Counterfactual Evidence Fields for Multimodal Foundation Models

Anonymous ACL submission

Abstract

Multimodal foundation models can answer correctly while relying on evidence that would not justify the answer. We introduce **Counterfactual Evidence Fields**, CEF-MFM, to bind each reasoning step to evidence slots that support, falsify, or leave the step ambiguous. Support slots must preserve the answer, falsification slots must make it fail when removed or swapped, and ambiguity slots must lower confidence when evidence is insufficient. We prove that these roles form a minimal decomposition for enforcing sufficiency, necessity, and calibrated uncertainty. A joint counterfactual objective trains the field with an evidence-calibrated decoder. On eight multimodal reasoning benchmarks, CEF-MFM improves same-backbone accuracy by up to $+1.6$ points and yields larger evidence gains: $+8.7$ faithfulness, $+13.6$ deletion sensitivity, $+9.4$ swap success, $+5.8$ distractor robustness, and lower calibration error, from .173 to .153. Mechanism analyses show a field-level triple dissociation, a deletion–swap asymmetry, and a controlled advantage over a single-verifier-head ablation. The results shift multimodal evaluation from answer correctness alone to whether the model fails when decisive evidence is removed. The code is available in `https://anonymous.4open.science/r/CEF_MFM-0F7E`.

1 Introduction

A multimodal foundation model can answer a chart question correctly while relying on a decorative legend instead of the decisive data bar (Figure 1). This *right answer, wrong evidence* failure is common: high accuracy often coexists with weak dependence on the evidence that makes an answer necessary (Turpin et al., 2023; Lanham et al., 2023). Standard benchmarks rarely expose this gap, but deployment requires confidence to fall when decisive evidence is removed.

The problem is representational. Existing methods lack an object that binds each reasoning step to the evidence whose removal should make that step fail. Attention maps show where a model looks, not which evidence would invalidate a conclusion (Jain and Wallace, 2019). Verifiers (Zhang et al., 2025), tree search (Wang et al., 2025b), and activation steering (Su et al., 2025) improve answer paths or suppress hallucination, but do not force the model to depend on the right evidence during training.

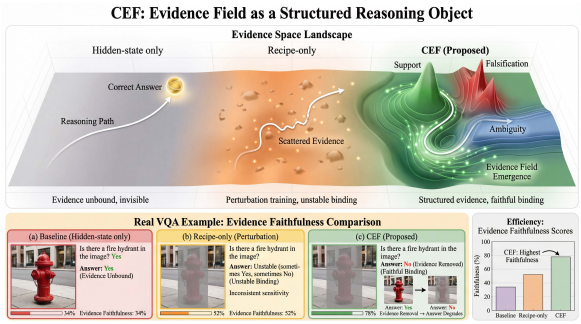

Figure 1: Standard multimodal pipelines can answer correctly while using non-decisive evidence. CEF-MFM couples each reasoning step to support, falsification, and ambiguity slots, so removing decisive evidence degrades the answer in a controlled way.

We introduce the **counterfactual evidence field**. For each step s_k, the field $E_k = (\alpha_k, \beta_k, \gamma_k)$ assigns multimodal slots to support, falsification, and ambiguity roles. The support field sustains a step, the falsification field identifies evidence whose removal should collapse it, and the ambiguity field marks evidence that should lower confidence. Section 3 shows that this is the minimal partition for jointly enforcing sufficiency, necessity, and calibrated uncertainty.

We instantiate the framework as **CEF-MFM**: a unified evidence slot bank, a step–evidence coupler, a falsification engine for targeted delete/swap/distract perturbations, and an evidence-calibrated decoder (Figure 2). Three contributions follow:

1. We formalize the **counterfactual evidence**

field and prove that its three-way decomposition is the minimal sufficient structure for enforcing sufficiency, necessity, and calibration; see Proposition 1.

2. We develop a **joint counterfactual objective** whose loss terms map directly to these conditions, and verify the predicted *triple dissociation* under ablation.

