MEA: A REWARD-DRIVEN MULTI-AGENT SYSTEM FOR FAITHFUL MODEL EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent years have seen the employment of a plethora of machine learning (ML) models in high-stakes domains, but they remain largely opaque to the practitioners who act on their predictions. While post-hoc explanation methods offer a lens into this model behavior, wielding them effectively demands expertise most domain experts lack: navigating high-dimensional outputs, selecting the best explanations, and synthesizing evidence across disparate tools. To this end, we present MEA, a multi-agent framework that removes the explanation knowledge barrier entirely: a *Proposer* agent selects and configures explanation tools based on the question and modality, while an *Actor* agent is optimized end-to-end against faithfulness, transforming the outputs into natural language explanations grounded in model behavior across tabular, text, and vision modalities. Further, we introduce diverse question types spanning feature attribution, counterfactual reasoning, and spurious feature detection, each paired with a perturbation-based faithfulness metric. We find that frontier LLMs systematically produce unfaithful explanations. By optimizing against faithfulness rewards augmented with a modality-adaptive penalty, MEA consistently outperforms post hoc explainers, agentic, and closed-source baselines across six datasets, with reward-driven optimization yielding faithfulness gains of $+28\%$ (tabular), $+21\%$ (text), and $+34\%$ (vision) over the untrained backbone. More broadly, our findings suggest that AI agents themselves can serve as a scalable, adaptable interface to ML explainability, opening a path toward natural-language explainability that generalizes beyond the fixed, single-purpose tools that have long defined the field.

1 INTRODUCTION

While machine learning (ML) models are increasingly deployed in high-stakes domains, such as healthcare, finance, and legal decision-making Rajpurkar et al. (2022); Cary Jr et al. (2023); Cao & You (2024); Flores et al. (2016), their opacity has become a critical concern for practitioners, regulators, and end users, hindering their trustworthiness. This gap between predictive performance and human interpretability is a central challenge that explainability research seeks to address.

Prior works have introduced a plethora of post-hoc explanation methods (Ribeiro et al., 2016; Sundararajan et al., 2017; Selvaraju et al., 2019; Lundberg & Lee, 2017; Pichery, 2014; Smilkov et al., 2017; Springenberg et al., 2015) that probe trained models to generate explanations. However, each method implements a fixed procedure yielding a single output type, *i.e.,* an attribution vector, a heatmap, or a set of super-pixels, independent of what the user actually wants to know. Further, deploying them correctly requires *selecting among competing methods, interpreting high-dimensional outputs, and synthesizing results across tools*, imposing a prohibitive knowledge barrier for domain practitioners without ML expertise.

Conversational explanation frameworks (CEF) address this barrier by allowing users to interrogate models in natural language Slack et al. (2023); Vanbrabant et al. (2025); Kim et al. (2025a); He & Martens (2026), mapping free-form queries to structured explanation operations without requiring users to select or configure individual methods. However, these systems share three limitations that remain unaddressed. First, CEF are largely *training-free*: they rely on hand-crafted parsing rules or prompt engineering to route user queries to explanation tools, *without any learned component that adapts to explanation quality or user intent*, resulting in unfaithful explanations. Second, CEF have

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.

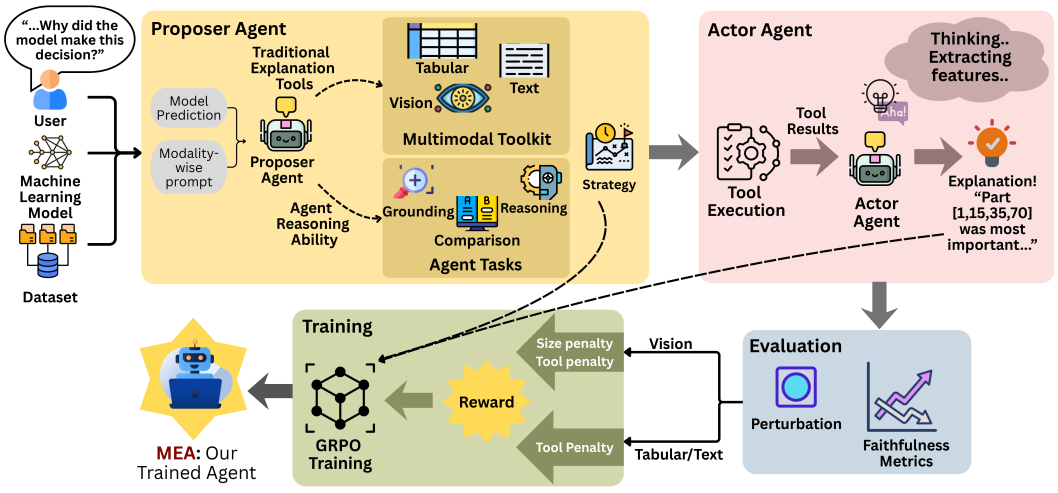


Figure 1: **Overview of MEA.** Given a user question about model behavior, the Proposer agent reasons over the model prediction and input modality to select an explanation strategy, choosing from a multimodal (tabular, text, and vision) toolkit and agent reasoning tasks (grounding, reasoning, and comparison), the Actor agent then executes the strategy, processes raw explanation outputs, and produces a structured natural-language explanation. The explanation is evaluated via input perturbation against faithfulness metrics. In training, both agents are jointly optimized end-to-end via GRPO Shao et al. (2024), using a reward combining faithfulness and tool penalties.

not been evaluated across large-scale datasets across different modalities. Kroeger et al. (2024) found that LLM-generated explanations fall behind post-hoc explainers and are susceptible to alphabetical bias, questioning whether LLMs can genuinely outperform existing explainers. Third, existing CEF lack a principled, quantitative evaluation framework: they either assess quality through single explanation types (Li et al., 2023; Rawal et al., 2025) or focus on simple linear models, where generated explanations are treated as final outputs with no verification step against model behavior.

Present work. We address the above gaps by introducing MEA, a novel multi-agent framework (see Fig. 1) for *prediction-level explainability across diverse modalities*. MEA comprises of a Proposer agent that selects and configures an appropriate combination of tools given a question and data modality, and an Actor agent that executes the tools and synthesizes a natural-language explanation. To scale across diverse explanation difficulty and data modalities, MEA optimizes a multi-modal agent backbone end-to-end via reinforcement learning, with faithfulness scores to directly incentivize explanations that are verifiably grounded in model behavior rather than merely plausible. Since explanation complexity and scale differ fundamentally across modalities, we further introduce a *modality-adaptive reward* that pairs a universal tool-count penalty with modality-specific constraints, preventing reward hacking while encouraging concise, targeted explanations uniformly across all modalities. Further, we introduce the first comprehensive, model- and modality-agnostic question **taxonomy**, comprising diverse question types that spans feature attribution, counterfactual reasoning, and spurious feature detection with their respective faithfulness metrics to quantify alignment between the agent's explanation and the model's actual behavior via perturbation-based testing. To enable large-scale training and evaluation, we construct MEA-BENCH, a dataset covering *three data modalities (tabular, text, and vision), six model architectures, and six datasets, comprising 13K training and 3.5K test instances*.

Our main contributions are as follows: 1) we present MEA, a reward-driven multi-agent framework for model-grounded explanations across tabular, text, and vision modalities, which optimizes explanation evidence against perturbation-based faithfulness rewards, and synthesizes this evidence into natural-language explanations; 2) we introduce MEA-BENCH, comprising diverse faithfulness-verified question types, **12** dataset-model combinations, and 13k training and 3.5k test instances across three modalities; 3) we demonstrate that frontier agents systematically hallucinate on explainability tasks and address this by optimizing MEA end-to-end via reinforcement learning,

using faithfulness as the reward augmented by a modality-adaptive penalty that prevents reward hacking across modalities; and 4) our trained MEA outperforms post-hoc explainers, closed-source frontier models, and advanced agentic frameworks, yielding average faithfulness gains of **+28%** (tabular), **+21%**(text), and **+34%** (vision) over the untrained backbone.

2 RELATED WORKS

Post-hoc Explainability. Post-hoc explainers attribute predictions to input features without modifying the model. Representative approaches include LIME (Ribeiro et al., 2016), which fits a locally linear surrogate around the instance, and SHAP (Lundberg & Lee, 2017), which provides Shapley-value-based attributions with consistency guarantees. For vision, Grad-CAM (Selvaraju et al., 2019) and gradient techniques (Sundararajan et al., 2017; Smilkov et al., 2017) supply pixel-level attribution signals.

Research Gap. Existing faithfulness evaluation (Li et al., 2023; Zhang et al., 2025) is conducted *offline* as a post-hoc assessment; no prior work adapts it into an *online training signal* to optimize an agentic framework across a comprehensive, multi-question, multi-modal setup.

Conversational and Agentic XAI. Early interactive works framed explanations as dialogues: TalkToModel (Slack et al., 2023) enables natural-language exploration of model behavior, and ECHO (Vanbrabant et al., 2025) grounds answers in computed attributions rather than parametric knowledge. FAX (Kim et al., 2025b) integrates agentic tool use with faithfulness evaluation, though it targets open-ended questions and relies on an LLM-as-a-judge pipeline. A parallel line applies multimodal agents to *mechanistic* interpretability of internal model representations, including MAIA (Shaham et al., 2025), OPENMAIA (Camuñas et al., 2025), and NEURONSCOPE (Liu et al., 2026); however, these are **not** designed for the prediction-level questions domain practitioners ask.

Research Gap. No existing work jointly optimizes a multi-agent pipeline end-to-end with RL across diverse modalities and question types. Our framework addresses this via a Proposer–Actor structure trained with GRPO, using faithfulness as the reward signal.

3 OUR FRAMEWORK: MEA

Problem Formulation. We study natural-language explanation generation for black-box classifiers. Given a trained classifier $f : \mathcal{X} \to \mathcal{Y}$, a question q from a predefined taxonomy $\mathcal{Q}$, and input instance(s) $x \in \mathcal{X}$, MEA generates a structured explanation e that answers q while remaining faithful to f's decision process. MEA supports three modalities (vision, text, and tabular data), where x is an image, token sequence, or feature vector respectively, and localizes evidence accordingly: a bounding box, token spans, or feature keys, identifying the subset of x that accounts for f's behavior with respect to q.

XAI Question Taxonomy. Drawing on prior work (Vanbrabant et al., 2025; Li et al., 2023), we define ten question types in three semantic categories (Appendix Table 3): 1) *Feature Attribution* asks which input parts drove or suppressed a prediction; 2) *Counterfactual* probes prediction sensitivity to hypothetical modifications; and 3) *Spurious Feature* examines whether the model relies on irrelevant cues. We exclude Q8 and Q9 from training and use them as *out-of-distribution* (OOD) evaluation questions, since these are only defined over *misclassified* instances ($\hat{y} \neq y^*$), which limits eligible training samples and destabilizes gradient estimation; we treat them as a held-out probe of generalization to unseen question types without task-specific supervision.

