EAG-CAP: EVIDENCE-AWARE GROUNDED CAPTION REVISION WITH VISUAL EVIDENCE SCREENING AND CAPTION-DRIVEN GROUNDING RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Grounded vision-language models connect generated descriptions to image regions, but dense grounded proposals mix useful localized content with unsupported or overly specific claims. Conservative visual evidence screening can clean these proposals, yet it introduces a precision–coverage trade-off: valid caption-relevant regions may be pruned together with noisy ones. We present **EAG-Cap**, a training-free framework that treats proposal phrases as candidate evidence-bearing claims rather than trusted facts. EAG-Cap screens their localized visual support, composes a concise caption under structured evidence constraints, and then uses that final caption to identify groundable concepts that still lack compatible spatial support. Caption-driven grounding recovery selectively re-queries only these missing concepts and merges recovered regions with retained grounding. On the full 5,000-image COCO val2017 split, EAG-Cap obtains 0.839 CIDEr and 18.01% CHAIRs, reducing hallucination relative to Florence-2 Detailed (29.62% CHAIRs) while retaining more annotated content than similarly concise baselines. More importantly, recovery nearly restores the caption-conditioned coverage of raw proposals (90.97% vs. 90.96%) while retaining 84.54% Precision@0.5, compared with 61.45% for raw proposals, and reducing boxes per image from 19.75 to 4.91. A controlled text-only rewrite explains part of the language behavior but provides no caption-aligned grounding output. These results characterize EAG-Cap as a coupled evidence-revision and grounding-recovery framework rather than a caption filter alone.

1 INTRODUCTION

Image captioning has progressed from encoder–decoder systems Vinyals et al. (2015); Xu et al. (2015) to large-scale vision-language models Li et al. (2022a; 2023a). Grounded generation additionally associates phrases with image regions Xiao et al. (2024); Peng et al. (2023), making descriptions more inspectable by exposing where their visual support is claimed to originate.

Grounding, however, does not imply faithfulness. A plausible region may be paired with an incorrect category, overly specific attribute, identity, brand, text, or unsupported relation (Figure 1). We therefore treat grounded proposals as candidate claims and screen their groundable content against the image, rather than interpreting a plausible box as proof of an entire phrase.

Conservative screening creates a second problem: aggressive rejection can remove valid concepts and yield high box precision but incomplete support for the final caption. Grounded caption revision therefore couples faithfulness–informativeness in language with precision–coverage in grounding. Its output should suppress unsupported claims while spatially supporting the groundable concepts it expresses, without exhaustively detecting every object.

Reference-based metrics measure similarity, while CHAIR and POPE diagnose unsupported objects Rohrbach et al. (2018); Li et al. (2023b). Text-only rewriting lacks direct visual evidence, and open-vocabulary grounding Kamath et al. (2021); Li et al. (2022b); Liu et al. (2023b) can localize queries but does not decide which claims belong in the final sentence. We study the complementary setting of post-hoc grounded caption revision: producing an auditable output whose expressed visual

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.

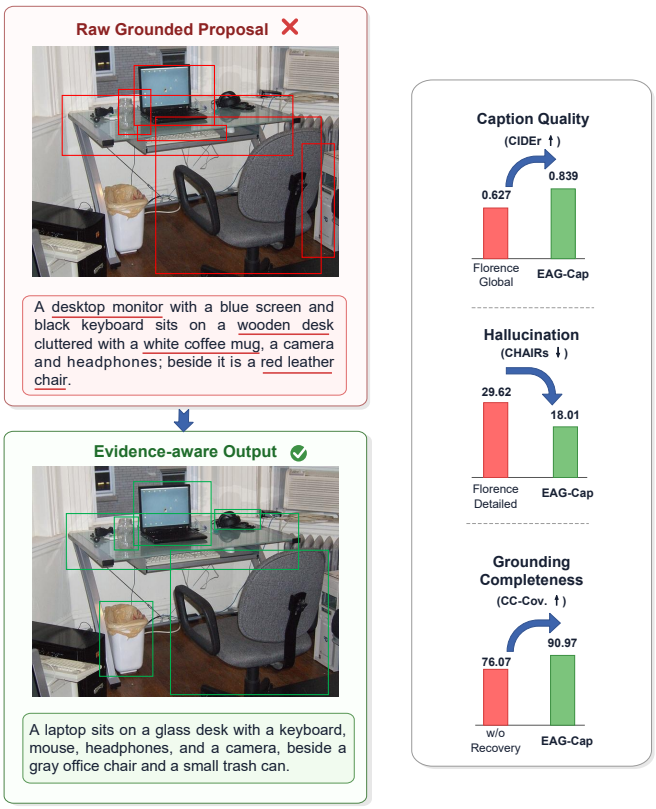

Figure 1: Motivation and effect of EAG-Cap. Raw grounded proposals may attach unsupported or overly specific phrases to plausible regions. EAG-Cap screens localized support for candidate claims, composes an evidence-aware caption, and recovers missing caption-relevant grounding.

concepts retain explicit spatial support. Caption quality alone does not test this evidence-selection and caption-grounding consistency problem.

