KESTREL: GROUNDED SELF-REFINEMENT FOR HALLUCINATION MITIGATION IN LVLMS

Anonymous authors
Paper under double-blind review

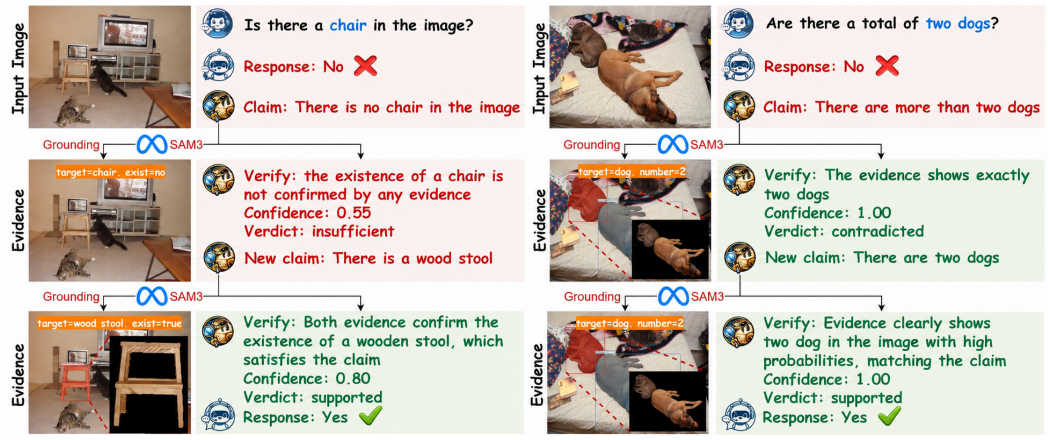


Figure 1: `Kestrel` progressively corrects hallucinated Stronger LVLM answers by integrating an external grounding agent with iterative self-improvement. At each round, the model grounds the current claim with explicit visual and textual evidence, conducts claim-level verification, and conservatively refines the response, yielding a final answer that is both more reliable and more interpretable.

ABSTRACT

Large vision-language models (LVLMs) have become increasingly strong and perform better than earlier ones, yet their hallucination issues are not completely solved. Previous training-free hallucination reduction methods are mostly developed for earlier LVLMs and rely on correcting internal representation, which we show that they can bring no or negative gain when applied on stronger LVLMs. On the other hand, introducing complementary external visual evidence can provide additional grounding beyond the LVLM itself, but its potential unreliability motivates conservative evidence verification and refinement rather than direct correction. Therefore, we propose `Kestrel`, a training-free framework for stronger LVLM hallucination mitigation that operationalizes conservative external-evidence intervention through explicit visual grounding and evidence-verified iterative self-refinement. In detail, `Kestrel` first collects explicit visual evidence and converts tool outputs into structured textual evidence. Second, to take full advantage of this evidence, `Kestrel` verifies it via the same LVLM judge for evidence checking, then iteratively self-refines answers based on verified evidence to reduce the risk of over-correction. Extensive experiments on stronger LVLMs show that `Kestrel` consistently improves performance across hallucination benchmarks. In particular, `Kestrel` improves Qwen3-VL by an average of +3.82% on POPE and +3.87% on MME-Hallucination, and improves InternVL3.5 by an average of +3.49% and +2.69%, respectively. The grounding agent contributes a relative gain of +2.14%, while the integrated self-refinement module contributes +2.23% on POPE. To conclude, our `Kestrel` effectively alleviates the hallucination in stronger LVLMs with transparent verification traces for diagnosis and analysis.

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.

1 INTRODUCTION

Recent advances in large-scale pretraining (Radford et al., 2021; Alayrac et al., 2022; Zhai et al., 2023) and multimodal instruction tuning (Liu et al., 2023; Dai et al., 2023) have substantially improved the capabilities of large vision-language models (LVLMs) (Grattafiori et al., 2024; Bai et al., 2025; Wang et al., 2025) on multimodal understanding and reasoning tasks. Nevertheless, hallucinations remain prevalent in multimodal tasks (Li et al., 2023c; Wang et al., 2023; Sun et al., 2024), making reliable hallucination mitigation for stronger LVLMs an increasingly important problem.

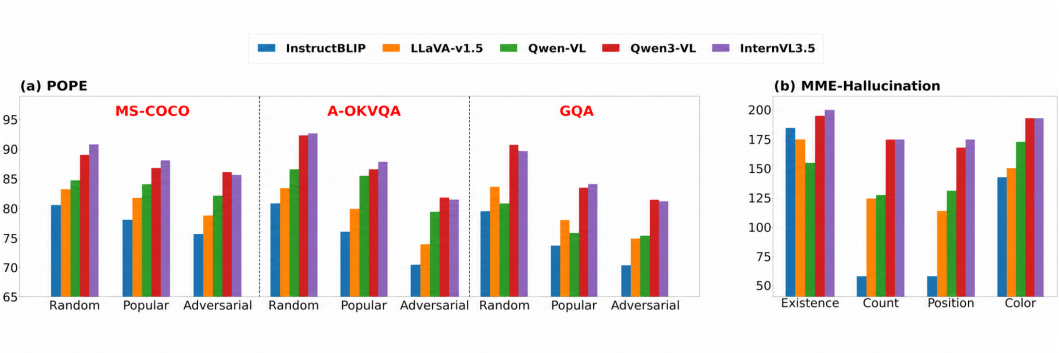


Figure 2: **Hallucination performance across different stage LVLMs.** Earlier LVLMs (InstructBLIP (Dai et al., 2023), LLaVA-v1.5 (Liu et al., 2024), and Qwen-VL (Bai et al., 2023)) exhibit substantially lower baseline performance than stronger LVLMs (Qwen3-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025)) across POPE and MME-Hallucination, suggesting that many hallucination issues prevalent in earlier LVLMs have already been alleviated as the LVLM stronger.

In earlier LVLMs, hallucinations were more frequent, leaving substantial room for such interventions to improve model predictions. However, as illustrated in Fig. 2, many hallucination issues prevalent in earlier models have been largely alleviated by more capable LVLMs such as Qwen3-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025) (over InstructBLIP/LLaVA-v1.5/Qwen-VL), Compared with these earlier LVLMs, Qwen3-VL achieves average relative improvements of 9.25% on POPE and 39.59% on MME-Hallucination, while InternVL3.5 achieves 9.66% and 41.82%, respectively., leaving fewer but more challenging residual hallucinations to be addressed.

We focus on training-free hallucination mitigation methods, as training these LVLMs is prohibitively expensive, making inference-time mitigation a more practical alternative (Zhao et al., 2024; Sicong Leng, 2023). Most existing methods (Zhang et al., 2025; Sicong Leng, 2023; Huang et al., 2024; Yin et al., 2024; Zhou et al., 2025) were developed and evaluated on earlier LVLMs, spanning interventions on decoding dynamics (Sicong Leng, 2023), model-internal representations (Sicong Leng, 2023; Woo et al., 2024; Huang et al., 2024; 2025b) and post-hoc correction and external verification (Yin et al., 2024; Zhang et al., 2025). However, experiments suggest that these methods do not necessarily remain effective on stronger LVLMs (Tab. 1 and Tab. 2). Decoding-based interventions show limited or inconsistent gains on stronger LVLMs, suggesting that aggressive inference-time perturbation may have less room to improve already strong predictions. Ablation (Tab. 3) shows that relying solely on the LVLM's own self-refinement yields only marginal gains, indicating that model-internal correction alone is insufficient in stronger LVLM setting. External visual evidence can instead provide complementary grounding beyond LVLM itself, which can improve the correction of such persistent errors. Nevertheless, external verification and post-hoc correction methods (Yin et al., 2024; Zhang et al., 2025) may still be insufficiently robust when external feedback is used directly for prediction revision (Tab. 3). Moreover, introducing an additional LVLM as the verifier increases model dependency and system complexity (Thakur et al., 2025; Huang et al., 2025a). These limitations motivate a design that incorporates external visual evidence while using the same LVLM throughout verification and refinement, with conservative evidence assessment and iterative updates to reduce erroneous corrections.

Motivated by the reduced effectiveness of existing training-free mitigation methods on stronger LVLMs and above limitations, we propose `Kestrel` (see Fig. 1), a training-free framework built around a conservative external-evidence intervention principle: rather than directly overriding strong LVLM predictions with potentially noisy tool outputs, `Kestrel` verifies and progressively accumu-

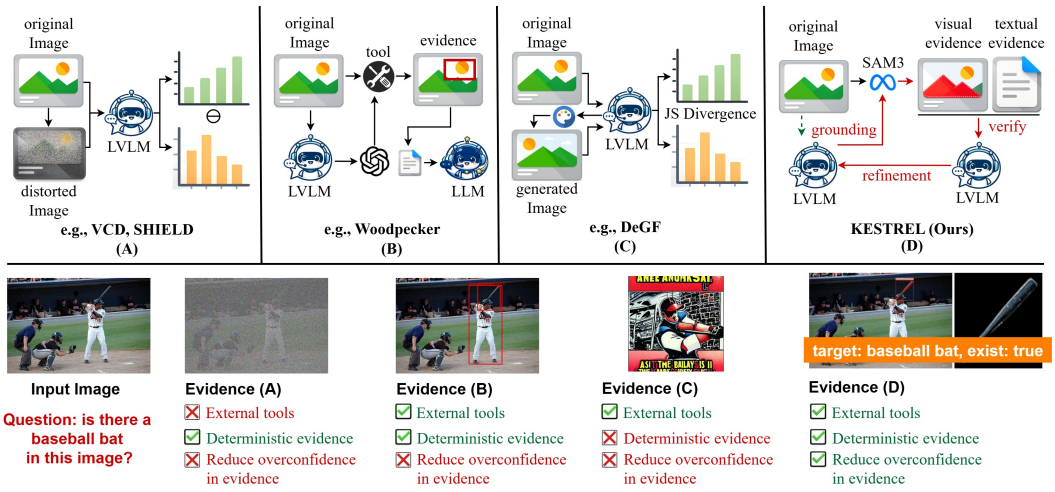


Figure 3: `Kestrel` **vs. prior training-free hallucination mitigation methods.** By combining an external grounding agent with iterative self-improvement, `Kestrel` collects explicit visual evidence and further converts tool outputs into structured textual evidence for verification. The key distinction is therefore not whether external evidence is introduced, but how such evidence is allowed to influence an already stronger LVLM prediction. Compared with prior correction pipelines that use external feedback more directly during prediction revision, `Kestrel` explicitly interposes claim-level verification and evidence gating before external evidence is allowed to trigger an answer change.