3.1 MEA OVERVIEW

MEA operates as a two-stage framework on a shared multimodal LLM backbone θ. First, the *Proposer Agent* π_θ^{Prop} receives question q, input $\mathbf{x}$, and prediction $\hat{y} = f(\mathbf{x})$, and produces a structured strategy S specifying which XAI tools and autonomous reasoning tasks to invoke (Appendix A.2). Then, the *Actor Agent* π_θ^{Act} executes S, processes the raw tool outputs, and

synthesizes an explanation e:

$$S \sim \pi_\theta^{\text{Prop}}(S \mid q, \mathbf{x}, \hat{y}, \mathcal{T}_m), \tag{1}$$

$$e \sim \pi_\theta^{\text{Act}}(e \mid q, \mathbf{x}, \hat{y}, S, \{o_t\}_{t \in \mathcal{T}_{\text{sel}}}), \tag{2}$$

where $\mathcal{T}_m$ is the tool set for modality m, $\mathcal{T}_{\text{sel}} \subseteq \mathcal{T}_m$ is the Proposer's selected subset, and o_t is tool t's summarized output.

3.2 Proposer Agent: Intent-Aware Strategy Formation

Prior conversational systems rely on fixed pipelines or hand-crafted routing rules with no learned component adapting to question semantics. To address this limitation, we introduce the trainable Proposer Agent π_θ^{Prop}, which performs strategy formation before any tool is executed. Given question type Q_i, modality m, and prediction $\hat{y} = f(\mathbf{x})$, it generates $S = (s_{\text{type}}, \mathcal{T}_{\text{sel}}, \mathcal{A}_{\text{auto}})$:

$$S \sim \pi_\theta^{\text{Prop}}(S \mid q, \mathbf{x}, \hat{y}, \mathcal{T}_m), \tag{3}$$

where $s_{\text{type}} \in \{\texttt{autonomous}, \texttt{tools}, \texttt{hybrid}\}$ selects the reasoning mode, $\mathcal{T}_{\text{sel}} \subseteq \mathcal{T}_m$ specifies which tools to invoke, and $\mathcal{A}_{\text{auto}}$ enumerates grounding/reasoning subtasks for the Actor. An ablation on the Proposer's strategy modes (Appendix A.9) confirms both tool use and autonomous reasoning are independently load-bearing, and that MEA learns to combine them more effectively than the base agent.

3.3 Actor Agent: Tool-Grounded Explanation Synthesis

Even when a well-formed strategy is available, raw tool outputs are not directly consumable as explanations. Kroeger et al. (2024) show that LLMs systematically bypass tool evidence in favor of parametric knowledge, producing plausible but unfaithful conclusions. To address this, the Actor Agent π_θ^{Act} operates in two steps enforcing a structural dependency on attribution evidence.

Step 1: Tool Execution and Summarization. For each tool $t \in \mathcal{T}_{\text{sel}}$, the Actor invokes t against f and $\mathbf{x}$, obtaining heatmaps, per-token, or feature importance scores, which are converted into structured text summaries (execution status, natural-language description, key statistics) rather than passed as raw tensors. To standardize granularity, the Actor returns the top 25% of features/tokens for tabular/text inputs, and aggregates top-1% attribution coordinates into a bounding box for vision.

Step 2: Evidence-Grounded Explanation Generation. The per-tool summaries are assembled into a structured prompt over which the Actor reasons to identify the most relevant features:

$$e \sim \pi_\theta^{\text{Act}}(e \mid q, \mathbf{x}, \hat{y}, S, \{o_t\}_{t \in \mathcal{T}_{\text{sel}}}), \quad e \in \mathcal{E}_m, \tag{4}$$

where $\mathcal{E}_m$ is the modality-appropriate output space: a bounding box, token span, or feature list.

3.4 Perturbation-based Faithfulness Evaluation

Evaluating explanation correctness without ground-truth labels is difficult. Prior work relies on LLM-as-a-judge pipelines susceptible to the same hallucinations they assess (Kroeger et al., 2024; Rawal et al., 2025), or evaluates a single explanation type in isolation (Li et al., 2023; Kim et al., 2025b); none adapts faithfulness evaluation into an online training signal across a multi-question, multi-modal setup. We address this with a unified perturbation-based protocol: explanation e identifying region r is faithful iff masking r produces the model behavior change q predicts. For each Q_i, metric m_i is computed by masking r in $\mathbf{x}$ and re-querying f, where $\mathbf{x}^{-r}$ is the masked input (zero-filled for tabular, grey-filled for vision, span-deleted for text) and $P(y \mid \mathbf{x})$ the predicted class probability.

Feature Attribution (Q1-Q4). Q1-Q2 form a complementary pair: masking the most responsible region should drop confidence; masking the least responsible should not:

$$\begin{aligned} m_1 &= \max\big(0,\ P(\hat{y} \mid \mathbf{x}) - P(\hat{y} \mid \mathbf{x}^{-r})\big), \\ m_2 &= 1 - \big|P(\hat{y} \mid \mathbf{x}) - P(\hat{y} \mid \mathbf{x}^{-r})\big|. \end{aligned} \tag{5}$$

Q3 checks whether masking the distinctive region collapses the top-1/top-2 margin; Q4 measures cross-instance contrastiveness via $\mathrm{sim}(F_A, F_B)$, requiring different attributions for differently-predicted instances.

Counterfactual (Q5–Q7). Q5 verifies the predicted direction of change agrees with the observed probability shift; Q6-Q7 measure whether the proposed modification raises the target-class probability and whether the predicted new class becomes likely:

$$m_6 = \max\big(0,\ P(\hat{y} \mid \mathbf{x}^{+\Delta}) - P(\hat{y} \mid \mathbf{x})\big)\,,$$
$$m_7 = P(\hat{y}' \mid \mathbf{x}^{-r}), \tag{6}$$

where $\mathbf{x}^{+\Delta}$ is the agent-proposed modification and $\hat{y}'$ the predicted class.

Spurious Features. Q8 measures the normalized gain in correct-class probability after masking:

$$m_8 = \max(0,\ \min(1,\ D(\mathbf{x}, r)))\,,$$
$$D(\mathbf{x}, r) = \frac{P(y^* \mid \mathbf{x}^{-r}) - P(y^* \mid \mathbf{x})}{1 - P(y^* \mid \mathbf{x})}, \tag{7}$$

normalized by remaining headroom for fair cross-instance comparison. Q9 averages m_8 over N misclassified instances; Q10 mirrors Q4's contrastive intuition on misclassified instances. Full definitions are in Appendix A.3.

3.5 End-to-End Training

Relying exclusively on raw faithfulness scores frequently precipitates reward hacking, typically via two degenerate strategies: trivial probability-drop maximization (*e.g.,* masking an entire image) and indiscriminate tool invocation, which inflates faithfulness without isolating informative features (Shao et al., 2024). Since naive objectives converge on these unfaithful traces, we employ GRPO (Guo et al., 2025) to jointly optimize both agents end-to-end. Each episode captures two transitions: strategy generation by the Proposer and explanation synthesis by the Actor.

We formalize MEA's training as a two-step Markov Decision Process (MDP) over a single episode, with the Proposer and Actor as two sequential decision points sharing backbone θ.

State (Proposer step). $s^{\mathrm{Prop}} = (q,\ \mathbf{x},\ \hat{y},\ \mathcal{T}_m)$: the question, input instance, frozen-classifier prediction, and the modality-compatible tool set (Sec. 3.3, Eq. 3).

Action (Proposer). The structured strategy $S = (s_{\mathrm{type}},\ \mathcal{T}_{\mathrm{sel}},\ \mathcal{A}_{\mathrm{auto}})$ (Sec. 3.3). $\mathcal{T}_{\mathrm{sel}}$ and $\mathcal{A}_{\mathrm{auto}}$ are combinatorial subsets rather than a single discrete choice and is regarded as a structured action space.

Transition. Deterministic: each selected tool $t \in \mathcal{T}_{\mathrm{sel}}$ is executed against the frozen classifier f and input $\mathbf{x}$, producing summarized outputs $\{o_t\}_{t \in \mathcal{T}_{\mathrm{sel}}}$ (Sec. 3.3, Step 1). No stochasticity is introduced by the environment; the only randomness in the episode comes from the two agents' own sampling.

State (Actor step). $s^{\mathrm{Act}} = (q,\ \mathbf{x},\ \hat{y},\ S,\ \{o_t\}_{t \in \mathcal{T}_{\mathrm{sel}}})$: the Proposer's emitted strategy together with the summarized tool outputs produced by the transition above.

Action (Actor). The explanation $e \in \mathcal{E}_m$ (Eq. 4).

Reward. $R = m_i(e)$, the faithfulness metric for question type Q_i (Sec. 3.4), augmented by the modality-adaptive size/tool penalties (Eq. 9 and the tool penalty in Sec. 3.3), evaluated strictly after the Actor's final output. No intermediate reward is given at the Proposer's transition.

Shared parameterization. Both $\pi_\theta^{\mathrm{Prop}}$ and $\pi_\theta^{\mathrm{Act}}$ route through the *same* LoRA-adapted backbone θ; the two roles are differentiated entirely by distinct system prompts (Appendix A.14, Figs. 12, 11) rather than separate adapters or a mode flag. Because R is terminal and both transitions belong to the same episode, the group-normalized advantage

$$\hat{A}_k = \frac{R_k - \mathrm{mean}\left(\{R_j\}_{j=1}^{G}\right)}{\mathrm{std}\left(\{R_j\}_{j=1}^{G}\right)}, \tag{8}$$

is applied identically to every token in both the Proposer's and the Actor's transitions, implementing the joint credit assignment described in Sec. 3.3.

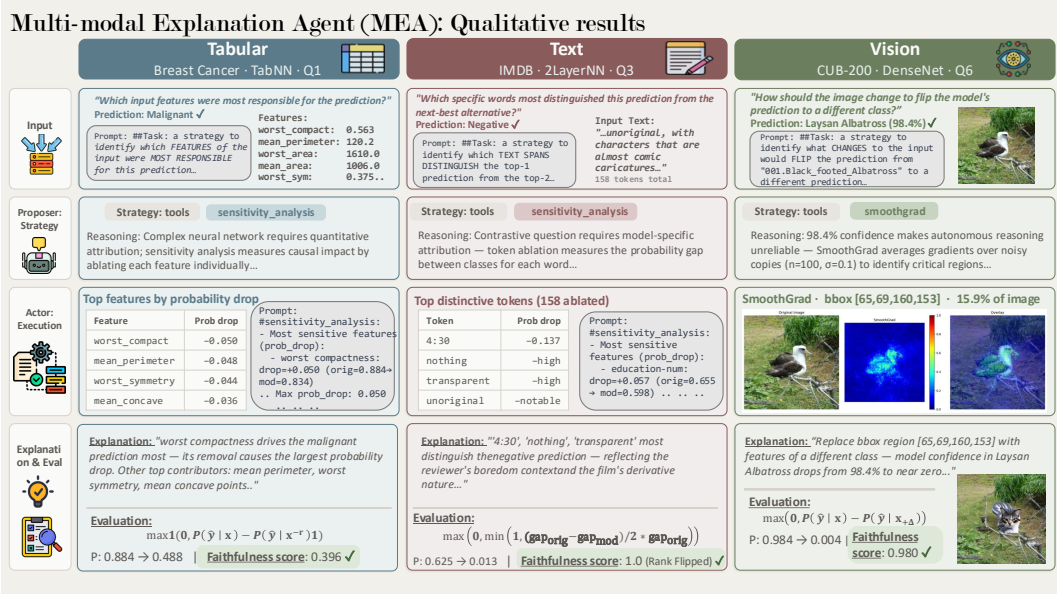


Figure 2: **Case study of MEA across three modalities.** Each column shows a complete pipeline execution for MEA: the input, Proposer's strategy and tool, Actor's execution output, and final explanation with faithfulness evaluation.

