retains much of the full-calibration performance, while the 25% regime shows more visible degradation on sensitive benchmarks. Overall, SCCI remains reasonably robust to moderate reductions in labeled calibration data, while the benefit of additional calibration examples is task and model dependent.

G Sensitivity of Non-Atomicity Diagnostics

The threshold ϵ in Eqs. 3 and 5 controls how aggressively near-zero removal utilities are treated as neutral. Unlike the routing hyperparameters ρ and β, ϵ is used only for post-hoc diagnosis and does not affect SCCI inference. We therefore examine whether the two signatures of demonstration non-atomicity—component utility disagreement and hidden component harm—remain visible after progressively excluding small-magnitude counterfactual effects. We evaluate $\epsilon \in \{0, 0.02, 0.05, 0.10, 0.20\}$ across all six benchmarks and both LVLMs.

Component utility disagreement. For each threshold, we recompute $R_{\text{dis}}^{(\epsilon)}$ using the component removal utilities defined in Sec. 3.2. Increasing ϵ makes the criterion more conservative because an image–text pair contributes only when the two utilities have opposite signs and both lie outside the

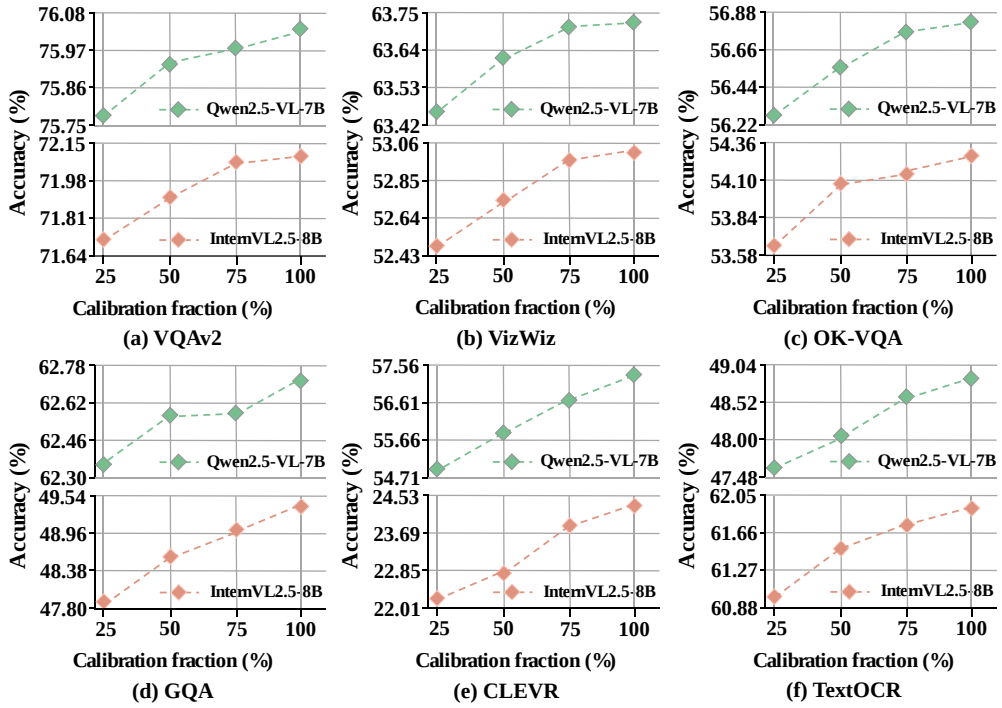


Figure 11: **Sensitivity to calibration-set size.** SCCI accuracy under 25%, 50%, 75%, and 100% calibration fractions across six benchmarks and two LVLMs. Performance generally improves with more calibration data, while moderate reductions retain comparable accuracy in most settings.

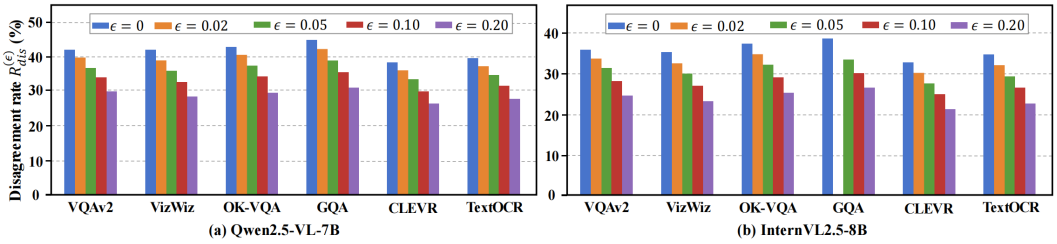


Figure 12: **Sensitivity of component utility disagreement to the diagnostic threshold ϵ.** Larger ϵ values treat near-zero component utilities as neutral, imposing a stricter mixed-sign criterion. The disagreement rate decreases but remains non-zero across both LVLMs and all six benchmarks.