lates grounded evidence, and revises a prediction only when the resulting evidence is sufficiently reliable. Specifically, `Kestrel` first decomposes the question into verifiable claim-level targets (e.g., existence, color, count, and position), and then invokes SAM3 (Carion et al., 2025) around each target to collect segmentation overlays, bounding boxes, target crop-and-zoom views, and text evidence derived from this collected visual evidence, all of which are collated as structured evidence items with citation identifiers. The framework then performs claim-level verification with verdicts and outputs confidence-aware verification results, forming an auditable evidence chain. To regulate the potential over-correction, we further introduce an evidence-gated update scheme into the iteration: the framework progressively supplements and strengthens claim-level evidence through multiple rounds of verification and revision, and permits answer flips only when evidence strength, confidence, and evidence coverage jointly satisfy predefined criteria. These designs preserve the training-free methods while improving the interpretability, robustness, and decision stability of hallucination mitigation.

Experiments on stronger LVLMs demonstrate that `Kestrel` consistently mitigates hallucinations, including cases where many existing training-free methods provide limited or negative gains. On POPE (Li et al., 2023c), `Kestrel` improves Qwen3-VL and InternVL3.5 by an average relative gain of 3.82% and 3.49%, respectively; it also surpasses prior training-free mitigation baselines by 1.56% and 1.73% on average under the same backbones. On the more challenging MME-Hallucination (Fu et al., 2023), `Kestrel` improves Qwen3-VL and InternVL3.5 by 3.87% and 2.69%, respectively. It further outperforms OPERA (Huang et al., 2024) by 2.24% on Qwen3-VL and VCD (Sicong Leng, 2023) by 2.69% on InternVL3.5, with consistent gains observed across different stronger LVLMs. These results establish `Kestrel` as an effective hallucination mitigation method for stronger LVLMs.

Our main contributions are summarized as follows:

- We reveal an emerging challenge for hallucination mitigation on stronger LVLMs: methods that were effective on earlier models do not necessarily remain effective on stronger ones.
- We propose `Kestrel`, a training-free framework that operationalizes conservative external-evidence intervention for stronger LVLMs. Instead of directly using external grounding to overwrite model predictions, `Kestrel` converts visual grounding into structured evidence, verifies and accumulates this evidence across multiple refinement rounds, and revises predictions only when the evidence is sufficiently reliable.
- Extensive experiments on Qwen3-VL and InternVL3.5 demonstrate that `Kestrel` consistently improves hallucination performance, where many existing training-free methods yield limited or negative gains, while maintaining interpretable verification traces and low over-correction.

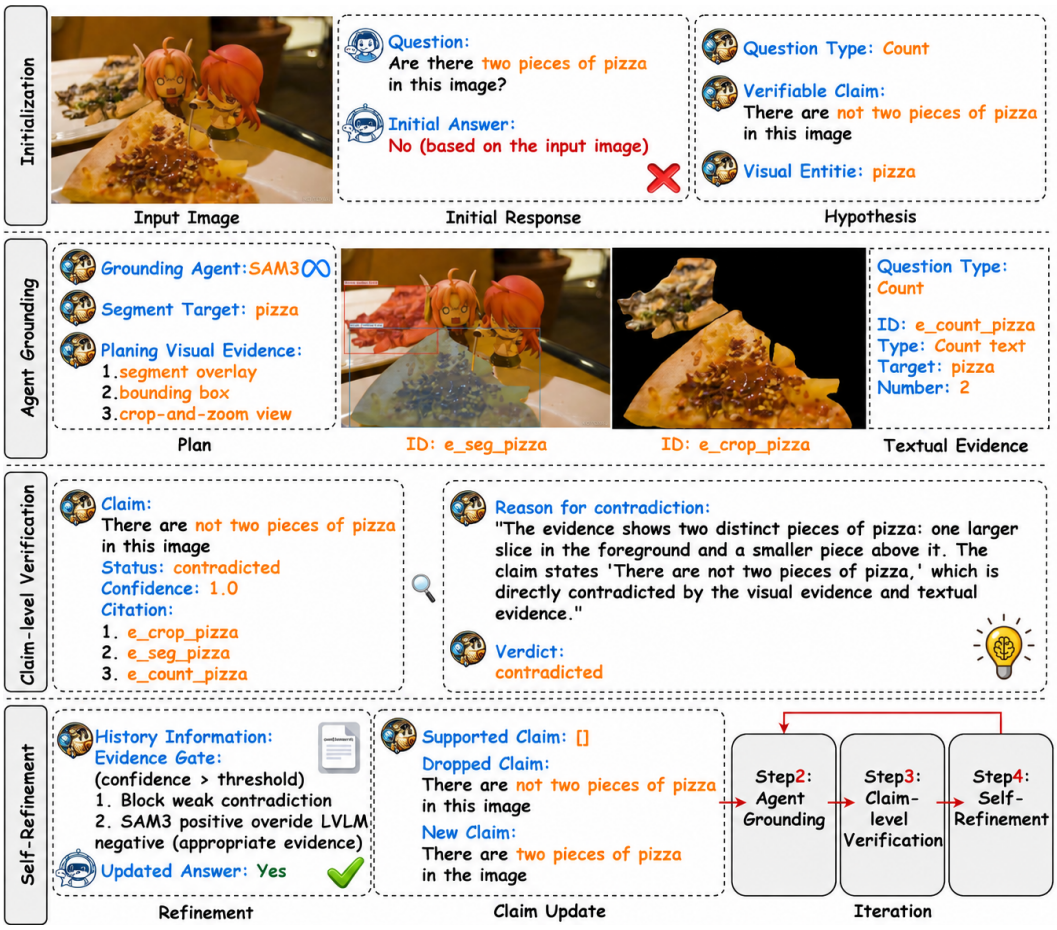


Figure 4: Overview of `Kestrel`. Given an image-question pair, `Kestrel` follows a four-stage training-free pipeline for LVLM hallucination mitigation: **(1) Initialization**, which obtains an initial answer and rewrites it into question-aligned verifiable claims with associated visual entities and claim types; **(2) Agent Grounding**, which invokes an external SAM3-based grounding agent to collect explicit visual evidence (e.g., segmentation overlays, boxes, and crop-and-zoom views) and convert them into structured textual evidence; **(3) Claim-level Verification**, which verifies each claim against the cited evidence to produce claim-wise verdicts, confidence scores, and a top-level verification decision; and **(4) Self-Refinement**, which performs evidence-gated answer updating based on the current and previous verification traces.

2 METHOD

We propose `Kestrel`, a training-free framework for mitigating hallucination with explicit visual grounding agent and structured evidence-driven self-refinement at test time, targeting on stronger LVLMs such as Qwen3-VL and InternVL3.5. Given an image $\mathcal{I}$ and corresponding question $\mathcal{Q}$, `Kestrel` iteratively follows a four-step pipeline: **(i)** initialization, **(ii)** agent grounding, **(iii)** claim-level verification, and **(iv)** self-refinement (see Fig. 4).

2.1 INITIALIZATION

`Kestrel` first queries the LVLM to obtain an initial answer $\hat{\mathcal{A}}^{(0)}$. To support claim-level verification, `Kestrel` converts $\mathcal{Q}$ into a small set of *verifiable claims* that directly correspond to the question. Concretely, `Kestrel` rewrites the question-answer decision into visually checkable claims, each anchored to one or two concrete visual entities. These entities serve as the detection targets for the grounding agent. Meanwhile, based on the verifiable attributes required by the question, we categorize the extracted claims (e.g., `existence`, `color`, `count`, `position`) to route subsequent agent grounding.

2.2 AGENT GROUNDING

To obtain explicit, inspectable grounding evidence, `Kestrel` invokes an external visual grounding agent built on **SAM3** promptable concept segmentation (Carion et al., 2025).

Visual evidence. SAM3 takes the visual entities in the claims as concept prompts and returns the matched instances, from which `Kestrel` collects explicit visual evidence, including: **(i)** segmentation overlays for transparent localization, **(ii)** instance bounding boxes (derived from SAM3 masks) to support geometry-based reasoning, and **(iii)** crop-and-zoom views around predicted instances to reduce ambiguity for attribute inspection (e.g., color) and local details.

Structured textual evidence. To make agent outputs directly usable for claim verification and auditable diagnosis, `Kestrel` derives textual evidence from visual evidence via LVLM for each claim type: **(i)** for `existence`, we convert the predicted instances into an existence statement by checking whether the number of matched instances is greater than zero; **(ii)** for `count`, we report the instance count computed from the number of predicted masks; **(iii)** for `color`, we generate a concise color observation conditioned on the masked crop-and-zoom and the full image; **(iv)** for `position`, we convert SAM3 geometry into text by deriving coarse spatial cues from the union bounding box and, when two entities are involved, computing their relative relation from the corresponding bounding-box centers. Each textual evidence item is paired with a citation identifier and can be referenced during verification and answer revision.