Size Penalty (Vision). To foreclose the large-mask shortcut, we penalize large bounding boxes:

$$m_i^{\text{vision}} = \max(0,\ m_i - \lambda_{\text{size}} \cdot \rho)\,, \tag{9}$$

where $\rho \in [0, 1]$ is the bounding-box-to-image area ratio and $\lambda_{\text{size}}=0.3$. For tabular/text, the fixed top-25% output constraint (Sec. 3.3) already governs scope, making an explicit size penalty unnecessary.

Tool Penalty. To discourage indiscriminate tool invocation, we apply an ℓ_1 tool-count penalty across all modalities: $r = \max\left(0,\ \min\left(1,\ m_i - \lambda_{\text{tool}} \cdot \frac{n_{\text{tools}}}{n_{\max}}\right)\right)$, where n_{tools} is the number of invoked tools/tasks, $n_{\max}=9$, and $\lambda_{\text{tool}}=0.05$. The coefficient is kept small to avoid suppressing gradients on question types with low pre-training scores (*e.g.,* tabular Q1/Q8: 0.01/0.04 at initialization). These penalties let the Proposer learn informative tool selection while the Actor learns faithful grounding, each reinforcing the other via shared gradients.

4 EXPERIMENTS

We address the following key research questions: RQ1) Does MEA generates faithful explanation across all three modalities? RQ2) Does MEA outperform widely used post hoc explainers? RQ3) Is the prompting style and agentic architecture necessary for MEA's performance? and RQ4) Does MEA generalize to OOD explanation questions, unseen datasets, and novel model architectures?

4.1 DATASETS AND EXPERIMENTAL SETUP

We first describe datasets designed to study the faithfulness of explanations generated by MEA and then outline experimental setup. Figure 2 illustrates MEA's end-to-end pipeline on representative samples from each modality, showing how the Proposer selects appropriate tools and the Actor synthesizes faithful explanations across diverse question types.

Dataset. We construct a multi-modal benchmark for analyzing agent behavior across tabular, text, and vision modalities (Appendix A.4), generated using predictions from ML models trained on standard supervised learning tasks. For each dataset-model pair, we generate structured questions (Table 3) covering feature attribution, counterfactual analysis, and spurious feature identification. *Tabular*: Adult Census Income (Becker & Kohavi, 1996) and Breast Cancer (Zwitter & Soklic,

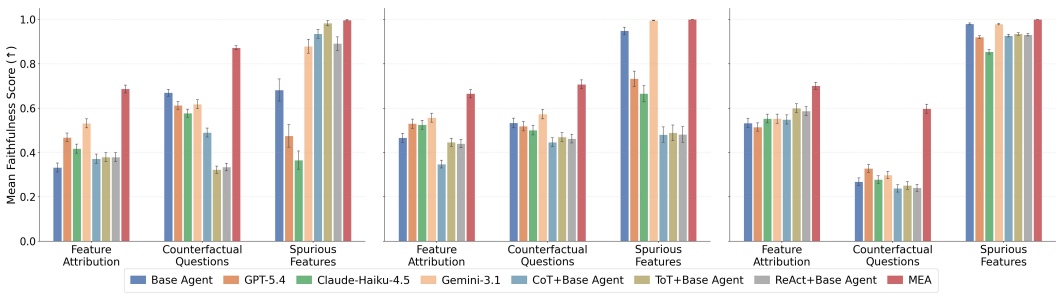


Figure 3: **MEA's performance.** Faithfulness scores of MEA vs. baselines across tabular (left), text (middle), and vision (right) modalities. Results show that MEA outperform all baselines, including closed-source agents and techniques like CoT, ToT, and ReACT.

1988), evaluated with a two-layer network and TabNet; *Text*: SNLI (Bowman et al., 2015) and IMDb (Maas et al., 2011), evaluated with a lightweight neural network and a CNN; and *Vision*: CUB-200-2011 (Welinder et al., 2010) and STL-10 (Coates et al., 2011), evaluated with ResNet-50 (He et al., 2015) and DenseNet-201 (Huang et al., 2018). Each instance contains the original input, ground-truth label, model prediction, and a natural-language analytical prompt.

Training Setup. We fine-tune Qwen3.6-35B-A3B (Qwen Team, 2026b) using LoRA (r=32) (Hu et al., 2021) for 3 epochs, following Shao et al. (2024) for GRPO hyperparameters (learning rate 1×10^{-5}, batch size 32, KL coefficient 0). We generate K=4 rollouts per question at T=1.0 (K=8 for harder types), with advantages group-normalized over each question's K rollouts; evaluation uses T=0.0. For training across modalities, we compare *Sequential* (tabular→text→vision, each stage initialized from the previous checkpoint) against *Hybrid* (all modalities interleaved per batch). Sequential training suffers from catastrophic forgetting. Performance on earlier modalities degrades and behavioral diversity collapses on the final modality (Appendix A.5), so we adopt Hybrid training for all main results. We also ablate on a 4B-parameter backbone (Appendix A.10), confirming faithfulness gains transfer to smaller models; computational cost details are in Appendix A.13.

Baselines. To isolate the impact of collaborative, iterative refinement, we benchmark MEA against four prompting paradigms sharing identical model infrastructure, execution environments, and XAI toolsets: *Zero-Shot* (control), *Chain-of-Thought (CoT)* (Wei et al., 2023), *ReAct* (Yao et al., 2023b), and *Tree of Thoughts (ToT)* (Yao et al., 2023a). We further compare against standard post-hoc explainers (Ribeiro et al., 2016; Lundberg & Lee, 2017; Smilkov et al., 2017; Selvaraju et al., 2019; Springenberg et al., 2015; Pichery, 2014; Sundararajan et al., 2017), and against specialized counterfactual generators, DiCE (Mothilal et al., 2020) and Growing Spheres (Laugel et al., 2017).

Evaluation. Following Agarwal et al. (2022), we evaluate explanation quality through *faithfulness*: whether the agent's claimed explanation is verified by the target model's behavior under perturbation. All metrics (Sec. 3.4) follow a common *perturb-and-measure* protocol: given an explanation identifying a region r (feature keys for tabular, text spans for text, a bounding box for vision), we perturb r to obtain x', re-run the prediction, and score it against the corresponding metric. Perturbations are modality-specific: vision regions are gray-filled (or inpainted with Stable Diffusion (Rombach et al., 2022) for counterfactuals), text spans are deleted or replaced with agent-provided counterfactual text, and tabular features are mean-filled or set to agent-specified values.

4.2 RESULTS

Here, we discuss experimental results that answer key questions (RQ1-RQ4) highlighted before.

RQ1) MEA improves explanation faithfulness across all modalities. A key challenge in XAI is that different explanation needs (feature attribution, counterfactual reasoning, and spurious feature detection) traditionally require separate tools. MEA addresses this by serving as a unified framework handling diverse question types across all three modalities without modality-specific engineering. Fig. 3 shows MEA achieves the highest faithfulness scores across all three modalities. On tabular data, MEA scores 0.855 on counterfactual questions, outperforming the best baseline (Base Agent, 0.670) by **28%**, and 0.997 on spurious features vs. Gemini-3.1's 0.879. On text, MEA scores 0.708

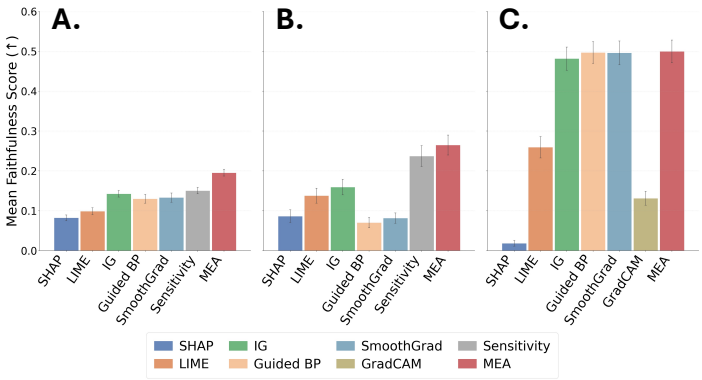


Figure 4: **MEA vs. Explainers.** Mean faithfulness scores of identifying most salient input features across **(A)** tabular, **(B)** text, and **(C)** vision modalities. MEA outperforms all baselines on tabular and text inputs, and matches the best-performing method on vision inputs.

on counterfactual questions vs. Gemini-3.1's 0.574 (**+23%**). The gap is largest on vision counterfactual questions, where MEA achieves 0.598 vs. GPT-5.4's 0.328 (**+82%**). Across all three modalities, MEA attains a perfect or near-perfect spurious feature detection score (0.997/1.000/1.000). To verify these similarities reflect genuine statistical equivalence rather than noise, we ran two-sided Welch's t-tests (WELCH, 1947) on all targeted comparisons: on spurious-feature questions (Q10), MEA achieves ceiling performance (1.000) on text and vision, and the strongest baselines (Gemini-3.1: 0.996 text, Base Agent: 0.981 vision) differ by at most 0.02–significant only due to MEA's zero variance at ceiling (Appendix A.12). Appendix A.11 further shows this faithfulness tracks the target model's decision logic, not the LLM's own priors.

These results confirm that MEA acts as a `one-stop-shop` delivering strong, reliable faithfulness across all ten question types and all three modalities. MEA also outperforms all closed-source models and alternative agentic frameworks (CoT, ToT, ReAct) on all question types across all modalities, confirming its role as a unified solution for diverse explainability needs.

RQ2) MEA outperforms post-hoc explainers. Standard post-hoc explainers (LIME, SHAP, GradCAM, etc.) are architecturally single-purpose, so a like-for-like comparison is well-defined only for Q1: *Which parts of the instance are most responsible for the model's prediction?* This restriction itself illustrates one of MEA's motivations: no single tool can address the diverse explanation needs spanned by MEA's ten-question taxonomy. We therefore compare MEA against two *disjoint* families of specialized baselines, each matched to the question type it targets: general-purpose attribution explainers on Q1, and specialized counterfactual generators on Q6. *Feature Attribution* Following the same output standardization used in MEA (top-25% of features for tabular/text, top-1% pixels as a bounding box for vision), Fig. 4 shows MEA outperforms the best post-hoc explainer on tabular (0.195 vs. 0.150; **+30%**) and text (0.265 vs. 0.237; **+12%**), and matches the best single-tool baseline on vision (0.499 vs. 0.497; $p = 0.944$, Appendix A.12). *Flip Prediction* We further compare MEA against two specialized counterfactual generators, DiCE and Growing Spheres. Neither has a meaningful notion of searching over text spans or image regions, so this comparison is restricted to *tabular* data, using the same test instances, preprocessor, and metrics as MEA's evaluation (Table 1). MEA achieves substantially higher faithfulness throughout (avg. target-class probability 0.886–0.894 vs. 0.499–0.631), and comparable flip-success rates. Where MEA's flip-success rate is slightly lower, we show the gap is not statistically significant (Fisher's exact, all $p \geq 0.23$, Appendix A.12)

Together, these results show that MEA, despite operating as a single **general-purpose agent** across all ten question types, is competitive with or superior to specialized single-purpose explainers on both feature attribution and counterfactual generation.