EAG-Cap is organized around proposal and claim preparation, visual evidence screening, evidence-aware caption composition, and caption-driven grounding recovery. It converts localized support into structured evidence states that constrain revision of a concise caption anchor, then re-queries final-caption concepts lacking retained support. Thus evidence screening determines what may influence the caption, while the resulting caption determines what requires spatial support.

Full COCO val2017 experiments expose both trade-offs. Screening and caption-use filtering raise precision from 61.45% to 88.17% but reduce caption-conditioned coverage from 90.96% to 76.07%. Recovery restores coverage to 90.97% while retaining 84.54% precision with roughly one quarter as many boxes as raw proposals. EAG-Cap is also as concise as Florence-2 Global and Kosmos-2 while retaining more annotated content; a controlled text-only rewrite produces no caption-aligned grounding output.

Our contributions are threefold:

- We formulate grounded caption revision as coupled faithfulness–informativeness and precision–coverage problems, identifying screening-induced over-pruning as a central failure mode.
- We map noisy candidate claims into structured evidence states and compose a concise caption under explicit support constraints.

- We use the final caption as a grounding contract and selectively recover missing regions. On COCO, this raises caption-conditioned coverage from 76.07% to 90.97%, while retaining 84.54% Precision@0.5 versus 61.45% for raw proposals.

2 Related Work

2.1 Image Captioning and Evaluation

Captioning has progressed from encoder–decoder models Vinyals et al. (2015); Xu et al. (2015) to object-aware and pretrained architectures Anderson et al. (2018); Li et al. (2022a; 2023a). COCO, Flickr30k, and NoCaps Lin et al. (2014); Young et al. (2014); Agrawal et al. (2019) are evaluated with reference metrics Papineni et al. (2002); Banerjee & Lavie (2005); Lin (2004); Vedantam et al. (2015); Anderson et al. (2016) and embedding-based scores Hessel et al. (2021); Zhang et al. (2020), none of which establishes whether localized claims have valid support.

2.2 Grounded Vision-Language Models

Instruction-tuned vision-language models Li et al. (2023a); Dai et al. (2023); Liu et al. (2023a); Bai et al. (2023) support broad multimodal generation, while Florence-2 and Kosmos-2 Xiao et al. (2024); Peng et al. (2023) emit phrase-level regions. EAG-Cap uses such outputs as proposals, then separates localized evidence screening, language revision, and final caption-aligned grounding.

GLaMM couples generation with pixel-level grounding Rasheed et al. (2024), while COCONut-PanCap studies fine-grained grounded captioning with panoptic understanding Deng et al. (2025). They provide richer settings for future evaluation beyond COCO and Flickr30k-Entities.

2.3 Open-Vocabulary Grounding and Region-Level Supervision

Visual Genome, DenseCap, Flickr30k-Entities, and referring-expression datasets provide region-level supervision Krishna et al. (2017); Johnson et al. (2016); Plummer et al. (2015); Yu et al. (2016). Open-vocabulary models Kamath et al. (2021); Li et al. (2022b); Liu et al. (2023b) localize natural-language queries beyond fixed categories. EAG-Cap uses grounding first to screen candidate support and later to recover support selected by the final caption.

2.4 Hallucination Evaluation and Correction

CHAIR and POPE expose unsupported content missed by reference similarity Rohrbach et al. (2018); Li et al. (2023b), while Woodpecker and LURE revise decomposed model outputs Yin et al. (2023); Zhou et al. (2023). FaithScore evaluates atomic image facts Jing et al. (2024), and recent hyper-detailed correction work jointly considers factuality and descriptive coverage Lee et al. (2025). EAG-Cap addresses the complementary grounded-output contract: retained caption concepts should preserve spatial support, exposing a precision–coverage failure not measured by text-only factuality scores.

3 Method

3.1 Problem Formulation and Overview

Given an image I, an off-the-shelf grounded proposal model returns image-level captions and localized phrase–box proposals:

$$\mathcal{G}^{\text{img}} = \{g_j\}_{j=1}^{M}, \qquad \mathcal{R} = \{(p_i, b_i)\}_{i=1}^{N}, \tag{1}$$

where g_j is an image-level description and (p_i, b_i) is a candidate phrase with its region. EAG-Cap produces a final caption y and caption-aligned grounding $\mathcal{G}_y$. The latter should spatially support visually groundable concepts actually expressed in y, rather than exhaustively detect all image content.