3. We introduce an **evidence-centric evaluation protocol** covering deletion sensitivity, swap success, distractor robustness, faithfulness, localization, and ECE, then isolate the decomposition effect with a single-verifier-head control.

Across eight benchmarks, CEF-MFM improves accuracy by $+0.9$ to $+1.6$ over the strongest same-backbone baseline and yields larger evidence-centric gains, including $+8.7$ faithfulness and $+13.6$ deletion sensitivity. The small accuracy gain and large evidence gain show that the method changes the mechanism behind correct answers rather than only their frequency.

2 Related Work

Reasoning verification and faithfulness. Visu-oThink (Wang et al., 2025b), GFlowVLM (Kang et al., 2025), MM-Verify (Sun et al., 2025), and Argus (Man et al., 2025) improve multimodal reasoning through search, flow modeling, or verification. Critic-V (Zhang et al., 2025), ROD-MLLM (Yin et al., 2025), and activation steering (Su et al., 2025) target reliability, while MMDT (Xu et al., 2025a) exposes brittleness under perturbation. Modern backbones (Wang et al., 2024; Liu et al., 2023; Chen et al., 2024c; Yao et al., 2024; OpenAI, 2023; Google DeepMind, 2023; Lu et al., 2024a; Chen et al., 2024b; Dong et al., 2024; Hurst et al., 2024; Grattafiori et al., 2024) have also raised accuracy. These methods improve answer paths or outputs, but do not constrain which evidence the answer depends on. SPARCL (Li and Li, 2025), Arasteh et al. (2026), DeFacto (Xu et al., 2025b), and MCIL (Zhou et al., 2024) use perturbations or counterfactual training, but treat evidence as a single signal rather than separating support from falsification.

Causal attribution and counterfactual explanation. Causal mediation (Vig et al., 2020), causal scrubbing (Chan et al., 2023), and counterfactual explanations (Goyal et al., 2019) probe or perturb models after prediction. RFEval (Han et al., 2026), FaithEval (Ming et al., 2024), and related work (Lanham et al., 2023; Turpin et al., 2023; Wu et al., 2025; Chen et al., 2024a; Mathew et al., 2021; Masry et al., 2022; Mathew et al., 2022; Wang et al., 2025a) show that coherent reasoning can still be unfaithful. These lines are close in spirit but post-hoc. CEF-MFM makes evidence dependence a training-time primitive: removing decisive support should make the answer fail, while inserting distractors should leave it stable.

3 Method

We formalize *evidence dependence* and derive the architecture from it.

3.1 Evidence Dependence: Three Conditions

Let $x = (x^{(1)}, \ldots, x^{(M)})$ be a multimodal context, q the query, and $p_\theta(y \mid x, q)$ the answer model. Evidence slots $\mathcal{Z} = \{z_1, \ldots, z_N\}$ project image regions, text spans, and chart tokens into a shared latent space. Reasoning is segmented into steps $s_1, \ldots, s_K$.

A reasoner exhibits *faithful evidence dependence* at step s_k with respect to a subset $D_k \subseteq \mathcal{Z}$ when three conditions hold simultaneously:

(S) Sufficiency. Restricting the input to D_k preserves the answer: $\hat{y}(D_k, q) = \hat{y}(\mathcal{Z}, q)$.

(N) Necessity. Removing or replacing D_k within the full context causes the answer distribution to diverge: $D_{\mathrm{JS}}(p_\theta(y \mid x, q) \,\|\, p_\theta(y \mid x^{\setminus D_k}, q)) \geq m$ for some margin $m > 0$.

(C) Calibrated uncertainty. When no subset of $\mathcal{Z}$ jointly satisfies (S) and (N), decoder confidence at step s_k is bounded above by a function of the evidential ambiguity.

Maximum-likelihood training only optimizes $\log p_\theta(y \mid x, q)$. It can preserve the answer distribution while reshuffling the link between slots and internal states, so S, N, and C remain unenforced. A model can therefore exploit spurious correlations (Turpin et al., 2023). The counterfactual evidence field turns these conditions into trainable objects.

