BEYOND CONFIDENCE: BASELINE-RELATIVE ANSWER REVISION FOR FROZEN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A better sampled answer does not necessarily justify replacing a vision-language model's first response. We study revision over a frozen candidate pool as two supervised decisions: selecting a proposal and estimating its value relative to the baseline. An evidence-aware ranker (FER) combines generation, visual-intervention and region-conditioned evidence; a Pairwise Utility Gate (UG) predicts the official-score change from replacement. Image-group-separated calibration follows each benchmark's split structure. The final policy improves the evaluated OK-VQA subset by $+1.198$ points and VQAv2 by $+0.6145$; GQA yields a statistically inconclusive $+0.064$. FER supplies the OK-VQA gain, while UG improves on FER on GQA and VQAv2. Matched schema controls, gain–harm frontiers, oracle headroom and cross-backbone results show where the decision layer helps and where it remains limited. The VLM is frozen, but ranking and acceptance require labeled supervision; the selected operating points favor score rather than sparse intervention.

1 INTRODUCTION

Large vision-language models (VLMs) can sample a better answer than their first response, yet replacing that response indiscriminately can overwrite an already correct answer. Figure 1 illustrates interacting generation uncertainty, incomplete visual support and mismatch with question-relevant regions. A high-likelihood answer need not be the best-supported one, and no single confidence value resolves these sources of error.

Revision therefore requires two judgments: which candidate is best supported, and whether it is better than retaining the baseline. Figure 2 separates proposal quality from replacement value. Even a strong proposal can be harmful when the baseline is stronger. We study this distinction with a two-stage supervised decision layer over fixed VLM outputs. FER ranks a cached candidate pool using multi-source evidence; UG estimates the selected proposal's official-score advantage over the baseline.

The contribution is an explicit baseline-relative learning target and a controlled empirical study of proposal selection and acceptance, not a new backbone or a claim that revision acceptance itself is new. We evaluate image-group-separated calibration, matched feature schemas, oracle headroom, gain–harm trade-offs and cross-backbone transfer. The results distinguish gains supplied by the ranker from those supplied by the gate and retain inconclusive and negative outcomes. "Frozen" refers to the VLM; both decision stages use labels. Acceptance chooses between answers, but does not imply infrequent changes: the reported thresholds maximize score.

2 RELATED WORK

VLM reliability and decoding. Modern VLMs combine visual encoders and autoregressive language models (Li et al., 2022; 2023; Dai et al., 2023; Liu et al., 2023; Bai et al., 2023; 2025). VCD contrasts original and distorted images (Leng et al., 2024); OPERA penalizes attention patterns and rolls back generation (Huang et al., 2024); other correction decoders strengthen visual information (Yin et al., 2025; Wang et al., 2024). We instead hold the backbone and sampled pool fixed.

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.

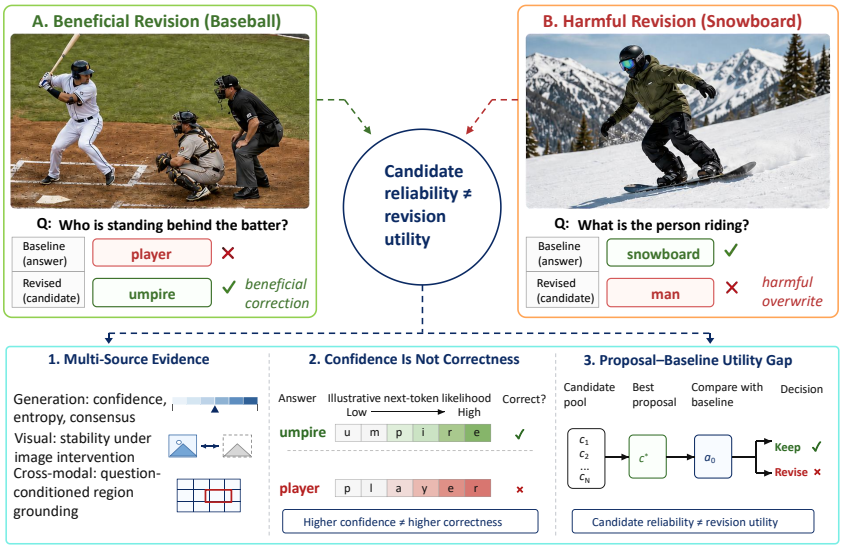


Figure 1: Why a VQA model can revise incorrectly despite a better available candidate. Generation uncertainty, visual ambiguity and cross-modal mismatch can interact. Likelihoods and distribution sketches are illustrative.