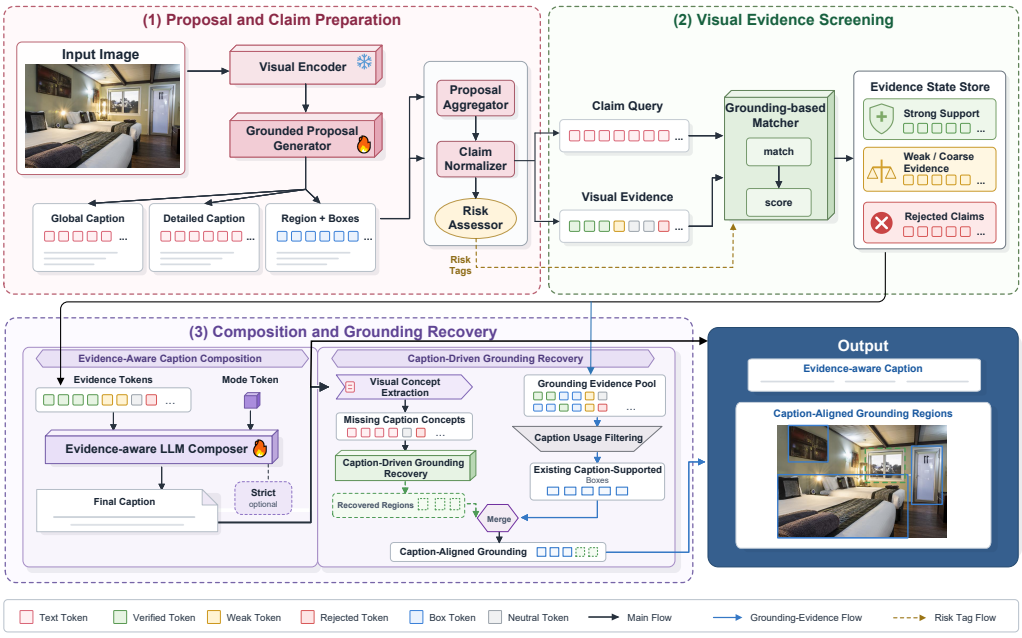

Figure 2: EAG-Cap overview. Proposal and claim preparation produces candidate visual claims; visual evidence screening organizes localized support in the Evidence State Store; evidence-aware composition generates the final caption. Visual concepts in that caption then drive recovery of missing regions, which are merged with Existing Caption-Supported Boxes to form Caption-Aligned Grounding.

The task has two coupled goals. First, revision should improve textual faithfulness without reducing the output to an uninformative object list. Second, grounding should balance precision with caption-conditioned completeness. The proposal set may be rich but noisy, so neither an image-level caption nor a localized phrase is assumed correct merely because it was generated or assigned a plausible box.

3.2 Proposal and Claim Preparation

The proposal stage collects a global caption, a detailed caption, object detections, and dense region captions. Our implementation uses Florence-2 Xiao et al. (2024), although the framework only requires image-level descriptions and region phrase–box candidates. The global caption supplies a concise linguistic anchor, whereas detailed and regional outputs expose additional candidate content.

For each proposal phrase, claim preparation constructs a calibrated visual claim $c_i = \phi(p_i)$ and risk tag $r_i = \rho(p_i)$. The preprocessor derives a conservative *safe query* and, when permitted, a more descriptive *specific query*; punctuation and simple plurals are normalized, while nonvisual wording is removed. Hard-risk patterns cover uncertainty, subjective wording, identity/title expressions, and other poorly verifiable semantics, and a hard-risk claim cannot be upgraded to the specific query. The resulting claim registry stores calibrated claims, risk tags, query forms, source proposals, and candidate boxes. It specifies what should be screened, but does not yet treat any claim as supported evidence or require a dataset-specific object vocabulary.

3.3 Visual Evidence Screening

For each calibrated claim, an open-set grounding model searches the image independently of the proposal box:

$$(\hat{b}_i, s_i) = V(I, c_i), \qquad z_i \in \{\texttt{strong}, \texttt{weak}, \texttt{coarse}, \texttt{rejected}\}. \tag{2}$$

Here $\hat{b}_i$ is the independently localized evidence region, s_i is its grounding score, and z_i is its discrete evidence state. The screening operator tests localized visual support for groundable claim content rather than certifying the full semantic truth of arbitrary attributes, relations, identities, counts, brands, or text. Semantics that cannot be reliably localized or are considered high-risk are conservatively assigned weak or rejected evidence states. These records form the **Evidence State Store**, an auditable structured interface that connects localized visual support with downstream caption composition. Strong grounding-supported evidence can directly constrain caption content; weak or coarse evidence is used cautiously; rejected claims are prohibited. Accepted records form the Grounding Evidence Pool $\mathcal{B}_{\mathcal{E}} = \{(c_i, \hat{b}_i, s_i) : z_i \neq \texttt{rejected}\}$. We instantiate V with GroundingDINO Liu et al. (2023b), using a box threshold of 0.30 and a text threshold of 0.25. An accepted detection with normalized area above 0.65 is marked `coarse`; otherwise, a score of at least 0.50 is `strong`, and a lower accepted score is `weak`. Unsupported or hard-risk unmatched claims are `rejected`.

Visual evidence screening is deliberately conservative. It separates grounding-supported content from proposal-only language and improves localization precision, but it can also reject valid regions because of query mismatch, granularity, or confidence. This trade-off motivates the recovery stage in Section 3.5 rather than treating the screened pool as a complete account of image content.