3.2 Counterfactual Evidence Field

For each step s_k we define a triple of distributions over slots,

$$E_k = (\alpha_k, \beta_k, \gamma_k) \in \Delta^N \times \Delta^N \times \Delta^N, \quad (1)$$

where Δ^N is the N-simplex. Each component maps to one condition:

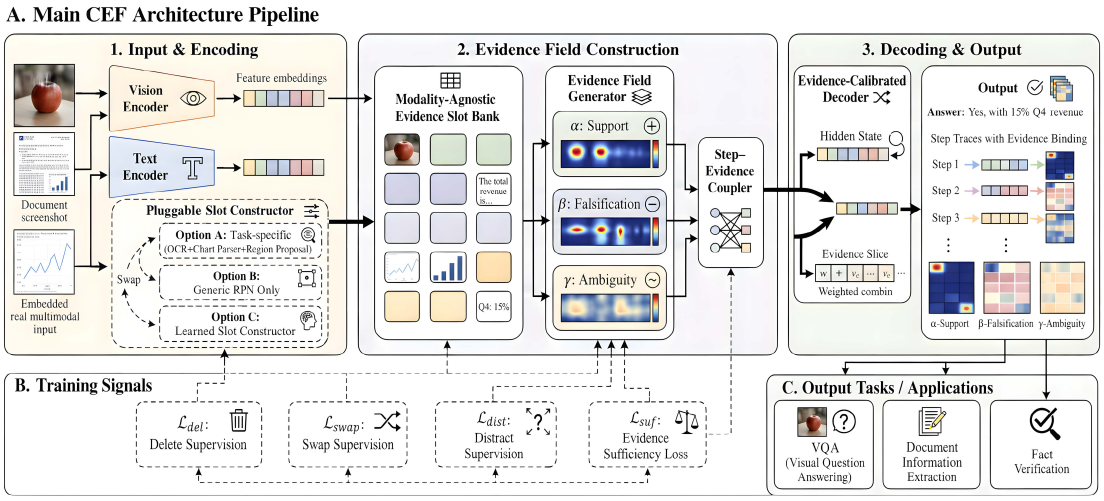

Figure 2: CEF-MFM architecture. Inputs are projected into a shared slot bank, reasoning steps are coupled to support, falsification, and ambiguity fields, perturbations are generated from falsification slots, and decoding is modulated by evidence-calibrated confidence.

- α_k, *support*: slots whose retention should recover the answer.
- β_k, *falsification*: slots whose removal or replacement should change the answer.
- γ_k, *ambiguity*: slots that signal uncertainty and should suppress confidence.

Two distributions may seem enough, since the slots that sustain a step often look like those whose removal should harm it. The proposition shows otherwise.

Proposition 1 (Minimality of the three-way decomposition)**.** *Let $\mathcal{L}_{\text{ans}} + \mathcal{L}_{\text{S}} + \mathcal{L}_{\text{N}} + \mathcal{L}_{\text{C}}$ be any loss whose minimization implies that the trained model satisfies (S), (N), (C) at every step s_k. Then $\mathcal{L}_{\text{S}}, \mathcal{L}_{\text{N}}, \mathcal{L}_{\text{C}}$ must depend on three distinct slot-level distributions: any two-distribution reduction permits a model that satisfies the corresponding two conditions while violating the third.*

Sketch. Optimizing only α_k can preserve decisive evidence under masking but cannot distinguish losing decisive evidence from losing redundant copies, so N can fail. Optimizing α_k and β_k can satisfy S and N but leaves confidence unconstrained when evidence is insufficient, so C can fail. The ambiguity field γ_k is therefore required. Appendix A gives the full argument.

Thus α, β, and γ are not arbitrary attention heads; they are the minimal structure for jointly enforcing the three conditions.

