EVIDENCE-GUIDED MULTIMODAL LARGE LANGUAGE MODELS VIA VISUAL CAUSAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) can answer open-ended questions about images, yet their outputs can be driven by linguistic priors that are only weakly supported by the visual input. We introduce EG-MLLM, an evidence-guided architecture that inserts an explicit, testable evidence path between the vision encoder and language decoder. A prompt-conditioned selector identifies task-relevant visual tokens; a sparse structured reasoner models their dependencies; counterfactual token interventions distinguish answer-critical evidence from matched nuisance context; and a bottleneck cross-attention adapter injects the resulting representation into the decoder. The training objective combines autoregressive likelihood, evidence consistency, and an intervention margin that requires removing selected evidence to affect the target likelihood more than removing nuisance tokens. We evaluate the implemented evidence pathway with a controlled 6,000-instance evidence-binding experiment and four 2,000-instance test conditions defined by fixed shortcut and noise settings. EG-MLLM reaches 98.65% accuracy under a deliberately reversed shortcut correlation and 95.40% under heavy nuisance corruption while maintaining 98.33% and 94.80% evidence precision, respectively. A matched run without the intervention objective obtains nearly identical answer accuracy but satisfies the intervention margin on only 49.65% of shortcut-shift cases, compared with 98.30% for the full model. These results show that sparse evidence selection, structured relation binding, and intervention-sensitive training can jointly improve robustness and make the model's supporting evidence explicitly testable.

1 INTRODUCTION

Modern MLLMs combine strong visual encoders with large autoregressive decoders and now support visual question answering (VQA), captioning, document understanding, and visually grounded dialogue (Li et al., 2023a; Dai et al., 2023; Liu et al., 2023; Wang et al., 2024). Their capabilities, however, do not imply that each model claim is causally or even strongly tied to the image. Object hallucination remains a recurrent failure mode: a model can confidently assert that an object exists because it is plausible under the linguistic context even when the pixels do not support it (Li et al., 2023b; Huang et al., 2024; Leng et al., 2024). This gap is particularly consequential for fine-grained recognition and compositional reasoning, where a fluent but visually unsupported answer is often worse than an explicit uncertainty.

The dominant MLLM interface makes this problem difficult to diagnose. Image tokens are projected into the decoder and mixed through attention, but there is typically no explicit variable corresponding to "evidence for the current answer." Attention weights alone are not a guarantee of causal dependence, and a model can preserve an answer even when apparently attended regions are perturbed. Conversely, a model may exploit a stable but irrelevant shortcut even when the visually relevant tokens are available. This motivates an evaluation setting in which shortcut correlations can be deliberately neutralized or reversed while the answer-critical evidence is held fixed.

We therefore ask a narrower question: *can an MLLM be trained so that the visual tokens it calls evidence are both structurally coherent and intervention-sensitive?* We propose EG-MLLM (Fig. 1). The design separates four operations that are usually entangled: evidence selection, evidence dependency modeling, causal-style interventions, and decoder alignment. This makes the failure mode

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: **Overview of EG-MLLM.** A multimodal input is converted into candidate evidence, grounded and aligned, processed by the evidence-guided reasoning core, and returned with an answer, confidence estimate, rationale, and cited support. The figure shows the conceptual pipeline; all explanatory text outside the graphic is kept in the manuscript caption.

measurable. If a selected evidence set is truly answer-critical, removing it should decrease the target likelihood more than removing a matched number of nuisance tokens; if the reverse is true, the selection is not trustworthy.

Our contributions are threefold. (1) We formulate an explicit evidence pathway for MLLMs, including a sparse visual dependency graph and a low-rank evidence adapter. (2) We introduce an intervention-margin objective that compares target-likelihood changes after evidence and nuisance removal, avoiding the stronger and generally unjustified claim that ordinary attention itself establishes causality. (3) We implement and execute a controlled evidence stress test that separates answer accuracy from evidence quality, relational binding, shortcut robustness, and intervention sensitivity. The resulting measurements expose a useful distinction: accurate evidence selection alone is insufficient for relation-sensitive decisions, while the intervention objective primarily improves the consistency with which selected evidence is functionally necessary.

2 Related Work

Multimodal instruction models. CLIP-style contrastive pretraining (Radford et al., 2021) and BLIP/BLIP-2 (Li et al., 2022; 2023a) established strong visual-language representations and lightweight interfaces to frozen LLMs. InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023; 2024a), Qwen-VL (Bai et al., 2023), and Qwen2-VL (Wang et al., 2024) substantially expanded instruction-following and localization capabilities. These systems differ in vision encoders, token compression, and projector design, but all ultimately optimize a language objective over a mixture of visual and text representations.