3.4 Evidence-Aware Caption Composition

The composer revises the global caption under the Evidence State Store. In the default balanced policy, the global caption is the linguistic anchor and the detailed caption is optional context rather than trusted text. Strong grounding-supported evidence may correct or lightly enrich the sentence, rejected claims cannot be mentioned, and weak or coarse evidence is admitted only when low-risk and visually plausible. The objective is a concise natural caption that suppresses risky detail without collapsing to an object list. We instantiate the composer C with Qwen-3B-Instruct using deterministic decoding. It operates on the structured evidence state together with the global-caption anchor and requires no task-specific training. We write

$$y = C(g_{\text{global}}, g_{\text{detailed}}, \mathcal{E}, m), \tag{3}$$

where $\mathcal{E}$ is the Evidence State Store and m is the generation policy. An optional stricter policy is available for applications that prioritize conservatism over descriptive richness; it is not the focus of the main experiments.

3.5 Caption-Driven Grounding Recovery

Caption Usage Filtering over the Grounding Evidence Pool yields a precise but potentially incomplete grounding set. Because the final caption determines which visual concepts ultimately require spatial support, EAG-Cap closes the revision loop by querying only caption concepts that lack compatible retained grounding. Under the deterministic *supplement* policy, it first retains the caption-supported boxes selected from the screened pool:

$$\mathcal{G}_{\text{exist}}(y) = \{\hat{b}_i : (c_i, \hat{b}_i, s_i) \in \mathcal{B}_{\mathcal{E}}, c_i \text{ is compatible with } y\}. \tag{4}$$

Visual Concept Extraction obtains short noun/entity phrases from y, augmented by the composer's recorded used claims and phrase-level evidence explicitly contained in the caption. Phrases are lowercased, punctuation and determiners are removed, simple singular/plural variants and a small set of generic lexical equivalents are normalized, and each query is limited to four words. Semantically redundant variants are collapsed, with at most 12 queries per image. A query is compatible with existing grounding when their normalized tokens match exactly or share the same lexical head with token containment or high overlap; raw substring matching is not used. Let $\mathcal{Q}_{\text{miss}}(y)$ contain the Missing Caption Concepts for which no such existing match is found. Caption-Driven Grounding Recovery re-queries only these concepts with GroundingDINO:

$$\mathcal{G}_{\text{recover}}(y) = \text{Ground}(I, \mathcal{Q}_{\text{miss}}(y)). \tag{5}$$

We use the same 0.30 box and 0.25 text thresholds as evidence screening and retain the highest-confidence valid Recovered Region for each missing query. Supplementation preserves Existing Caption-Supported Boxes except for exact duplicate removal. Query-level synonym variants are

removed before inference, and exact duplicates after merging are identified by normalized phrase tokens and box coordinates rounded to three decimals. Caption-Aligned Grounding is

$$\mathcal{G}_y = \mathrm{Dedup}(\mathcal{G}_{\mathrm{exist}}(y) \cup \mathcal{G}_{\mathrm{recover}}(y)) \,. \tag{6}$$

Recovery uses only the input image, final EAG-Cap caption, existing evidence, and pretrained GroundingDINO. It never accesses reference captions, GT labels or boxes, or evaluation annotations. Thus it restores caption-aligned support without turning the task into exhaustive detection or adding learned parameters.

3.6 Training-Free Inference

All models are frozen, and no task-specific training is required. The complete inference procedure is summarized in Algorithm 1 in Appendix A.1.

4 Experiments

4.1 Experimental Setup

Our main COCO evaluation uses the full 5,000-image COCO val2017 split (COCO-5K) Lin et al. (2014). We retain Flickr30k-Entities1000, a 1,000-image Flickr30k-Entities test subset Plummer et al. (2015), for phrase–region evaluation. A fixed COCO-2K subset is used only for controlled diagnostics, including text-only Qwen rewriting and threshold analysis; its values are never mixed with COCO-5K results without an explicit label. COCO supports caption metrics, CHAIR, and category-aware grounding diagnostics, whereas Flickr30k-Entities supplies phrase-level entity regions. We use Florence-2 as the grounded proposer, GroundingDINO as the open-set evidence screener and recovery model, and a local Qwen-3B-Instruct composer. Unless stated otherwise, EAG-Cap denotes the default balanced policy with a global-caption anchor and deterministic supplement recovery. Generation uses temperature 0.0. All experiments run on one NVIDIA RTX 3090 GPU with 24GB memory.

EAG-Cap trades additional inference cost for evidence auditing and caption-aligned grounding. On a 100-image profiling subset using one RTX 3090, the full pipeline requires 9.41 s/image with 6.21 GiB peak incremental GPU memory, compared with 0.79 s/image and 3.62 GiB for Kosmos-2. Proposal generation (3.77 s/image), caption composition (3.27 s/image), and evidence screening (1.83 s/image) dominate latency. Caption-driven recovery adds 0.50 s/image, about 5.4% of total runtime. EAG-Cap is therefore substantially slower than a single-pass baseline and prioritizes auditable caption–grounding consistency rather than single-pass efficiency; Appendix A.2 reports the full breakdown.