2.3 CLAIM-LEVEL VERIFICATION

Given the claims and corresponding structured evidence items, `Kestrel` performs **claim-level verification** using LVLM-as-a-judge. The LVLM itself is instructed to base its decision only on the provided evidence and to cited the corresponding evidence. For each claim, the verifier outputs: **(i)** a verdict (*supported* / *contradicted* / *insufficient*), **(ii)** a confidence score, and **(iii)** a short reasoning that must cite the relevant evidence. We then consolidate claim-wise judgments into a top-level verification verdict for the current answer: it is labeled as *contradicted* if any claim is confidently refuted with cited evidence, as *supported* only when all claims are confidently supported, and as *insufficient* otherwise. The resulting verification trace constitutes an explicit, evidence-grounded audit trail, enabling interpretable analysis of when hallucinations arise and how corrections are triggered.

2.4 SELF-REFINEMENT

A key challenge in stronger LVLMs is that most initial predictions are already correct, leaving less room for aggressive intervention and making erroneous revisions particularly undesirable. Therefore, external evidence should not automatically act as a correction signal. `Kestrel` instead adopts an evidence-gated self-refinement strategy that treats prediction revision as a conservative evidence-selection process: an answer is revised only when verification provides sufficiently reliable signals, i.e., high-confidence claim-level judgments supported by cited evidence. Otherwise, `Kestrel` preserves the current answer $\hat{\mathcal{A}}^{(i)}$ (where i denotes i th iteration) and continues to strengthen the evidence in subsequent rounds rather than immediately overriding the current prediction.

Importantly, the self-refinement is **stateful**: the revision step conditions not only on the current verification results but also on prior rounds' claims, evidence, and decisions. Based on the verification trace, `Kestrel` updates the answer to obtain $\hat{\mathcal{A}}^{(i+1)}$ and proposes a new set of claims for the next iteration, prioritizing claims that remain uncertain or are implicated by contradictions. This iterative process progressively strengthens evidence and stabilizes decision-making, while remaining training-free.

`Kestrel` repeats the refinement process for several rounds, and stops early when the answer stabilizes under consistently supportive verification, or when additional iterations no longer yield stronger evidence. The final output is the answer together with its claim-level verification traces.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate `Kestrel` on 1) **POPE** (Li et al., 2023c) in three source datasets (MS-COCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson & Manning, 2019)), each with random (Rand.), popular (Pop.), and adversarial (Adv.) subsets, and 2) **MME-Hallucination** (Fu et al., 2023) evaluates fine-grained hallucination across existence, count, position, and color. Additional CHAIR (Rohrbach et al., 2018) results for `Kestrel` are reported in Appendix F.

LVLM backbones. We evaluate `Kestrel` with capable open-weight LVLMs, e.g., **Qwen3-VL 8B** (Bai et al., 2025) and **InternVL3.5 8B** (Wang et al., 2025).

Baselines. We compare against Qwen3-VL agent (with zoom-in tool)[1] and training-free LVLM hallucination mitigation baselines: **decoding-based methods**, including **VCD** (Sicong Leng, 2023), **OPERA** (Huang et al., 2024), **RITUAL** (Woo et al., 2024), and **HALC** (Chen et al., 2024b); **model-internal methods**, including **LogicCheckGPT** (Wu et al., 2024), **Clearsight** (Yin et al., 2025), **CausalMM** (Zhou et al., 2025), and **MemVR** (Zou et al., 2025); **external-evidence and post-hoc correction methods**, including **Woodpecker** (Yin et al., 2024) and **DeGF** (Zhang et al., 2025); and **agentic methods**, including the training-free **Qwen3-VL agent**. All baselines are evaluated with the same target LVLM backbone for a controlled comparison. For Woodpecker and LogicCheckGPT, all LVLM-based components except the visual verification module are replaced with the corresponding backbone. We use official implementations whenever available and follow the method-specific hyperparameters recommended by the original papers or repositories, without tuning on the test sets.

Implementation details. We set the maximum number of self-refinement iterations to $K=3$ and stop early when the answer is stable with two consecutive *supported* verification verdicts. For the SAM3 grounding agent, we use a confidence threshold of 0.5 (with a recheck threshold of 0.35 when needed for existence claim). In self-refinement, we use confidence thresholds in the range $[0.82, 0.90]$ for different claim types, so that answer revisions are triggered only by sufficiently confident verification signals. All experiments are conducted on NVIDIA A100 GPUs. Details of the corresponding prompt template are provided in the Appendix D.

3.2 RESULTS AND DISCUSSION

Results on POPE. Tab. 1 compares `Kestrel` with baselines under two capable LVLM backbones (Qwen3-VL and InternVL3.5) across three evaluation sources (MS-COCO, A-OKVQA, and GQA). While many existing training-free methods have reported positive gains on earlier LVLMs (e.g., InstructBLIP (Dai et al., 2023), LLaVA-v1.5 (Liu et al., 2024), and Qwen-VL (Bai et al., 2023)), their effectiveness is substantially less consistent when applied to stronger LVLM backbones, with several methods even degrading the original predictions on average (e.g., ClearSight (Yin et al., 2025) decreases Qwen3-VL (Bai et al., 2025) by 5.57 points, while CausalMM (Zhou et al., 2025) decreases InternVL3.5 (Wang et al., 2025) by 6.33 points.). In contrast, `Kestrel` exhibits the most consistent gains on the more challenging popular and adversarial splits, while remaining competitive on random sets, indicating robust hallucination mitigation across backbones and data sources.

With **Qwen3-VL**, `Kestrel` achieves the best performance on the more challenging **MS-COCO** popular and adversarial splits, while remaining competitive on the random split. On **A-OKVQA**, `Kestrel` reaches **91.73**% (Pop.) and **86.76**% (Adv.), surpassing decoding-based baselines such as OPERA (89.50%/83.86%) and VCD (87.86%/82.00%). On **GQA**, `Kestrel` further improves to **90.33**% (Pop.) and **86.27**% (Adv.), outperforming OPERA (87.11%/83.30%) and VCD (85.76%/81.93%) by a clear margin. With **InternVL3.5**, we observe a similar trend, delivering the largest improvements over the LVLM backbone and the strongest overall performance.

Compared with decoding-centric methods that reshape token distributions at inference time (e.g., VCD (Sicong Leng, 2023) and OPERA (Huang et al., 2024)) and post-hoc correction pipelines such as Woodpecker (Yin et al., 2024), `Kestrel` externalizes grounding into explicit evidence and performs

[1] https://github.com/QwenLM/Qwen-Agent/blob/main/examples/cookbook_think_with_images.ipynb

Table 1: **Results on POPE (Li et al., 2023c) benchmark**. Higher (↑) accuracy indicates better performance. The best results are **bolded**, and the second-best are underlined.

Backbone	Method	MS-COCO			A-OKVQA			GQA		
		Rand. ↑	Pop. ↑	Adv. ↑	Rand. ↑	Pop. ↑	Adv. ↑	Rand. ↑	Pop. ↑	Adv. ↑
InstructBLIP	Baseline (Dai et al., 2023)	80.71	78.22	75.84	80.91	76.19	70.71	79.65	73.87	70.56
LLaVA-v1.5	Baseline (Liu et al., 2024)	83.29	81.88	78.96	83.45	79.90	74.04	83.73	78.17	75.08
Qwen-VL	Baseline (Bai et al., 2023)	84.73	84.13	82.26	86.67	85.56	79.57	80.97	75.99	75.46
Qwen3-VL	Baseline (Bai et al., 2025)	89.00	86.92	86.20	92.36	86.67	81.87	90.70	83.63	81.50
	Qwen3-VL agent (Bai et al., 2025)	91.03	88.06	86.13	92.87	85.20	78.03	91.41	81.81	78.30
	VCD (Sicong Leng, 2023)	90.40	88.80	87.41	93.53	87.86	82.00	91.56	85.76	81.93
	RITUAL (Woo et al., 2024)	86.20	83.67	82.27	87.67	83.50	77.76	86.86	82.30	78.23
	OPERA (Huang et al., 2024)	90.50	88.83	87.50	**93.76**	89.50	83.86	91.80	87.11	83.30
	HALC (Chen et al., 2024b)	90.40	87.23	86.27	93.33	88.63	83.03	91.17	86.53	81.87
	LogicCheckGPT (Wu et al., 2024)	77.47	76.60	76.40	79.03	77.30	73.57	77.30	75.47	73.30
	Clearsight (Yin et al., 2025)	89.10	85.33	83.43	87.86	82.70	77.20	81.16	71.70	70.26
	CausalMM (Zhou et al., 2025)	91.80	87.73	86.20	93.33	87.93	82.33	91.67	86.73	82.73
	MemVR (Zou et al., 2025)	91.76	88.80	86.76	93.26	87.31	81.50	**92.11**	86.06	81.96
	Woodpecker (Yin et al., 2024)	89.97	88.03	87.10	93.23	88.90	83.33	91.27	86.27	82.77
	DeGF (Zhang et al., 2025)	90.33	88.16	86.90	92.96	87.70	82.61	91.13	83.79	82.00
	`Kestrel`$_{(ours)}$	91.53	**89.30**	**87.53**	93.46	**91.73**	**86.76**	91.67	**90.33**	**86.27**
InternVL3.5	Baseline (Wang et al., 2025)	90.77	88.10	85.73	92.67	87.83	81.53	89.77	84.10	81.31
	VCD (Sicong Leng, 2023)	91.35	89.22	87.60	92.87	89.73	83.73	**91.60**	85.07	83.37
	RITUAL (Woo et al., 2024)	91.60	89.03	87.48	93.71	89.75	83.90	91.39	85.18	83.29
	OPERA (Huang et al., 2024)	91.53	89.18	87.55	93.55	89.79	83.81	91.45	85.20	83.31
	HALC (Chen et al., 2024b)	**93.50**	84.60	80.83	92.43	77.27	78.67	90.00	73.57	69.23
	LogicCheckGPT (Wu et al., 2024)	79.96	77.00	76.06	78.06	75.79	72.63	75.46	73.13	71.27
	Clearsight (Yin et al., 2025)	92.43	86.16	82.07	92.93	80.90	74.55	90.13	79.06	74.26
	CausalMM (Zhou et al., 2025)	92.88	83.27	80.03	91.93	76.80	68.23	89.77	73.13	68.81
	MemVR (Zou et al., 2025)	92.95	86.90	82.90	93.66	82.93	75.96	90.76	80.51	75.63
	Woodpecker (Yin et al., 2024)	91.20	89.11	87.50	**93.73**	89.80	84.00	91.43	85.16	83.26
	DeGF (Zhang et al., 2025)	91.43	89.12	87.37	93.39	89.68	84.11	91.48	85.06	83.20
	`Kestrel`$_{(ours)}$	91.27	**89.27**	**88.10**	93.57	**91.80**	**87.13**	91.57	**89.87**	**86.53**

