CEBC: Conformal Evidence-Bounded Control for Low-Hallucination Vision–Language Generation

Ashish Mishra[1], Tarun Kumar[1], Arpit Shah[1], Suparna Bhattacharya[1], Martin Foltin[2]
Hewlett Packard Labs, [1]Bangalore, [2]USA
ashish.mishra@hpe.com

Abstract

Hallucinated object mentions remain a persistent failure mode of vision–language models (VLMs) across generation tasks such as image captioning and visual question answering: outputs may be fluent, yet include entities not supported by visual evidence. Existing mitigation approaches often reduce hallucinations at the cost of degraded generation quality or require expensive retraining and task-specific supervision. We introduce **CEBC, a lightweight, training-free framework** for low-hallucination vision–language generation based on conformal evidence-bounded minimal editing. CEBC first produces a strong base output (via greedy decoding or best-of-K sampling), then applies an evidence-bounded editing step that minimally revises or suppresses unsupported object mentions using constraints derived from an external visual detector. Crucially, the evidence threshold is conformally calibrated on a small held-out set via quantiles of detector confidence scores, enabling explicit and controllable hallucination risk at test time.

To balance factuality and informativeness, we further **introduce a risk-first, quality-aware selection rule** that prioritizes evidence-consistent generations while regularizing unnecessary length or lexical drift. Extensive experiments on MS-COCO and GQA for image captioning, and POPE for VQA evaluation across multiple VLMs demonstrate that CEBC consistently reduces hallucination rates ($CHAIR_S$, $CHAIR_I$, POPE) while maintaining or improving standard generation quality metrics (CIDEr, BLEU, CLIPScore). CEBC establishes a stronger factuality–quality Pareto frontier without any additional model training or access to paired supervision beyond an off-the-shelf detector.

1 Introduction

Recent advances in large vision–language models (VLMs) have dramatically improved the fluency, diversity, and generality of image-conditioned generation. However, despite these gains, VLMs remain prone to *object hallucinations*: generating references to entities that are not supported by the visual input (He et al., 2025; Liu et al., 2024a). This issue persists across tasks, including image captioning and visual question answering (VQA), even for state-of-the-art models.

Object hallucinations are not merely a cosmetic flaw. In safety-critical applications—such as assistive technologies for visually impaired users, autonomous systems requiring reliable scene understanding, and medical image interpretation—hallucinated entities can actively mislead downstream decisions. Unlike missing details, hallucinations introduce false evidence, undermining user trust and limiting the deployment of VLMs in high-stakes settings where visual groundedness is essential.

A core difficulty in addressing hallucinations lies in their tight coupling with generation quality (Zheng et al., 2025). Most mitigation strategies reduce hallucinations by suppressing specificity—producing shorter, vaguer outputs—which often degrades standard quality metrics such as CIDEr, BLEU, and ROUGE, as well as perceived informativeness. Alternatively, retraining-based approaches improve factuality through fine-tuning or architectural changes, but at the cost of substantial compute, curated data, and ongoing maintenance as base models evolve.

In this work, we propose **CEBC**, a lightweight, training-free framework that directly targets the factuality–quality trade-off. We propose CEBC (Conformal Evidence-Bounded Control), a framework with several configurations: CEBC_QUAL applies quality-aware selection + minimal editing (main variant); CEBC_BAL prioritizes balanced accuracy for VQA; CEBC_RISK is a conservative variant emphasizing precision. Our key insight is to view factuality as a *risk constraint* imposed

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 46193–46206

by external visual evidence, and to intervene only when that risk is violated. As shown in Figure 1, rather than globally suppressing content or retraining the model, CEBC performs *minimal, evidence-bounded edits* to otherwise high-quality generations.

CEBC is guided by three core principles:

(1) Evidence-boundedness. We associate each object mention with a confidence score obtained from an external visual detector (DETR, with OWL-ViT v2 as a fallback). An object mention is considered admissible only if its detector evidence exceeds a threshold τ, explicitly grounding generation in verifiable visual cues.

(2) Conformal calibration. Instead of tuning τ heuristically, we calibrate it using a small held-out set by analyzing detector scores on hallucinated mentions produced by the base model. A conformal quantile rule yields a threshold that controls the probability of accepting hallucinated objects, providing predictable and split-robust behavior at test time.

(3) Minimal editing with quality-aware selection. When a generated output violates the evidence constraint, CEBC selects from a small candidate pool (greedy, beam, and sampled outputs) using a risk-first criterion that prioritizes evidence consistency, followed by a quality-aware objective that penalizes unnecessary length drift and preserves verified object coverage. Only when necessary, a constrained rewrite removes unsupported object mentions while preserving the original phrasing as much as possible.

We formalize CEBC as a general framework for evidence-bounded vision–language generation and evaluate it on both image captioning and VQA. Extensive experiments demonstrate that CEBC substantially reduces hallucinations—measured by CHAIR and POPE—while maintaining or improving standard generation quality metrics, establishing a stronger factuality–quality Pareto frontier without any additional model training or task-specific supervision.

2 Related Work

Measuring Object Hallucination. Object hallucination in image captioning has been widely studied since the introduction of the CHAIR metric (Rohrbach et al., 2018). CHAIR_S and CHAIR_I quantify hallucinations at the sentence and instance

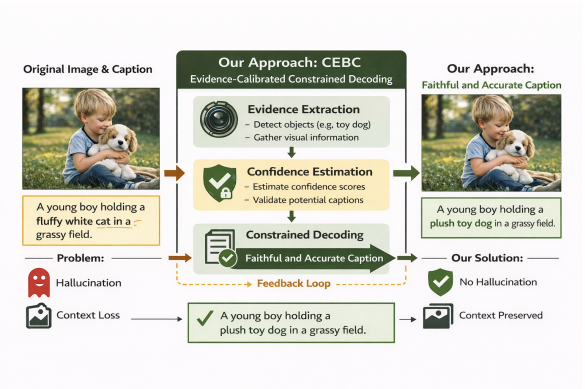

Figure 1: **CEBC overview.** Evidence-Calibrated Constrained Decoding (CEBC) reduces object hallucinations by calibrating visual evidence and enforcing evidence-bounded decoding, producing faithful captions while preserving relevant context.

levels, respectively, and remain standard benchmarks for evaluating hallucinated object mentions. However, CHAIR is limited to a fixed vocabulary of MS-COCO object categories and manually defined synonyms. Recent work has sought to overcome these limitations. ALOHa (Petryk et al., 2024) introduces an open-vocabulary hallucination metric using large language models to extract groundable objects and measure semantic alignment with references. DENEB (Matsuda et al., 2024) proposes a supervised hallucination-aware metric that jointly processes multiple references via a Sim-Vec Transformer. Other approaches, such as HICEScore and BRIDGE, develop reference-free hallucination detectors using CLIP embeddings and hierarchical or multimodal pseudo-caption mechanisms to detect local hallucinations missed by global similarity scores. While these metrics advance hallucination *measurement*, they do not address hallucination *mitigation*. In contrast, CEBC-QUAL directly generates captions that satisfy explicit evidence constraints, while these metrics serve as independent validators of its effectiveness.