Caption quality is measured with BLEU-4, METEOR, ROUGE-L, CIDEr, SPICE, and CLIPScore Papineni et al. (2002); Banerjee & Lavie (2005); Lin (2004); Vedantam et al. (2015); Anderson et al. (2016); Hessel et al. (2021); CHAIRs and CHAIRi Rohrbach et al. (2018) assess COCO object hallucination. Grounding Precision@0.5 is the fraction of evaluable predicted boxes correctly localized under semantic compatibility. Caption-conditioned coverage (CC-Cov.) restricts the denominator to evaluable GT concepts explicitly mentioned by the generated caption, and caption-conditioned IoU (CC-IoU) averages their best compatible IoUs. We report micro CC-Cov. in the main tables. For COCO, boundary-aware CHAIR aliases map caption mentions and grounding phrases to categories; each caption-mentioned category is one diagnostic unit and may match any compatible annotated instance rather than requiring a singular mention to cover every same-category instance. For Flickr30k-Entities, normalized phrase matching requires token containment or sufficient token overlap together with lexical-head compatibility. Appendix F provides the complete matching rules. CC-Cov. is not standard detection recall. All-GT coverage is retained only as a secondary diagnostic because captions are not expected to mention every annotated object.

4.2 Caption Quality and Faithfulness

Table 1 shows complementary trade-offs rather than a uniform ranking. Kosmos-2 is a strong concise baseline with higher CIDEr and CLIPScore, while Florence-2 Global has the lowest object hallucination rate. Relative to Florence-2 Detailed, EAG-Cap raises CIDEr from 0.005 to 0.839 and lowers

Dataset	Method	BLEU-4 ↑	METEOR ↑	ROUGE-L ↑	CIDEr ↑	SPICE ↑	CLIPScore ↑	CHAIRs ↓	CHAIRi ↓
COCO-5K	Kosmos-2	0.240	0.266	0.485	**0.916**	0.202	**0.2808**	15.14	10.47
	Florence-2 Global	**0.289**	0.261	**0.534**	0.627	0.170	0.2616	**12.55**	**8.62**
	Florence-2 Detailed	0.096	0.217	0.303	0.005	0.184	0.2675	29.62	17.47
	EAG-Cap	0.260	**0.288**	0.517	0.839	**0.227**	0.2629	18.01	11.68
Flickr30k-E1000	Kosmos-2	0.214	0.245	0.470	0.701	0.164	**0.2897**	–	–
	Florence-2 Global	**0.303**	**0.261**	**0.535**	**0.882**	**0.187**	0.2656	–	–
	Florence-2 Detailed	0.089	0.222	0.307	0.019	0.183	0.2751	–	–
	EAG-Cap	0.224	0.251	0.475	0.638	0.185	0.2680	–	–

Table 1: Caption quality, alignment, and hallucination results on full COCO val2017 and Flickr30k-E1000. Grounding recovery occurs after caption generation and therefore does not alter these values. Bold and underlined entries are the best and second-best results within each dataset, respectively.

Method	Words	Unique words	Content words	Obj. mentions	Mention cov. (macro)	Mention cov. (micro)
Florence-2 Global	13.739	10.276	6.917	1.267	0.4867	0.3957
Florence-2 Detailed	44.178	30.256	21.343	1.891	0.5878	0.5333
Kosmos-2	13.030	11.408	6.462	1.284	0.4763	0.3928
EAG-Cap	13.580	11.721	**7.712**	**1.470**	**0.5236**	**0.4436**

Table 2: Caption informativeness diagnostics on COCO-5K. Object mentions and mention coverage use lexical matches to annotated COCO categories; they are transparent content diagnostics, not learned semantic metrics. Bold marks the best value among the similarly concise methods (Florence-2 Global, Kosmos-2, and EAG-Cap).

CHAIRs from 29.62% to 18.01%, while retaining more annotated content than similarly concise baselines (Table 2). The following grounding analysis tests whether caption-mentioned concepts also retain spatial support.

4.3 Informativeness and Faithfulness Trade-off

At 13.58 words, EAG-Cap is similar in length to Florence-2 Global and Kosmos-2 but contains more unique content words and covers more annotated categories (Table 2). Florence-2 Detailed obtains higher lexical coverage with 44.18 words but also 29.62% CHAIRs, so mention coverage must be interpreted with hallucination and grounding quality.

4.4 Grounding Precision and Coverage

Table 3 isolates where coverage is lost and recovered. Open-set evidence screening raises Precision@0.5 from 61.45% to 83.92%, but CC-Cov. drops from 90.96% to 80.68%. Caption Usage Filtering further raises precision to 88.17% while reducing coverage to 76.07%. Caption-driven recovery reverses this over-pruning: the full method reaches 90.97% CC-Cov. and 0.844 CC-IoU, nearly matching raw proposals (90.96% and 0.844), while retaining 84.54% precision and reducing boxes per image from 19.75 to 4.91. The result exposes the screening-induced precision–coverage trade-off and explains why precision alone is insufficient for caption-aligned grounding.

Direct grounding of Florence-2 Global is a strong control, with 83.86% precision, 85.13% CC-Cov., and 0.793 CC-IoU. EAG-Cap attains comparable precision (84.54%) but higher CC-Cov. (90.97%) and CC-IoU (0.844), suggesting a benefit from combining final-caption queries with retained proposal evidence rather than re-grounding from scratch.