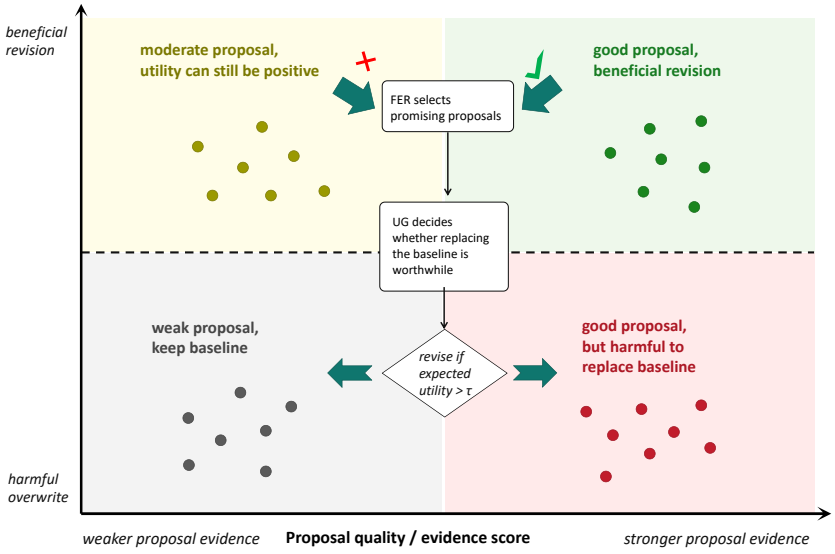


Figure 2: Proposal quality and revision utility are distinct. FER uses proposal-side evidence; UG compares the selected answer with the baseline. Arrows denote decision flow, not movement of points.

Verification and revision. Self-consistency aggregates repeated generations (Wang et al., 2023). Woodpecker verifies visual claims before rewriting (Yin et al., 2023), and Volcano learns self-feedback with revision acceptance or rejection (Lee et al., 2024). MMRefine evaluates self-correction (Paik et al., 2025); risk-aware prompting selectively invokes verification (Huang et al., 2026). Weak self-verification can reinforce errors (Xu et al., 2024; Zhang et al., 2024). Our distinction is external official-score utility estimation over a fixed pool, not the general separation of generation and acceptance (Supplementary Table S18).

Uncertainty and selection. Calibration aligns confidence and accuracy (Guo et al., 2017); selective prediction studies abstention or deferral (Geifman & El-Yaniv, 2017), with uncertainty affected

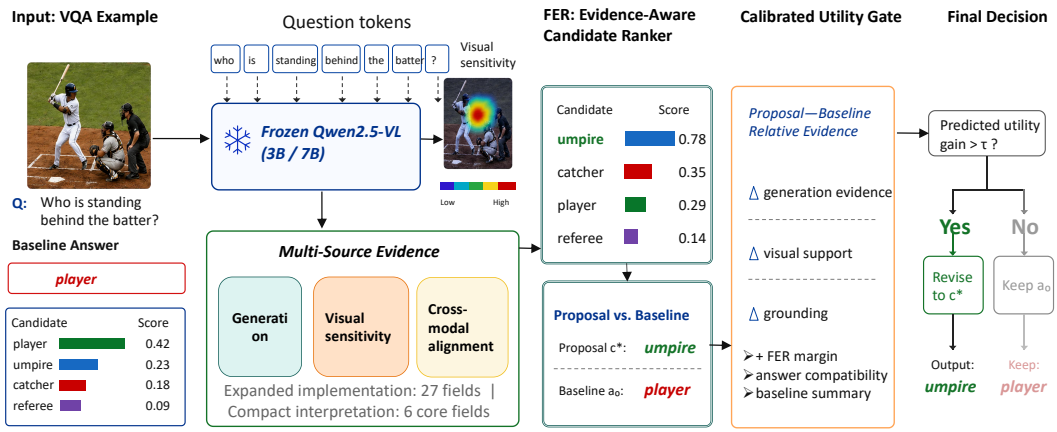


Figure 3: Evidence-guided revision. FER ranks fixed candidates; UG estimates baseline-relative replacement utility. Ranking and gating are supervised; VLM weights are unchanged. Scores and visual diagnostics are schematic.