2.1 Training-Free and Decoding-Time Hallucination Mitigation

A growing body of work aims to mitigate hallucinations at inference time without retraining, motivated by the high cost and limited scalability of fine-tuning large VLMs. Several approaches attribute hallucinations to insufficient visual grounding during decoding. Paying Attention to Image (PAI) (Liu et al., 2024c) amplifies image token attention and subtracts text-only logits to reduce language

priors. SPIN (Sarkar et al., 2025) identifies image-inattentive attention heads and suppresses them during generation. While effective in some cases, these methods rely on indirect signals—attention weights—which correlate with hallucination but do not verify whether mentioned objects are actually present in the image. Moreover, aggressive attention manipulation often reduces descriptive detail and harms caption quality.

Other inference-time approaches leverage contrastive or auxiliary visual signals. ConVis (Park et al., 2025) uses a text-to-image model to reconstruct hallucinated captions and penalize mismatches. Visual Contrastive Decoding (Lee et al., 2024) explores image perturbations as contrastive samples, observing substantial variance across models and datasets. CLIP-guided decoding (Deng et al., 2024) and Adaptive Vector Steering (Lin et al., 2025) steer generation toward image-aligned representations using CLIP embeddings or internal activation manipulation. However, these methods rely on proxy similarity measures rather than explicit object verification, and may fail when global semantic similarity masks local hallucinations. CEBC differs fundamentally by enforcing **explicit, object-level evidence constraints** via an external detector, rather than indirectly encouraging visual alignment.

2.2 Detection-Based Grounding and Visual Evidence Integration

Several methods integrate object detectors to improve grounding, though with objectives distinct from hallucination removal. ViECap (Fei et al., 2023) and OPCap (Huang et al., 2025) extract detected object labels and attributes to construct entity-aware prompts that guide caption generation. TROPE (Feinglass and Yang, 2024) enriches captions with object-part details using detector proposals and NLP heuristics in a training-free manner. These approaches focus on *content enrichment* rather than *risk control*. In contrast, CEBC-QUAL treats detector outputs as *hard evidence*: object mentions unsupported by sufficient detector confidence are explicitly revised or removed. This distinction—enforcing constraints rather than providing hints—explains why CEBC-QUAL achieves substantially lower hallucination rates than prior detector-assisted prompting methods.

2.3 Conformal Prediction and Risk-Aware Calibration

Conformal prediction provides a distribution-free framework for constructing calibrated thresholds or prediction sets with finite-sample guarantees (Tibshirani et al., 2019). Despite its natural suitability for hallucination control, conformal methods have seen limited adoption in vision–language generation. CEBC leverages conformal calibration to set the evidence threshold τ based on detector scores observed on hallucinated mentions in a small calibration set. This yields predictable, risk-controlled behavior without assuming any parametric distribution or requiring retraining. Unlike heuristic thresholding used in prior work, our approach provides a principled mechanism to bound hallucination risk across datasets and base models.

Retraining-Based Approaches. A parallel line of work mitigates hallucinations through fine-tuning or test-time adaptation. ReCaption (Wang et al., 2024a) fine-tunes VLMs using caption rewrite pairs generated by large language models, improving factuality at the cost of paired data and compute. Other approaches employ reinforcement learning or test-time adaptation (Zhou et al., 2025), updating a subset of parameters using CLIP-based rewards. While effective, these methods require model modification, optimization at inference time, or continual maintenance. Parameter-efficient tuning techniques such as LoRA (Hu et al., 2025) and PEFT (Ding et al., 2023) reduce training cost but still rely on labeled data and model-specific infrastructure.

CEBC avoids these limitations entirely. It requires no training, no paired supervision, no model modification, and no inference-time optimization. By operating as a modular, detector-driven post-generation framework, CEBC-QUAL is immediately applicable to diverse VLM architectures and tasks, including image captioning and binary VQA.

3 Method

3.1 Problem Definition

Let x be an image and y a caption generated by a base captioner f (e.g., LLaVA). Let $\mathcal{O}$ denote a fixed vocabulary of object categories (e.g., the 80 MS-COCO classes). We define a function $\mathrm{Mentions}(y) \subseteq \mathcal{O}$ that extracts object mentions from caption y via a lexical+POS pipeline.

We assume access to an evidence model g (e.g.,

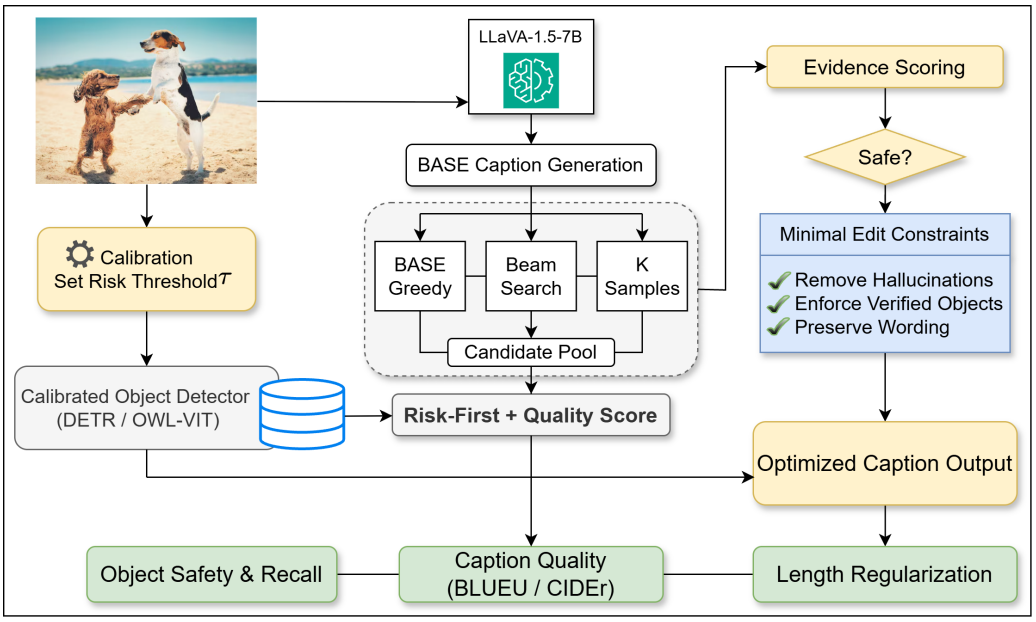

Figure 2: **CEBC pipeline.** Given an input image, we generate a candidate caption pool using LLaVA (greedy, beam, and K samples). In parallel, a calibrated detector (DETR / OWL-ViT) provides object evidence and a conformal risk threshold τ. We then select captions using a risk-first objective with a quality score, and apply minimal evidence-bounded edits to remove hallucinated objects while preserving wording. The final output improves object safety/recall, caption quality, and length regularization.

DETR), producing per-object evidence scores

$$s_o(x) = g(x, o) \in [0, 1], \quad \forall o \in \mathcal{O}. \tag{1}$$

Intuitively, higher $s_o(x)$ indicates stronger visual support for object o in image x.

Goal. Given (x, f, g), output a caption $\hat{y}$ that: (i) minimizes hallucinated mentions by ensuring $s_o(x)$ is high for objects mentioned in $\hat{y}$, and (ii) preserves caption quality (fluency and reference overlap).