3.3 Joint Counterfactual Training Objective

The total loss augments answer likelihood with four auxiliary terms:

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \lambda_1 \mathcal{L}_{\text{sup}} + \lambda_2 \mathcal{L}_{\text{fal}} + \lambda_3 \mathcal{L}_{\text{amb}} + \lambda_4 \mathcal{L}_{\text{step}}, \tag{2}$$

with $\mathcal{L}_{\text{ans}} = -\sum_t \log p_\theta(y_t \mid y_{<t}, x, q)$.

Support sufficiency. Let $\text{top}_r(\alpha_k)$ retain the r highest-weighted slots and mask the rest. Then

$$\mathcal{L}_{\text{sup}} = -\sum_{k=1}^{K} \log p_\theta\big(y \mid \text{top}_r(\alpha_k), q\big). \tag{3}$$

This requires a sparse α_k to suffice for a correct prediction.

Falsification sensitivity. For a perturbation operator π acting on the top-β_k slots from Section 3.4.3, let $x^{\pi,k}$ denote the perturbed input:

$$\mathcal{L}_{\text{fal}} = \sum_{k=1}^{K} \max\big(0,\, m - D_{\text{JS}}(p_\theta(y \mid x, q) \,\|\, p_\theta(y \mid x^{\pi,k}, q))\big), \tag{4}$$

with margin m=0.3. The model is rewarded for being appropriately fragile to slots it marks as falsification-critical. $\mathcal{L}_{\text{sup}}$ alone cannot enforce this because sufficiency may rely on supersets containing spurious slots.

Ambiguity calibration. When γ_k concentrates on specific uncertainty sources, decoder confidence

should not exceed a ceiling:

$$\mathcal{L}_{\text{amb}} = \sum_{k=1}^{K} \max\big(0,\, \text{conf}_k - \tau\big) \cdot \mathbf{1}[\mathcal{H}(\gamma_k) < \epsilon], \tag{5}$$

where conf_k is the decoder's top-token probability, τ is the ceiling, ϵ is an entropy threshold, and $\mathcal{H}$ is Shannon entropy. The indicator activates only under concentrated ambiguity, avoiding a uniform confidence cap.

Step coherence. Adjacent reasoning steps that address different sub-problems should shift their evidence support; steps revisiting the same evidence should keep it. We use a gated divergence,

$$\mathcal{L}_{\text{step}} = -\sum_{k=1}^{K-1} w_k \cdot D_{\text{JS}}(\alpha_k \| \alpha_{k+1}), \tag{6}$$

$$w_k = \sigma(u_k^\top W_g\, u_{k+1}), \tag{7}$$

where u_k is the step representation and W_g is learned. The gate is high for sub-problem transitions and low for evidence-persistent steps, avoiding artificial drift. Removing $\mathcal{L}_{\text{step}}$ reduces step consistency by 5.5 points; the gated variant improves faithfulness by 1.2 points over an ungated one, as shown in Appendix F.

Constraint interaction. S rewards answer preservation under masking, while N rewards answer change under perturbation. The model can satisfy both only by learning which slots are decisive under retention and which are decisive under intervention. These sets can differ under redundant evidence.

3.4 Architectural Realization

3.4.1 Evidence Slot Bank

Let $h_i^{(m)} \in \mathbb{R}^{d_m}$ denote the encoder state for the i-th unit from modality m, such as an image patch, OCR span, chart element, table cell, or temporal segment. Each unit is projected into a shared d-dimensional slot space, $z_i = P_{m(i)}\, h_i^{(m(i))}$, where $P_{m(i)} \in \mathbb{R}^{d \times d_{m(i)}}$ is learned. The slot bank is the grounding interface between heterogeneous perception and the evidence field.