neutral interval $[-\epsilon, \epsilon]$. Hence, $R_{\text{dis}}^{(\epsilon)}$ is non-increasing with ϵ by construction. As shown in Fig. 12, the disagreement rate decreases across all benchmarks as the neutral margin grows. At $\epsilon = 0$, the Qwen2.5-VL-7B values coincide with the strict sign-disagreement statistics in Fig. 2. Importantly, disagreement remains non-zero under positive margins on both LVLMs, indicating that the observed image–text conflicts are not explained solely by utilities arbitrarily close to zero.

Hidden component harm. We next vary the same threshold in the hidden-harm criterion $H_{n,i}^{(\epsilon)}$. Increasing ϵ simultaneously requires a stronger harmful component effect and a stronger beneficial whole-demonstration effect, making $R_{\text{hch}}^{(\epsilon)}$ increasingly conservative. As shown in Fig. 13, the hidden-harm rate decreases as weak cross-granularity reversals are filtered out, while remaining non-zero across all six benchmarks for both LVLMs. This indicates that hidden component harm is not restricted to borderline cases in which either the component-level or whole-demonstration utility is nearly zero; instead, a subset of demonstrations exhibits reversals with non-trivial counterfactual magnitude across diverse tasks and distinct model-specific utility distributions.

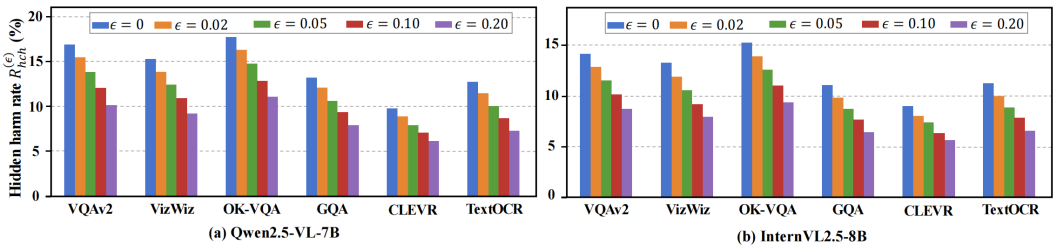


Figure 13: **Sensitivity of hidden component harm to the diagnostic threshold ϵ.** Larger ϵ requires stronger component-level harm and whole-demonstration benefit. The hidden-harm rate decreases under this stricter criterion but remains non-zero across both LVLMs and all six benchmarks.

Overall, both diagnostics decrease monotonically as the neutral margin becomes more conservative, while remaining observable beyond $\epsilon = 0$. These results indicate that demonstration non-atomicity is not merely an artifact of near-zero counterfactual utilities and further support reasoning at component granularity rather than relying only on whole-demonstration attribution.

H Additional Intervention Analysis

We further analyze SCCI's intervention behavior along two complementary dimensions: when component intervention helps or harms, and how a selected component should be modified. We first stratify queries by *vanilla MICL* correctness to characterize the rescue–damage trade-off, then compare physical deletion with superficial masking under matched component selection.

H.1 Intervention Outcomes by Vanilla MICL Correctness

We first examine when SCCI intervention is beneficial by partitioning evaluation queries according to the correctness of the *vanilla MICL* prediction, i.e., the prediction from the unchanged 4-shot context. Let $c_n^{\text{base}} \in \{0, 1\}$ denote the binary correctness indicator assigned using the same answer-normalization and correctness rule as in our evaluation pipeline. We define the *base-correct* subset as $\{n : c_n^{\text{base}} = 1\}$ and the *base-wrong* subset as $\{n : c_n^{\text{base}} = 0\}$. This partition is fixed before evaluating SCCI.

Within each subset, we report the change in benchmark-specific accuracy after applying SCCI relative to vanilla MICL. The base-wrong subset measures SCCI's ability to recover initially incorrect predictions, whereas the base-correct subset measures the risk of disturbing already correct predictions. For benchmarks with soft VQA scoring, the binary stratification criterion differs from the benchmark accuracy metric; hence, subgroup-level vanilla scores need not be exactly 100% or 0%.

Table 9: **Accuracy change stratified by vanilla MICL correctness.** The two subsets are fixed according to the vanilla prediction, and all entries report ΔAcc. (pp) after applying SCCI within the corresponding subset. SCCI improves the base-wrong subset in every setting, while changes on the base-correct subset are non-positive.