RQ3) MEA is robust to different prompt styles and outperforms critic-augmented agents. Here, we first test sensitivity to question phrasing by evaluating three paraphrased variants of the question taxonomy (see A.7) and compare mean faithfulness against the original. Next, we investigate whether augmenting a base agent with a Critic (Shinn et al., 2023), which receives the original strategy and explanation, critiques them, and regenerates improved outputs, can close the gap with MEA. Results in Appendix A.7 Fig. 7 show that paraphrase variants score within $< 1\%$ of the

Table 1: MEA vs. counterfactual explainers on Q6 (flip prediction), aggregated across model architectures per dataset. Flip = flip success; Faith. = avg. faithfulness. Best per column in **bold**.

	Adult Census		Breast Cancer	
Method	Flip	Faith.	Flip	Faith.
MEA	**100.0%**	**0.894**	95.5%	**0.886**
DiCE	**100.0%**	0.631	**97.7%**	0.580
Growing Spheres	98.4%	0.499	81.8%	0.412

Table 2: OOD faithfulness on unseen datasets/architectures, averaged across question types per group. Base: Proposer–Actor without GRPO; MEA: full GRPO-tuned system.

	Tabular		Text		Vision	
Group	Base	MEA	Base	MEA	Base	MEA
Feat. Attr.	0.447	**0.783**	0.428	**0.563**	**0.621**	0.583
Counterfactual	0.422	**0.499**	**0.860**	0.770	0.246	**0.521**
Spurious Feat.	0.264	**0.349**	0.639	**0.778**	0.416	**0.624**
Overall	0.379	**0.581**	0.625	**0.689**	0.448	**0.576**

original across all modalities and question groups (*e.g.,* tabular counterfactual: paraphrase 0.934 vs. MEA 0.939; vision feature attribution: 0.686 vs. 0.692), confirming that MEA is robust to prompt rephrasing. MEA also consistently outperforms a base agent augmented with an inference-time Critic (Agent+Reflection) across all three modalities, with the largest gains on counterfactual questions: **+36%** on text (0.876 vs. 0.644) and **+30%** on vision (0.569 vs. 0.439), and Agent+Reflection even *degrading* performance relative to the untrained base agent on tabular spurious features (Appendix A.8, Fig. 8). Appendix A.12 shows paraphrase robustness tests results, which demonstrate no significant performance degradation across 8 of 9 question-type×modality conditions ($p > 0.05$ in all but vision counterfactual). These results demonstrate that RL-based end-to-end training provides substantially stronger and more reliable gains than inference-time critique.

RQ4) MEA generalizes to OOD datasets and model architectures. We evaluate MEA on unseen datasets and model architectures across all modalities (Yelp Review Polarity dataset Zhang et al. (2016) with a BERT-based classifier Devlin et al. (2019) for text; German Credit dataset Dua & Graff (2019) with a three-layer neural network for tabular; CIFAR-10 Krizhevsky et al. (2009) with ResNet-18 He et al. (2015) for vision). Results in Table 2 show that MEA achieves the highest overall faithfulness on all three modalities (0.581 vs. 0.379 on tabular, 0.689 vs. 0.625 on text, 0.576 vs. 0.448 on vision). The largest gains appear on tabular inputs ($+$**53%**), with feature attribution alone rising from 0.447 to 0.783 (**+75%**). On vision, counterfactual faithfulness improves from 0.246 to 0.521 (**+112%**). MEA only underperforms the base agent on text counterfactual questions (0.770 vs. 0.860) and vision feature attribution (0.583 vs. 0.621), which we attribute to further hyperparameter tuning needed for better modality-specific trade-offs. MEA also generalizes to concept-level explanation, a question type outside the trained taxonomy evaluated via representation-space CAV ablation, outperforming the base agent by $\sim$18% in mean faithfulness. Overall, these results confirm that MEA **generalizes** to unseen datasets and model architectures, and to explanation mechanisms beyond input perturbation (see Appendix Table 6 for detailed results, Table 7 for concept-level results, and Appendix A.6 for an ablation study of OOD question types).

5 CONCLUSION

We presented MEA, a multi-agent framework for faithful, prediction-level explanation across tabular, text, and vision modalities, and MEA-BENCH, a model- and modality-agnostic benchmark of 16k instances each paired with a faithfulness metric. Optimizing MEA end-to-end via RL with a modality-adaptive reward yields an agent that consistently outperforms post-hoc explainers, frontier models, and advanced agentic frameworks by $+28\%$ (tabular), $+21\%$ (text), and $+34\%$ (vision) faithfulness, showing **learning to explain** is more principled than prompting to explain. MEA is currently scoped to classification tasks and a fixed tool library, with a perturbation-based protocol sensitive to distributional shift like others in its class; extending to other tasks, tool libraries, and faithfulness signals is promising future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing and debugging code for implementing supplementary experiments, to draft sections of the paper, and to assist with editing the paper for readability, wording, and formatting. We have not used generative AI tools to generate synthetic datasets, formulate mathematical claims, design our core research methodology, or interpret experimental results.

We have reviewed all AI-assisted work: all AI-generated code was tested and verified for correctness by the authors, all AI-drafted text was reviewed and revised by the authors against the underlying experimental results before inclusion, and all AI-suggested text edits were similarly checked before inclusion. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

MEA is intended to make ML model behavior more accessible to domain practitioners in high-stakes settings such as healthcare, finance, and legal decision-making. However, faithfulness does not equal causal correctness: MEA will faithfully explain biased or discriminatory patterns without flagging them, so users should not mistake a faithful explanation of a biased model for a safe one, and MEA should support human judgment rather than replace it. Its counterfactual and spurious-feature capabilities could also be repurposed to craft adversarial inputs, so we encourage appropriate access controls before deployment in sensitive systems. Since MEA requires query access to the target classifier, any deployment on personal data (e.g., medical or financial records) must comply with applicable privacy regulations (HIPAA, GDPR). All datasets used are publicly available, de-identified, and widely used in prior ML research; no human subjects were involved.

REPRODUCIBILITY STATEMENT

We provide extensive details to support reproducibility of our results. The full architecture and training procedure of MEA are described in Sec. 3.3 and formalized in Appendix 3.5. All GRPO hyperparameters (learning rate, batch size, KL coefficient, rollout count K, and temperature), LoRA configuration (r=32), and the backbone model are reported under Training Setup and further detailed in Appendix A.13. Complete definitions of all ten faithfulness metrics (m_1–m_{10}) are given in Sec. 3.4 and Appendix A.3, and the full question taxonomy is provided in Appendix A.1 (Table 3). The datasets, model architectures, and dataset statistics used to construct MEA-BENCH are described under Dataset and Appendix A.4, and the sequential/hybrid training strategies compared in our ablations are detailed in Appendix A.5. All prompt templates used by the Proposer and Actor agents, as well as the reasoning-mode preambles for the CoT/ReAct/ToT baselines and the Critic agent prompt, are provided verbatim in Appendix A.14 and Appendix A.8. Statistical testing procedures (Welch's and paired t-tests) and significance thresholds are described in Appendix A.12.We will release code and MEA-BENCH upon publication.

REFERENCES

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. *Advances in neural information processing systems*, 35:15784–15799, 2022.

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. *arXiv preprint arXiv:2308.12966*, 2023.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference, 2015. URL https://arxiv.org/abs/1508.05326.

Josep Lopez Camuñas, Christy Li, Tamar Rott Shaham, Antonio Torralba, and Agata Lapedriza. OpenMAIA: a multimodal automated interpretability agent based on open-source models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=KitDRi76It.

Kai Cao and Haifeng You. Fundamental analysis via machine learning. *Financial Analysts Journal*, 80(2):74–98, 2024.

Michael P Cary Jr, Anna Zink, Sijia Wei, Andrew Olson, Mengying Yan, Rashaud Senior, Sophia Bessias, Kais Gadhoumi, Genevieve Jean-Pierre, Demy Wang, et al. Mitigating racial and ethnic bias and advancing health equity in clinical algorithms: a scoping review: scoping review examines racial and ethnic bias in clinical algorithms. *Health Affairs*, 42(10):1359–1368, 2023.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/coates11a.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Dheeru Dua and Casey Graff. Uci machine learning repository. http://archive.ics.uci.edu/ml, 2019.

Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38, 2016.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Yifan He and David Martens. An agentic approach to generating xai-narratives, 2026. URL https://arxiv.org/abs/2603.20003.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. URL https://arxiv.org/abs/1608.06993.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pp. 2668–2677. PMLR, 2018.

Haechang Kim, Hao Chen, Can Li, and Jong Min Lee. Talktoagent: A human-centric explanation of reinforcement learning agents with large language models, 2025a. URL https://arxiv.org/abs/2509.04809.

Jaechang Kim, Sunung Mun, Seungjoon Lee, Jaewoong Cho, and Jungseul Ok. Towards faithful agentic XAI, 2025b. URL https://openreview.net/forum?id=uGPbHJ0MSe.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. In-context explainers: Harnessing llms for explaining black box models, 2024. URL https://arxiv.org/abs/2310.05797.

Thinking Machines Lab. Tinker, 2025. URL https://thinkingmachines.ai/tinker/.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.

Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. $\mathcal{M}^4$: A unified XAI benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=6zcfrSz98y.

Weiqi Liu, Yongliang Miao, Haiyan Zhao, Yanguang Liu, and Mengnan Du. Neuronscope: A multi-agent framework for explaining polysemantic neurons in language models, 2026. URL https://arxiv.org/abs/2601.03671.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020.

C. Pichery. Sensitivity analysis. *Encyclopedia of Toxicology*, 12 2014. doi: 10.1016/B978-0-12-386454-3.00431-0.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022.

Kaivalya Rawal, Zihao Fu, Eoin Delaney, and Chris Russell. Evaluating model explanations without ground truth. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 3400–3411. ACM, 2025. doi: 10.1145/3715275.3732219. URL http://dx.doi.org/10.1145/3715275.3732219.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL https://arxiv.org/abs/1602.04938.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent, 2025. URL https://arxiv.org/abs/2404.14394.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Talktomodel: Explaining machine learning models with interactive natural language conversations, 2023. URL https://arxiv.org/abs/2207.04154.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. URL https://arxiv.org/abs/1706.03825.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015. URL https://arxiv.org/abs/1412.6806.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

Sebe Vanbrabant, Gilles Eerlings, Gustavo Alberto Rovelo Ruiz, and Davy Vanacken. Echo: Enhancing conversational explainable ai through tool-augmented language models. *Proc. ACM Hum.-Comput. Interact.*, 9(4), June 2025. doi: 10.1145/3734191. URL https://doi.org/10.1145/3734191.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

B. L. WELCH. The generalization of 'student's' problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 01 1947. ISSN 0006-3444. doi: 10.1093/biomet/34.1-2.28. URL https://doi.org/10.1093/biomet/34.1-2.28.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023a. URL https://arxiv.org/abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023b. URL https://arxiv.org/abs/2210.03629.

Shichang Zhang, Tessa Han, Usha Bhalla, and Himabindu Lakkaraju. Towards unified attribution in explainable ai, data-centric ai, and mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.18887.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification, 2016. URL https://arxiv.org/abs/1509.01626.

Matjaz Zwitter and Milan Soklic. Breast Cancer. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C51P4M.