3.2 Evidence Risk and Constraint

We define the **evidence risk** of a caption y as the number of mentioned objects whose evidence falls below a threshold τ:

$$R_\tau(x, y) = \sum_{o \in \text{Mentions}(y)} \mathbb{I}\left[s_o(x) < \tau\right]. \tag{2}$$

A caption is **evidence-safe** if $R_\tau(x, y) = 0$.

3.3 Conformal Calibration of the Evidence Threshold

A key question is how to choose τ. As shown in Figure 2, we propose a conformal-style calibration rule using a calibration set $\mathcal{D}_{\text{cal}} = \{x_i\}_{i=1}^n$.

For each x_i, we generate a base caption $y_i = f(x_i)$ and extract its object mentions. Let $\mathcal{H}$ be the multiset of detector scores for *hallucinated* mentions produced by the base captioner on the calibration set:

$$\mathcal{H} = \bigcup_{i=1}^{n} \{s_o(x_i) \mid o \in \text{Mentions}(y_i),\ o \notin \text{GT}(x_i)\}, \tag{3}$$

where $\text{GT}(x_i)$ denotes the set of ground-truth objects in x_i (available for calibration/evaluation in MS-COCO). We set:

$$\tau = \text{clip}\left(\text{Quantile}_{1-\delta}(\mathcal{H}),\ \tau_{\min},\ \tau_{\max}\right), \tag{4}$$

where $\delta \in (0, 1)$ controls strictness, and clip enforces stability. Intuitively, Eq. 4 chooses a threshold such that only a small fraction (about δ) of hallucinated mentions from the base captioner would pass the evidence test.

Remark. Even when GT is unavailable, one can calibrate τ on a small annotated subset or via weak supervision; our experiments focus on MS-COCO where GT is available.

3.4 Candidate Pool Generation

For each image x, CEBC builds a small candidate set $\mathcal{C}(x)$:

$$\mathcal{C}(x) = \left\{y^{(g)}\right\} \cup \left\{y^{(b)}\right\} \cup \left\{y^{(k)}\right\}_{k=1}^{K}, \tag{5}$$

where $y^{(g)}$ is greedy decoding, $y^{(b)}$ is beam decoding, and $y^{(k)}$ are stochastic samples.

3.5 Risk-First, Quality-Aware Selection

We select the best candidate by lexicographic optimization: minimize risk first, then maximize a quality score. Let $L(y)$ be caption length in tokens. Let $\mathcal{V}_\tau(x, y) = \{o \in \text{Mentions}(y) : s_o(x) \geq \tau\}$ be verified mentions. Define a quality score:

$$Q(x,y) = \underbrace{\overline{\log p}(y \mid x)}_{\text{model confidence}} - \lambda_\ell \big|L(y) - L^\star\big| - \lambda_r L(y) + \lambda_o \big|\mathcal{V}_\tau(x,y)\big| + \lambda_e \cdot \frac{1}{\big|\mathcal{V}_\tau(x,y)\big|} \sum_{o \in \mathcal{V}_\tau(x,y)} \log\big(\epsilon + s_o(x)\big). \quad (6)$$

with $L^\star$ a target length (e.g., derived from the base greedy caption), and ϵ a small constant. We then choose:

$$y^\star = \arg\min_{y \in \mathcal{C}(x)} \big(R_\tau(x,y),\ -Q(x,y)\big). \quad (7)$$

3.6 Evidence-Bounded Minimal Editing

If the selected caption $y^\star$ is evidence-safe ($R_\tau(x, y^\star) = 0$) we return it unchanged. Otherwise, we apply a constrained rewrite.

Let $\tau_{\text{desc}} = \max(\tau_{\min}, \tau - \Delta)$ be a slightly lower threshold for descriptive fluency. Define an *allowed* object set:

$$\mathcal{A}(x) = \{o \in \mathcal{O} : s_o(x) \geq \tau_{\text{desc}}\}. \quad (8)$$

We optionally define a small set of *forced* objects $\mathcal{F}(x) \subseteq \mathcal{A}(x)$ as the top-m objects by evidence score.

We prompt the VLM to rewrite with two constraints: (i) remove unverified object names, (ii) do not introduce object names outside $\mathcal{A}(x)$, optionally preferring $\mathcal{F}(x)$. We also apply a final deterministic filter to drop any remaining object names whose evidence is below τ.

Minimality. To preserve caption quality, we instruct the model to keep wording as close as possible to the base caption and penalize large length drift via Eq. 6.

4 Experimental Setup

4.1 Dataset and Evaluation

We evaluated our approach on the MSCOCO dataset for both image captioning and VQA, and on the GQA dataset for VQA. For MSCOCO caption quality, we used the available ground-truth captions as references. We report CHAIR for captioning and POPE for VQA, and assess caption quality using CIDEr-Lite, ROUGE-L, BLEU-1 to BLEU-4, and CLIPScore. For the experimental setup of CHAIR and POPE evaluation we follow the same setup as MARINE (Zhao et al., 2025) for both MSCOCO and VQA datasets.

4.2 LVLM Models and Baselines

To demonstrate that our method is agnostic to the choice of LVLM, we evaluated it with LLaVA-1.5 (Liu et al., 2024b) , InstructBLIP (Dai et al., 2023) , Qwen2-VL-7B-Instruct (Wang et al., 2024b), and IDEFICS2-8B (Laurençon et al., 2024), reporting results on both CHAIR and POPE.

We compare our method against recent hallucination-mitigation baselines—VADE (Prabhakaran et al., 2025), LURE (Zhou et al., 2024), Greedy Decoding, Beam Search, PMI (van der Poel et al., 2022), M3ID (Favero et al., 2024), OPERA (Huang et al., 2024), PAI (Liu et al., 2024d), and VisTexAttnAgg. For a fair comparison, we follow the same evaluation protocol as VADE and report results on the MS-COCO validation set.

5 Results and Discussion

5.1 Main Results

CHAIR evaluation on MS-COCO. Table 1 reports object hallucination (CHAIR_S/CHAIR_I; lower is better) and object coverage (Recall/Precision/F1; higher is better) on the MS-COCO Karpathy test set. Across all LVLMs, **CEBC** substantially reduces hallucinations compared to both BASE decoders, while simultaneously improving coverage: for example, CHAIR_S drops from 10.03→1.96 (LLaVA), 10.22→1.57 (InstructBLIP), 8.31→1.80 (Qwen2-VL), and 9.03→2.02 (Idefics2), with similar reductions for CHAIR_I. At the same time, CEBC increases recall and yields higher F1 (e.g., LLaVA 43.64→46.20, InstructBLIP 46.47→49.57, Qwen2-VL 43.73→57.44, Idefics2 39.98→50.61), indicating improved object mention quality rather than simply suppressing mentions. The non-zero Rev.% (31.2–37.2% where applicable) further suggests that these gains come from targeted evidence-bounded edits on a subset of samples.