by distribution shift (Ovadia et al., 2019). MBR and semantic uncertainty aggregate utility or answer equivalence (Freitag et al., 2022; Kuhn et al., 2023). Here the decision is between two concrete answers, rather than confidence calibration alone.

3 Problem Formulation and Method

3.1 Proposal Quality and Replacement Value

For image–question pair $x = (I, Q)$, a frozen VLM produces baseline a_0 and ordered pool $\mathcal{C}(x) = (c_1, \ldots, c_K)$. Let $\mathcal{S}(a; x) \in [0, 1]$ be the official per-example score. A proposal policy selects $c^\star$; an acceptance policy $g \in \{0, 1\}$ determines the final answer:

$$\Delta(c_i, a_0; x) = \mathcal{S}(c_i; x) - \mathcal{S}(a_0; x), \qquad \widehat{a} = \begin{cases} c^\star, & g = 1, \\ a_0, & g = 0. \end{cases} \tag{1}$$

Positive, negative and zero Δ denote beneficial, harmful and neutral proposals. Actual score change is $g\Delta(c^\star, a_0; x)$: KEEP has zero change regardless of the rejected proposal. If no candidate improves the baseline, neither ranker nor gate can create the missing answer.

3.2 Multi-Source Candidate Evidence and Ranking

Figure 3 summarizes the pipeline. Each pool has $K = 5$ records; duplicate normalized answers remain separate, and the baseline is not forcibly inserted. Cached prompt, whole-image and region probes supply evidence, not new pool candidates. Each probe response is compared with all candidates. Same-backbone policies share pools, normalization and evidence.

FER has 27 serialized slots covering confidence, entropy, consensus, prompt stability, visual sensitivity and region support, including transformed, aggregate and auxiliary views. They are not 27 independent mechanisms. FER-6 summarizes them as confidence, negative entropy, vote share, prompt consistency, visual support and region alignment. Supplementary Sections G–H retain the full inventory and mapping.

For a length-T candidate, average log probability is $\bar{\ell}(c) = T^{-1} \sum_t \log p_t(y_t)$ and entropy is $H(c) = -T^{-1} \sum_t \sum_v p_t(v) \log p_t(v)$. The archived confidence scalar $C(c)$ is stored separately; we do not equate it with sequence probability or $\exp \bar{\ell}(c)$. Plots use $-H$ for a support-positive direction. Vote share is normalized duplicate count. Visual support is agreement under blur, center crop, grayscale, JPEG, central occlusion and resizing. Region grounding combines crop agreement, blur agreement and mask disagreement as $0.40R_{\text{crop}} + 0.30R_{\text{blur}} + 0.30R_{\text{mask}}$. These probes are diagnostics, not internal attention or causal explanations.

A gradient-boosted regressor is supervised by official candidate score:

$$s_i^{\mathrm{FER}} = F_\theta(\mathbf{e}_i), \qquad c^\star = \arg\max_{c_i \in \mathcal{C}(x)} s_i^{\mathrm{FER}}. \tag{2}$$

Historical FER-6 and FER-27 use different training data; their deployment comparison is not a dimensionality ablation. Matched controls retrain both on identical rows and folds (Supplementary Section H).