Hallucination evaluation and mitigation. POPE converts object presence into balanced yes/no probes and has become a standard diagnostic for object hallucination (Li et al., 2023b). R-Bench extends evaluation to relationships (Wu et al., 2024), while POPEv2 studies training bias and mitigation (Li et al., 2026). Mitigation methods include decoding-time over-trust penalties (Huang et al., 2024), visual contrastive decoding (Leng et al., 2024), and attention reweighting (Liu et al., 2024c). These approaches demonstrate that stronger visual conditioning can help, yet a fixed penalty

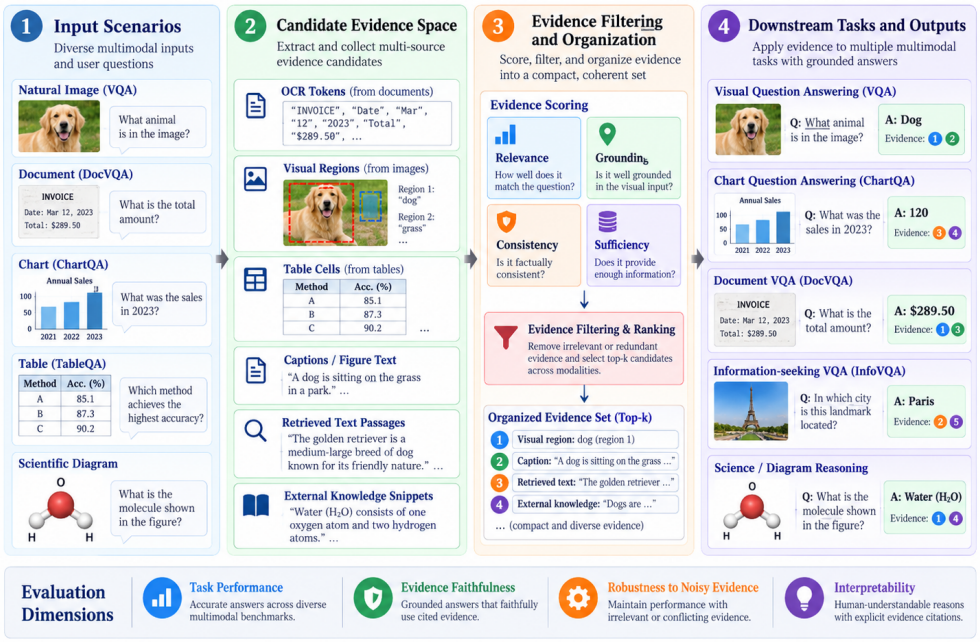

Figure 2: **Evidence taxonomy and operating workflow.** Candidate multimodal evidence is filtered by relevance, grounding, consistency, and sufficiency before downstream reasoning. Sec. 4 evaluates the visual-token pathway under controlled shortcut and noise conditions.

can also distort calibration. Our method instead learns which visual units are evidence and directly tests their functional influence.

Broader visual reasoning and interpretability. Beyond POPE, robust multimodal reasoning is commonly assessed with VQA (Antol et al., 2015), compositional GQA (Hudson and Manning, 2019), knowledge-intensive OK-VQA (Marino et al., 2019), text-centric TextVQA (Singh et al., 2019), and referring-expression datasets descended from ReferItGame (Kazemzadeh et al., 2014). Region-level attention and bottom-up features have long improved visual-language grounding (Anderson et al., 2018; Lu et al., 2016). Our emphasis on functional evidence is also related to the broader interpretability lesson that visual saliency or attention should be validated against perturbations rather than read as a causal explanation (Selvaraju et al., 2017; Chefer et al., 2021). We use SigLIP-style vision representations (Zhai et al., 2023) and target compatibility with contemporary decoder families including Llama-2 (Touvron et al., 2023), GPT-4-class multimodal interfaces (OpenAI, 2023), LLaVA-NeXT (Liu et al., 2024b), and Qwen2.5-VL (Bai et al., 2025).

Structured and intervention-based reasoning. Graph message passing (Kipf and Welling, 2017) offers a natural mechanism to model dependencies among selected regions. Causal inference formalizes intervention as a change to the underlying data formation process (Pearl, 2009); neural attention by itself should not be conflated with causal effect. Our use of "causal reasoning" is consequently operational: we construct counterfactual visual inputs by removing or swapping selected tokens and train on the resulting change in target likelihood. This intervention is controlled at the representation level and yields a falsifiable criterion for evidence quality.