4.5 Cross-Dataset Grounding Analysis

Flickr30k-Entities shows the same trade-off (Table 4): recovery raises CC-Cov. from 44.07% to 78.66% while precision changes from 92.73% to 87.39%. Hybrid obtains 87.21% P@0.5, 78.82% CC-Cov., and 0.730 CC-IoU, compared with 87.39%, 78.66%, and 0.729 for supplement. Their near-equivalent behavior motivates the simpler supplement policy as the default (Appendix G).

Stage	Boxes/img	P@0.5 ↑	CC-Cov. ↑	CC-IoU ↑
Raw Proposals	19.751	61.45	90.96	**0.844**
Evidence Pool	4.160	83.92	80.68	0.751
w/o Grounding Recovery	2.313	**88.17**	76.07	0.708
EAG-Cap (Full)	4.907	84.54	**90.97**	**0.844**

Table 3: Stage-wise grounding analysis on full COCO val2017. All rows use the same final EAG-Cap caption for caption-conditioned evaluation. Visual evidence screening and Caption Usage Filtering improve precision but over-prune caption-relevant regions; recovery restores their coverage. P@0.5 and CC-Cov. are percentages.

Grounding stage	P@0.5 ↑	CC-Cov. ↑	CC-IoU ↑
Raw Proposals	52.97	60.09	0.572
Screened Evidence Pool	91.04	47.18	0.433
w/o Grounding Recovery	**92.73**	44.07	0.404
EAG-Cap (Full)	87.39	**78.66**	**0.729**

Table 4: Flickr30k-E1000 grounding-stage analysis with a fixed final-caption denominator. Visual evidence screening improves precision but removes support; supplement recovery restores missing caption concepts. CC-Cov. and P@0.5 are percentages.

4.6 Does EAG-Cap Go Beyond Text-Only Rewriting?

With the same deterministic language model, EAG-Cap obtains 0.868 versus 0.612 CIDEr and 17.79% versus 19.26% CHAIRs (Table 5). The text-only control explains part, but not all, of the caption behavior and produces no caption-aligned grounding output. It does not by itself establish that every pipeline component is necessary.

Method	BLEU-4 ↑	METEOR ↑	ROUGE-L ↑	CIDEr ↑	SPICE ↑	CHAIRs ↓	CLIPScore ↑
Qwen Rewrite (Global + Detailed)	0.191	0.278	0.480	0.612	0.222	19.26	**0.2682**
EAG-Cap	**0.264**	**0.290**	**0.523**	**0.868**	**0.230**	**17.79**	0.2640

Table 5: Controlled text-only rewriting diagnostic on the fixed COCO-2K subset. Both rows use the same local Qwen-3B-Instruct checkpoint and deterministic decoding; the rewrite baseline receives no image, boxes, claims, risk tags, or evidence states.

4.7 Robustness to Grounding Thresholds

On the fixed COCO-2K diagnostic subset, lowering the GroundingDINO box threshold from the default 0.30 to 0.25 slightly increases caption-conditioned coverage at a modest precision cost, whereas raising it to 0.35 slightly increases precision while reducing coverage. Varying the text threshold from 0.20 to 0.30 produces little change in the tested configuration. We therefore retain box threshold 0.30 and text threshold 0.25 for all main experiments and do not tune either threshold using evaluation annotations.

4.8 Qualitative Analysis

Figure 3 illustrates the complementary roles of visual evidence screening and caption-driven grounding recovery. Screening removes unsupported proposal content, while recovery restores spatial support for caption-relevant concepts after composition.

5 Limitations

EAG-Cap depends on its proposer, composer, and grounding model. Localized support does not prove arbitrary attributes, relations, counts, identities, brands, or text, and COCO/Flickr primarily

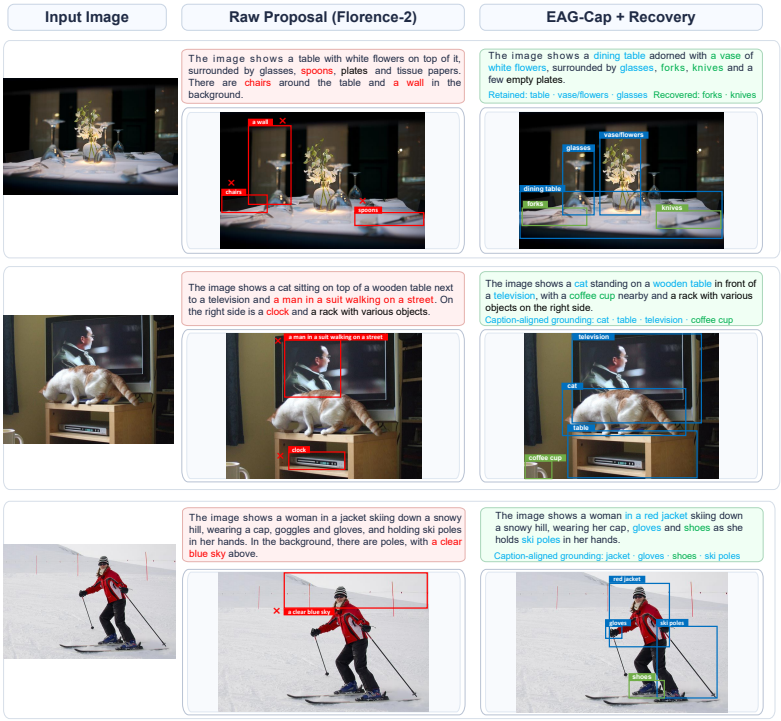

Figure 3: Qualitative examples of EAG-Cap. Red boxes and phrases denote unsupported proposal content removed by visual evidence screening; solid blue boxes indicate retained caption-supported grounding, and dashed green boxes denote regions restored by caption-driven grounding recovery.