3.3 Pairwise utility gate

UG has its own 28-field input, not a 27-dimensional candidate difference with extra features appended. Six answer summaries (confidence, negative entropy, consistency, grounding, stability and likelihood) contribute proposal values, baseline values and signed differences. Two ranker summaries and eight controls complete the input:

$$\mathbf{z} = \Pi[s^\star; m_{\mathrm{top2}}; \mathbf{t}(c^\star); \mathbf{t}(a_0); \mathbf{t}(c^\star) - \mathbf{t}(a_0); \mathbf{u}] \in \mathbb{R}^{28}. \tag{3}$$

Here m_{top2} is the leading-versus-runner-up FER gap, and Π restores serialized order. Controls comprise vote support, compatibility, normalized-answer equality, token-length support, baseline uncertainty and three answer-form indicators: $2 + 6 + 6 + 6 + 8 = 28$. Supplementary Table S20 gives the ordered fields.

$$\widehat{\Delta} = G_\psi(\mathbf{z}) \approx \mathbb{E}[\Delta(c^\star, a_0; x) \mid \mathbf{z}], \qquad g = \mathbf{1}[\widehat{\Delta} > \tau]. \tag{4}$$

UG uses squared-error supervision, median imputation, standardization and histogram gradient boosting; settings are retained in Supplementary Section M. The threshold uses labeled calibration data separated from evaluation image groups.

For a fixed proposal, the standard Bayes rule accepts when conditional expected gain is positive; plug-in regret is bounded by $\mathbb{E}|\widehat{\mu} - \mu|$ (Devroye et al., 1996). Supplementary Proposition S1 gives the proof. This characterizes the learning target, not a finite-sample guarantee or the performance of nonzero calibrated thresholds.

4 Experimental setup and evaluation protocol

We evaluate Qwen2.5-VL-3B-Instruct on the archived 3,789-question OK-VQA validation subset (Marino et al., 2019), GQA test-dev (12,576) (Hudson & Manning, 2019), and full VQAv2 validation (214,354) (Goyal et al., 2017). OK-VQA and VQAv2 use official soft scores; GQA uses exact match. The OK-VQA result is not for the entire public split.

Image-group separation has benchmark-specific constructions. OK-VQA uses five-fold cross-fitting; a row's image group is excluded from fitting its predictor and threshold. GQA fits/calibrates UG on 8,000 validation examples and deploys to question- and image-disjoint test-dev. VQAv2 has 36,896 labeled calibration questions on 7,581 images: all 39,996 validation questions on those images receive image-excluded OOF predictions, and the remaining 174,358 use the full-calibration gate. This is a hybrid estimate, not pure OOF. Full provenance is in Supplementary Tables S2–S3.

Thresholds are -0.10 for OK-VQA, -0.02 for GQA, and $0, 0, 0, 0, 0.02$ across VQAv2 OOF folds, with 0 on its disjoint complement. Coverage counts accepted proposals, not necessarily changed answer strings. We report official score, W2C/C2W transitions and 10,000 paired image-group bootstrap resamples (seed 20260722); macro changes are unweighted. These are supervised, score-oriented operating points. Unlabeled deployment, full inference latency and separately reported VQAv2 stratum scores are not established here.

5 Results

5.1 Same-backbone performance

Table 1 separates retention, proposal-only revision and calibrated acceptance. FER improves OK-VQA but degrades GQA and VQAv2. FER+UG reaches 60.5067 on OK-VQA ($+1.1982$, 95%

Table 1: Frozen Qwen2.5-VL-3B comparison. Cells give score / change from Base Keep under Section 4's evaluation constructions. Dashes are unreported results.

Method	Selection / gate	OK-VQA	GQA	VQAv2	Macro Δ
Base Keep	baseline	59.31/+0.00	60.86/+0.00	80.01/+0.00	+0.000
Self-consistency (Wang et al., 2023)	majority vote	58.29/−1.02	60.20/−0.66	–	–
Temp. scaling (Guo et al., 2017)	calibrated confidence	59.54/+0.23	61.05/+0.19	–	–
MBR consensus (Freitag et al., 2022)	consensus utility	55.45/−3.86	60.54/−0.32	–	–
Semantic uncertainty (Kuhn et al., 2023)	semantic clusters	55.63/−3.67	60.59/−0.27	–	–
VCD proxy (Leng et al., 2024)	visual contrast	52.25/−7.06	60.77/−0.10	80.42/+0.40	−2.253
FER-27	rank; always revise	60.56/+1.25	59.98/−0.88	79.88/−0.13	+0.079
FER-27+UG	calibrated acceptance	**60.51/+1.20**	**60.93/+0.06**	**80.63/+0.61**	**+0.625**

Table 2: Paired intervals and calibrated acceptance signals. Panel (b) shares proposals, folds and threshold-selection rules within each dataset, but does not match UG's multivariate supervised capacity.