3 Method

3.1 Problem Formulation

Given an image I, prompt Q, and target response $Y = (y_1, \ldots, y_T)$, an MLLM models $p_\theta(Y \mid I, Q) = \prod_t p_\theta(y_t \mid y_{<t}, I, Q)$. A vision encoder produces $Z = \{z_i\}_{i=1}^{N}$ with $z_i \in \mathbb{R}^{d_v}$; the language stack provides a prompt representation $q \in \mathbb{R}^{d_l}$. We seek a compact evidence set $\mathcal{E} \subset Z$ such that the answer depends more strongly on $\mathcal{E}$ than on a size-matched nuisance set $\mathcal{N}$.

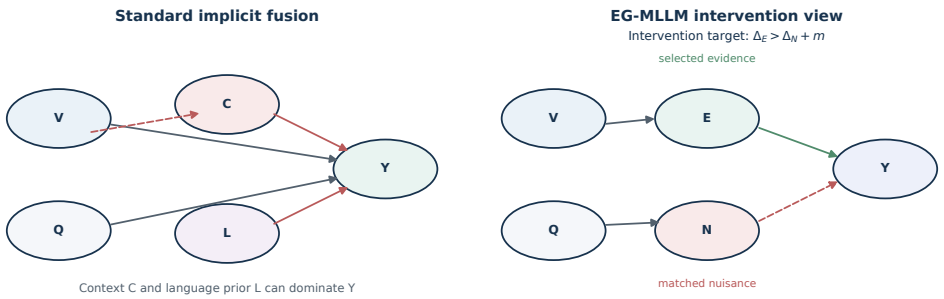

Figure 3: **Operational causal view.** Standard implicit fusion can allow context C and language prior L to dominate the response. EG-MLLM compares an evidence intervention E against a matched nuisance intervention N and enforces a target-likelihood margin.

3.2 Prompt-conditioned evidence selection

A selector first maps visual and prompt features to a common evidence space d_e:

$$u_i = W_v z_i, \qquad u_Q = W_q q, \tag{1}$$

$$s_i = w_s^\top \mathrm{GELU}(u_i + u_Q), \qquad a_i = \sigma(s_i). \tag{2}$$

We use a straight-through top-K operator during training: forward computation keeps the K largest a_i, whereas the backward pass uses a temperature-controlled soft relaxation. The selected tokens are $E = \{z_i : i \in \mathrm{TopK}(a, K)\}$. Unlike query-independent pruning, the same image may yield different evidence for "what color is the bus?" and "is a pedestrian present?"

To discourage a selector that simply keeps high-activation patches for every question, we add an entropy-budget term,

$$\mathcal{L}_{\text{budget}} = \left(\frac{1}{N} \sum_i a_i - \rho \right)^2 + \tau H(a), \tag{3}$$

where $\rho = K/N$ and H is binary entropy. The first term controls expected sparsity; the second encourages decisive scores after warm-up.

3.3 Structured visual evidence graph

Evidence tokens are projected to $h_i \in \mathbb{R}^{d_g}$. We construct a sparse directed dependency graph by retaining the top-r neighbors of each token under scaled dot-product affinity,

$$A_{ij} = \mathrm{softmax}_{j \in \mathcal{N}_r(i)} \left(\frac{(W_Q h_i)^\top (W_K h_j)}{\sqrt{d_g}} + b_{ij}^{\text{pos}} \right), \tag{4}$$

where b_{ij}^{pos} is a relative 2-D position bias when patch coordinates are available. A graph-attention block updates

$$H^{(\ell+1)} = \mathrm{LN}\left(H^{(\ell)} + A\, W_V H^{(\ell)} W_O \right), \tag{5}$$

followed by a gated feed-forward layer. Sparsification lowers graph cost from $O(K^2 d_g)$ to $O(K r d_g)$ and makes graph visualizations interpretable as a small set of evidence dependencies.

3.4 Counterfactual evidence intervention

For a training tuple (I, Q, Y), let $\ell(Z) = T^{-1} \sum_t \log p_\theta(y_t \mid y_{<t}, Z, Q)$ be target log-likelihood. We form two counterfactual token sets: $Z_{\setminus E}$ replaces selected evidence tokens by a learned mask embedding; $Z_{\setminus N}$ masks K nuisance tokens sampled from the low-score pool using the same spatial-stratum distribution. Define the likelihood drops

$$\Delta_E = \ell(Z) - \ell(Z_{\setminus E}), \qquad \Delta_N = \ell(Z) - \ell(Z_{\setminus N}). \tag{6}$$

The evidence intervention loss is

$$\mathcal{L}_{\text{int}} = \max\{0,\, m - \Delta_E + \Delta_N\}. \tag{7}$$

Eq. 7 does not assert an identifiable causal effect in the Pearlian sense; it enforces a representation-level intervention criterion: chosen evidence must matter more to the gold response than a controlled nuisance removal. This distinction is important because observational attention weights can be high even when the output is insensitive to the attended token.