3.4.2 Step–Evidence Coupling

The decoder's chain-of-thought output is segmented into semantic units by a single-layer MLP boundary detector over consecutive hidden states; within each segment, hidden states are mean-pooled to produce $u_k \in \mathbb{R}^d$. Three independent

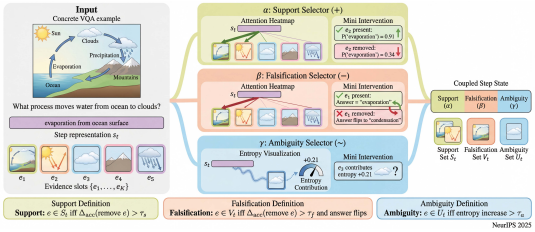

Figure 3: Step–evidence coupling from each reasoning step to support, falsification, and ambiguity distributions.

bilinear selectors then compute the evidence field components (Figure 3),

$$\alpha_{k,i} = \text{softmax}_i(u_k^\top W_\alpha\, z_i), \tag{8}$$

$$\beta_{k,i} = \text{softmax}_i(u_k^\top W_\beta\, z_i), \tag{9}$$

$$\gamma_{k,i} = \text{softmax}_i(u_k^\top W_\gamma\, z_i), \tag{10}$$

with $W_\alpha, W_\beta, W_\gamma \in \mathbb{R}^{d \times d}$. Softmax normalizes each component, while independent projections let the three roles attend to different evidence aspects. The coupler conditions the *forward decoding path* on E_k, making evidence dependence part of computation rather than retrospective attribution.

Slots need not be exclusive across fields. Separation comes from the losses: $\mathcal{L}_{\text{sup}}$ rewards retaining high-α slots, $\mathcal{L}_{\text{fal}}$ rewards degradation under high-β perturbation, and $\mathcal{L}_{\text{amb}}$ penalizes overconfidence under concentrated γ. Section 4.3 confirms that scrambling each field causes a distinct degradation pattern.

3.4.3 Falsification Engine

The engine converts β_k into three perturbation operators (Figure 4):

- **Delete**, π_{del}: zero the embeddings of the top-3 high-β_k slots to test dependence on decisive evidence presence.
- **Swap**, π_{swap}: replace a high-β_k slot with a contradictory alternative to test dependence on evidence content.
- **Distract**, π_{dist}: inject a correlated but non-decisive slot to test robustness to inert evidence.

Each candidate is oracle-validated: swaps must change the gold prediction, while distractors must not. Validation rates are 72% for swaps and 81% for distractors. Appendix I gives details, and Appendix I.1 shows stability under three alternative generators.

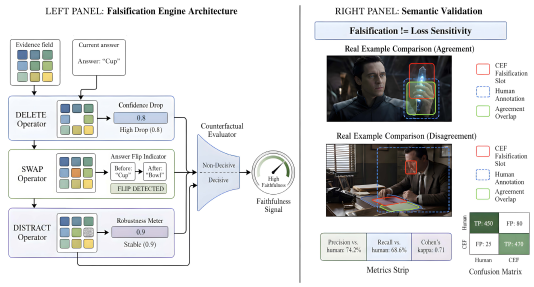

Figure 4: Falsification engine. High-β_k evidence is deleted, swapped, or paired with distractors to test whether the answer depends on decisive slots.

3.4.4 Evidence-Calibrated Decoding

The final architectural step connects E_k to decoder confidence. Because α_k and γ_k are normalized distributions, the relevant signal is their entropy. We define a scalar confidence gate

$$g_k = \sigma(a\,\mathcal{H}(\gamma_k) - b\,\mathcal{H}(\alpha_k) + c), \qquad (11)$$

with learnable a, b, c. Concentrated support and diffuse ambiguity open the gate toward confident decoding; the opposite regime makes generation conservative. Confidence is therefore tied to the local evidential state. Appendix H formalizes training and inference.

4 Experiments

We test four claims: CEF-MFM improves same-backbone accuracy, produces larger evidence-centric gains, shows the predicted *triple dissociation*, and explains its gains through mechanism analyses.

4.1 Setup

CEF-MFM uses Qwen2-VL-7B (Wang et al., 2024) with a frozen vision encoder and lightweight slot, coupling, and calibration heads. Each example has up to 32 evidence slots over image regions, OCR spans, chart elements, and short textual facts. Models train for 12 epochs on 4×A100 GPUs with up to 6 reasoning steps and 4 perturbation probes per step. Loss weights are $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)=(1.0, 0.5, 0.3, 0.2)$, selected on a held-out split and fixed across benchmarks. All numbers are mean ± std over 3 seeds.