(a) Paired image-group bootstrap

Dataset	Base	FER-27+UG / Δ	Paired 95% CI
OK-VQA	59.309	60.507 / +1.198	[0.517, 1.874]
GQA	60.862	60.926 / +0.064	[−0.297, 0.421]
VQAv2	80.012	80.627 / +0.6145	[0.550, 0.678]

(b) Calibrated acceptance signals

Signal	OK-VQA	GQA	VQAv2	Macro	Cov.
Conf. gap	+0.317	+0.080	+0.088	+0.162	53.9%
Entropy	+0.818	-0.103	+0.011	+0.242	27.1%
FER margin	+1.251	+0.000	+0.011	+0.421	34.3%
Pairwise UG	+1.198	+0.064	+0.6145	+0.625	90.3%

CI $[0.5169, 1.8744]$), 60.9256 on GQA ($+0.0636$, $[-0.2966, 0.4208]$) and 80.6269 on VQAv2 ($+0.6145$, $[0.5496, 0.6784]$). GQA's gain over Base Keep is inconclusive. Coverage is 98.839%, 90.768% and 81.376%; W2C/C2W counts are 122/73, 255/247 and 3,633/2,337. Macro gain is $+0.625$.

Adding UG to FER changes score by -0.0528, $+0.9462$ and $+0.7462$, respectively. FER supplies the OK-VQA gain. Recovery from a weaker proposal-only policy and improvement over Base Keep are distinct comparisons.

5.2 Acceptance signals and schema control

FER margin is strongest on OK-VQA, confidence gap slightly higher on GQA, and UG strongest on VQAv2 and in the macro average (Table 2). No signal dominates all datasets. Matched training gives FER-6 a selector CV mean of 66.0790 (SE 0.5778), versus FER-27's 66.2685 (0.6325); end-to-end differences are small and dataset-dependent. This controls schema size, not removal of evidence families. FER-27 is retained for deployment continuity (Supplementary Section H).

5.3 Qualitative outcomes, frontiers and headroom

Figure 4 shows why plausible proposal evidence alone does not justify replacement; Supplementary Section H.1 adds twelve success/failure cases. With FER fixed, Figure 5 separates gain and harm as acceptance varies. Mean calibrated coverage is 90.33%: high coverage is not evidence that all

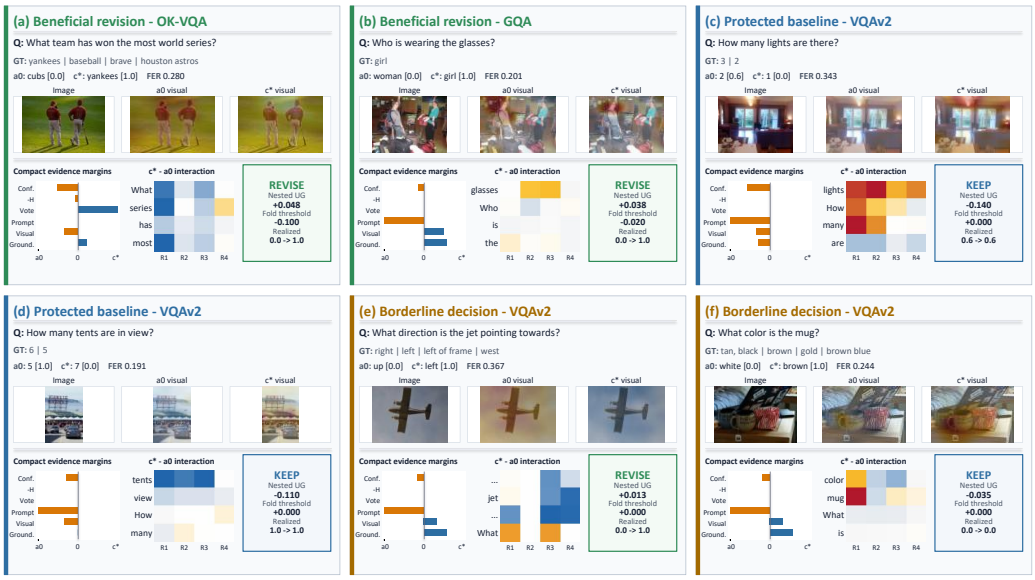


Figure 4: Beneficial revisions, protected baselines and borderline decisions. Six compact cues are a projection, not the full gate input. Maps are post-hoc diagnostics. "Nested UG"/"Fold threshold" are legacy labels; each row follows its dataset protocol.