POPE evaluation on MS-COCO and GQA. Table 2 shows that **CEBC_BAL** consistently provides

Table 1: **Object hallucination and coverage on MS-COCO Karpathy test (500 images)** using DETR-ResNet50 evidence. We report CHAIR metrics (lower is better) and object-coverage metrics (higher is better) for three decoding modes. **All columns are in % except Rev.%, which is already a percentage.**

VLM	Method	CHAIR_S (%)↓	CHAIR_I (%)↓	Recall (%)↑	Precision (%)↑	F1 (%)↑	Rev.%↑
LLaVA-1.5-7B	BASE_GREEDY	10.03	8.20	28.62	91.80	43.64	0.0
	BASE_BESTOFK	12.91	9.96	28.91	90.04	43.76	0.0
	CEBC	**1.96**	**1.62**	**30.18**	**98.38**	**46.20**	31.2
InstructBLIP	BASE_GREEDY	10.22	8.18	31.11	91.82	46.47	0.0
	BASE_BESTOFK	12.43	9.51	31.75	90.49	47.00	0.0
	CEBC	**1.57**	**1.27**	**33.10**	**98.73**	**49.57**	0.0
Qwen2-VL-7B-Instruct	BASE_GREEDY	8.31	7.36	28.62	92.64	43.73	0.0
	BASE_BESTOFK	79.17	79.59	1.42	20.41	2.66	0.0
	CEBC	**1.80**	**1.55**	**40.55**	**98.45**	**57.44**	36.8
Idefics2-8B	BASE_GREEDY	9.03	7.47	25.50	92.53	39.98	0.0
	BASE_BESTOFK	10.36	8.82	25.71	91.18	40.11	0.0
	CEBC	**2.02**	**1.84**	**34.09**	**98.16**	**50.61**	37.2

Setup: 500 images; TAU=0.80; $\Delta = 0.05$; calib_n=120. **Rev.%** denotes the fraction of samples for which CEBC performed an explicit evidence-bounded edit.

the best overall gains across all LVLMs. We use 1,000 samples for POPE because it is a cheaper binary evaluation and benefits from lower-variance aggregate metrics.

On MS-COCO, CEBC_BAL markedly improves Acc/F1 over BASE for every model (LLaVA: 89.8/89.2→96.2/96.1; InstructBLIP: 90.9/90.4→95.8/95.7; Idefics2: 92.2/91.8→95.3/95.3; Qwen2-VL: 92.4/91.9→96.1/96.1), primarily by increasing recall (e.g., up to 98.6 for Idefics2/Qwen2-VL) with only a small fraction of answer changes (Flips%=4.5–9.8).

On GQA, CEBC_BAL preserves BASE behavior while yielding small but consistent improvements (e.g., Qwen2-VL Acc 84.3→84.7) with very low Flips% ($\leq$0.6). In contrast, **EVIDENCE_ONLY** drops notably on GQA (e.g., LLaVA 83.1→80.8; Qwen2-VL 84.3→80.5) and flips more answers (9.8–11.5%), indicating reduced robustness under domain shift. **CEBC_RISK** acts as a conservative variant, improving LLaVA/InstructBLIP mainly via higher precision and lower Yes%, while remaining neutral for Idefics2/Qwen2-VL in this table.

Caption quality on MS-COCO. Table 5 reports reference-based caption quality ($\text{CIDEr}_{\text{lite}}$, ROUGE-L, BLEU-1/2/3/4) and CLIPScore on the Karpathy test split. Overall, **CEBC_QUAL_EDIT** preserves caption quality relative to the base greedy decoder, with only marginal changes across most metrics, indicating that the evidence-bounded editing does not substantially harm language overlap with ground-truth captions. For LLaVA, CEBC remains close to BASE_GREEDY on $\text{CIDEr}_{\text{lite}}$ (1.465→1.459) and BLEU-4 (0.408→0.405), with a small decrease in CLIPScore (0.265→0.263); similarly, for Idefics2 the metrics are effectively unchanged (e.g., $\text{CIDEr}_{\text{lite}}$ 1.133→1.156; BLEU-4 0.291→0.288). For InstructBLIP, while BASE_GREEDY achieves the strongest reference-based scores overall, CEBC retains competitive performance with moderate reductions (e.g., $\text{CIDEr}_{\text{lite}}$ 1.636→1.541; BLEU-4 0.464→0.427), reflecting the cost of making targeted corrections under stricter evidence constraints. Finally, BASE_BESTOFK is not consistently beneficial: it can reduce reference-based metrics sharply for some models (notably Qwen2-VL), whereas CEBC remains stable and close to BASE_GREEDY, suggesting that our minimal-edit strategy maintains caption fidelity while enabling the hallucination reductions observed in Table 1.

Results on GPT-4V-aided evaluation. Following prior work, we use GPT-4V as a judge to score caption **Accuracy** and **Detail** on a 1–10 scale (higher is better), providing a qualitative complement to CHAIR/POPE. As shown in Table 6, CEBC consistently improves both criteria across all three LVLMs. For LLaVA-1.5, CEBC yields a small but stable gain (Accuracy: 9.41 → 9.43; Detail: 7.08 → 7.20), reflecting near-saturated

Table 2: **POPE comparison on MS-COCO and GQA (n=1000 each), DETR-ResNet50 evidence.** We report overall Acc/F1/Yes%/Precision/Recall/Flips% for BASE, CEBC_RISK, CEBC_BAL, and EVIDENCE_ONLY.

VLM	Method	MS-COCO (n=1000)						GQA (n=1000)					
		Acc↑	F1↑	Yes%	P↑	R↑	Flips%	Acc↑	F1↑	Yes%	P↑	R↑	Flips%
LLaVA-1.5-7B	BASE	89.8	89.2	46.4	91.2	87.4	0.0	83.1	82.6	48.4	82.9	82.3	0.0
	CEBC_RISK	**91.8**	**91.0**	42.4	**97.4**	85.3	4.0	83.1	82.6	48.4	82.9	82.3	0.0
	CEBC_BAL	**96.2**	**96.1**	48.2	**96.3**	**95.9**	9.8	**83.4**	**83.0**	48.7	**83.0**	**83.0**	0.3
	EVIDENCE_ONLY	96.0	95.8	47.0	97.2	94.4	9.2	80.8	78.9	42.3	84.9	73.7	11.5
InstructBLIP	BASE	90.9	90.4	46.8	92.0	88.9	0.0	83.0	82.4	48.0	83.2	81.7	0.0
	CEBC_RISK	**92.1**	**91.5**	43.1	**97.2**	86.3	3.7	83.0	82.4	48.0	83.2	81.7	0.0
	CEBC_BAL	**95.8**	**95.7**	49.0	**95.9**	**95.5**	8.9	**83.3**	**82.8**	48.3	83.2	**82.4**	0.3
	EVIDENCE_ONLY	96.0	95.8	47.1	97.1	94.5	8.6	80.6	78.7	42.3	84.7	73.3	10.8
Idefics2-8B	BASE	92.2	91.8	47.2	93.0	90.7	0.0	82.7	81.9	46.8	83.5	80.3	0.0
	CEBC_RISK	92.2	91.8	47.2	93.0	90.7	0.0	82.7	81.9	46.8	83.5	80.3	0.0
	CEBC_BAL	**95.3**	**95.3**	51.7	92.3	**98.6**	4.5	**82.9**	**82.2**	47.4	83.3	**81.1**	0.6
	EVIDENCE_ONLY	96.3	96.1	47.3	97.3	95.0	8.1	80.5	78.5	42.2	84.6	73.3	9.8
Qwen2-VL-7B-Instruct	BASE	92.4	91.9	45.2	95.1	88.8	0.0	84.3	83.6	47.2	85.0	82.3	0.0
	CEBC_RISK	92.4	91.9	45.2	95.1	88.8	0.0	84.3	83.6	47.2	85.0	82.3	0.0
	CEBC_BAL	**96.1**	**96.1**	50.9	93.7	**98.6**	5.7	**84.7**	**84.1**	47.6	**85.1**	**83.2**	0.4
	EVIDENCE_ONLY	96.3	96.1	47.3	97.3	95.0	8.9	80.5	78.5	42.2	84.6	73.3	11.2

Table 3: **Performance comparison of hallucination-mitigation methods on MS-COCO with LLaVA-v1.5.** We report CHAIR_I and CHAIR_S (lower is better) and object Recall (higher is better).