Table 3: Overview of the ten explanation questions organised by category.

ID	Category	Question
Feature Attribution - which parts of the input drove or suppressed the prediction?		
Q1	Most Responsible	Which parts of the instance are most responsible for the model's prediction?
Q2	Least Responsible	Which parts of the instance are least responsible for the model's prediction?
Q3	Distinctive	Which specific parts of the instance distinguish its prediction from the next-best alternative?
Q4	Contrastive	Why are instances A and B given different predictions?
Counterfactual - how sensitive is the prediction to hypothetical modifications?		
Q5	Mask Prediction	If we mask a certain part of this instance, would the prediction change?
Q6	Flip Prediction	How should the instance change to flip the model into a different prediction?
Q7	Change Prediction	If we remove or change one important part of the instance, how would the prediction change?
Spurious Feature - does the model rely on irrelevant cues?		
Q8	Irrelevant Parts	Is there any irrelevant part in this instance that causes the model's wrong prediction?
Q9	Shared Feature	What shared feature makes misclassified instances A, B, C, ... difficult for the model?
Q10	Similar-Different	Why are instances A and B assigned different predictions despite apparent similarity?

A Appendix

The Appendix provides supplementary material supporting the main paper, including the complete question taxonomy (A.1), details of the XAI tools employed (A.2), faithfulness evaluation metrics (A.3), dataset statistics (A.4), descriptions of the sequential and hybrid training procedures (A.5), ablation studies on out-of-distribution (OOD) question types, unseen datasets/architectures, and concept-level explanation generalization (A.6), prompt robustness (A.7), the critic agent (A.8), the Proposer's strategy modes (A.9), backbone scale transfer (A.10), a sanity check on a classifier with a synthetic shortcut (A.11), statistical significance analyses (A.12), computational cost details (A.13) and prompt templates for Actor and Proposer Agents (A.14).

A.1 Question Taxonomy

The detailed question definition and category is showed in Table 3.

A.2 XAI Tools

To support explanation generation, our framework provides a library of established attribution methods spanning all three modalities:

Gradient-based. GradCAM Selvaraju et al. (2019) (layer-wise gradient attribution, vision only), Integrated Gradients Sundararajan et al. (2017), Guided Backpropagation Springenberg et al. (2015), SmoothGrad Smilkov et al. (2017) (noise-averaged gradient estimation), and Sensitivity Analysis Pichery (2014) (input-gradient saliency, text and tabular only).

Perturbation-based. LIME Ribeiro et al. (2016) (Local Interpretable Model-agnostic Explanations with superpixel or token-level segmentation for vision, and token-level perturbation for text) and SHAP Lundberg & Lee (2017) (Shapley Additive exPlanations via marginal contribution estimation).

A.3 Faithfulness Metrics

Here gap_{orig}, gap_{mod} are the original gap between top-1 and top-2 predicted classes and the gap after perturbation; y^* is the ground-truth class; $\hat{y}$ is the predicted class after perturbation; δ is the predic-

Table 4: Summary of evaluation metrics corresponding to different explanation faithfulness questions.

Question	Metric m_i	Intuition
Q1	$\max\big(0, P(\hat{y} \mid \mathbf{x}) - P(\hat{y} \mid \mathbf{x}^{-r})\big)$	Masking the most responsible region should drop predicted-class confidence
Q2	$1 - \big\lvert P(\hat{y} \mid \mathbf{x}) - P(\hat{y} \mid \mathbf{x}^{-r})\big\rvert$	Masking the least responsible region should leave confidence unchanged
Q3	$\max\left(0, \min\left(1, \frac{\text{gap}_{\text{orig}} - \text{gap}_{\text{mod}}}{2 * \text{gap}_{\text{orig}}}\right)\right)$	The top-1/top-2 margin should collapse after masking the distinctive region
Q4	$1 - \text{sim}(F_A, F_B)$	The agent should attribute different features to each instance
Q5	$\begin{cases} \max(0,\ \delta) & \hat{c} = 1 \text{ (predicts change)} \\ \max(0,\ 1 - \lvert\delta\rvert) & \hat{c} = 0 \text{ (predicts stable)} \end{cases}$	Probability change should agree in direction with the agent's yes/no answer
Q6	$\begin{cases} \max(0, P(\hat{y} \mid \mathbf{x}) - P(\hat{y} \mid \mathbf{x}_{+\Delta})) & \text{vision} \\ \max(0, \min(1, P(y_t \mid \mathbf{x}_{+\Delta}))) & \text{text/tab} \end{cases}$	Applying the proposed change should raise the probability of the target class
Q7	$P(\hat{y} \mid \mathbf{x}^{-r})$	The agent's predicted new class should have high probability after masking
Q8	$\max\left(0, \min\left(1, \frac{P(y^* \mid \mathbf{x}^{-r}) - P(y^* \mid \mathbf{x})}{1 - P(y^* \mid \mathbf{x})}\right)\right)$	Masking the spurious region should improve correct-class probability
Q9	$\frac{1}{N}\sum_{i=1}^{N} \max\left(0, \min\left(1, \frac{\Delta_i}{1 - P(y_i^* \mid \mathbf{x}_i)}\right)\right)$	Improvement should hold across all N misclassified instances
Q10	$1 - \text{sim}(F_{\text{corr}}, F_{\text{wrong}})$	Features attributed to the correct vs. wrong instance should differ

tion probability drop after perturbation; and $F_{\text{correct}}, F_{\text{wrong}}$ are attribution maps for the correctly and incorrectly classified instances respectively.

A higher metric score indicates a more faithful explanation. For binary metrics (Q3, Q5, Q6, Q7), the score is 1 if the condition holds and 0 otherwise. For Q6, tabular and text inputs are evaluated against a specific target class y_t, whereas vision inputs are evaluated by whether the prediction changes to any alternative class, reflecting the substantially larger label space in vision tasks.

The detailed metric calculation method and intuition is showed in Table 4.

A.4 Dataset Statistics

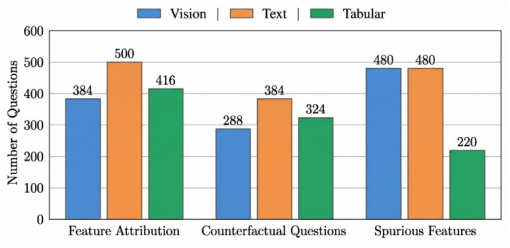


Figure 5: Dsitribution of data across modalities.

Due to the heterogeneous prediction behavior exhibited across models and datasets, the number of generated instances varies across categories (**Q1-Q10**). Different question types require specific prediction configurations, such as correct predictions, informative misclassifications, contrastive prediction pairs, or shared behavioral patterns across models. Consequently, the final instance count for each category is determined by the availability of meaningful and semantically valid examples satisfying these construction criteria. This naturally results in varying sample counts across question types, while ensuring that each generated instance reflects authentic model behavior and preserves the fidelity of the underlying prediction distributions.

A.5 Sequential and Hybrid Training

We compare two multi-modal GRPO training schedules, both using Qwen3-VL-30B-A3B-InstructBai et al. (2023). *Sequential* training fine-tunes one modality at a time, carrying each checkpoint forward to the next phase. *Hybrid* training interleaves all three modalities within every batch, grouped by question type via Q-type stratification.

Sequential Training

Phase 1 - Tabular. Fine-tuning on tabular data across 3 epochs drives the mean batch reward from 0.3 to 0.6. The resulting checkpoint achieves a tabular faithfulness score of 0.632.

Phase 2 - Text. Continuing from the tabular checkpoint drives the batch reward from 0.4 to 0.6 on text examples. However, re-evaluating the new checkpoint on the *tabular* test set reveals measurable catastrophic forgetting. The degradation happens across question types: Q3 drops from 0.326 to 0.264 (-0.06) and Q6 from 0.780 to 0.728 (-0.05). Both Q-types demand accurate numerical reasoning over feature rankings or feature value intervals – skills reinforced during tabular training and partially overwritten as the gradient signal shifts to text-span operations.

Phase 3 - Vision. Vision training for 200 steps shows a markedly weaker learning signal: the batch reward rises only from 0.512 to 0.578, compared to the steep trajectories in Phases 1 and 2 (Figure 6, left). We attribute the limited learning to two compounding effects. First, by Phase 3 the model's policy is already tightly peaked around tabular/text output formats, leaving insufficient entropy to explore the vision-specific bounding-box. Second, the flat reward curve indicates that rollout diversity has collapsed: all sampled trajectories produce similar rewards, so the group-normalized advantage $\hat{A}_k \approx 0$ for most rollouts, suppressing the gradient signal.

Hybrid Training

Hybrid training on all three modalities simultaneously runs for 902 steps across 3 epochs, with the batch reward rising from 0.313 to 0.594 (Figure 6, right). The slower convergence relative to any single-modality phase reflects the wider diversity of the joint training distribution: the model must maintain a flexible policy across tabular, text, and vision inputs simultaneously.

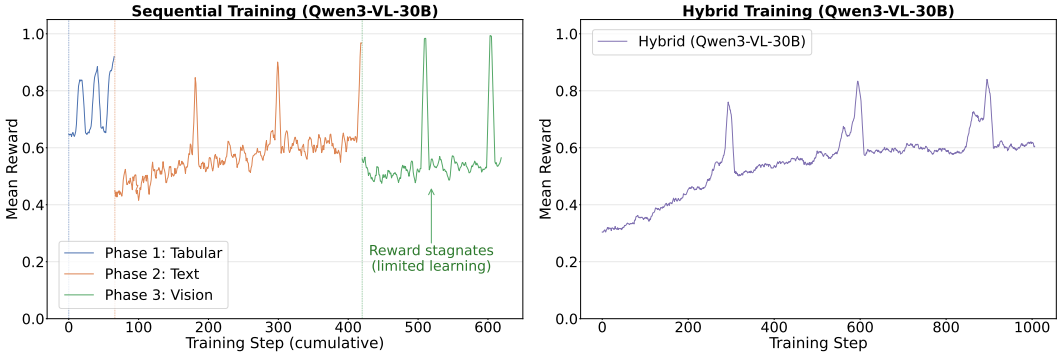


Figure 6: **Training reward curves. (Left)** *Sequential training:* Vertical dashed lines mark phase boundaries (Tab $\rightarrow$ Text $\rightarrow$ Vision); the vision reward barely rises, indicating collapsed rollout diversity. **(Right)** *Hybrid training (all modalities interleaved):* The recurring spikes are an artifact of the sampling seed, which concentrates relatively easy questions towards the end of each epoch.

A.6 Ablation Study: OOD Generalization

We evaluate MEA's out-of-distribution generalization along three axes: held-out question types within the proposed taxonomy, unseen datasets and model architectures, and a held-out question type entirely outside the taxonomy (concept-level explanation).

Held-out question types. We use spurious feature questions as OOD question types, as MEA is never trained on them (see Sec. 3). Results in Table 5 show that MEA achieves higher overall faithfulness on tabular and text modalities, with particularly notable gains on Q8 for text. On vision inputs, the base agent slightly outperforms MEA, suggesting that generalization to spurious feature reasoning remains challenging for vision modality without explicit training signal.

Table 5: OOD faithfulness scores on held-out question types Q8 and Q9. Best per row in **bold**.