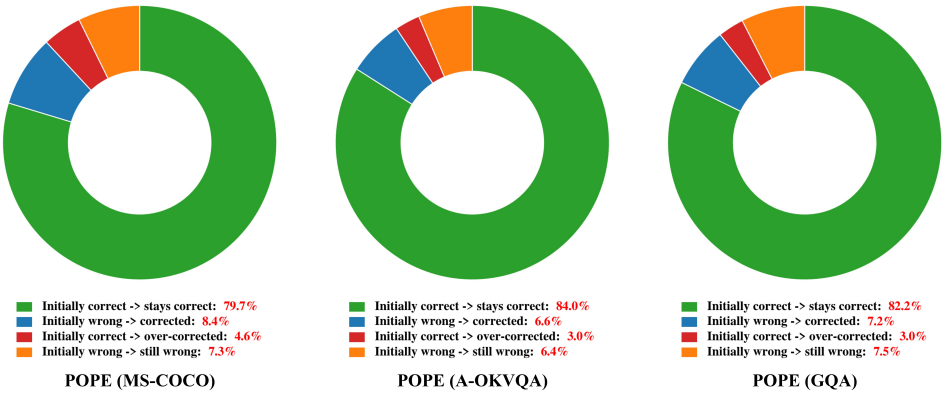


Figure 5: **Prediction transition statistics with Qwen3-VL before and after refinement.** Prediction are categorized into four types: correctly preserved, error corrected, over-corrected, and incorrectly preserved. The results show that the refinement process is conservative, retaining most originally correct predictions while correcting a portion of erroneous ones, with limited over-correction. Zoom in for a better view.

claim-level verification with conservative, evidence-gated updates. This design better preserves correct predictions while correcting residual hallucinations on stronger LVLMs. Fig. 5 further shows that the gains on POPE mainly come from correcting initially incorrect predictions, while most initially correct predictions are preserved after refinement. The relatively low rate of over-correction suggests that our conservative gating mechanism effectively suppresses unnecessary revisions. More broadly, this correction–preservation trade-off empirically illustrates the central motivation behind `Kestrel`: effective hallucination mitigation on stronger LVLMs is not simply a matter of correcting more predictions, but of selectively correcting residual errors without destabilizing the much larger set of already-correct predictions.

Table 2: **Results on MME-Hallucination (Fu et al., 2023) benchmark.** Higher scores (↑) indicate better performance. The best results are **bolded**, and the second-best are underlined.

Backbone	Method	Object-level		Attribute-level		MME Score ↑
		Existence ↑	Count ↑	Position ↑	Color ↑	
InstructBLIP	Baseline (Dai et al., 2023)	185.00	58.00	58.00	143.00	444.00
LLaVA-v1.5	Baseline (Liu et al., 2024)	175.67	124.67	114.00	151.00	565.33
Qwen-VL	Baseline (Bai et al., 2023)	155.00	127.67	131.67	173.00	587.33
Qwen3-VL	Baseline (Bai et al., 2025)	195.00 (+0.00)	175.00 (+0.00)	168.33 (+0.00)	193.33 (+0.00)	731.66 (+0.00)
	Qwen3-VL agent (Bai et al., 2025)	**200.00** (+5.00)	181.67 (+6.67)	168.33 (+0.00)	193.33 (+0.00)	743.33 (+11.67)
	VCD (Sicong Leng, 2023)	195.00 (+0.00)	180.00 (+5.00)	168.33 (+0.00)	193.33 (+0.00)	736.66 (+5.00)
	RITUAL (Woo et al., 2024)	195.00 (+0.00)	180.00 (+5.00)	168.33 (+0.00)	193.33 (+0.00)	736.66 (+5.00)
	OPERA (Huang et al., 2024)	195.00 (+0.00)	180.00 (+5.00)	168.33 (+0.00)	**200.00** (+6.67)	743.33 (+11.67)
	HALC (Chen et al., 2024b)	195.00 (+0.00)	173.33 (−1.67)	153.33 (−15.00)	**200.00** (+6.67)	721.66 (−10.00)
	LogicCheckGPT (Wu et al., 2024)	195.00 (+0.00)	173.33 (−1.67)	148.33 (−20.00)	190.00 (−3.33)	706.66 (−25.00)
	Clearsight (Yin et al., 2025)	195.00 (+0.00)	173.33 (−1.67)	158.33 (−10.00)	190.00 (−3.33)	716.66 (−15.00)
	CausalMM (Zhou et al., 2025)	195.00 (+0.00)	168.33 (−6.67)	153.33 (−15.00)	190.00 (−3.33)	706.66 (−25.00)
	MemVR (Zou et al., 2025)	195.00 (+0.00)	168.33 (−6.67)	158.33 (−10.00)	195.00 (+1.67)	716.66 (−15.00)
	Woodpecker (Yin et al., 2024)	195.00 (+0.00)	173.33 (−1.67)	168.33 (+0.00)	195.00 (+1.67)	731.66 (+0.00)
	DeGF (Zhang et al., 2025)	195.00 (+0.00)	181.67 (+6.67)	166.67 (−1.66)	188.33 (−5.00)	731.67 (+0.01)
	`Kestrel`$_{(ours)}$	**200.00** (+5.00)	**186.67** (+11.67)	**180.00** (+11.67)	193.33 (+0.00)	**760.00** (+28.34)
InternVL3.5	Baseline (Wang et al., 2025)	**200.00** (+0.00)	175.00 (+0.00)	175.00 (+0.00)	193.33 (+0.00)	743.33 (+0.00)
	VCD (Sicong Leng, 2023)	**200.00** (+0.00)	175.00 (+0.00)	175.00 (+0.00)	193.33 (+0.00)	743.33 (+0.00)
	RITUAL (Woo et al., 2024)	195.00 (−5.00)	175.00 (+0.00)	175.00 (+0.00)	193.33 (+0.00)	738.33 (−5.00)
	OPERA (Huang et al., 2024)	195.00 (−5.00)	173.33 (−1.67)	175.00 (+0.00)	**195.00** (+1.67)	738.33 (−5.00)
	HALC (Chen et al., 2024b)	**200.00** (+0.00)	170.00 (−5.00)	168.33 (−6.67)	190.00 (−3.33)	728.33 (−15.00)
	LogicCheckGPT (Wu et al., 2024)	195.00 (−5.00)	168.33 (−6.67)	158.33 (−16.67)	180.00 (−13.33)	701.66 (−41.67)
	Clearsight (Yin et al., 2025)	195.00 (−5.00)	163.33 (−11.67)	161.66 (−13.34)	180.00 (−13.33)	699.99 (−43.34)
	CausalMM (Zhou et al., 2025)	**200.00** (+0.00)	170.00 (−5.00)	166.66 (−8.34)	180.00 (−13.33)	716.66 (−26.67)
	MemVR (Zou et al., 2025)	195.00 (−5.00)	168.33 (−6.67)	161.66 (−13.34)	180.00 (−13.33)	704.99 (−38.34)
	Woodpecker (Yin et al., 2024)	**200.00** (+0.00)	166.67 (−8.33)	161.67 (−13.33)	186.67 (−6.66)	715.01 (−28.32)
	DeGF (Zhang et al., 2025)	195.00 (−5.00)	175.00 (+0.00)	168.33 (−6.67)	188.33 (−5.00)	726.66 (−16.67)
	`Kestrel`$_{(ours)}$	**200.00** (+0.00)	**186.67** (+11.67)	**181.67** (+6.67)	**195.00** (+1.67)	**763.34** (+20.01)

Results on MME-Hallucination. Tab. 2 reports MME-Hallucination results decomposed into object-level (Existence, Count) and attribute-level (Position, Color) subsets. Tab. 2 shows that the effectiveness of existing training-free mitigation methods varies substantially on stronger LVLM backbones, whereas `Kestrel` maintains consistent improvements. Overall, `Kestrel` achieves the best **MME Score** under Qwen3-VL, reaching **760.00** (+28.34), outperforming baselines such as OPERA (743.33 +11.67) and VCD/RITUAL (736.66 +5.00). These results suggest that `Kestrel` 's evidence-driven, claim-level verification and conservative update rule particularly strengthen object existence/count and spatial reasoning, yielding improvements across multiple hallucination-sensitive categories. We also observe slightly larger variance for `Kestrel` on some subsets, which is expected when grounding and verification are performed with stochastic components; nevertheless, the mean improvement remains consistent and substantial over all compared methods.