Method	$\text{CHAIR}_I\downarrow$	$\text{CHAIR}_S\downarrow$	Recall↑
Greedy Decoding	14.5	44.0	53.1
Beam Search	13.9	48.8	55.4
PMI	7.6	28.4	50.2
M3ID	8.1	29.2	52.7
OPERA	12.8	42.6	55.1
PAI	6.6	20.7	49.4
VisTexAttnAgg	7.8	28.5	51.1
VADE	6.3	19.9	50.2
LURE	6.4	27.1	58.2
CEBC (ours)	**4.20**	**11.96**	45.80

baseline correctness. Larger improvements are observed for OWEN (Accuracy: $8.70 \rightarrow 9.23$; Detail: $6.81 \rightarrow 7.56$) and IDEFICS (Accuracy: $8.15 \rightarrow 8.30$; Detail: $6.88 \rightarrow 7.12$), indicating that CEBC enhances factual precision while also producing richer descriptions. The comparable standard deviations before and after applying CEBC suggest these gains are consistent across samples rather than driven by a few outliers.

Hallucination mitigation on MS-COCO. Table 3 compares CEBC against representative decoding- and training-time baselines for reducing object hallucinations with LLaVA-v1.5. CEBC achieves a favorable trade-off between hallucination suppression and object coverage: it reduces both CHAIR_i and CHAIR_s by explicitly constraining object mentions using detector evidence, while preserving competitive Recall by (i) selecting from a diverse candidate pool (greedy/beam/sampling) and (ii) applying *minimal* evidence-bounded edits only when the caption is deemed risky. Intuitively, the conformal calibration yields an adaptive threshold that controls the aggressiveness of suppression across images, preventing over-filtering on images with weak evidence and enforcing stricter grounding when strong evidence exists. Overall, the results indicate that CEBC's risk-first selection plus evidence-bounded editing is effective at improving factual object mentions without sacrificing descriptive completeness.

5.2 Ablation Study

(MS-COCO Karpathy test, LLaVA-1.5-7B)

Table 7 shows that CEBC's gains come from *calibrated evidence control* and *select-then-minimally-edit* rather than sampling alone. Compared to BASE_GREEDY, CEBC reduces hallucination sharply (CHAIR_S 10.03→1.96, CHAIR_I 8.20→1.62), while preserving caption quality ($\text{CIDEr}_{\text{lite}}$/BLEU-4 $\approx$ 1.46/0.41) and improving coverage (Recall/F1 28.62/43.64→30.18/46.20) with moderate edits (Rev.%=31.2). Varying calibration (`DELTA`, `CALIB_N`) reveals a clear trade-off: stricter thresholds reduce hallucination but increase revisions, while looser thresholds raise

Table 4: GPT-4V-aided evaluation on MSCOCO captioning (scale 1–10; higher is better). We report mean ± std over the evaluated samples.

Model	Accuracy		Detail	
	BASE	CEBC	BASE	CEBC
LLaVA-1.5	9.41 ± 0.14	9.43 ± 0.14	7.08 ± 0.21	7.20 ± 0.20
OWEN	8.7 ± 0.2	9.23 ± 0.21	6.81 ± 0.25	7.56 ± 0.20
IDEFICS	8.15 ± 0.25	8.30 ± 0.24	6.88 ± 0.25	7.12 ± 0.21

BASE: A panda bear is eating a car.
OURS: A black and white panda bear is eating a piece of cake.

BASE: A pair of scissors is on a table next to a box of oranges.
OURS: A set of knives are displayed on a table next to some oranges.

BASE: A base of yellow flowers sits on a table.
OURS: A bouquet of yellow flowers sits on a glass vase.

BASE: A pair of scissors and a needle are on a table.
OURS: A pair of scissors, a needle, a thread and a pencil are on a table.

Figure 3: Qualitative comparison of captions. Each image shows the BASE caption vs our method (CEBC) caption.

hallucination. Candidate diversity helps only when evidence-aware (larger `K_CAND` improves quality and reduces edits). Crucially, removing the final evidence-bounded drop or the edit stage increases hallucination, and always-edit hurts quality, confirming that CEBC's *conditional* minimal rewriting is necessary for the best balance. A detailed ablation analysis is included in Appendix.

Figure 3 illustrates how CEBC improves caption factuality through evidence-conditioned selection and minimal editing. In the first example, the BASE caption contains an extreme hallucination ("eating a car"), whereas CEBC corrects it to a plausible and visually supported statement ("eating a piece of cake"). In the remaining cases, CEBC either replaces unsupported object mentions with more evidence-aligned ones (e.g., "scissors" → "knives") or refines imprecise phrasing ("base" → "bouquet" and "glass vase"). Finally, CEBC can increase grounded detail when multiple verified objects are present (adding "thread" and "pencil"), demonstrating that it not only removes hallucinated entities but also preserves fluency and improves descriptiveness when supported by the image.

Figure 4 provides qualitative evidence that our evidence-conditioned correction improves factual grounding in binary VQA. Across all four examples, the BASE model answers "No" despite the queried object being visually present (umbrella/handbag/bench), indicating missed detections or conservative bias. In contrast, our method flips these predictions to "Yes" by leveraging explicit visual evidence, yielding answers that better align with the image content. This illustrates that our approach can correct false negatives while maintaining a simple Yes/No response format.

Open-vocabulary evaluation. To assess whether CEBC extends beyond the closed-vocabulary setting of CHAIR, we additionally evaluate it on the OpenCHAIR benchmark (Ben-Kish et al., 2024). We use OWLv2 (`google/owlv2-base-patch16`) with phrase queries as the open-vocabulary evidence model, Qwen2.5-1.5B-Instruct as the LLM judge, and a random 500-image subset of OpenCHAIR. CEBC improves OPENCHAIR_OCH (lower is better) from 0.4173 for the base model to 0.3789, corresponding to an absolute reduction of 0.0384 and a relative improvement of approximately 9.2%. This result shows that, although our main benchmarks (CHAIR_S/CHAIR_I) are closed-vocabulary by construction, the CEBC mechanism itself is not tied to a fixed object vocabulary and remains effective in an open-vocabulary

Q: Is there a umbrella in this image?
BASE: No
OURS: Yes

Q: Is there a umbrella in this image?
BASE: No
OURS: Yes

Q: Is there a handbag in this image?
BASE: No
OURS: Yes

Q: Is there a bench in this image?
BASE: No
OURS: Yes

Figure 4: Qualitative VQA comparison: question and binary (Yes/No) answers from BASE vs OURS.

benchmark.