Modality	Question Type	Base Agent	MEA
Tabular	Q8	0.132	**0.184**
	Q9	**0.024**	0.019
	Overall	0.081	**0.106**
Text	Q8	0.186	**0.353**
	Q9	0.066	**0.075**
	Overall	0.116	**0.186**
Vision	Q8	**0.068**	0.050
	Q9	**0.070**	0.023
	Overall	**0.069**	0.034

Unseen datasets and model architectures.. Table 6 reports faithfulness on entirely unseen dataset-model combinations for each modality. MEA achieves the highest overall faithfulness across all three modalities, though it underperforms the base agent on text counterfactual and vision feature attribution, which we attribute to further hyperparameter tuning needed for better modality-specific trade-offs.

Table 6: OOD faithfulness scores on different datasets and model architectures. Best performance is **bolded**.

Modality	Question Group	Base Agent	MEA
Tabular	Feature Attribution	0.447	**0.783**
	Counterfactual	0.422	**0.499**
	Spurious Features	0.264	**0.349**
	Overall	0.379	**0.581**
Text	Feature Attribution	0.428	**0.563**
	Counterfactual	**0.860**	0.770
	Spurious Features	0.639	**0.778**
	Overall	0.625	**0.689**
Vision	Feature Attribution	**0.621**	0.583
	Counterfactual	0.246	**0.521**
	Spurious Features	0.416	**0.624**
	Overall	0.448	**0.576**

Concept-level explanation generalization. MEA's faithfulness protocol (Sec. 3.4) is fundamentally tied to *local input-region perturbation*: masking a feature subset, token span, or bounding box. To test generalization beyond input-perturbation and beyond the trained question taxonomy entirely, we introduce a held-out question type, **Q11** ("Which concept was most responsible for the model's prediction?"), on CUB-200-2011, which asks the agent to name a semantic concept (*e.g.,* "bill shape: hooked seabird") rather than a region.

Faithfulness is scored by training a **Concept Activation Vector** (CAV) (Kim et al., 2018), which is a linear probe on the model's penultimate-layer representation for each concept, and orthogonally projecting that direction out of the activation ($h' = h - (h \cdot \hat{v})\hat{v}$) before re-running only the classifier head; the probability drop $\max(0, P_{\text{original}} - P_{\text{modified}})$ is the faithfulness score, and the entire

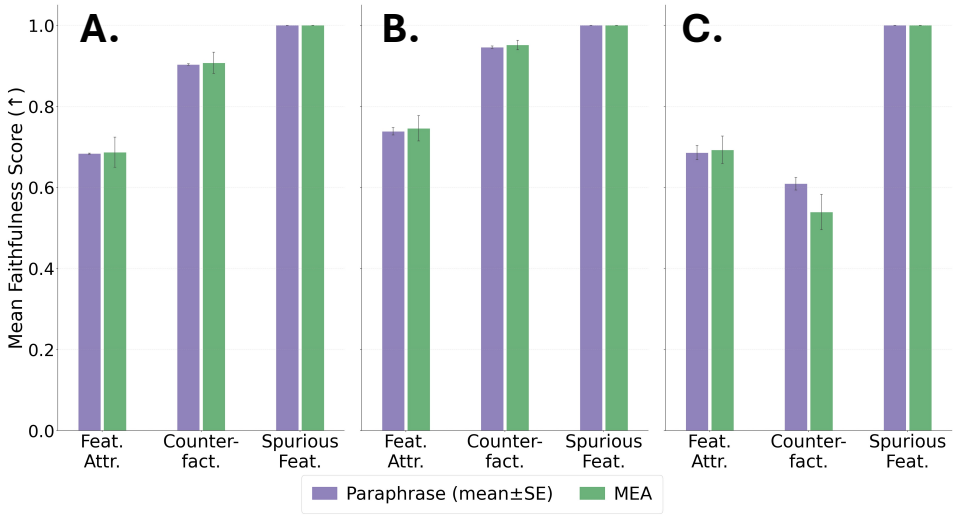


Figure 7: **Ablation on prompt robustness.** Mean faithfulness scores of MEA evaluated on three paraphrased variants of the question taxonomy across **(A)** tabular, **(B)** text, and **(C)** vision modalities. Minimum variance indicates that MEA is robust to changes in question phrasing.

measurement happens in representation space. The Proposer/Actor pipeline still uses MEA's existing tool library to decide which concept to name (Appendix A.2); only the evaluation mechanism differs from Q1–Q10.

Table 7shows that, across 120 questions (`cub_resnet` + `cub_densenet`), MEA scores ∼18% higher mean concept faithfulness than the base agent despite never being trained on this question type. Together, these results indicate MEA's thought-action-observation loop generalizes across ablation targets (input region *or* representation direction).

Table 7: Base agent vs. MEA on Q11 (concept attribution), 120 questions.

Metric	Base Agent	MEA
Mean faithfulness	0.0316	**0.0373**
Median faithfulness	0.0150	**0.0218**
Max faithfulness	0.2398	**0.2650**
Class-flip rate (ablation changes prediction)	5/120	4/120

A.7 Ablation Study: Prompt Robustness

To assess prompt robustness, we generate three semantically-equivalent paraphrases of every question in the taxonomy. Dynamic values (feature names, row IDs, class labels, masked tokens) are extracted from the original question via regular expressions and re-injected verbatim into each paraphrase template, preserving semantic equivalence. Q1-Q4 receive three lexically distinct variants; Q5-Q7 and Q10 share a single uniform paraphrase across all three runs (variant-independent templates). Table 8 lists the question-field phrasings for each type, illustrated with tabular-modality examples. Due to computational constraints, paraphrase evaluation is conducted on one dataset per modality.

A.8 Ablation Study: Critic Agent

The Critic Agent augments the base agent pipeline with an inference-time reflection step. After the Actor Agent produces a strategy and explanation and receives a faithfulness score, the Critic receives a structured prompt and returns two critiques: a *Proposer Reflection* (feedback on tool selection) and an *Actor Reflection* (feedback on explanation quality). The actor agent then regenerates

Table 8: Three paraphrase variants of each question type (*Q-field only*; tabular modality shown). Dynamic values such as feature names, row IDs, and class labels (denoted {*placeholder*}) are extracted from the original question via regex and re-injected verbatim at generation time.

Q-Type	Variant 1	Variant 2	Variant 3
Q1	What was the most influential feature driving the model's prediction?	Which feature played the biggest role in this prediction?	What did the model rely on most heavily to reach this conclusion?
Q2	Which input feature had the least effect on the model's decision?	Which feature mattered least to the model's prediction?	Which feature is least relevant to the model's conclusion?
Q3	What features were most critical in setting this prediction apart from the runner-up class?	Which features tipped the balance toward this prediction over the runner-up?	What distinguishes this prediction from the second most likely alternative?
Q4	What differences in the inputs account for the model giving instances A and B different predictions?	Why does the model give different predictions to these two instances?	What input differences lead to contrasting predictions for these two instances?
Q5	Would the model's output change if a specific feature were removed from this input?	If a particular feature were excluded, would the model's prediction remain the same?	How sensitive is the model's output to the presence of a specific feature?
Q6	What modifications to this instance would cause the model to output a different class?	How should this input be changed to flip the model's prediction to a different class?	What changes to this instance would lead the model to predict a different outcome?
Q7	How would the model's output be affected if {*feature*} were altered or removed?	What would happen to the model's prediction if {*feature*} were modified or excluded?	In what way would the model's output change if {*feature*} were no longer present?
Q10	What explains the model's success on {*correct instances*} but its failure on {*wrong instance*}?	Why does the model correctly classify {*correct instances*} but fail on {*wrong instance*}?	What distinguishes {*wrong instance*} from {*correct instances*} such that the model predicts it incorrectly?

its strategy and explanation conditioned on this feedback. Fig. 8 shows that MEA consistently outperforms Agent+Reflection despite the added inference-time step: on text counterfactual questions, Agent+Reflection scores 0.644 while MEA reaches 0.876 (**+36%**); on vision counterfactual questions, 0.439 vs. 0.569 (**+30%**). Strikingly, on tabular spurious features the critic step even *degrades* performance (0.514) relative to the base agent (0.677), while MEA achieves 0.997. These results demonstrate that RL-based end-to-end training provides substantially stronger and more reliable gains than inference-time critique.

Prompt. Figure 10 shows the complete prompt structure passed to the Critic Agent. Each section is populated with runtime values extracted from the pipeline outputs.

Output Schema. The Critc returns two feedbacks, separated by the markers `PROPOSER_REFLECTION` and `ACTOR_REFLECTION`. Table 9 describes each field. The Proposer Reflection is consumed by the Proposer Agent to revise its tool selection; the Actor Reflection is consumed by the Actor Agent to revise its explanation.

A.9 ABLATION STUDY: PROPOSER STRATEGY MODE

The Proposer's strategy schema $S = (s_{\text{type}}, \mathcal{T}_{\text{sel}}, \mathcal{A}_{\text{auto}})$ (Sec. 3.3) allows $s_{\text{type}} \in \{\texttt{autonomous}, \texttt{tools}, \texttt{hybrid}\}$, but the main results always let the Proposer choose freely among the three. To isolate how much autonomous reasoning contributes beyond tool use alone, we force the Proposer into each single-mechanism regime – `tool_only` (external XAI tools alone,

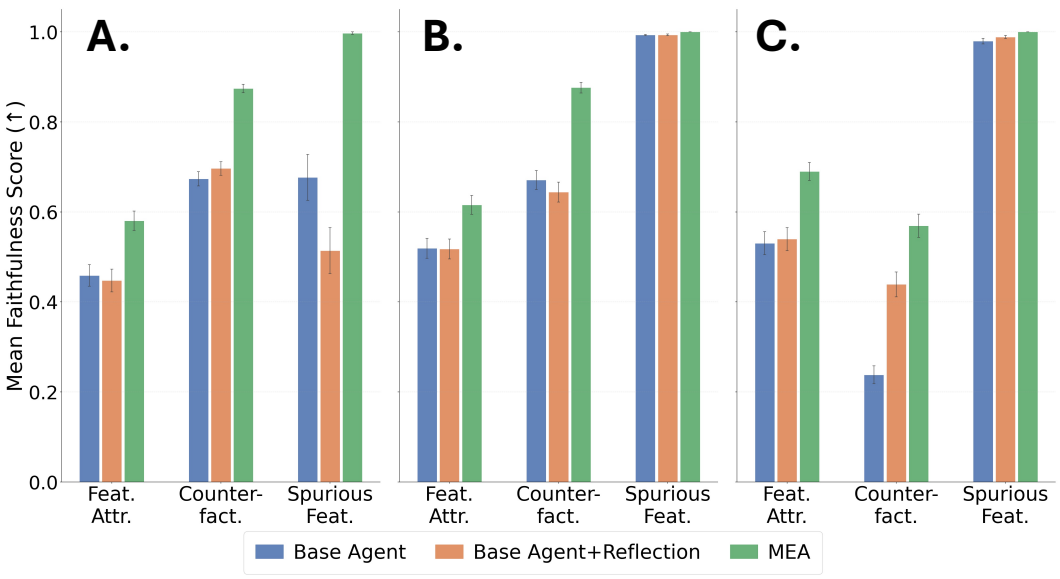


Figure 8: **Ablation on critic agent.** Mean faithfulness scores across three que1on categories for the base agent, base agent augmented with a critic agent, and MEA, evaluated on **(A)** tabular, **(B)** text, and **(C)** vision modalities. MEA consistently outperforms both baselines, demonstrating that RL-based training yields stronger improvements than inference-time critique.