Tab. 2 presents the results of `Kestrel` with the InternVL3.5 backbone (Wang et al., 2025) on MME Hallucination. Notably, `Kestrel` still achieves further improvement over such a strong backbone, showing that our framework remains effective even when the base model already performs at a high level. This result is particularly meaningful, as reducing hallucinations becomes increasingly difficult as the backbone grows stronger. Moreover, among the methods compared in Tab. 2, `Kestrel` is the only one that consistently improves InternVL3.5, further highlighting the advantage of our grounded evidence-based refinement framework.

3.3 Ablation Study

Designs. Tab. 3 reports the ablation results of `Kestrel` on POPE-COCO. Naive self-refinement improves the baseline by only +0.07 points (87.37→87.44), while external grounding alone yields only a +0.13-point gain (87.37→87.50). These results distinguish `Kestrel` from simply combining an external grounding model with self-refinement: neither component alone explains the final improvement. Substantial gains emerge only when external evidence is incorporated through structured claim-level verification and controlled prediction revision, supporting our central design principle that external evidence should be verified rather than directly acted upon. Specifically, incorporating the grounding agent into verification-guided refinement improves performance from 87.58 to 88.61 (+1.03 points). Ablating structured textual evidence or claim-level verification from the full pipeline decreases performance by 0.84 and 0.40 points, respectively, further demonstrating the importance of

Table 3: **Ablation study of core design components.** Higher scores indicate better performance.

Setting	Grounding Agent	Structured Textual Evidence	Claim-level Verification	Evidence-gated Update	Self-Refinement	History-Aware	POPE MS-COCO
baseline	✗	✗	✗	✗	✗	✗	87.37
s1	✗	✗	✗	✗	✓	✗	87.44
s2	✓	✗	✗	✗	✗	✗	87.50
s3	✗	✗	✓	✓	✓	✓	87.58
s4	✓	✗	✓	✓	✓	✓	88.61
s5	✓	✓	✗	✓	✓	✓	89.05
s6	✓	✓	✓	✗	✓	✓	89.30
s7	✓	✓	✓	✓	✗	✗	89.16
s8	✓	✓	✓	✓	✓	✗	89.32
s9	✓	✓	✓	✓	✓	✓	**89.45**

Table 4: **Ablation of visual evidence types.** Contribution of different visual evidence forms, including segmentation, boxes, and crop-zoom views. Higher scores indicate better performance.

visual	segment	box	crop-zoom	MME Score
baseline	–	–	–	731.66
s1	✗	✓	✓	738.33
s2	✓	✗	✓	743.33
s3	✓	✓	✗	740.00
s4	✓	✓	✓	760.00

Table 5: **Ablation of structured textual evidence.** We study the contribution of different textual evidence types used for claim verification. Higher scores indicate better performance.

textual	exist	count	color	position	MME Score
s1	✗	✓	✓	✓	755.00
s2	✓	✗	✓	✓	751.67
s3	✓	✓	✗	✓	753.33
s4	✓	✓	✓	✗	748.33
s5	✓	✓	✓	✓	760.00

converting external grounding into structured evidence and verifying it before revision. Removing the evidence-gated update further reduces performance by 0.15 points, confirming the role of conservative update control in suppressing unreliable corrections. Finally, history-aware refinement provides an additional +0.13-point gain, with the full `Kestrel` achieving the best overall performance of 89.45.

Evidence. The evidence ablations (Tab. 4 and Tab. 5) further highlight the complementary roles of both visual and textual evidence in `Kestrel`. For visual evidence, removing any single component consistently degrades performance relative to the full setting, showing that segmentation overlays (segment), bounding boxes (box), and crop-zoom views (crop-zoom) all contribute to the final result. Among them, removing segmentation or crop-zoom causes a larger drop than removing boxes, suggesting that precise region localization and enlarged local inspection are particularly important for reliable claim verification. For structured textual evidence, ablating any evidence type also leads to performance degradation, confirming that existence, count, color, and position evidence all provide useful support for refinement. Notably, removing position or count evidence causes a relatively larger drop, indicating that these attributes benefit more from explicit structured evidence. Overall, the best performance is achieved when all visual and textual evidence types are used together, demonstrating that the different evidence forms are complementary rather than redundant.

4 Conclusion

We present `Kestrel`, a training-free framework for hallucination mitigation in stronger LVLMs. Experiments highlight a changing mitigation regime: as LVLM backbones become stronger, many existing training-free methods provide limited or even negative gains, while residual hallucinations become increasingly important to correct without disrupting already-correct predictions. Motivated by this setting, `Kestrel` follows a conservative external-evidence intervention principle, in which grounded evidence is structured, verified, and accumulated before it is allowed to trigger prediction revision. Experiments on Qwen3-VL and InternVL3.5 show that this design consistently improves hallucination-sensitive performance on POPE and MME-Hallucination, while maintaining a favorable correction–preservation trade-off. Ablation studies further show that neither naive self-refinement nor external grounding alone explains the gains, supporting the importance of verified and controlled evidence use. Overall, `Kestrel` suggests that effective hallucination mitigation for stronger LVLMs requires not only stronger external grounding, but also careful control over when external evidence should override strong model predictions.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, including improving clarity and readability, refining phrasing, and translation.

ETHICS STATEMENT

All authors of this work have read and commit to adhering to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The main algorithmic pipeline is specified in Section 2. The experimental section reports the benchmarks, backbones, baselines, iteration limits, confidence thresholds, stopping rules, hardware, latency, and memory requirements used for evaluation. The appendix provides the complete prompt templates, constrained output formats, additional ablations, qualitative cases, and seed-robustness results needed to reproduce and verify the main findings. The full code can be found in the *Supplementary Material*.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2, 14

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 14

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 2, 6, 7, 8

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 2, 6, 7, 8, 14, 16, 20

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719. 3, 5

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models, 2024a. URL https://arxiv.org/abs/2402.11684. 14

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023a. URL https://arxiv.org/abs/2311.12793. 14

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024b. 6, 7, 8

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023b. 14

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 2, 6, 7, 8, 14

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3, 6, 8, 14

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 14

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024. 14

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5880–5895, 2025a. 2, 15

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024. 2, 3, 6, 7, 8, 14

Yiyang Huang, Liang Shi, Yitian Zhang, Yi Xu, and Yun Fu. Shield: Suppressing hallucinations in lvlm encoders via bias and vulnerability defense. *arXiv preprint arXiv:2510.16596*, 2025b. 2, 14

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019. 6

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27036–27046, 2024. 14

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27228–27238, 2024. 14

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 *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a. 14

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023b. 14

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023c. 2, 3, 6, 7, 14

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 *European conference on computer vision*, pp. 740–755. Springer, 2014. 6, 16

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 14

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024. 2, 6, 7, 8, 14

Xinyu Lyu, Beitao Chen, Lianli Gao, Hengtao Shen, and Jingkuan Song. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *Advances in Neural Information Processing Systems*, 37:122811–122832, 2024. 14

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 14

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. 2, 14

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, 2018. 6, 19, 20

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022. 6

Guanzheng Chen Xin Li Shijian Lu Chunyan Miao Lidong Bing Sicong Leng, Hang Zhang. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023. URL https://arxiv.org/abs/2311.16922. 2, 3, 6, 7, 8, 14

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024. 2, 14

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pp. 404–430, 2025. 2, 15

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 2, 14

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024. 14

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2, 6, 7, 8, 14, 20

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. 2024. 2, 6, 7, 8, 14

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*, 2024. 6, 7, 8

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14625–14634, 2025. 6, 7, 8

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. *Science China Information Sciences*, 67(12):220105, 2024. 2, 6, 7, 8, 14

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023. 2

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2502.06130*, 2025. 2, 6, 7, 8, 15

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pp. 196–213. Springer, 2024. 14

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. *arXiv preprint arXiv:2402.08680*, 2024. 2

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 6, 7, 8

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *Forty-second International Conference on Machine Learning (ICML)*, 2025. 6, 7, 8

TECHNICAL APPENDICES

A RELATED WORK

A.1 LARGE VISION-LANGUAGE MODELS

Large vision-language models (LVLMs) have advanced rapidly through large-scale multimodal pretraining and instruction tuning, achieving strong performance across multimodal understanding and reasoning tasks. Representative paradigms include CLIP-style vision-language pretraining (Radford et al., 2021), Flamingo-style few-shot multimodal modeling (Alayrac et al., 2022), and BLIP-2-style modular alignment between frozen vision encoders and LLMs (Li et al., 2023a). Subsequent LVLMs, including LLaVA (Liu et al., 2023; 2024), InstructBLIP (Dai et al., 2023), OpenFlamingo (Awadalla et al., 2023), CogVLM (Wang et al., 2024), Kosmos-2 (Peng et al., 2023), and more recent models (Bai et al., 2025; Wang et al., 2025; Grattafiori et al., 2024; Chen et al., 2024a; 2023a) such as Qwen3-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025), further demonstrate the rapid progress of multimodal alignment and visual instruction tuning. This progression has substantially improved visual understanding and reduced many failure modes prevalent in earlier LVLMs. Nevertheless, faithful grounding between model responses and image content remains imperfect, particularly in fine-grained perception and reasoning, leaving hallucination as a persistent challenge even for stronger LVLMs.

A.2 HALLUCINATION IN LVLMS

Hallucination remains a persistent challenge in LVLMs, where model responses may be inconsistent with or insufficiently supported by the visual input. Early studies (Li et al., 2023c) primarily focus on object-level hallucinations, such as predicting non-existent objects, as evaluated by POPE (Li et al., 2023c). Subsequent benchmarks extend hallucination evaluation to finer-grained errors in attributes, counts, and relations, including AMBER (Wang et al., 2023), MME (Fu et al., 2023), HallusionBench (Guan et al., 2024), MMHal-Bench (Sun et al., 2024), and THRONE (Kaul et al., 2024). Collectively, these studies show that hallucination is heterogeneous and closely related to failures in visual grounding and multimodal reasoning. While stronger LVLMs have substantially reduced many hallucination errors observed in earlier models, the remaining failures persist across fine-grained perception and reasoning settings, motivating mitigation methods that can reliably verify model predictions against explicit visual evidence.