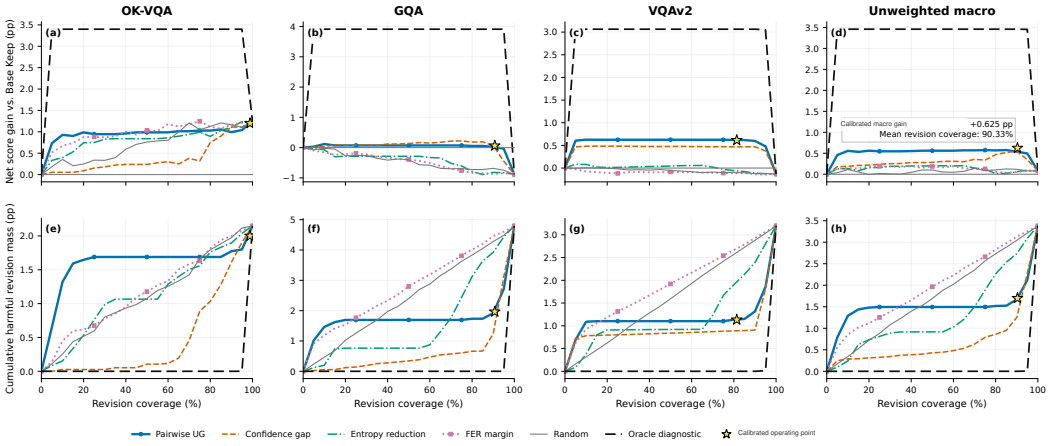


Figure 5: Revision utility frontiers with FER fixed. Rows show cumulative score gain and harmful mass versus coverage. Stars mark calibrated points; oracle ordering is diagnostic only. Columns give the three datasets and their unweighted macro average.

accepted proposals are beneficial. The baseline-inclusive pool oracle scores 72.125, 71.326 and 87.276 on OK-VQA, GQA and VQAv2; the final policy recovers 9.3%, 0.6% and 8.5% of available headroom. Both ranking and acceptance leave substantial room below the ceiling. Figure 6 reports historical confidence-stratified behavior. These curves use differently trained artifacts and cannot establish incremental evidence-family value; matched schema controls are separate.

5.4 Cross-backbone validation

On Qwen-7B OK-VQA, FER gains $+1.362$; UG retains $+1.288$ (CI $[0.412, 2.180]$) and reduces C2W from 156 to 148 at 98.76% coverage. Frozen transfer gains $+0.612$ on GQA and $+0.684$ on VQAv2-10K. BLIP-2's large improvements accompany low baselines and wide oracle headroom. LLaVA's OK-VQA/GQA changes are inconclusive; VQAv2-10K is slightly negative (-0.197, CI