test object- and entity-level content. Recovery cannot restore information absent from the final caption or concepts the grounding model cannot localize. CC-Cov. is a diagnostic rather than standard detection recall. The sequential pipeline also costs more than single-pass captioning, and we do not claim computational parity. Our baselines do not exhaust modern VLMs, grounded systems, or correction methods; broader backbones and richer panoptic or hyper-detailed benchmarks remain future work.

6 Conclusion

EAG-Cap treats grounded proposals as candidate evidence rather than trusted facts. Screening raises localization precision but can over-prune caption-relevant support. Evidence-aware composition determines what should be expressed, and the resulting caption identifies concepts that still require grounding. Caption-driven recovery closes this loop while retaining a compact grounding set, exposing and addressing the precision–coverage trade-off in conservative grounded-caption revision.

AI Use Statement

Generative AI tools were used for research assistance and manuscript drafting and editing. All AI-assisted outputs were reviewed and verified by the authors, who take responsibility for the final work.

REFERENCES

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. NoCaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8948–8957, 2019.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, pp. 382–398. Springer, 2016.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning, 2023.

Xueqing Deng, Linjie Yang, Qihang Yu, Ali Athar, Chenglin Yang, Xiaojie Jin, Xiaohui Shen, and Liang-Chieh Chen. COCONut-PanCap: Joint panoptic segmentation and grounded captions for fine-grained understanding and generation. In *Advances in Neural Information Processing Systems*, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, 2024.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4565–4574, 2016.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790, 2021.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Saehyung Lee, Seunghyun Yoon, Trung Bui, Jing Shi, and Sungroh Yoon. Toward robust hyperdetailed image captioning: A multiagent approach and dual evaluation metrics for factuality and coverage. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33815–33832, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022b.

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

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pp. 74–81, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection, 2023b.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. GLaMM: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, 2015.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2048–2057. PMLR, 2015.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*, pp. 69–85. Springer, 2016.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations*, 2020.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models, 2023.

A Implementation Details

A.1 Inference Procedure

Algorithm 1 EAG-Cap Inference

Input: image I
Obtain image-level captions $\mathcal{G}^{\text{img}}$ and region proposals $\mathcal{R}$.
Convert proposals into calibrated visual claims and risk tags.
Screen candidate claims for localized visual support.
Build structured claim-level evidence states $\mathcal{E}$.
Compose final caption y under evidence constraints.
Select reliable existing caption-supported boxes $\mathcal{G}_{\text{exist}}(y)$.
Extract visually groundable final-caption concepts.
Identify concepts without compatible existing grounding.
Re-ground the missing concepts to obtain $\mathcal{G}_{\text{recover}}(y)$.
Merge and deduplicate existing and recovered regions.
Output: caption y, grounding $\mathcal{G}_y$, suppressed claims.

A.2 Efficiency Analysis

We profile 100 images on one NVIDIA RTX 3090. Full EAG-Cap requires 941.071 s wall-clock time in total, or 9.4107 s/image. Table 6 reports per-stage latency and peak incremental GPU allocation.

Stage	Latency (s/img)	Peak incremental GPU memory (MiB)
Florence proposal generation	3.7670	1673
Claim preparation	0.0127	0
Visual evidence screening	1.8328	2731
Caption composition	3.2727	6357
Grounding refinement	0.0155	0
Caption usage filtering	0.0051	0
Caption-driven grounding recovery	0.5048	2731
Full EAG-Cap	9.4107	6357
Kosmos-2	0.7864	3707

Table 6: Efficiency profiling on 100 images using one NVIDIA RTX 3090.

Memory denotes peak incremental GPU allocation measured by the profiling procedure; 0 MiB indicates no measurable additional GPU allocation for lightweight CPU or post-processing stages. The pretrained proposal generator, evidence screener, and LLM composer dominate runtime. Recovery accounts for approximately 5.4% of full-pipeline latency.

B Claim Preparation and Risk Tagging

The preprocessor derives a conservative *safe query* and, when permitted, a more descriptive *specific query* from each proposal phrase. It normalizes punctuation and simple plurals, removes nonvisual terms, and prefers recognized compound heads before falling back to filtered content tokens. Hard-risk patterns cover uncertainty, subjective language, dietary claims, and identity/title expressions; soft-risk patterns flag fine-grained food terms and named-entity-like phrases. Hard-risk claims cannot be upgraded to a specific query. The claim registry retains both query levels, removed surface claims, risk tags, source boxes, and an operation hint for downstream auditing.