Evaluation covers **compositional image reasoning** with ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2024), MathVista (Lu et al., 2024b), RealWorldQA (xAI, 2024), and SEED-style subsets (Li et al., 2023); **document and chart reasoning** with DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), and ChartQA (Masry et al., 2022); and **counterfactual diagnostics** built from targeted deletion, slot swapping, and distractor insertion on 500 held-out test instances per benchmark.

Five baselines share the backbone, data, and schedule: CoT VLM, Verifier-based VLM (Zhang et al., 2025), Rationale-supervised VLM, Grounding-augmented VLM, and an Attention-map proxy that replaces the coupler with raw cross-attention. Appendices J–K give implementation details, perturbation protocol, and external references.

4.2 Main Results

Evidence-centric metric definitions. **Faithfulness**: $\frac{1}{K}\sum_k \mathrm{Prec@5}(\alpha_k, \mathcal{A}_k^*)$, where $\mathcal{A}_k^*$ is the human-annotated decisive set. **Deletion sensitivity**: $\mathbb{E}[\mathrm{JSD}(p(y \mid x, q) \,\|\, p(y \mid x^{\mathrm{del}}, q))]$, scaled to $[0, 100]$. **Swap success**: $\mathbb{E}[\mathbf{1}(\hat{y} \neq \hat{y}^{\mathrm{swap}})]$. **Distractor robustness**: $\mathbb{E}[\mathbf{1}(\hat{y} = \hat{y}^{\mathrm{dist}})]$. **Localization**: IoU between top-5 support slots and annotated evidence regions. **ECE** (Guo et al., 2017) uses 15 confidence bins.

Accuracy–evidence asymmetry. Table 1 shows accuracy gains of +0.9 to +1.6 over the strongest baseline per benchmark. Table 2 shows larger gains: +8.7 faithfulness, +13.6 deletion sensitivity, +9.4 swap success, and −0.020 ECE. This asymmetry matches the theory. Standard accuracy can be saturated by dataset priors and surface co-occurrence (Turpin et al., 2023); evidence metrics isolate cases where reasoning must depend on decisive evidence.

Verifier comparison. The Verifier-based VLM is the strongest accuracy baseline, with 74.9 on RealWorldQA and 73.2 on MathVista, but trails CEF-MFM by 8.7 faithfulness points. The verifier re-ranks answers and rejects obvious errors; it does not constrain which evidence supports the answer. Closing the faithfulness gap requires changing the evidence representation, not only the final answer distribution.

4.3 Ablations and Triple Dissociation

Triple dissociation. Proposition 1 predicts that each field primarily controls one condition. Table 3 confirms this: scrambling α_k most hurts faithfulness by −7.8, scrambling β_k most hurts deletion sensitivity by −12.4, and scrambling γ_k most hurts

Table 1: Controlled same-backbone comparison using Qwen2-VL-7B with identical data and schedule. Mean ± std over 3 seeds. External reference models are in Appendix K.

Method	RealWorldQA	MMStar	MMBench	SEEDBench	ScienceQA	MMT-Bench	MathVista	MathVerse
CoT VLM	70.1±0.3	60.7±0.4	80.7±0.3	74.7±0.2	73.4±0.5	60.4±0.3	61.4±0.6	25.8±0.7
Verifier-based VLM	74.9±0.4	56.2±0.5	82.8±0.3	76.5±0.3	74.5±0.4	62.0±0.3	73.2±0.5	32.9±0.8
Rationale-supervised	72.3±0.3	59.1±0.4	81.4±0.2	75.3±0.3	74.1±0.4	61.2±0.4	68.7±0.5	29.4±0.6
Grounding-augmented	73.1±0.3	60.9±0.3	82.0±0.3	75.8±0.2	75.2±0.4	61.5±0.3	70.1±0.4	30.8±0.7
Attention-map proxy	71.8±0.4	58.4±0.5	81.1±0.3	74.9±0.3	73.9±0.5	60.8±0.4	66.3±0.6	28.1±0.8
CEF-MFM (Ours)	**76.2**±0.3	**62.4**±0.4	**84.1**±0.2	**78.1**±0.3	**76.1**±0.4	**63.5**±0.3	**74.8**±0.4	**34.2**±0.6

Table 2: Evidence-centric evaluation (mean ± std, 3 seeds). Lower is better for ECE; higher for all others.