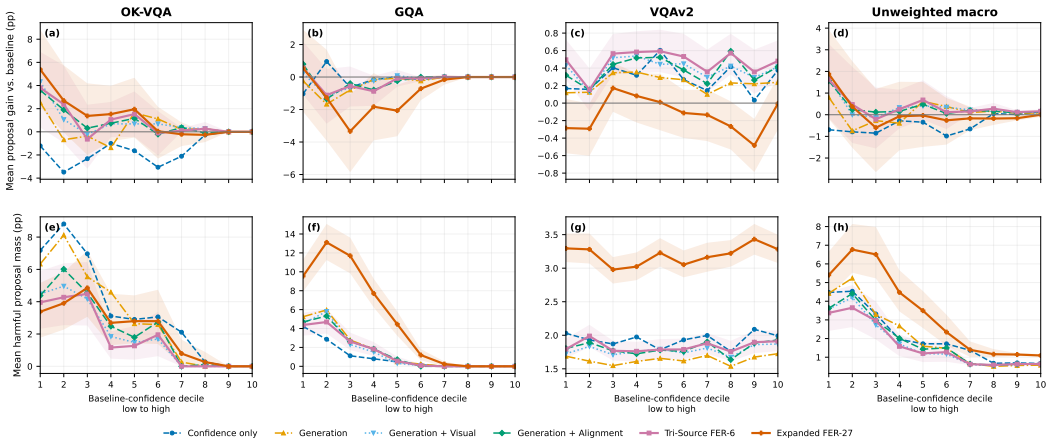


Figure 6: Historical confidence-stratified selectors: proposal gain and harmful mass by baseline-confidence decile. Differently trained FER-6/FER-27 curves are descriptive, not a matched evidence-family or dimensionality ablation.

Table 3: Cross-backbone scores. Qwen-7B uses cross-fitting on OK-VQA and frozen transfer on GQA/VQAv2; other families use model-specific calibration. VQAv2 uses a frozen 10K subset for Qwen and fresh 10K subsets for BLIP-2/LLaVA.

Backbone	Dataset	Base	Final	Δ
Qwen2.5-VL-7B	OK-VQA	53.138	54.426	+1.288
	GQA	59.876	60.488	+0.612
	VQAv2-10K	79.035	79.719	+0.684
BLIP-2/OPT-2.7B	OK-VQA	20.781	26.651	+5.870
	GQA	13.573	29.548	+15.975
	VQAv2-10K	19.564	42.091	+22.527
LLaVA-1.5-7B	OK-VQA	63.927	64.001	+0.074
	GQA	60.528	60.671	+0.143
	VQAv2-10K	79.034	78.837	-0.197

$[-0.410, 0.028]$). Table 3 retains absolute-score context; Supplementary Section I gives all intervals and transitions. Portability does not imply uniformly positive transfer.

6 Discussion and Conclusion

The evidence supports dataset-dependent benefits from separating proposal selection and baseline-relative acceptance. It does not establish superiority over an equally expressive, supervision-matched joint ranker, necessity of each evidence family, or uniform gains from the expanded schema. Most candidate-pool headroom remains unrecovered. Target-domain labels are needed; calibration-size curves, separate VQAv2 stratum results and full pipeline latency remain unreported. Missing historical decoder settings limit exact pool regeneration, despite auditable frozen outputs. The standard Bayes argument characterizes the target, not these empirical gaps. Within these boundaries, supervised revision improves OK-VQA and VQAv2 over Base Keep and recovers most FER degradation on GQA without changing VLM weights.

AI Use Statement

Generative AI assisted planning discussions, coding and offline audits, writing, figure preparation, mathematical exposition and LaTeX editing. Reported measurements come from recorded artifacts, not synthesized results. The authors remain responsible for verifying code, citations and results.

Ethics and reproducibility statements

Benchmark annotations may be noisy or biased, and individual revisions can be harmful despite aggregate gains; safety-critical suitability is not established. The independently compiled supplement retains proofs, feature definitions, provenance, controls and all extended figures/results. Missing decoder settings limit exact regeneration. The source package is not a substitute for experimental code, weights or full caches.

References

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.

Shuai Bai et al. Qwen2.5-vl technical report, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, New York, 1996. doi: 10.1007/978-1-4612-0711-5.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 2022.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

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.

Yuang Huang, Yafeng Zhang, and Zilan Yu. Risk-aware selective prompting for hallucination mitigation in large vision-language models. *arXiv preprint arXiv:2605.28123*, 2026. doi: 10.48550/arXiv.2605.28123.

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*, 2019.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 391–404, 2024. doi: 10.18653/v1/2024.naacl-long.23.

Sicong Leng, Hang Zhang, Guanzhong Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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

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*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

Gio Paik, Geewook Kim, and Jinbae Im. Mmrefine: Unveiling the obstacles to robust refinement in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26883–26904, 2025. doi: 10.18653/v1/2025.findings-acl.1378.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: Llm amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15474–15492, 2024. doi: 10.18653/v1/2024.acl-long.826.

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.

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.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. Small language models need strong verifiers to self-correct reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15637–15653, 2024. doi: 10.18653/v1/2024.findings-acl.924.