C VISUAL EVIDENCE SCREENING DETAILS

GroundingDINO receives the calibrated query with box and text thresholds 0.30 and 0.25. For each query, the highest-scoring valid detection is retained. The evidence builder labels a detection `coarse` when its normalized area exceeds 0.65; otherwise scores at least 0.50 are `strong` and lower accepted scores are `weak`. Unsupported or hard-risk unmatched claims are rejected. These labels describe degrees of localized grounding support rather than proof of the complete phrase semantics. The composer-facing Evidence State Store groups the records as strong support, weak support, coarse context, and rejected or unsupported evidence.

D EVIDENCE-AWARE CAPTION COMPOSER PROMPT

The default balanced-policy system prompt states the following operational rules: use the Florence-2 global caption as the primary sentence anchor if it is not unsafe; preserve its wording and structure as much as possible; use the detailed caption only as optional context; use strong grounding-supported evidence to correct or lightly enrich the caption; add at most one additional supported object, attribute, or action; do not turn the caption into an object list; prefer one natural COCO-style sentence of 8–18 words; do not mention rejected evidence or entries in `do_not_mention`; use weak/coarse evidence only when low-risk and directly image-supported; do not mention exact text, logos, identities, brands, counts, or fine-grained categories without strong support; and keep an already safe, visually plausible global caption with minimal edits. The output must be a COCO-style caption rather than a detailed scene report and must be strict JSON with fields `caption`, `used_evidence`, `suppressed_claims`, and `revision_operations`. Generation is deterministic at temperature 0.

Text-only controls. The detailed-only baseline prompt asks the same Qwen checkpoint to rewrite the detailed caption into one concise, natural, visually plausible caption, remove redundant, uncertain, or overly specific details, introduce no new objects, attributes, relations, identities, brands, text, counts, or other information, and return only the rewritten caption. The global-plus-detailed prompt asks for one concise caption using the global caption as the primary linguistic anchor and the detailed caption only for compatible additional information; it removes uncertain or overly specific detail, forbids information absent from both inputs, and returns only the final caption. Neither baseline receives an image, boxes, claims, risk tags, or evidence states.

E CAPTION-DRIVEN GROUNDING RECOVERY DETAILS

Surface queries are extracted around determiners and phrase starters while relation/action words and punctuation terminate a candidate. Candidates are augmented with the composer's `used_evidence`, `used_claims`, and `caption_claims`, plus phrase-level evidence whose normalized tokens are contained in the caption. Normalization lowercases, splits hyphens, removes punctuation and leading determiners, trims terminal `nearby`/`indoors`/`outdoors`, and truncates queries to four words. It handles irregular plurals (people, men, women, children, teeth, and feet) and regular `-ies`, `-ches`/`-shes`, `-ses`/`-xes`/`-zes`, and final `-s` forms. The fixed generic equivalences are man/woman/boy/girl/child/people $\mapsto$ person, television/monitor $\mapsto$ tv, sofa $\mapsto$ couch, bike $\mapsto$ bicycle, and cellphone $\mapsto$ phone. Candidate queries must be token-contained in the final caption.

Two phrases are compatible when normalized token sets are equal, or when they share a lexical head and one set contains the other or their overlap covers at least 75% of the smaller set. Compatible query variants are collapsed before inference. Supplementation queries GroundingDINO only when no compatible existing caption grounding is available, then selects its top-scoring valid box. The final list preserves every Existing Caption-Supported Box and appends Recovered Regions. Exact duplicates are removed using normalized phrase tokens and coordinates rounded to three decimals; the implementation does not use GT-guided selection or IoU-based NMS. The maximum is 12 caption queries per image.

F Grounding Metric and Matching Details

All-GT Coverage measures support for all evaluable annotated content, whereas Caption-Conditioned Coverage measures only annotated concepts explicitly mentioned by the generated caption; neither is standard detection recall. For COCO, boundary-aware CHAIR aliases map caption mentions and grounding phrases to categories. Each mentioned category is one diagnostic unit and may match any compatible annotated instance, avoiding the assumption that a singular mention refers to every same-category instance. For Flickr30k-Entities, punctuation and stopwords are removed, simple plurals are normalized, and phrase matching requires token containment or at least 50% overlap relative to the smaller phrase. Caption mentions prefer full normalized phrase containment and otherwise require a matching lexical head. Images with no caption-mentioned evaluable unit are excluded from the macro average but retained in the diagnostic counts.

G Alternative Recovery Policy

The hybrid policy queries every extracted caption concept and compares a new box with compatible existing grounding. If their IoU is at least 0.70, it keeps the higher-confidence candidate, preferring existing evidence when a score is unavailable; on disagreement it also retains the existing box. On Flickr30k-E1000, hybrid obtains 87.21% Precision@0.5, 37.21% All-GT Coverage, 78.82% caption-conditioned micro coverage, 81.03% macro coverage, and 0.730 caption-conditioned average IoU. Supplement obtains 87.39%, 36.98%, 78.66%, 80.83%, and 0.729, respectively. The small difference motivates the simpler supplement policy used by the full method.

H Optional Strict Policy

An optional strict composer uses only strong primary evidence and broad grounding-supported context, suppressing weak, proposal-only, and risky claims. It is intended for conservative applications rather than reference-metric optimization and is not included in the main comparisons.