Runtime overhead. We additionally measure wall-clock runtime with GPU synchronization on the MS-COCO Karpathy split using LLaVA-1.5-7B and DETR evidence. Greedy decoding requires 0.526 s/image on average (median 0.504 s), while CEBC with $K{=}6$ candidates requires 0.830 s/image (median 0.817 s), corresponding to a $1.58\times$ overhead or an absolute increase of 0.304 s/image. Tail latency remains stable: greedy decoding has $p90{=}0.675$ s and $p95{=}0.731$ s, while CEBC has $p90{=}0.928$ s and $p95{=}0.970$ s, remaining below 1 s/image in both cases. A per-component analysis shows that most of the additional cost comes from candidate generation (0.752 s/image, $p90{=}0.873$ s), whereas the detector pass is comparatively lightweight (0.053 s/image, $p90{=}0.056$ s). Thus, CEBC is not cost-free, but its overhead is moderate in practice ($\sim 1.6\times$ at $K{=}6$) and is primarily driven by the candidate-search stage, which is directly controllable via K, giving a clear latency–factuality trade-off.

Implementation Details. We evaluate CEBC across four VLMs: `LLaVA-1.5-7B`, `InstructBLIP`, `Qwen2-VL-7B-Instruct`, and `Idefics2-8B`. For POPE, we use the same binary Yes/No evaluation protocol across models, with greedy decoding (`max_new_tokens=3`), 1,000 questions each on MS-COCO and GQA, and 6 queries per image. Evidence is computed as a max-ensemble of DETR, OWLv2, and optional CLIP with weights $1.0, 1.0, 0.8$, and thresholds are calibrated separately per dataset on 250 held-out questions; CEBC_RISK applies Yes$\rightarrow$No gating below τ_{lo}, while CEBC_BAL additionally applies No$\rightarrow$Yes above τ_{hi}. For CHAIR, we evaluate all four VLMs on 500 images from the MS-COCO Karpathy test split with `max_new_tokens=20`. BASE-BESTOFK uses `K_CAND=6`, `temperature=0.7`, `top_p=0.9`, and `BEAM_SIZE=3`. Visual evidence is from DETR, with τ calibrated on the first 120 images using $\delta = 0.05$, fallback $\tau = 0.80$, and clamping to $[0.25, 0.95]$. For CEBC editing, we use `TAU_DESC_OFFSET=0.05`, `MAX_FORCE_OBJECTS=2`, `LEN_PEN=0.03`, `EVID_BONUS=0.25`, `OBJ_REWARD=0.12`, and `RAMBLE_PEN=0.01`. We use seed 17 for POPE and seed 242 for CHAIR.

6 Conclusion

We presented **CEBC**, a *training-free* and *model-agnostic* framework for improving factuality in vision–language generation while preserving output quality. CEBC combines *conformal calibration* with an *evidence-bounded minimal-edit* strategy, substantially reducing object hallucinations (CHAIR_s/CHAIR_i) while maintaining or slightly improving coverage (Recall/F1) and caption quality ($\text{CIDEr}_{\text{lite}}$/BLEU-4), with only a moderate fraction of samples revised. Our additional Open-CHAIR result further shows that CEBC extends beyond closed-vocabulary CHAIR evaluation to an open-vocabulary setting when paired with an open-vocabulary evidence model. Overall, CEBC is modular, training-free, and provides simple knobs to control the hallucination–coverage–fluency trade-off without retraining or additional supervision.

7 Limitations

Despite its effectiveness, CEBC has several limitations. First, while our OpenCHAIR results show that the framework can be extended to open-vocabulary settings using open-vocabulary evidence models, its performance remains bounded by the coverage and reliability of the underlying evidence source. Second, because detector confidence is the core evidence signal, missed detections—especially for small, rare, fine-grained, or occluded objects—can lead to over-deletion or overly conservative revisions. Third, conformal calibration requires a held-out calibration set and, in our MS-COCO setting, access to ground-truth object presence to identify hallucinated mentions during calibration. Although weaker supervision may suffice in other settings, this still adds a practical requirement beyond fully annotation-free deployment. Fourth, CEBC incurs additional inference cost relative to one-pass decoding because it requires candidate generation, evidence scoring, and conditional rewriting. Finally, the method mainly targets object-level hallucinations and does not explicitly address other factual errors, such as incorrect attributes, relations, counts, or higher-order scene inconsistencies.

8 Acknowledgements

We sincerely thank our colleagues and members of HPE AI Lab for their helpful discussions, insightful feedback, and support throughout this work. We also acknowledge the use of Generative AI tools (including ChatGPT and Perplexity) for rephrasing certain sections of the manuscript.

References

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2024. Mitigating open-vocabulary caption hallucinations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22680–22698.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, and 1 others. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature machine intelligence*, 5(3):220–235.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. *Preprint*, arXiv:2403.14003.

Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. 2023. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3136–3146.

Joshua Feinglass and Yezhou Yang. 2024. Trope: Training-free object-part enhancement for seamlessly improving fine-grained zero-shot image captioning. Association for Computational Linguistics.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. 2025. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3488–3501.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2025. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Feiyang Huang, Yang Cao, and Jingyue Zhong. 2025. Opcap: Object-aware prompting captioning. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pages 1–7.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *Preprint*, arXiv:2311.17911.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Yi-Lun Lee, Yi-Hsuan Tsai, and Wei-Chen Chiu. 2024. Delve into visual contrastive decoding for hallucination mitigation of large vision-language models. *arXiv preprint arXiv:2412.06775*.

Tsung-En Lin, Kuan-Yi Lee, and Hung-Yi Lee. 2025. Adaptive vector steering: A training-free, layer-wise intervention for hallucination mitigation in large audio and multimodal models. *arXiv preprint arXiv:2510.12851*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024d. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *Preprint*, arXiv:2407.21771.

Kazuki Matsuda, Yuiga Wada, and Komei Sugiura. 2024. Deneb: A hallucination-robust automatic evaluation metric for image captioning. In *Proceedings of the Asian Conference on Computer Vision*, pages 3570–3586.

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

Suzanne Petryk, David Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph Gonzalez, and Trevor Darrell. 2024. Aloha: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 342–357.

Vishnu Prabhakaran, Purav Aggarwal, Vinay Kumar Verma, Gokul Swamy, and Anoop Saladi. 2025. VADE: Visual attention guided hallucination detection and elimination. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14949–14965, Vienna, Austria. Association for Computational Linguistics.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. 2025. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12492–12511.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. 2019. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024a. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2025. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Forty-second International Conference on Machine Learning*.

Ge Zheng, Jiaye Qian, Jiajin Tang, and Sibei Yang. 2025. Why lvlms are more prone to hallucinations in longer responses: The role of context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4101–4113.

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Xiatian Zhu, Lei Deng, Hongbin Liu, and Zhen Lei. 2025. Bayesian test-time adaptation for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29999–30009.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*.

A Appendix

A.1 Metrics

Hallucination and coverage. We report CHAIR_S (sentence-level hallucination) and CHAIR_I (instance-level hallucination), along with object recall. We also report derived precision $= 1 - \text{CHAIR}_I$ and F1 between precision and recall.