A.3 TRAINING-BASED HALLUCINATION MITIGATION

Early approaches improve faithfulness by redesigning instruction data or supervision signals so that models better distinguish grounded from ungrounded responses. For example, robust visual instruction tuning (Chen et al., 2023b) introduces hallucination-oriented supervision, while HACL (Jiang et al., 2024) uses contrastive learning to separate grounded and hallucinated representations. Reflective instruction tuning (Zhang et al., 2024) further improves reliability by adding rationale supervision. Alignment-based methods, such as factually augmented RLHF (Sun et al., 2024) and Silkie (Li et al., 2023b), incorporate preference or factual signals during post-training, and HIO (Lyu et al., 2024) strengthens token-level contrastive learning around hallucinated content. Overall, training-based methods are effective, but they usually require additional annotations, synthetic data, preference collection, or repeated optimization, leading to higher training cost and deployment complexity.

A.4 TRAINING-FREE HALLUCINATION MITIGATION

Training-free hallucination mitigation aims to improve LVLM reliability at inference time without updating model parameters. A major line of work intervenes directly on decoding dynamics or internal representations, including visual contrastive decoding (Sicong Leng, 2023), random image transformations (Woo et al., 2024), over-trust penalty and retrospection allocation (Huang et al., 2024), and representation-level visual enhancement (Huang et al., 2025b). These approaches are lightweight and broadly applicable, but largely rely on model-internal signals and therefore may inherit the same biases that underlie persistent hallucinations. Another line of work introduces post-hoc correction or external verification. Woodpecker (Yin et al., 2024) performs multi-stage hallucination correction

using external evidence, while DeGF (Zhang et al., 2025) leverages generative feedback for iterative refinement. Although such methods introduce additional information beyond the original LVLM prediction, external evidence may be noisy or incomplete, and introducing additional LVLMs for verification can confound the source of performance gains with verifier capability (Thakur et al., 2025; Huang et al., 2025a). Moreover, many of these methods were primarily developed and validated on earlier LVLMs, leaving their effectiveness on stronger contemporary backbones less clear. In contrast, `Kestrel` introduces an explicit visual grounding model as an external evidence source while retaining the same target LVLM for verification and refinement, avoiding performance gains that could otherwise be confounded by introducing an additional verifier LVLM. It further performs claim-level verification and evidence-gated iterative refinement to reduce unreliable corrections under noisy evidence.



Figure 6: **Qualitative results of** `Kestrel`**.** We compare the VQA responses from the regular baseline and our method based on Qwen3-VL. Zoom in for a better view.

B Qualitative Analysis

In Fig. 6, we present results of hallucination mitigation across existence, count, color, and position. The initial LVLM answers are inconsistent with the image content, but are corrected by `Kestrel` through external evidence from the grounding agent and multi-round self-refinement. The visual

Table 6: **Efficiency comparison.** We report the average inference latency per instance, peak GPU memory, and POPE accuracy on MS-COCO. Experiments are conducted on a single NVIDIA A100 GPU. Lower latency/memory is better, while higher POPE accuracy is better.

Method	Average Latency ↓	GPU Memory ↓	Checked Case	POPE (MS-COCO) ↑
Qwen3-VL (Bai et al., 2025)	0.78 s (×1.00)	17428 MB (×1.00)	9000	87.37
`Kestrel` (1st iteration)	12.00 s (×15.38)	21472 MB (×1.23)	9000	89.28
`Kestrel` (2nd iteration)	8.54 s (×10.94)	21296 MB (×1.22)	4978	**89.45**
`Kestrel` (3rd iteration)	7.12 s (×9.12)	21274 MB (×1.22)	477	<u>89.34</u>
`Kestrel`	18.75 s (×24.03)	21472 MB (×1.23)	9000	**89.45**

evidence provides explicit support for target localization and attribute verification, which helps recover missed objects, disambiguate object counts, reject unsupported color predictions, and correct erroneous spatial relations. Based on these grounded cues, the iterative refinement procedure progressively updates the answer through claim-level verification, yielding more reliable corrections than direct one-step revision. These results demonstrate that explicit grounding and conservative iterative refinement work together to effectively reduce hallucinations across diverse perception scenarios. More visual cases can be found in Appendix H.

C Efficiency

On MS-COCO (Lin et al., 2014), Tab. 6 quantifies the computational overhead introduced by explicit grounding and iterative verification. Compared with single-pass Qwen3-VL (Bai et al., 2025) inference, the full `Kestrel` pipeline increases end-to-end latency from 0.78,s to 18.75,s (×24.03), while peak GPU memory increases only moderately from 17.4,GB to 21.5,GB (×1.23). The additional latency mainly arises from external grounding, claim-level verification, and iterative refinement. However, early stopping substantially reduces the number of cases entering later rounds: all 9,000 samples are processed in the first iteration, whereas only 4,978 and 477 proceed to the second and third iterations, respectively. Consistent with this behavior, the best POPE accuracy is already achieved after the second iteration, while the third iteration produces only marginal changes. These results indicate that most refinement gains are obtained in the early rounds, while later iterations are invoked only for a small subset of unresolved cases, providing a practical trade-off between reliability and inference cost.

D Prompt Template

We provide the prompt templates used in the three core stages of `Kestrel`: Initialization (Fig. 7), claim-level verification (Fig. 8), and self-refinement (Fig. 9). These prompts follow the main evidence-grounded iterative framework, in which `Kestrel` first rewrites the original question into a concrete and visually verifiable claim, then verifies the claim against the collected evidence, and finally updates the answer conservatively through refinement. To improve reproducibility, all prompts adopt constrained JSON outputs with explicit field definitions. The initialization prompt produces a question-aligned claim for grounding, the verification prompt returns evidence-based verdicts and confidence scores, and the self-refinement prompt updates the answer while proposing a new claim for the next round.

Prompt template for Initialization

messages=[{"role":"user","content":[{"type":"image","image":sample["image"]},
{"type":"text","text": You are given an image and a Yes/No question. Determine the answer and output one verifiable claim.
Task requirements: - answer must be exactly "Yes" or "No". - output exactly one claim in verifiable_claims. - claim fields: id, type, text, targets. - text must be concrete and visually checkable.
Target rules: - if type = position: targets contain 1–2 short object phrases; for two-object relation, use [subject, anchor]. - otherwise: targets contain exactly one object noun from question; do not include color, number, or position words.
Return JSON only: { "answer":"Yes|No", "verifiable_claims":[{"id":"c1", "type":"sample["expected_claim_type"]", "text":"...", "targets":sample["example_targets"]}] }
Question: "sample["question"]" PrevSummary (optional): "sample["prev_summary"]" <image> "}]}]

Yes-Guard Template (optional, only when initial answer = Yes):
messages=[{"role":"user","content":[{"type":"image","image":sample["image"]},
{"type":"text","text": Check whether the Yes/No question is true in this image.
Return JSON only: { "answer":"yes|no|unclear", "confidence":"high|medium|low", "reason":"one short sentence" }
Rules: - use image evidence only; - yes: clearly true; no: clearly false; otherwise unclear; - be conservative when evidence is weak.
Question: "sample["question"]" Target hint: "sample["target_hint"]" <image> "}]}]

Figure 7: Prompt template for initialization stage.

Prompt template for claim-level verification

messages=[{"role":"user","content":[sample["evidence_content"], {"type":"text","text": You are a strict verifier. Judge each claim using ONLY the provided evidence items.
How to use evidence: - Each item has an EvidenceID and a Type. - Use only EvidenceID values that appear in the given evidence. - Typical IDs: - seg_overlay: e_seg_{tkey} - count_text: e_count_{tkey} - count_compare_text: e_countcmp_{tkey} - count_vision_text: e_countvis_{tkey} - count_vision_compare_text: e_countviscmp_{tkey} - crop_zoom: e_crop_{tkey} - color_text: e_color_{tkey} - position_text: e_pos_{tkey} - position_relation_text: e_posrel_{claim_id}
Judging rules: - For each claim, choose exactly one status: supported | contradicted | insufficient. - supported: evidence clearly confirms the claim. - contradicted: evidence clearly refutes the claim. - insufficient: evidence is missing or ambiguous. - Do NOT use common sense.
Top-level verdict: - contradicted: at least one claim is strongly contradicted. - supported: all claims are strongly supported. - otherwise: insufficient.
Return JSON only: { "verdict":"supported|contradicted|insufficient", "checked":[{ "claim_id":"...", "status":"supported|contradicted|insufficient", "confidence":0.0, "why":"...", "citations":["EvidenceID1","EvidenceID2"] }] }
Input: Question: "sample["question"]" Claims: sample["claims_json"] PrevVerdict (optional): sample["prev_verdict_json"] "}]}]

Figure 8: Prompt template for claim-level verification stage.