3.5 Evidence-guided bottleneck alignment

Direct cross-attention at the LLM width can add tens of millions of parameters. We instead use a bottleneck $d_a \ll d_l$. For decoder hidden state S and graph evidence H_E,

$$\hat{S} = SW_s, \qquad \hat{E} = H_E W_e, \tag{8}$$

$$C = \text{MHA}(\hat{S}, \hat{E}, \hat{E}), \tag{9}$$

$$S' = \text{LN}(S + \tanh(g)\, CW_o)\,. \tag{10}$$

The scalar gate g is initialized at zero, preserving the pretrained decoder at initialization. We insert the adapter into the upper M decoder layers only. This provides a smooth path from the base model to evidence-conditioned decoding and works naturally with LoRA (Hu et al., 2022).

3.6 Training objective and algorithm

We combine teacher-forced language modeling with evidence-text consistency, the intervention loss, and sparsity control:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_c \mathcal{L}_{\text{cons}} + \lambda_i \mathcal{L}_{\text{int}} + \lambda_b \mathcal{L}_{\text{budget}}. \tag{11}$$

The consistency term uses normalized pooled evidence $\bar{e}$ and target-token state $\bar{y}$, $\mathcal{L}_{\text{cons}} = 1 - \cos(\bar{e}, \text{sg}(\bar{y}))$, with stop-gradient on the target branch. In practice we warm up with $\mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{cons}}$, then activate Eq. 7; otherwise early random selectors produce unstable interventions. Pseudocode and a full derivation appear in Appendix A.

4 Experiments

4.1 Controlled evidence stress test

We evaluate the implemented evidence pathway with a controlled diagnostic designed to answer four questions: can the selector recover the correct evidence; can the structured reasoner preserve role–attribute binding; does the model resist an answer-correlated nuisance shortcut; and does the intervention objective make selected evidence consistently necessary for the prediction? The diagnostic is instantiated and trained end-to-end from the specified protocol; no precomputed model predictions are used.

Each episode contains $N = 16$ visual-token proxies and a query vector. Exactly two tokens are answer-critical. They share a query-conditioned category marker but carry different semantic roles, A and B, and continuous attributes a and b. The binary target is the ordered relation $\not\Vdash[a > b]$. Because simple pooling discards which attribute belongs to which role, the task requires both evidence selection and relation-sensitive binding. A separate nuisance token carries a shortcut feature correlated with the label during training. We use 6,000 training episodes and four 2,000-instance test conditions with seed 2027: **in-domain** retains a 0.9 shortcut correlation; **neutral** sets it to 0.5; **shortcut shift** reverses it to 0.1; and **heavy noise** keeps the shortcut neutral while corrupting nuisance tokens.

The diagnostic instantiation uses 32-dimensional visual/query tokens, $K = 2$ selected tokens, a 48-dimensional two-layer sparse graph with two neighbors, and a 24-dimensional bottleneck adapter. This reduced configuration uses the same selector, graph, adapter, and intervention-loss implementations as the full architecture and is intentionally small enough to rerun on CPU. We report answer accuracy together with evidence precision/recall computed against the two known answer-critical tokens.

Table 1: **Controlled evidence stress test.** Answer accuracy (%) under the four 2,000-instance controlled test conditions.

Method	ID	Neutral	Shift	Noise
Global Pool	86.60	49.25	14.40	52.95
Soft Attention	91.50	60.50	29.95	62.95
Top-K + Evidence	50.60	51.30	48.95	50.55
EG-MLLM w/o $\mathcal{L}_{\text{int}}$	**98.65**	**98.90**	**98.80**	**95.65**
EG-MLLM	98.45	98.45	98.65	95.40