Model	Subset (ΔAcc., pp)	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Qwen2.5-VL-7B	Base-correct	-0.95	-2.38	-2.86	-1.65	-6.54	-1.10
	Base-wrong	+5.61	+9.65	+8.64	+4.88	+16.13	+7.34
InternVL2.5-8B	Base-correct	-0.47	-1.80	-1.29	-0.71	-13.16	0.00
	Base-wrong	+28.71	+17.28	+10.33	+27.26	+9.88	+9.52

As shown in Table 9, SCCI exhibits a consistent sign asymmetry between the two subsets. The base-wrong subset improves in every model–benchmark setting, showing that component intervention can recover initially incorrect predictions. In contrast, the base-correct subset shows either a negative change or no change, indicating that intervention can also disturb already successful predictions. The magnitude of this rescue–damage trade-off is model and benchmark dependent: InternVL2.5-8B shows particularly large gains on base-wrong VQAv2 and GQA, whereas CLEVR exhibits the

largest degradation on base-correct queries for both LVLMs. This asymmetry motivates SCCI's selective routing and abstention: intervention can be highly corrective, but indiscriminate intervention risks damaging effective contexts. Having established *when* intervention can help or hurt, we next examine *how* a selected component should be modified.

H.2 Physical Deletion versus Superficial Masking

We next examine whether structural deletion is necessary or whether attenuating a selected component while preserving its prompt position is sufficient. To isolate the intervention primitive, we construct two controlled variants using the same raw-evidence selection rule, $a_n^{\mathrm{raw}} = \arg\max_{a \in \mathcal{A}_n}^{\mathrm{TB}} \Delta_n(a)$. Because $\varnothing$ has score zero, the original context is retained when no component has positive evidence. *Gray/Placeholder* neutralizes the selected component while preserving its prompt position, whereas *Physical Deletion* removes it entirely. Full SCCI instead applies the reliability-aware router in Eq. 11 before deletion. Thus, the controlled pair isolates the intervention operation, while comparison with SCCI additionally reflects selective routing and abstention.

As shown in Table 10, under matched raw-evidence selection, Physical Deletion outperforms Gray/Placeholder in 10 of the 12 model–benchmark settings. The pattern suggests that removing a selected component is generally more effective than attenuating its content while preserving its prompt position. GQA is an exception for both LVLMs, indicating task-dependent differences in the benefit of deletion. Full SCCI further improves over Physical Deletion in all reported settings. Unlike the controlled deletion variant, SCCI accounts for evidence concentration, conflict calibration, and weak-evidence abstention. Thus, structural deletion alone is insufficient: effective intervention depends on both *how* the component is modified and *when* and *which* component is selected. Together with the correctness-stratified analysis, these results support SCCI's combination of physical component deletion and reliability-aware selective routing.

Table 10: **Superficial masking versus physical deletion.** The controlled variants use matched raw-evidence selection, whereas SCCI uses the full reliability-aware router. Accuracy (%).

Model	Method	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Qwen2.5-VL-7B	Gray/Placeholder	75.65	60.98	54.97	62.62	56.00	47.50
	Physical Deletion	75.84	61.83	55.78	62.10	56.76	48.21
	SCCI	**76.05**	**63.73**	**56.84**	**62.72**	**57.51**	**49.00**
InternVL2.5-8B	Gray/Placeholder	68.94	45.88	52.62	49.74	21.73	59.71
	Physical Deletion	70.31	47.41	53.38	49.06	22.84	60.81
	SCCI	**72.12**	**53.03**	**54.32**	**49.50**	**24.48**	**62.01**

I Additional Validation on Idefics2-8B

To assess SCCI's generalizability across LVLM backbones, we conduct additional experiments on Idefics2-8B (Laurençon et al., 2024), complementing the Qwen2.5-VL-7B and InternVL2.5-8B results. We use the same six benchmarks, 4-shot setting, and component-intervention and calibration protocol unless otherwise noted. Rather than repeating the full main evaluation, we focus on whether the key findings transfer to this additional backbone through overall performance, routing behavior, controlled removal comparisons, and robustness to demonstration sampling. The results provide additional evidence that SCCI's benefits extend beyond the two primary LVLMs.

Table 11: **Results on Idefics2-8B under the 4-shot setting.** Accuracy (%).

Method	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Vanilla MICL	77.43	34.74	53.71	51.58	43.49	28.96
SCCI	**78.91**	**39.60**	**54.86**	**54.12**	**47.00**	**31.00**

Overall performance. Table 11 compares SCCI with *Vanilla MICL* under the fixed 4-shot configuration used for the router, controlled-removal, and branch-wise analyses below. SCCI improves performance on all six benchmarks, with gains ranging from 1.15 to 4.86 percentage points. The improvements span visual question answering, compositional reasoning, and scene-text recognition

tasks, indicating that the benefit is not confined to a particular benchmark type. Together with the results on the two primary LVLMs, these improvements show that SCCI retains its effectiveness across multiple LVLM backbones and diverse vision-language tasks.