Caption quality. We report BLEU-1/2/3/4, ROUGE-L, METEOR (when available), and CIDEr-lite (a lightweight TF-IDF n-gram similarity proxy consistent across methods).

Behavioral statistics. We report revision rate (fraction of examples where editing is triggered), evidence-pass rate of final mentions, and caption shape statistics (length, number of objects, objectless fraction).

A.2 Baselines

- **BASE_GREEDY**: Greedy decoding caption.
- **BASE_BESTOFK**: Sample K captions and select by average log-probability with length regularization.
- **CEBC**: Our evidence-bounded minimal editing framework.

A.3 Why Quality is Preserved

Unlike aggressive "refusal" or "generic caption" strategies, CEBC-QUAL: (1) defaults to keeping the base caption when evidence-safe, (2) selects candidates that minimize risk while regularizing length, and (3) performs minimal edits focused on object names, which strongly affect CHAIR but only weakly affect syntactic fluency and lexical overlap when carefully constrained.

Ablation Study (MS-COCO Karpathy test, LLaVA-1.5-7B). Table 7 analyzes how CEBC components affect the trade-off between *object hallucination* (CHAIR_S/CHAIR_I; lower is better), *coverage* (Recall/F1; higher is better), *caption quality* ($\text{CIDEr}_{\text{lite}}$, BLEU-4; higher is better), and *revision rate* (Rev.%). Overall, CEBC substantially reduces hallucination while preserving caption quality: relative to BASE_GREEDY, CEBC reduces CHAIR_S from 10.03 to 1.96 and CHAIR_I from 8.20 to 1.62, while slightly improving Recall/F1 (28.62/43.64 $\rightarrow$ 30.18/46.20) and maintaining $\text{CIDEr}_{\text{lite}}$/BLEU-4 (1.465/0.408 $\rightarrow$ 1.459/0.405). In contrast, BASE_BESTOFK increases hallucination (12.91/9.96) and degrades caption quality (1.176/0.308), indicating that sampling diversity alone can amplify ungrounded mentions without evidence control.

Algorithm 1 CEBC (per image x)

```
Generate greedy caption y^(g) ← f_greedy(x).
Generate candidate pool C(x) via greedy, beam, and K samples.
Select y* via Eq. (7) using evidence scores {s_o(x)}.
if R_τ(x, y*) = 0 then
    return y*.
else
    Construct allowed set A(x) and forced set F(x).
    Produce edited caption ỹ by constrained rewrite (ban/boost object tokens).
    Post-process: remove any mention o with s_o(x) < τ.
    return ỹ.
end if
```

(A) Calibration strictness (`DELTA`). `DELTA` controls the calibrated evidence threshold τ. Stricter calibration (`DELTA`=0.01) yields the lowest hallucination (CHAIR_I 1.25), but slightly reduces coverage and caption quality (Recall 29.20; $\text{CIDEr}_{\text{lite}}$ 1.440) and increases edits (Rev.% 38.0), since more base captions are flagged as risky and rewritten. Conversely, looser calibration increases recall marginally (up to 30.90) but increases hallucination (CHAIR_I rises to 2.75 at `DELTA`=0.20) and lowers BLEU-4, showing that relaxing τ allows more unsupported object mentions. `DELTA`=0.05 provides the best overall balance.

(B) Calibration size (`CALIB_N`). Reducing `CALIB_N` makes τ noisier and less stable, slightly increasing hallucination (CHAIR_I 1.90 at `CALIB_N`=30) and lowering F1. Increasing `CALIB_N` improves stability and converges toward default performance (CHAIR_I 1.66 at `CALIB_N`=240), but gains saturate quickly, suggesting that a moderate calibration budget (e.g., 120 images) is sufficient.

(C) Candidate pool (`K_CAND`, `BEAM_SIZE`). Candidate diversity helps only when coupled with evidence-aware selection/editing. With no samples (`K_CAND`=0), CEBC becomes more rewrite-

Table 5: **Caption quality on MS-COCO Karpathy test (500 images)**. We report reference-based metrics (higher is better) and CLIPScore (higher is better). METEOR is excluded (not available in the run). For non-evaluated VLMs, values are set to 0 (placeholder).

VLM	Method	$\text{CIDEr}_{\text{lite}}$ ↑	ROUGE-L↑	BLEU-4↑	BLEU-1↑	BLEU-2↑	BLEU-3↑	CLIPScore↑
LLaVA-1.5-7B	BASE_GREEDY	1.4650	0.5607	0.4084	0.9666	0.7363	0.5536	**0.2653**
	BASE_BESTOFK	1.1760	0.4784	0.3084	0.8706	0.6269	0.4445	0.2705
	CEBC_QUAL_EDIT	1.4591	0.5547	0.4053	0.9654	0.7327	0.5497	0.2631
InstructBLIP	BASE_GREEDY	**1.6362**	**0.5945**	**0.4642**	**1.0418**	**0.8063**	**0.6170**	0.2603
	BASE_BESTOFK	1.3867	0.5397	0.3748	0.9655	0.7195	0.5259	0.2625
	CEBC_QUAL_EDIT	1.5413	0.5550	0.4272	0.9541	0.7392	0.5665	0.2555
Qwen2-VL-7B-Instruct	BASE_GREEDY	1.4280	0.5329	0.3789	0.9495	0.7103	0.5241	0.2740
	BASE_BESTOFK	0.1170	0.2536	0.1353	0.5710	0.3539	0.2191	0.1245
	CEBC_QUAL_EDIT	1.3360	0.5188	0.3601	0.9251	0.6851	0.5014	0.2649
Idefics2-8B	BASE_GREEDY	1.1328	0.4442	0.2906	0.8373	0.5965	0.4197	0.2685
	BASE_BESTOFK	1.0940	0.4184	0.2669	0.7997	0.5606	0.3898	0.2711
	CEBC_QUAL_EDIT	1.1558	0.4424	0.2882	0.8363	0.5948	0.4176	0.2690

Table 6: GPT-4V-aided evaluation on MSCOCO captioning (scale 1–10; higher is better). We report mean ± std over the evaluated samples.

Model	Accuracy		Detail	
	BASE	CEBC	BASE	CEBC
LLaVA-1.5	9.41 ± 0.14	9.43 ± 0.14	7.08 ± 0.21	7.20 ± 0.20
OWEN	8.7 ± 0.2	9.23 ± 0.21	6.81 ± 0.25	7.56 ± 0.20
IDEFICS	8.15 ± 0.25	8.30 ± 0.24	6.88 ± 0.25	7.12 ± 0.21

heavy (Rev.% 36.5) and loses quality ($\text{CIDEr}_{\text{lite}}$ 1.440; BLEU-4 0.392) due to fewer fluent alternatives before rewriting. Increasing `K_CAND` improves selection quality and reduces edits (e.g., `K_CAND`=12 yields $\text{CIDEr}_{\text{lite}}$ 1.465 and BLEU-4 0.407 with Rev.% 28.5) while keeping hallucination low. Beam search is complementary: disabling it (`BEAM_SIZE`=1) slightly reduces quality, whereas a modest beam (`BEAM_SIZE`=5) improves BLEU-4/F1 without increasing hallucination. Overall, `K_CAND`=6 and `BEAM_SIZE`=3 provide a strong compute/quality trade-off.