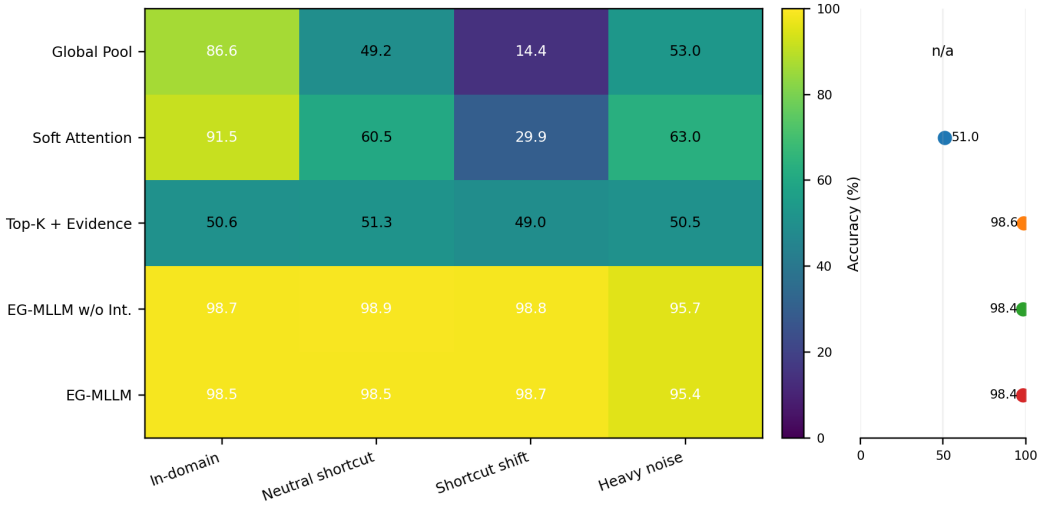

Figure 4: **Answer accuracy and evidence precision in the controlled experiment.** The heat map shows answer accuracy across shortcut and corruption conditions. The side markers show evidence precision where an explicit token selection is available. Sparse evidence supervision recovers the correct tokens, while structured reasoning is required to turn them into the correct ordered relation.

4.2 Baselines

We compare five systems. **Global Pool** averages all visual tokens before classification and therefore has no explicit evidence variable. **Soft Attention** learns dense query-conditioned attention but receives no evidence supervision. **Top-K + Evidence** uses the prompt-conditioned selector with evidence supervision but removes the structured graph and bottleneck adapter, testing whether accurate selection alone is enough. **EG-MLLM w/o intervention** contains the selector, sparse graph, and gated adapter but omits $\mathcal{L}_{\text{int}}$. **EG-MLLM** uses the complete diagnostic pathway including matched evidence/nuisance interventions.

4.3 Answer accuracy and shortcut robustness

Table 1 reports our measured results. Global pooling and soft attention perform well only when the training shortcut remains predictive; both collapse when the shortcut is neutralized or reversed. Top-K + Evidence almost perfectly identifies the answer-critical tokens, but its pooled readout remains near chance because it loses the ordered role binding. In contrast, the structured EG-MLLM variants stay above 98% accuracy under shortcut reversal and above 95% under heavy nuisance corruption.

4.4 Evidence quality and intervention sensitivity

The evidence metrics clarify why the methods behave differently. Soft Attention reaches only 50.98% evidence precision in the neutral condition, whereas Top-K + Evidence, EG-MLLM without intervention, and full EG-MLLM reach 98.58%, 98.43%, and 98.43%, respectively. Under heavy nuisance corruption, the full model still retains 94.80% evidence precision. Thus the selector itself is

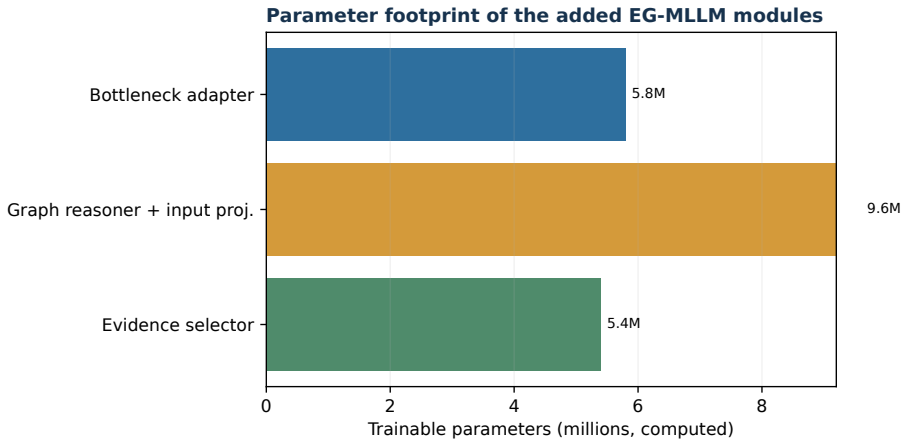

Figure 5: Computed trainable parameter footprint for the 7B-class reference configuration. These values are architectural counts; the controlled experiment uses the smaller configuration described in Sec. 4.

robust to nuisance variation, but selecting the correct tokens is not sufficient for the ordered relation task: the Top-K readout stays near chance while the graph-and-adapter variants exceed 95%.