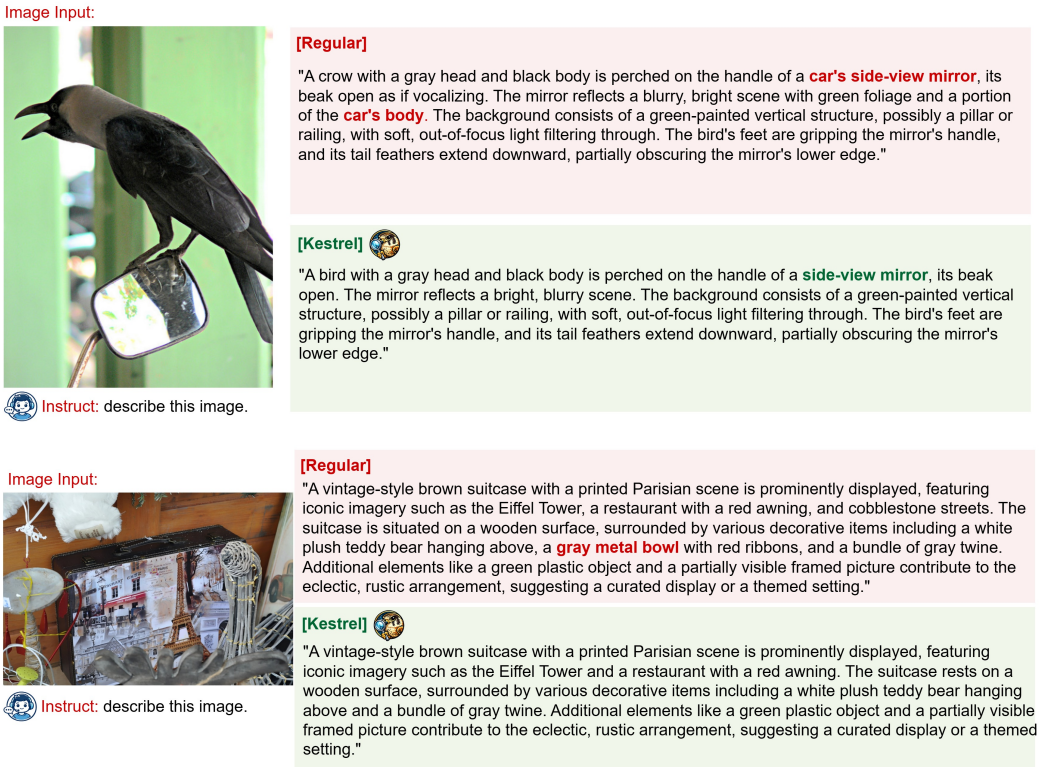


Figure 10: **Qualitative results on the CHAIR benchmark.** Image captions visualization results generated by the regular baseline and `Kestrel` based on Qwen3-VL.

Prompt template for Self-Refine

messages=[{"role":"user","content":[{"type":"image","image":sample["image"]}, {"type":"text","text": You are a cautious VQA assistant. Decide the final answer as a binary choice. HARD REQUIREMENT: - Answer must be exactly "Yes" or "No" (capitalized). - No punctuation, no explanation, no extra tokens. - You must output a concrete answer.
Use all available history: - previous rounds: hypothesis / verify / refine; - current round: hypothesis and verify details.
Also output one new_claim for the next round.
Claim constraints: - RELEVANCE: every new_claim must directly verify the Question's Yes/No. - claim.text must mention at least one key entity / attribute / relation from the Question.
Type-specific target rules: - if type = position: - targets contain 1–2 short object phrases from Question; - if two-object relation exists, use [subject, anchor]; - keep distinguishing modifiers when needed. - otherwise: - targets contain exactly one object noun from Question; - do not include color / number / position / quantifier words. - if a claim is not clearly usable to support/refute Question, it is invalid and must not be output.
Return JSON only: { "new_claims":[{"id":"c1", "type":"sample["expected_claim_type"]", "text":"...", "targets":sample["example_targets"], "priority":1}], "Answer":"YeslNo" }
Question: "sample["question"]" PreviousAnswer: "sample["A_t"]" RoundHistory: sample["round_history_json"] CurrentRoundContext: sample["current_round_context_json"] <image>"}]}]

Figure 9: Prompt template for self-refinement stage.

E HUMAN STUDY

To complement the automatic benchmark results, we conduct a human study to evaluate whether the responses produced by `Kestrel` are better aligned with human preference under evidence-grounded

judgment. To complement automatic benchmarks, we conduct a human preference study to assess whether `Kestrel` produces responses that are more reliable and more interpretable.

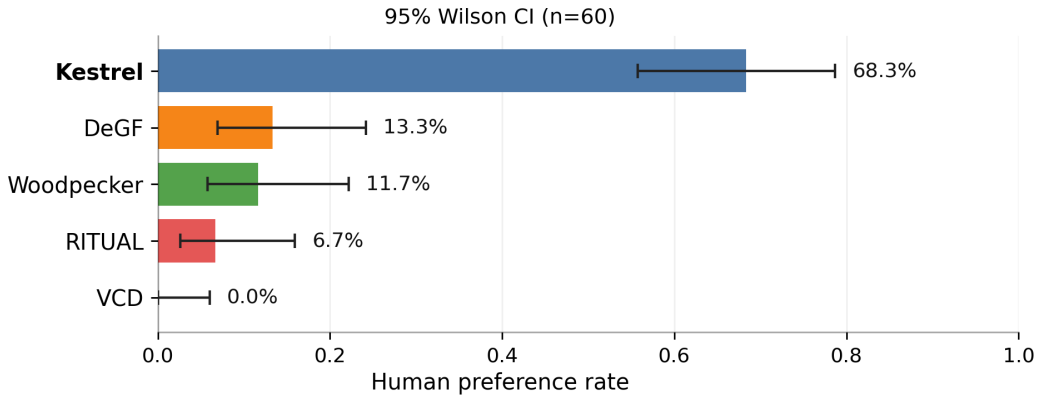


Figure 11: **Human preference comparison across methods.** Bars show the fraction of trials in which each method is preferred (n=60). Error bars denote 95% Wilson confidence intervals.

Setup. We sample 60 evaluation cases and compare five training-free mitigation methods: `Kestrel`, DeGF, Woodpecker, RITUAL, and VCD, which are also included in our main experiments. For each case, annotators are presented with the candidate outputs in randomized order and asked to choose the response they prefer based on grounding quality, factual consistency with the image, and overall answer reliability. We report the human preference rate of each method.

Results. As shown in Fig. 11, `Kestrel` is preferred in 41/60 cases (68.3%), substantially outperforming DeGF (13.3%), Woodpecker (11.7%), RITUAL (6.7%), and VCD (0.0%). This result indicates that `Kestrel` 's outputs are more consistent with human judgments of groundedness and answer quality. We attribute this advantage to `Kestrel` 's explicit evidence-backed verification process. Instead of relying only on decoding-time intervention, `Kestrel` externalizes grounding into structured visual and textual evidence, verifies question-aligned claims against citepd evidence, and updates the answer only when the evidence is sufficiently reliable.

F Caption Results

`Kestrel` shows clear gains on the CHAIR benchmark (Rohrbach et al., 2018) (see Tab. 7), indicating its effectiveness in reducing object hallucinations in image captioning. The improvement across both metrics suggests that it not only decreases the number of captions containing hallucinated objects, but also reduces the overall frequency of hallucinated object mentions. Fig. 10 demonstrates that `Kestrel` generalizes well to open-ended caption generation and improves the faithfulness of generated descriptions to the input image.

G Seed Robustness

To assess the stability of `Kestrel` under stochastic inference, we conduct a seed robustness study on the Qwen3-VL backbone using three different random seeds. As shown in Fig. 12, `Kestrel` exhibits highly consistent performance across seeds on POPE. Across the three seeds, the maximum performance variation is only 0.14, 0.05, and 0.17 points on MS-COCO, A-OKVQA, and GQA, respectively. A similar trend is observed on MME-Hallucination, where `Kestrel` achieves MME scores of 760, 755, and 750 under the three seeds, corresponding to a maximum variation of 10 points. These results indicate that `Kestrel` remains stable under the stochasticity introduced by the Qwen3-VL inference pipeline. We do not perform a seed sweep for InternVL3.5 because it is evaluated with deterministic decoding, under which repeated inference with different random seeds produces identical outputs. Therefore, varying the random seed does not affect its reported results.

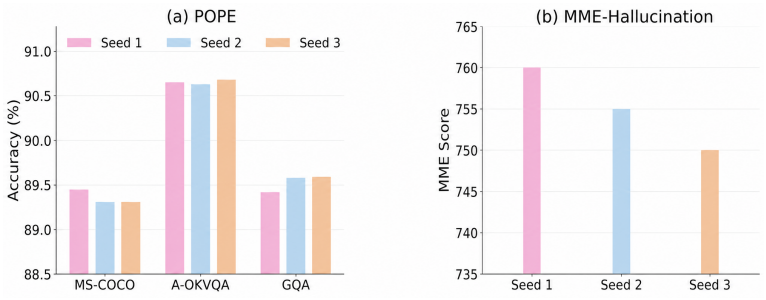


Figure 12: **Seed robustness of `Kestrel` on the Qwen3-VL backbone.** We report results across three random seeds on POPE and MME-Hallucination. `Kestrel` shows consistently stable performance across different seeds.

H More Case Studies

Fig. 13 and Fig. 14 present representative examples of `Kestrel` 's evidence-grounded self-refinement process. In each case, Qwen3-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025) initially produce an incorrect answer based on their first-pass visual impression, while `Kestrel` revises the prediction after introducing an explicit claim and verifying it against grounded evidence. The examples cover several queries, showing that the proposed framework is effective across different claim types. These cases highlight two key properties of `Kestrel`. First, the method can correct errors even when the initial response is confidently wrong, by converting free-form reasoning into claim-level verification. Second, the refinement is interpretable: each answer update is supported by explicit visual evidence, such as segmentation overlays, cropped regions, or structured text evidence describing existence or position. Overall, the qualitative results show that `Kestrel` improves faithfulness by grounding the refinement process in verifiable visual observations, rather than relying on unconstrained self-correction.

Table 7: **Results on Chair (Rohrbach et al., 2018) benchmark.** Lower scores (↓) indicate better performance. The best results are **bolded**, and the second-best are underlined.