Table 9: Output JSON schema of the critic agent.

Object	Field	Description
PROPOSER	`analysis`	Overall assessment of tool selection
	`tool_analysis`	Per-tool rating (*high/medium/low*) and recommendation (*keep/remove/replace*)
	`tools_to_keep`	Tools to retain in the revised strategy
	`tools_to_remove`	Tools to drop
	`tools_to_add`	New tools to include
	`strategy_suggestions`	Free-text suggestions to improve tool selection
ACTOR	`analysis`	Overall assessment of explanation quality
	`identified_issues`	Specific errors or weaknesses in the explanation
	`region_feedback`	Feedback on the accuracy of the identified region or feature
	`explanation_feedback`	Feedback on clarity and correctness
	`improvement_suggestions`	Free-text suggestions to improve the explanation

autonomous reasoning disabled) and `autonomous_only` (self-reasoning alone, zero tool calls) – for both the base agent and MEA, over a fixed 500-question sample stratified by (dataset × question type) and allocated across modalities in proportion to test-set size (tabular 150 / text 195 / vision 155; Q1–Q7, Q10). Isolation is enforced at two levels: the disallowed branch's schema field is omitted from the Proposer prompt entirely, and a post-hoc check strips any leaked fields and fails the sample if the required branch is empty – verified directly against execution logs (zero leaked autonomous tasks in `tool_only` runs, zero tool invocations in `autonomous_only` runs).

Tool use is the stronger single mechanism for both models, but autonomous reasoning with *zero* external tool calls already recovers 87.3% (base) / 91.3% (MEA) of tool-only faithfulness. Training improves the autonomous-only arm *more* than the tool-only arm ($+0.153$ vs. $+0.141$ mean faithfulness), narrowing the tool/autonomous gap after training. Most notably, for MEA letting the Proposer choose freely (0.748) outperforms *both* forced single-mechanism arms in every one of the three modalities (tabular $+0.006$, text $+0.018$, vision $+0.026$ over the better single mode), whereas for the base agent, free choice (0.516) is statistically indistinguishable from autonomous-only and

CoT Actor Preamble

```
## Reasoning Mode:  Chain-of-Thought
```

Before generating your explanation, reason through the evidence step-by-step:

1. **Understand the question**: What exactly is being asked?
2. **Analyze the XAI tool results**: What do the attribution maps, importance scores, and statistics tell us?
3. **Identify the key features/regions**: Based on the tool outputs, which specific parts of the input are most relevant?
4. **Synthesize your answer**: Combine the evidence into a clear, specific, and well-supported explanation.

Think step by step, then provide your final answer as valid JSON.

ReAct Actor Preamble

```
## Reasoning Mode:  ReAct (Reasoning + Acting)
```

Interleave your reasoning with the evidence to build your explanation:
For each piece of evidence from the XAI tools:
Thought: What does this result tell us about the model's decision?
Observation: Note the key values, features, or regions highlighted.
Reflection: How does this connect to answering the question?

After processing all evidence:
Final Thought: Synthesize all observations into a coherent answer.

Then provide your final answer as valid JSON.

ToT Actor Preamble

```
## Reasoning Mode:  Tree of Thought
```

Consider multiple possible interpretations of the evidence before settling on your answer:
Interpretation 1: What does the evidence suggest if you weight the highest-attribution features most heavily?
Interpretation 2: What alternative explanation could account for the same tool results?
Interpretation 3: Which interpretation is most consistent with ALL the evidence, not just the strongest signal?

Evaluate each interpretation against the evidence quality, specificity, and completeness. Select the strongest interpretation and provide your final answer as valid JSON.

Figure 9: Generic actor preambles prepended to the actor prompt for each reasoning baseline across all question types.

clearly below tools-only (0.589) – i.e., productively combining the two mechanisms is a capability training creates, not a free byproduct of offering both options. On Q6 (flip prediction) specifically, MEA's free-choice arm (0.822) exceeds both tool_only (0.766) and autonomous_only (0.782) taken individually on the identical instances, a direct in-model demonstration that the two mechanisms are complementary rather than redundant.

A.10 ABLATION STUDY: BACKBONE SCALE TRANSFER

To test whether MEA's training benefit transfers to a substantially smaller policy backbone, we repeat the Base-Agent-vs-MEA comparison on **Qwen3.5-4B** (Qwen Team, 2026a) (roughly an order of magnitude smaller in active parameters), in addition to training-reward curves on a second backbone Qwen3-VL-30B-A3B-Instruct in Appendix A.5. We train Qwen3.5-4B for a single epoch (300

Table 10: Mean faithfulness by model and Proposer strategy mode.

Model	Strategy Mode	n	Mean Faith.	Pass Rate
Base Agent	Autonomous only	493	0.515	48.5%
	Tools only	475	0.589	54.9%
	Free choice (hybrid)	2622	0.516	49.4%
MEA	Autonomous only	498	0.668	66.7%
	Tools only	498	0.731	74.7%
	Free choice (hybrid)	2654	**0.748**	**76.5%**

LoRA steps, rank 32) with the identical reward pipeline, under the same evaluation protocol used throughout the paper.

Table 11 shows that, on this smaller 4B backbone, MEA shows similar improvement as reported for the 30B & 35B backbones, suggesting that training improves faithfulness in every modality overall (tabular $0.419 \rightarrow 0.671$ ($+60\%$), text $0.403 \rightarrow 0.575$ ($+43\%$), vision $0.142 \rightarrow 0.508$ ($+258\%$)), demonstrating that faithfulness gain comes from the training procedure rather than being an artifact specific to a single, large backbone.

Table 11: Faithfulness by question category, 4B backbone. Best per row in **bold**.

Modality	Category	Baseline	Trained
Tabular	Feature Attribution	0.446	**0.578**
	Counterfactual	0.398	**0.763**
	Spurious Features	0.356	**0.933**
	Overall	0.419	**0.671**
Text	Feature Attribution	0.325	**0.534**
	Counterfactual	0.413	**0.480**
	Spurious Features	0.660	**0.960**
	Overall	0.403	**0.575**
Vision	Feature Attribution	0.297	**0.522**
	Counterfactual	0.028	**0.488**
	Spurious Features	1.000	0.955
	Overall	0.142	**0.508**

A.11 SANITY CHECK: SYNTHETIC SHORTCUT ON A BIASED CLASSIFIER

A concern with any LLM-based explanation agent is that it may be faithful to the target model only when the model's decisions already align with the LLM's own semantic priors, rather than to the target model's actual (possibly idiosyncratic) decision process. To test this directly, we trained a `TwoLayerNN` on Adult Census with an injected synthetic shortcut column, `verification_flag`, $\sim$90% correlated with the true label (noise rate 0.10, seed 42) – a feature whose dominance a domain expert or LLM prior would not anticipate for an income-prediction task. We independently verified the injected bias before any agent evaluation: overall test accuracy 0.939; accuracy on rows where the flag agrees with the true label (89.7% of rows) is 0.981, dropping to 0.575 on the 10.3% of rows where they disagree, and on those disagreement rows the model follows the flag (not the truth) 42.5% of the time; an ablation flip-rate test confirms `verification_flag` is $\sim 20\times$ more influential on the model's output than a comparable real feature (education-num flip rate 0.8% vs. 16.1%).

MEA wins or ties on every one of the 8 question types tested (Q1–Q7, Q10), with 5 of 8 significant at $p < 0.05$ (Table 12). Beyond the aggregate score, we checked whether each agent actually *names* the shortcut when it matters: on the 13/20 Q1 rows where `verification_flag` is verifiably the top driver by occlusion importance, MEA names it correctly in 12/13 vs. 9/13 for the base agent; on Q2 ("least responsible feature"), the base agent incorrectly calls `verification_flag` least responsible on 10/20 rows where it is actually dominant – a genuine faithfulness failure – while MEA does this 0/20 times. On a classifier whose decision logic is engineered to deviate

Table 12: Faithfulness score by question category on the biased classifier, paired t-test by (q_type, row_no).

Category	Base	MEA	Δ	p
Feature Attribution (Q1–4)	0.407	**0.626**	$+0.219$	$1.0e{-6}$
Counterfactual (Q5–7)	0.653	**0.941**	$+0.288$	$1.2e{-5}$
Spurious Features (Q10)	0.725	**1.000**	$+0.275$	0.0022

from human/LLM-intuitive reasoning, MEA's explanations remain grounded in the model's actual (biased) decision process rather than reverting to prior-knowledge-consistent reasoning. We scope this finding to the single bias-injection design and tabular modality tested here, and note it does not directly test vision-specific shortcuts (*e.g.,* watermarks) or amplified real sensitive attributes.

A.12 STATISTICAL SIGNIFICANCE TESTS

We report two-sided Welch's t-tests (unequal variance, unpaired) WELCH (1947) for comparisons against independent baselines, and paired t-tests for the paraphrase ablation (where scores are matched by sample ID across conditions). The significance level is $\alpha = 0.05$; "n.s." denotes $p > 0.05$. All values are mean $\pm$ standard error of the mean (SEM).

Vision Feature-Attribution Tool-Only Baselines.

Table 13 compares MEA (Q1, vision; $n = 96$, mean$= 0.499 \pm 0.029$) against six tool-only XAI baselines evaluated on the same vision feature-attribution task (Sensitivity had no evaluation data and is excluded). MEA is statistically equivalent to three gradient-based methods (IG, Guided Backprop, SmoothGrad) and significantly outperforms three others (SHAP, LIME, GradCAM).

Table 13: Welch's t-test: MEA vs. tool-only XAI baselines (vision, feature attribution Q1). MEA mean $= 0.499 \pm 0.029$ ($n = 96$).

Baseline	n	Mean $\pm$ SEM	Δ (MEA $-$BL)	t	p	Sig.
SHAP	59	0.018 ± 0.007	$+0.481$	16.26	< 0.001	*
LIME	96	0.259 ± 0.027	$+0.240$	6.09	< 0.001	*
Integrated Gradients	**96**	$\mathbf{0.481 \pm 0.029}$	$\mathbf{+0.018}$	**0.44**	**0.661**	**n.s.**
Guided Backprop	**96**	$\mathbf{0.497 \pm 0.027}$	$\mathbf{+0.003}$	**0.07**	**0.944**	**n.s.**
SmoothGrad	**96**	$\mathbf{0.496 \pm 0.030}$	$\mathbf{+0.003}$	**0.08**	**0.935**	**n.s.**
GradCAM	96	0.131 ± 0.017	$+0.368$	10.95	< 0.001	*

Spurious-Feature Faithfulness (Q10). Table 14 reports Welch's t-test results for the spurious-feature question type (Q10) across text and vision modalities. MEA achieves **ceiling performance** (1.000 ± 0.000, $n = 160$) on both modalities–every sample is scored correctly. All Welch's t-tests return $p < 0.05$; however, this should be interpreted in light of the ceiling effect: when one group has zero variance, any non-zero difference in means will be flagged as significant regardless of its practical magnitude. For the strongest baselines the actual performance gap is negligible–Gemini-3.1 reaches 0.996 on text (gap: $+0.004$) and Base Agent reaches 0.981 on vision (gap: $+0.019$)–indicating that top-performing models already saturate Q10 and MEA's advantage is marginal. The significant p-values for weaker baselines (e.g. Claude-Haiku-4.5: 0.665 text, 0.855 vision) do reflect a meaningful improvement. CoT, ToT, and ReAct baselines had no Q10 evaluation data available.