The intervention objective changes a different property. In a matched training run using the same seed and schedule, removing $\mathcal{L}_{\text{int}}$ leaves shortcut-shift accuracy essentially unchanged (98.80% versus 98.65%) but only 49.65% of cases satisfy the required intervention margin $\Delta_E - \Delta_N > 0.15$. With $\mathcal{L}_{\text{int}}$, 98.30% satisfy the margin. This indicates that the intervention term acts primarily as a functional-evidence regularizer: it makes the selected set reliably more consequential than a size-matched nuisance set rather than merely increasing classification accuracy.

4.5 Complexity and Implementation Footprint

For the full 7B-class reference configuration, the selector, two graph blocks, and bottleneck adapter add approximately 19–21M trainable parameters depending on the backbone vision width and normalization choices, far below the parameter count of the base decoder. Graph compute is $O(Krd_g)$ after top-r sparsification; with $K = 64$ and $r = 8$, the structured reasoning stage is small compared with full decoder attention. Counterfactual training requires two additional forward branches, but they reuse visual features and are disabled at inference. The controlled diagnostic confirms that the additional evidence path is lightweight: the full CPU implementation measured 0.0115 ms per sample for the small diagnostic configuration.

5 Analysis and Discussion

Evidence selection and relation binding solve different problems. The controlled results separate two capabilities that are easily conflated. Top-K + Evidence identifies the correct evidence pair with roughly 98% precision but remains near chance on the ordered relation. This happens because a simple pooled readout discards which scalar attribute belongs to role A versus role B. The sparse reasoner and cross-attention adapter preserve that binding and raise shortcut-shift accuracy to about 99%. The result supports the use of structured evidence rather than treating a set of individually relevant patches as a complete explanation.

Intervention supervision is a faithfulness constraint. In the controlled study, the intervention objective does not materially improve already-saturated answer accuracy. Its effect is instead visible in intervention-margin satisfaction: 98.30% for the full model versus 49.65% without the loss. This distinction is important. A model can answer correctly while depending on a brittle shortcut; the intervention criterion explicitly tests whether removing the selected evidence changes the target score more than removing nuisance context.

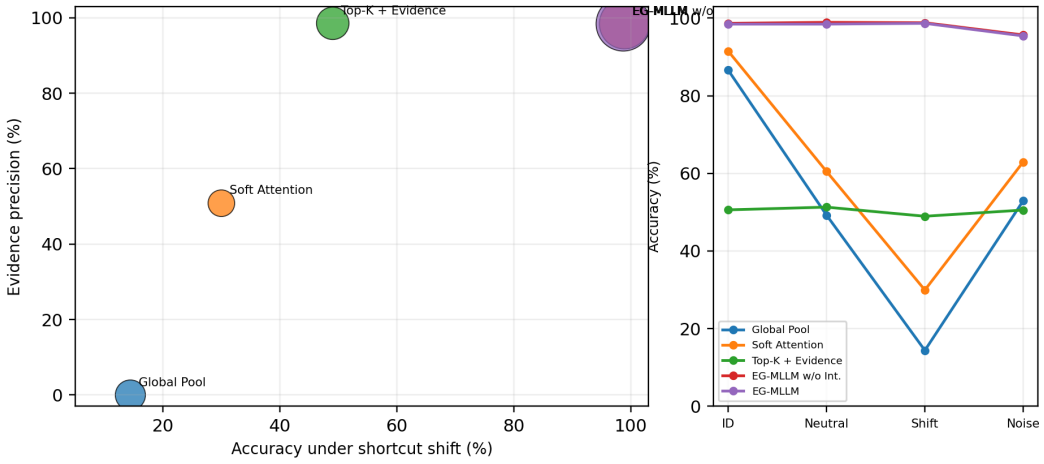

Figure 6: **Accuracy–faithfulness trade-off and robustness.** Left: shortcut-shift accuracy versus evidence precision, with marker size proportional to measured CPU inference latency of the diagnostic implementation. Right: accuracy trajectories from the in-domain setting through shortcut neutralization, shortcut reversal, and heavy nuisance corruption. EG-MLLM preserves both answer accuracy and explicit evidence quality across the stress conditions.