Backbone	Method	$CHAIR_S$ ↓	$CHAIR_i$ ↓
Qwen3-VL (Bai et al., 2025)	Baseline (Bai et al., 2025)	36.76	10.50
	`Kestrel`$_{(ours)}$	**33.57**	**9.73**

`Kestrel` introduces a multi-stage, iterative inference pipeline that includes claim decomposition, external grounding, evidence construction, claim-level verification, and self-refinement. Although early stopping significantly reduces the number of refinement rounds in practice, the overall inference latency remains substantially higher than a single-pass LVLM inference due to the additional tool calls and verification steps. This overhead may limit the applicability of the framework in latency-sensitive or large-scale deployment scenarios. In future work, we plan to explore more efficient strategies for evidence-driven verification, such as adaptive tool invocation based on uncertainty signals, prioritizing high-risk claims for grounding, and reusing intermediate evidence across iterations. Such improvements could substantially reduce the computational overhead while preserving the benefits of explicit grounding and iterative verification.

I Limitations and Future Work

Limitations. `Kestrel` introduces additional inference overhead due to external grounding, claim-level verification, and iterative refinement. Although early stopping substantially reduces the number of samples entering later rounds, the end-to-end latency remains considerably higher than single-pass LVLM inference, which may limit deployment in latency-sensitive scenarios. In addition, `Kestrel` relies on the quality of the external grounding evidence; missed detections, inaccurate localization, or ambiguous visual evidence may propagate to subsequent verification and refinement. Finally, our current evaluation focuses primarily on hallucinations related to existence, count, position, and color, and therefore does not fully characterize the effectiveness of `Kestrel` on broader reasoning-intensive or open-ended hallucination settings. Future work may explore adaptive tool invocation, more robust evidence aggregation, and broader evaluation across diverse hallucination types.

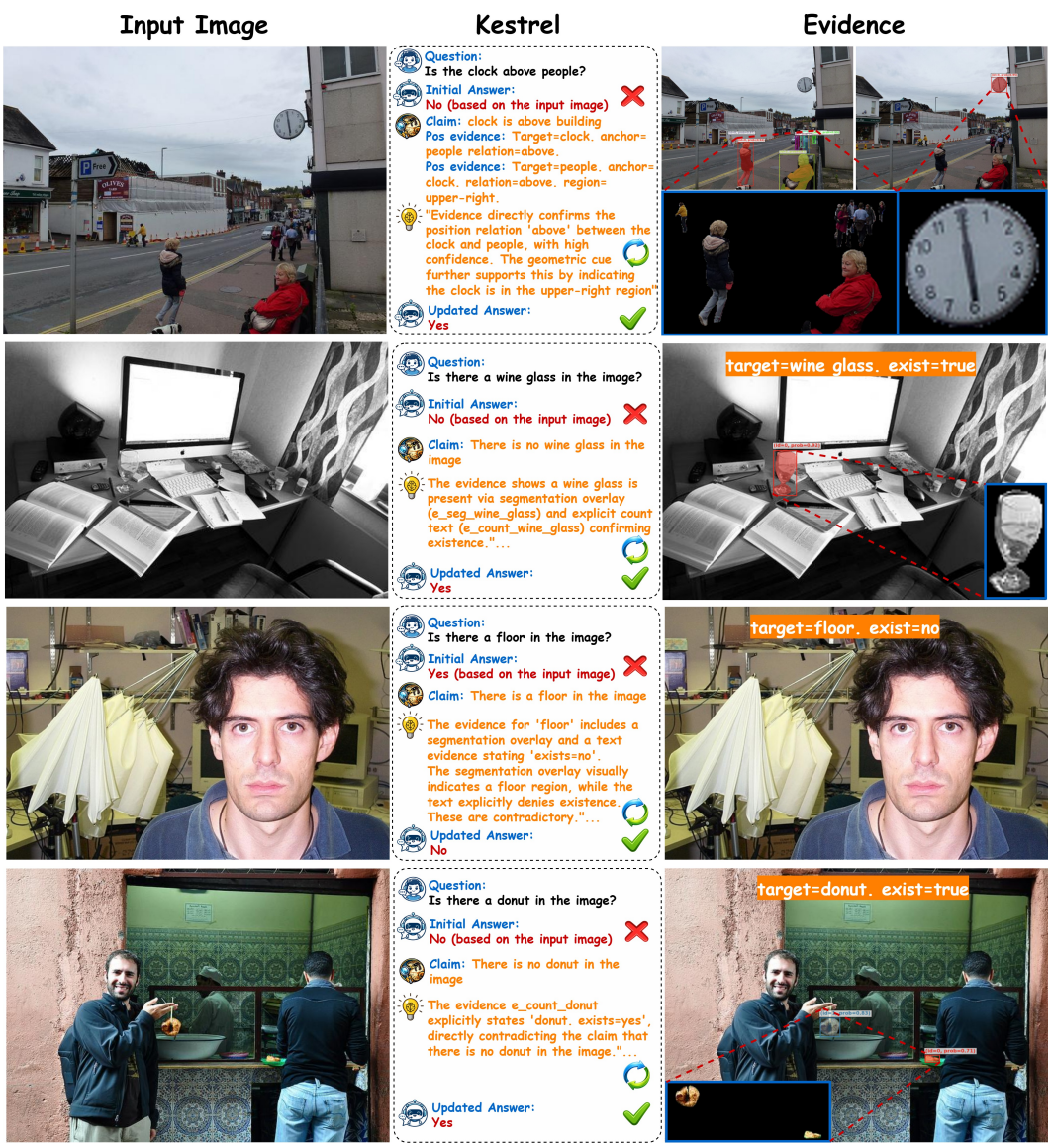


Figure 13: **Qualitative results of `Kestrel`.** We compare the VQA responses from the regular baseline and our method based on Qwen3-VL. Zoom in for a better view.

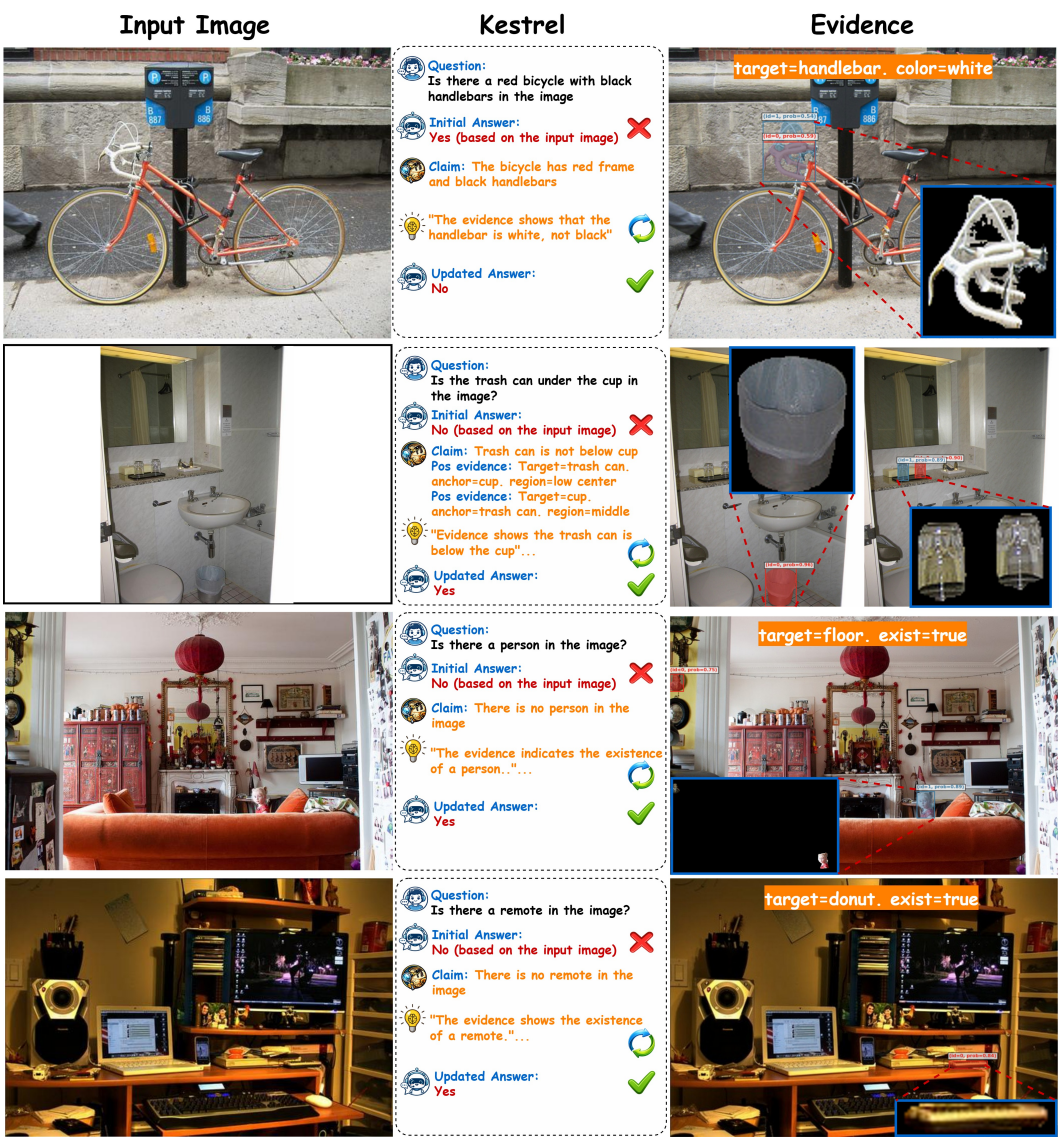


Figure 14: **Qualitative results of `Kestrel`.** We compare the VQA responses from the regular baseline and our method based on InternVL-3.5. Zoom in for a better view.