Counterfactual Flip-Success Significance (Q6). Table 1 reports flip-success rates per model architecture, and Table 15 reports two-sided Fisher's exact tests comparing MEA against each baseline on a per-model basis. No comparison reaches significance at $\alpha = 0.05$: even where a baseline's raw flip-success rate is nominally higher than MEA's (e.g., DiCE 100.0% vs. MEA 90.9% on `cancer_2layernn`), the difference is not statistically distinguishable given the sample size ($n = 22$). This confirms that MEA's substantially higher faithfulness scores (Table 1) are not offset by a genuine loss in flip-success reliability.

Paraphrase Robustness Ablation. Table 16 presents paired t-tests comparing Paraphrase (mean of three paraphrase runs, matched on common sample IDs) against MEA. In 8 of 9 question-

Table 14: Welch's t-test: MEA (1.000 ± 0.000, $n = 160$) vs. baselines on spurious-feature questions (Q10).

Modality	Baseline	n	Mean $\pm$ SEM	Δ (MEA $-$BL)	t	p	Sig.
Text	Base Agent	156	0.949 ± 0.017	$+0.051$	3.09	0.002	*
	GPT-5.4	157	0.733 ± 0.035	$+0.268$	7.67	< 0.001	*
	Claude-Haiku-4.5	158	0.665 ± 0.037	$+0.335$	9.11	< 0.001	*
	Gemini-3.1	160	0.996 ± 0.001	$+0.004$	2.92	0.004	*
Vision	Base Agent	148	0.981 ± 0.004	$+0.019$	5.15	< 0.001	*
	GPT-5.4	156	0.921 ± 0.006	$+0.079$	12.53	< 0.001	*
	Claude-Haiku-4.5	160	0.855 ± 0.011	$+0.145$	13.58	< 0.001	*
	Gemini-3.1	159	0.979 ± 0.003	$+0.021$	6.54	< 0.001	*

Table 15: Fisher's exact test: MEA vs. DiCE and Growing Spheres on flip success (Q6), per model architecture.

Model	**MEA**	**Baseline**	p
vs. DiCE			
`adult_2layernn`	31/31 (100.0%)	32/32 (100.0%)	1.000
`adult_tabnn`	32/32 (100.0%)	32/32 (100.0%)	1.000
`cancer_2layernn`	20/22 (90.9%)	22/22 (100.0%)	0.488
`cancer_tabnn`	22/22 (100.0%)	21/22 (95.5%)	1.000
vs. Growing Spheres			
`adult_2layernn`	31/31 (100.0%)	32/32 (100.0%)	1.000
`adult_tabnn`	32/32 (100.0%)	31/32 (96.9%)	1.000
`cancer_2layernn`	20/22 (90.9%)	17/22 (77.3%)	0.412
`cancer_tabnn`	22/22 (100.0%)	19/22 (86.4%)	0.233

type$\times$modality conditions, there is no statistically significant difference, confirming that MEA's performance is robust to prompt paraphrasing. The single significant difference (vision counterfactual, $p = 0.026$) shows Paraphrase outperforming MEA by 0.075, suggesting that this question type may benefit from additional training diversity.

Table 16: Paired t-test: Paraphrase vs. MEA (matched common samples). Δ = Paraphrase $-$ MEA.

Modality	Question Group	n_{pairs}	Paraphrase ($\mu \pm$ SEM)	MEA ($\mu \pm$ SEM)	Δ	t	p	Sig.
Tabular	Feature Attribution	106	0.683 ± 0.037	0.687 ± 0.037	-0.004	-0.63	0.532	n.s.
	Counterfactual	32	0.934 ± 0.023	0.939 ± 0.021	-0.005	-1.98	0.056	n.s.
	Spurious Features	17	1.000 ± 0.000	1.000 ± 0.000	0.000	–	–	n.s.
Text	Feature Attribution	126	0.738 ± 0.030	0.746 ± 0.031	-0.008	-0.83	0.407	n.s.
	Counterfactual	85	0.947 ± 0.012	0.952 ± 0.012	-0.005	-0.73	0.469	n.s.
	Spurious Features	37	1.000 ± 0.000	1.000 ± 0.000	0.000	–	–	n.s.
Vision	Feature Attribution	96	0.686 ± 0.034	0.693 ± 0.034	-0.007	-0.56	0.575	n.s.
	Counterfactual	**72**	$\mathbf{0.614 \pm 0.041}$	$\mathbf{0.539 \pm 0.044}$	$\mathbf{+0.075}$	**2.27**	**0.026**	*
	Spurious Features	40	1.000 ± 0.000	1.000 ± 0.000	0.000	–	–	n.s.

Note: "–" denotes that both conditions achieve identical ceiling performance (all scores $= 1.0$), making t-test inapplicable.

A.13 COMPUTATIONAL COST

Model architecture and parameters. MEA uses **Qwen3.6-35B-A3B** (Qwen Team, 2026b) as its backbone, a Mixture-of-Experts multi-modal language model with **35 billion total parameters** and approximately **3.5 billion parameters activated per forward pass** (A3B). During inference for evaluation, we serve the model via LoRA adapters with rank 32, adding $< 1\%$ additional parameters relative to the frozen base. Baseline comparisons using the un-finetuned backbone employed `Qwen3.6-35B-A3B`, while closed-source frontier baselines (Gemini-3.1-flash, GPT-5.4, Claude-Haiku-4.5) were accessed via their respective public APIs.

Critic Agent Prompt Template

System Role:

You are a critical analyst evaluating an XAI agent's strategy and explanation.

User Input:

Original Question
`question text / example context`

Question Context
Question Type: `Q1-10`
Modality: `tabular | text | vision`

Strategy Used
Strategy Type: `tool-based`
Selected Tools: `list of tool names`
Strategy Reasoning: `Proposer's reasoning for tool selection`

Generated Explanation

Output: `Actor's structured output dict`
Explanation: `natural language explanation`
Confidence: `Actor's self-reported confidence`

Faithfulness Evaluation
Score: `faithfulness score`
Metric: `metric name and formula`
P_{original}: `value` P_{modified}: `value`
⋮

Task Instruction:
Analyze the strategy and results, then provide **two separate JSON outputs**:

1. `PROPOSER_REFLECTION`: feedback on tool selection strategy.
2. `ACTOR_REFLECTION`: feedback on explanation generation.

Consider what went wrong, which tools were most or least helpful, and how both the strategy and explanation could be improved.

Figure 10: Complete prompt structure of the Critic Agent.

GRPO training. All reinforcement learning training was conducted through *Tinker* (Lab, 2025), a cloud-based training service that manages GPU allocation and distributed rollout execution internally; per-GPU-hour accounting is therefore not directly available from our training logs. The main hybrid training run ran for $\sim$93 GRPO steps with a group size of 128-256 rollouts per step processed per step, covering all three modalities jointly. Each step produced a mean episode length of ≈ 3.6 turns and ≈ 393 action tokens.

Post-hoc explainers. All post-hoc explainers and faithfulness metric computation, and ablations were run on a NVIDIA A100 GPUs (1 GPU per job, 16 CPU cores, 30 GB system RAM). Jobs were parallelized across datasets and question types (up to 4 concurrent jobs per modality). The full test-set evaluation for a single modality (3.5K instances across 4 dataset-model combinations and 10 question types) required approximately 8-12 A100 GPU-hours depending on modality complexity, with vision being the most expensive due to image preprocessing and SmoothGrad/GradCAM computation. Summing across the three modalities, primary test evaluations consumed an estimated 30-40 A100 GPU-hours. Ablation runs (prompt robustness, critic agent, OOD datasets) added approximately an additional 20-30 A100 GPU-hours. Total estimated evaluation compute is therefore on the order of 50-70 A100 GPU-hours.

A.14 Prompt Templates for Proposer and Actor Agent

Refer to Fig.11 for the Actor Agent's prompt template and Fig.12 for the Proposer Agent's prompt template.

Actor Agent Prompt Template

System Role:

You are an XAI expert. Based on the provided analysis, `execute task-specific directive for Q1-Q10`.

User Input:

Context & Prediction
Question (`Q1-10`): `specific question text`
Model Prediction: Class: `predicted class` Conf: `score`

Input Data
`raw input text / tabular row / image data`

XAI Analysis
Tool Results Summary:
`status and most influential features per tool`

Detailed Statistics:

- **Integrated Gradients**: `top positive/negative attributions`
- **SHAP**: `top positive/negative shap values`
- **Sensitivity**: `most sensitive prob_drops`

Autonomous Reasoning Results:
GROUNDING: `extracted features matching query`
REASONING: `semantic explanation for features`
COMPARISON: `alignment with model prediction`

Your Task
`Specific directive corresponding to Question Q1-10`

Required Output Format (JSON only)

```
{
"output": {
"modality_key": expected data structure
},
"explanation": "2-3 sentences explaining your reasoning."
}
```

Critical Requirements

- **For vision:** `bounding_box` as `[x_min, y_min, x_max, y_max]`
- **For text:** `spans` as a list of {`start_index, end_index`}
- **For tabular:** `feature_keys` as a list of column names
- **OUTPUT SIZE CONSTRAINT:** `size limit dictated by question type`
- Base your decision on the attribution analysis.

Final Instruction: Respond with ONLY JSON.

Figure 11: Prompt structure for the Actor Agent. Task directives, constraints, and JSON schema keys are populated dynamically at runtime based on the specific question type (Q1-Q10) and data modality.

Proposer Agent Prompt Template

System Role:

You are an AI explainability expert designing a strategy to answer a user's question about a machine learning model's prediction on `MODALITY`.

User Input:

Question & Data Context
Question: `user question text`
Input Content: `truncated input text, image, or tabular row`

Model Information
Model: `model name` Arch: `architecture`
Prediction: `class index & name` (Conf: `score`)
Top-K Predictions: `list of dicts containing top probabilities`

Available Methods

1. Autonomous Analysis: `instructions on using internal reasoning to ground/extract features`

2. External XAI Tools:

- `lime`: executes local surrogate explanation
- `integrated_gradients`: computes path attribution
- `shap`: computes shapley values
- `sensitivity_analysis`: leave-one-out ablation
- `guided_backprop`: modified ReLU gradients
- `smoothgrad`: averaged noisy saliency

Your Task
`Specific strategy design directive corresponding to Question Q1-10`

Required Output Format (JSON only)

```
{
"strategy_type": "autonomous" | "tools" | "hybrid",
"reasoning": "Explain strategy choice (2-3 sentences)",
"autonomous_tasks": [
{
"task_type": "grounding" | "reasoning" | "comparison",
"query": "Specific query for autonomous analysis",
"expected_output": "What should be extracted"
}
],
"tool_selection": {
"selected_tools": subset of available tools,
"reasoning": "Why these tools for this modality"
}
}
```

Final Instruction: Provide your strategy as a JSON object:

Figure 12: Prompt structure for the Proposer Agent. Task directives are populated dynamically based on the active question type (Q1-Q10), ensuring the strategy generated aligns with the specific objective.