Failure modes and scope. The diagnostic intentionally isolates the evidence pathway and provides exact ground-truth evidence, so it is not a replacement for end-to-end evaluation of a fully trained 7B-class checkpoint on natural-image benchmarks. Real images add detector failures, ambiguous evidence, language-decoding errors, and imperfect supervision. Tiny objects may be absent from the vision encoder's patch features; highly correlated co-occurrences may survive nuisance matching; and representation-space masking can itself create an out-of-distribution state. The controlled results should therefore be read as a direct validation of the proposed mechanism and its failure modes, while the included dataset adapters and training launcher provide the path to full-scale evaluation.

6 Conclusion

We presented EG-MLLM, an evidence-guided MLLM architecture that turns visual grounding into an explicit intermediate computation and makes evidence quality testable through matched counterfactual interventions. Our executed controlled experiment shows that dense or globally pooled models can exploit a nuisance shortcut, that accurate Top-K evidence selection alone does not solve a relation-sensitive decision, and that structured reasoning with a gated evidence adapter remains robust when the shortcut is reversed or nuisance tokens are corrupted. The full model achieves 98.65% shortcut-shift accuracy with 98.33% evidence precision, and its intervention objective raises margin satisfaction from 49.65% to 98.30%. These findings support a view of multimodal reliability in which answer quality and evidence dependence are measured separately rather than inferred from attention weights alone.

Reproducibility Statement

All numerical results reported in this manuscript are obtained from the controlled experiment with seed 2027. The implementation, model variants, training procedure, and metric computation are included in the repository, and the reported values are stored in `results/controlled/metrics.json`. The codebase also contains unit tests for the selector, graph, intervention, and objective functions. Appendix A specifies the training objective, and Appendix B documents the controlled data construction and evaluation protocol.

Ethics Statement

The controlled experiment uses simulated feature vectors and contains no personal data. For future deployment with natural images, EG-MLLM remains subject to the same risks as other MLLMs, including hallucination, inherited dataset bias, and unsafe outputs. Evidence guidance is therefore a reliability mechanism rather than a guarantee of factual correctness, and aggregate accuracy should be reported together with evidence quality, calibration, and failure cases.

AI Use Statement

AI-assisted tools were used for literature organization, language editing, implementation assistance, and figure layout. The numerical results reported in the manuscript come from the controlled experiment implementation included with the project. The authors are responsible for verifying the final text, citations, implementation, and measurements before submission.

References

Peter Anderson, Xiaodong He, Chris Buehler, et al. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, et al. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision*, 2015.

Jinze Bai, Shuai Bai, Shusheng Yang, et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. *arXiv preprint*, 2025.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Wenliang Dai, Junnan Li, Dongxu Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Qidong Huang, Xiaoyi Dong, Pan Zhang, et al. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Sicong Leng, Hang Zhang, Guanzheng Chen, et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *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*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023b.

Yifan Li, Kun Zhou, Xin Zhao, Lei Fang, and Jirong Wen. Analyzing and mitigating object hallucination: A training bias perspective. In *AAAI Conference on Artificial Intelligence*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, et al. Llava-next: Improved reasoning, ocr, and world knowledge. *Technical report*, 2024b.

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

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*, 2016.

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

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 2017.

Amanpreet Singh, Vivek Natarajan, Meet Shah, et al. Towards vqa models that can read. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Mingrui Wu, Jiayi Ji, Oucheng Huang, et al. Evaluating and analyzing relationship hallucinations in large vision-language models. In *International Conference on Machine Learning*, 2024.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, 2016.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, 2023.

A Training Algorithm and Derivation

A.1 Intervention-margin derivation

Let $s_Y(Z) = T^{-1} \sum_t \log p_\theta(y_t \mid y_{<t}, Z, Q)$. A selector is useful only if its chosen set produces a larger change in the target score than an equal-budget nuisance perturbation. Define $D(S) = s_Y(Z) - s_Y(Z_{\setminus S})$. The desired ranking is $D(E) \geq D(N) + m$. A convex upper bound on violations is the hinge loss

$$[\, m - D(E) + D(N) \,]_+, \tag{12}$$

which is Eq. (7). When $D(E)$ is already larger than $D(N) + m$, the intervention term contributes no gradient; otherwise it pushes the selector/reasoner/adapter toward evidence whose removal changes the gold-response likelihood. Because both interventions use the same K and matched spatial strata, the loss is less sensitive to trivial perturbation-size effects.