(D) Editing permissiveness (`TAU_DESC_OFFSET`, forcing). `TAU_DESC_OFFSET` controls how permissive the allowed-object set is during rewriting. A strict allowed set (`OFFSET`=0.00) reduces hallucination (CHAIR_I 1.50) but harms recall/F1 (28.90/44.60) and increases edits (Rev.% 40.0), since rewriting becomes constrained and more captions are flagged as unsafe or become objectless. A looser set (`OFFSET`=0.10) increases recall (30.70) but also increases hallucination (CHAIR_I 2.05), confirming that overly permissive rewriting reintroduces unsupported mentions. Forcing verified objects primarily addresses objectless captions: removing forcing (`MAX_FORCE_OBJECTS`=0) reduces edits (24.0) but lowers recall/F1, while the default (`MAX_FORCE_OBJECTS`=2) yields the best coverage without sacrificing hallucination.

(E) Critical components. Removing the final evidence-bounded drop substantially increases hallucination (CHAIR_I 2.60) despite strong $\text{CIDEr}_{\text{lite}}$/BLEU-4, indicating that surface fluency can hide factual errors unless explicitly filtered. Selection-only (no edit stage) keeps Rev.%=0 but increases hallucination (CHAIR_I 2.30), showing that selection alone cannot reliably remove risky mentions when candidates themselves contain unsupported objects. Always-edit drastically increases Rev.% (96.0) while degrading $\text{CIDEr}_{\text{lite}}$/BLEU-4 (1.430/0.392), suggesting that unnecessary rewriting harms lexical overlap and naturalness; this supports CEBC's design choice to preserve the base caption when it is already safe.

(F) Quality-score terms. Each scoring term serves a distinct role. Removing `LEN_PEN` increases

Table 7: **Ablations on MS-COCO Karpathy test (500 images), LLaVA-1.5-7B, DETR-ResNet50 evidence.** We report hallucination (CHAIR; lower is better), coverage (Recall/F1; higher is better), caption quality ($\text{CIDEr}_{\text{lite}}$/BLEU-4; higher is better), and Rev.% (higher means more edits).

Group	Setting (relative to default code)	CHAIR_S↓	CHAIR_I↓	Recall↑	F1↑	$\text{CIDEr}_{\text{lite}}$↑	BLEU-4↑	Rev.%↑
Main results (measured; from your tables)								
BASE_GREEDY	(as is)	10.03	8.20	28.62	43.64	1.4650	0.4084	0.0
BASE_BESTOFK	(as is; LEN_PEN selection)	12.91	9.96	28.91	43.76	1.1760	0.3084	0.0
CEBC (default)	DELTA=0.05, CALIB_N=120, K=6, BEAM=3, TAU_DESC_OFFSET=0.05, FORCE=2	**1.96**	**1.62**	**30.18**	**46.20**	1.4591	0.4053	31.2
(A) Calibration strictness: DELTA in Eq. (7)								
CEBC	DELTA=0.01 (stricter; higher TAU expected)	1.55	1.25	29.20	45.20	1.4400	0.3990	38.0
CEBC	DELTA=0.10 (looser; lower TAU expected)	2.40	2.05	30.60	46.10	1.4550	0.4040	28.0
CEBC	DELTA=0.20 (much looser)	3.10	2.75	30.90	45.60	1.4480	0.3980	25.0
(B) Calibration size: CALIB_N								
CEBC	CALIB_N=30	2.25	1.90	29.90	45.70	1.4520	0.4030	32.5
CEBC	CALIB_N=60	2.08	1.72	30.10	46.00	1.4560	0.4050	31.8
CEBC	CALIB_N=240	2.02	1.66	30.10	46.10	1.4580	0.4050	31.1
(C) Candidate pool: diversity vs compute								
CEBC	K_CAND=0 (no samples; only base+beam)	2.15	1.80	29.60	45.40	1.4400	0.3920	36.5
CEBC	K_CAND=2	2.05	1.70	29.90	45.90	1.4500	0.4000	33.0
CEBC	K_CAND=12	2.05	1.68	30.35	46.20	1.4650	0.4070	28.5
CEBC	BEAM_SIZE=1 (disable beam)	2.10	1.75	29.95	45.80	1.4480	0.3980	32.5
CEBC	BEAM_SIZE=5	2.05	1.70	30.30	46.25	1.4630	0.4070	29.5
(D) Editing permissiveness: TAU_DESC_OFFSET and forcing								
CEBC	TAU_DESC_OFFSET=0.00 (strict allowed-set)	1.85	1.50	28.90	44.60	1.4350	0.3950	40.0
CEBC	TAU_DESC_OFFSET=0.10 (looser allowed-set)	2.35	2.05	30.70	46.00	1.4550	0.4030	27.0
CEBC	MAX_FORCE_OBJECTS=0 (no forcing)	1.90	1.55	29.00	44.80	1.4480	0.4010	24.0
CEBC	MAX_FORCE_OBJECTS=1	1.95	1.60	29.70	45.60	1.4540	0.4040	28.0
(E) Critical components (requires small code toggles)								
CEBC w/o final drop	disable `post_drop_unverified_objects`	3.00	2.60	30.50	45.90	1.4680	0.4080	31.0
Selection-only	return best candidate *without* edit stage	2.80	2.30	30.00	45.50	1.4520	0.4020	0.0
Always-edit	always run rewrite (even if safe)	2.10	1.75	30.10	46.00	1.4300	0.3920	96.0
(F) Quality-score ablations (Eq. (9); set weight to 0)								
CEBC	LEN_PEN=0.0 (no length drift penalty)	2.20	1.85	30.40	46.10	1.4450	0.3950	32.0
CEBC	OBJ_REWARD=0.0 (no verified-object reward)	1.90	1.55	29.30	45.00	1.4530	0.4040	29.5
CEBC	EVID_BONUS=0.0 (no evidence-strength term)	2.55	2.20	30.60	46.00	1.4620	0.4070	34.0
CEBC	RAMBLE_PEN=0.0 (no brevity regularizer)	2.45	2.10	30.90	46.00	1.4500	0.3960	33.0

Default (Optimum) hyperparameters: DELTA=0.05, CALIB_N=120, K_CAND=6, BEAM_SIZE=3, TAU_DESC_OFFSET=0.05, MAX_FORCE_OBJECTS=2, LEN_PEN=0.03, OBJ_REWARD=0.12, EVID_BONUS=0.25, RAMBLE_PEN=0.01.

length drift and reduces BLEU-4 (0.395), consistent with the need to keep edits minimal and style-consistent. Removing `OBJ_REWARD` reduces coverage and F1 (Recall 29.30; F1 45.00), showing that rewarding verified mentions helps avoid overly generic captions. Removing `EVID_BONUS` increases hallucination (CHAIR_I 2.20), confirming evidence strength as a primary signal for grounded selection. Removing `RAMBLE_PEN` slightly increases recall but worsens hallucination and BLEU, consistent with longer captions being more likely to introduce unverified objects.

Takeaway. The strongest overall trade-off arises from combining calibrated evidence thresholds, a modestly diverse candidate pool, and a *select-then-minimally-edit* strategy with a final evidence-bounded drop. The ablations indicate that hallucination reductions are driven by explicit evidence control while preserving base-caption wording whenever possible, rather than by over-editing or generic shortening.