Table 12: **Router comparison on Idefics2-8B.** Accuracy (%) under the 4-shot setting.

Method	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
PCP	77.85	38.47	54.65	53.74	45.52	29.04
Gated PCP	78.62	38.67	54.67	52.50	45.97	30.03
SCCI	**78.91**	**39.60**	**54.86**	**54.12**	**47.00**	**31.00**

Router comparison. We next examine whether the benefit of reliability-aware routing transfers to Idefics2-8B. *PCP* acts directly on the largest raw target-free evidence, whereas *Gated PCP* additionally restricts intervention using the entropy-based reliability gate. As shown in Table 12, Gated PCP improves over PCP on five of the six benchmarks, while full SCCI achieves the highest accuracy on all six. The GQA exception shows that entropy gating alone is not uniformly beneficial, whereas the complete routing policy remains consistently effective.

Table 13: **Controlled removal comparisons on Idefics2-8B.** Accuracy (%) under matched intervention opportunities.

Method	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Random component	78.09	35.31	53.98	52.66	42.33	28.83
Whole-demo removal	78.38	36.84	54.29	53.17	43.96	29.59
SCCI	**78.91**	**39.60**	**54.86**	**54.12**	**47.00**	**31.00**

Controlled removal comparisons. Table 13 compares SCCI with random component removal and matched whole-demonstration removal. SCCI outperforms both controls across all six benchmarks. Its advantage over random component removal shows that the gains depend on query-conditioned component selection rather than arbitrary deletion, while its advantage over whole-demonstration removal supports the importance of operating at component granularity. Together, these results indicate that SCCI benefits from both selective component identification and fine-grained intervention rather than context removal alone.

Table 14: **Branch-wise performance on Idefics2-8B.** Accuracy (%) within the decisive and conflict branches. The flat branch is omitted because SCCI abstains and retains the original context by construction.

Branch	Method	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Decisive	Vanilla MICL	48.32	4.78	22.46	23.40	37.50	10.00
	SCCI	**53.09**	**23.46**	**28.84**	**32.60**	**50.00**	**17.50**
Conflict	Vanilla MICL	79.45	22.30	48.36	42.49	45.00	22.50
	SCCI	**80.23**	**30.57**	**48.81**	**46.09**	**47.50**	**23.75**

Routing behavior. We further examine SCCI within the two branches where component intervention is applied. As shown in Table 14, SCCI improves over *Vanilla MICL* in both the *decisive* and *conflict* branches across all six benchmarks. The *decisive* branch consistently exhibits larger gains, indicating that more concentrated counterfactual evidence is associated with stronger intervention benefits. The *conflict* branch nevertheless achieves positive improvements throughout, showing that calibrated intervention remains effective under less decisive evidence. We omit the *flat* branch because SCCI abstains and retains the original context by construction. This pattern is consistent with the routing behavior observed on the two primary LVLMs and further supports the transferability of SCCI's reliability-aware routing.

Robustness to demonstration sampling. We additionally evaluate Idefics2-8B across three independent demonstration-sampling seeds to assess sensitivity to the sampled context. Table 15 reports the mean and standard deviation across these seeds. SCCI achieves higher mean accuracy than *Vanilla MICL* on all six benchmarks, with mean gains ranging from 0.76 to 2.37 percentage points. These results show that the improvement remains positive when averaged across independently sampled demonstration contexts.

Table 15: **Robustness to demonstration sampling on Idefics2-8B.** Accuracy (%) across three independent sampling seeds is reported as mean$_{\pm \mathrm{std}}$.

Method	VQAv2	VizWiz	OK-VQA	GQA	CLEVR	TextOCR
Vanilla MICL	$77.33_{\pm 0.25}$	$35.30_{\pm 0.55}$	$54.09_{\pm 0.33}$	$51.69_{\pm 0.14}$	$46.67_{\pm 5.06}$	$28.33_{\pm 1.15}$
SCCI	$\mathbf{78.97}_{\pm 0.31}$	$\mathbf{37.67}_{\pm 1.85}$	$\mathbf{54.85}_{\pm 0.02}$	$\mathbf{53.73}_{\pm 0.38}$	$\mathbf{48.17}_{\pm 3.40}$	$\mathbf{30.00}_{\pm 1.00}$

The means in Table 15 need not match those in Table 11: the latter reports the fixed 4-shot demonstration configuration used for the controlled Idefics2-8B analyses, whereas the former averages over three independently sampled demonstration sets. The difference therefore reflects the sampling and aggregation protocol rather than a change in the method or inference configuration.