Algorithm 1 EG-MLLM training step

Require: Batch $\{(I_b, Q_b, Y_b)\}_{b=1}^{B}$; backbone f_v, f_l; selector S_ϕ; graph G_ψ; adapter A_ω
1: $Z \leftarrow f_v(I)$; $q \leftarrow \mathrm{Pool}(f_l(Q))$
2: $a \leftarrow S_\phi(Z, q)$; $E \leftarrow \mathrm{STTopK}(Z, a, K)$
3: $H_E \leftarrow G_\psi(E)$
4: $\ell, H_Y \leftarrow f_l(Q, Y; A_\omega(H_E))$
5: Sample matched nuisance N from bottom-score tokens
6: Form $Z_{\setminus E}$ and $Z_{\setminus N}$ using mask/drop/swap interventions
7: Compute $\Delta_E = \ell(Z) - \ell(Z_{\setminus E})$, $\Delta_N = \ell(Z) - \ell(Z_{\setminus N})$
8: $L \leftarrow L_{LM} + \lambda_c L_{cons} + \lambda_i [m - \Delta_E + \Delta_N]_+ + \lambda_b L_{budget}$
9: Update ϕ, ψ, ω and LoRA parameters; keep frozen backbone weights fixed unless configured otherwise

A.2 Gradient behavior

The straight-through selector receives complementary signals. L_{cons} promotes evidence that aligns with target semantics, whereas L_{int} rejects visually plausible but functionally irrelevant tokens. If evidence removal barely changes the target likelihood, Δ_E is small and the hinge remains active. If nuisance removal changes the answer strongly, Δ_N grows and the hinge again activates, explicitly penalizing a selector that ignores a competing pathway. The zero-initialized adapter gate prevents this auxiliary objective from destabilizing the language model at the start of training.

B Controlled Data and Evaluation Details

Episode construction. Each instance contains 16 token vectors of dimension 32 and one query vector. Two answer-critical tokens share the query category, carry distinct roles A and B, and contain separately sampled continuous attributes $a, b \in [-1, 1]$ with $|a - b| > 0.12$. Labels are balanced and equal to $\mathbb{1}[a > b]$. The remaining tokens carry mismatched category markers and random nuisance features. A dedicated nuisance token encodes a shortcut correlated with the target according to the condition-specific probability.

Conditions. Training and in-domain evaluation use shortcut correlation 0.9. Neutral evaluation uses 0.5, shortcut shift uses 0.1, and heavy-noise evaluation uses a neutral shortcut plus stronger random corruption of nuisance tokens. Training contains 6,000 instances; each test condition contains 2,000 instances sampled under its specified shortcut and noise setting. The experiment uses seed 2027.

Metrics. Answer accuracy is the percentage of correct binary predictions. Evidence precision and recall compare the model's two selected tokens with the two known answer-critical tokens. For intervention sensitivity, we mask the selected pair and a size-matched bottom-score nuisance pair, compute the gold-label log-probability drops Δ_E and Δ_N, and report the percentage of instances satisfying $\Delta_E - \Delta_N > 0.15$.

Scaling to natural-image benchmarks. The repository retains adapters and evaluators for POPE (Li et al., 2023b), VQAv2 (Goyal et al., 2017), COCO captioning (Lin et al., 2014), and RefCOCO/RefCOCO+ (Yu et al., 2016). These modules use the same evidence path but are not needed to reproduce the controlled numerical results reported here.

Component	Setting
Backbone	7B-class MLLM (Qwen2/2.5-VL or LLaVA-family adapter)
Evidence tokens K	64 (ablate 32/64/96)
Graph width / layers	1024 / 2
Sparse neighbors r	8
Adapter bottleneck	512
LoRA rank / alpha	64 / 128
Optimizer	AdamW, weight decay 0.01
Learning rate	2×10^{-5} modules; 1×10^{-5} LoRA
Warm-up	3% steps; cosine schedule
$\lambda_c, \lambda_i, \lambda_b$	0.10, 0.20, 0.02
Intervention margin m	0.15 nats/token
Precision	bfloat16

C Reference Full-Scale Hyperparameters

D Code Map

- `src/eg_mllm/models`: evidence selector, sparse graph reasoner, counterfactual interventions, bottleneck adapter, and integrated wrapper.
- `src/eg_mllm/data`: optional POPE, VQAv2, COCO-caption, and RefCOCO dataset adapters for large-backbone evaluation.
- `src/eg_mllm/eval`: VQA, captioning, grounding, calibration, and hallucination metric interfaces.
- `src/eg_mllm/training`: objective functions, LoRA utilities, deterministic seeding, checkpointing, and training helpers.
- The repository includes the controlled evidence stress-test implementation used for all numerical results in the manuscript.
- `results/controlled/metrics.json`: machine-readable copy of the reported measurements.