Method	Faith.↑	Del.↑	Swap↑	Dist.↑	Loc.↑	ECE↓
CoT VLM	58.2	41.7	36.5	61.1	52.9	.184
Rationale-sup.	60.1	43.8	38.2	62.4	54.1	.179
Grounding-aug.	61.7	44.6	37.9	63.8	55.8	.176
Verifier-based	63.4	46.9	40.3	64.7	56.4	.173
CEF-MFM	**72.1**	**60.5**	**49.7**	**70.5**	**62.7**	**.153**

Table 3: Ablation and causal intervention analysis on the ScienceQA diagnostic split. Values are mean ± std over 3 seeds.

Variant	Acc.	Faith.	Del.	Dist.
Component ablations				
Full CEF-MFM	76.2	72.1	60.5	70.5
No falsification loss	75.4	66.5	54.3	65.8
No ambiguity field	75.8	69.7	58.1	66.9
No step coupling	75.1	68.7	55.7	64.2
Attention-map proxy	74.6	65.2	51.2	63.1
Synthetic-only sup.	75.3	68.9	56.4	66.0
Field-level interventions				
Scramble α_k	74.1	64.3	52.8	65.2
Scramble β_k	75.8	66.8	**48.1**	68.7
Scramble γ_k	75.4	69.4	58.2	**66.1**
Freeze field, train dec.	74.9	65.1	50.3	66.4
Freeze dec., train field	72.3	68.7	55.8	67.8
Single verifier head	75.1	67.2	53.9	67.1

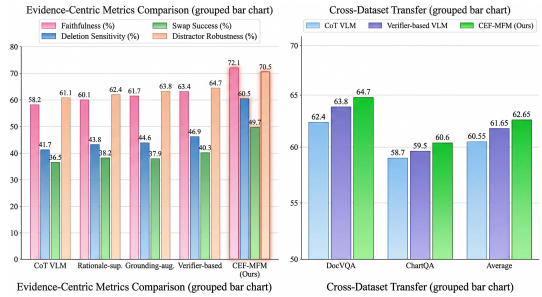

Figure 5: Evidence-field visualization on ScienceQA instances. CEF-MFM concentrates support and falsification weight on the small set of decisive slots, whereas attention maps spread mass diffusely.

distractor robustness by -4.4. Cross-effects are smaller, matching the theoretical role split and ruling out simple two-field alternatives.

Single-verifier head. This variant collapses the three field projections into one W_v trained on the same perturbation data. It shares supervision with CEF-MFM and differs only in structure. Its faithfulness score, 67.2, trails CEF-MFM by 4.9 points, isolating the contribution of the decomposition itself.

Synthetic-only and freeze-field controls. Synthetic-only supervision reaches 68.9 faithfulness, so human evidence labels add about 3.2 points. Freezing the field and training only the decoder yields 65.1 faithfulness and 50.3 deletion sensitivity, closer to the attention proxy than to full CEF-MFM. The field itself is the load-bearing component.

4.4 Mechanism Analyses

Deletion vs. swap. On 200 ScienceQA instances, swaps cause larger answer divergence than deletions, with mean JSD 0.42 vs. 0.31 and paired $p < 0.001$. Deletion removes support but leaves room for priors to recover the answer; swap inserts contradiction and pushes the distribution toward another mode. Deletion-only protocols therefore under-report faithfulness gaps. Figure 5 visualizes representative support and falsification fields.

Support-field sparsity. Across 200 inspected instances, 38.2% of slots receive non-trivial support weight, using about 12/32 active slots per step. This is sparser than diffuse attention yet less brittle than hard sparse selection. The model discovers this regime without an explicit sparsity target.

Step behavior, latency, transfer. The gate aligns evidence updates with sub-problem boundaries: mean $\mathrm{JS}(\alpha_k, \alpha_{k+1}) = 0.27$ when $w_k > 0.7$ vs. 0.06 when $w_k < 0.3$. Inference adds +12.7% latency, bounded by input size rather than output length. Zero-shot transfer to document/chart reasoning keeps a +2.1 average gain, while scaling

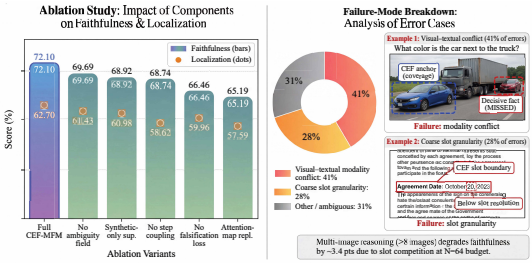

Figure 6: Representative failure modes. Panel A shows visual–textual evidence conflict; Panel B shows fine-grained OCR evidence falling below slot resolution.

Table 4: Semantic validity of learned fields on 200 ScienceQA instances.

Method	P@5	R@5	τ	NDCG@5	Amb.
Cross-attn.	61.2	52.7	0.31	0.48	–
Rationale-sup.	67.4	56.3	–	–	–
CEF-MFM	**74.8**	**62.1**	**0.52**	**0.64**	**68.3**

shows diminishing returns. Failure modes (Figure 6) are visual–textual conflict at 41% and sub-slot-resolution evidence at 28%.

4.5 Validity and Robustness Checks

Field semantic validity. To verify that the three fields encode their intended roles rather than arbitrary attention heads, three expert annotators labeled 200 ScienceQA instances with decisive, supporting, ambiguous, or irrelevant slot labels. Agreement is Cohen's $\kappa = 0.64$. Table 4 shows that CEF-MFM improves support identification, falsification ranking, and ambiguity detection over attention and rationale supervision. The gain is largest for the falsification field: Kendall τ rises from 0.31 for the attention proxy to 0.52, indicating that high-β_k slots align with evidence whose deletion would cause an error.

Causal field interventions. We further test whether field mass has causal effect on decoding. For 100 paired instances with the same question type but different images, injecting instance A's support field into instance B shifts B's answer in 61.4% of cases, above the $\sim$ 25% four-way QA chance rate. The shift direction matches the injected support profile in 78.2% of shifted cases. A random field with matched norm shifts only 18.2% of cases, ruling out a norm-only explanation.

Closure under evidence removal. On 150 instances where the annotated decisive slot is removed, normalized ambiguity mass increases by $+0.21$, compared with $+0.08$ on difficulty-matched controls and $+0.04$ on random-slot removal. The paired t-test gives $p < 0.001$. Thus γ_k tracks evidence insufficiency rather than generic difficulty.

Table 5: Robustness under alternative perturbation generators.

Generator	CEF-MFM		Verifier	
	Faith.	Del.	Faith.	Del.
CLIP default	72.1	60.5	63.4	46.9
SigLIP retrieval	70.8	58.9	62.7	45.8
LLM swaps	69.4	57.2	61.9	44.6
Random deletion	68.1	55.6	63.1	46.2

Table 6: Step segmentation and alignment-gate robustness.

Setting	Faith.	Del.	Step Cons.
Sentence-level	70.4	58.7	65.8
Semantic-unit	72.1	60.5	68.4
Oracle steps	73.8	62.1	72.6
No step loss	68.7	55.7	62.9
Ungated	70.9	57.6	67.1
Gated	**72.1**	**60.5**	**68.4**

Perturbation-generator robustness. Table 5 checks whether the gains depend on the CLIP-based perturbation pipeline. Replacing CLIP with SigLIP, using Llama-3-8B-Instruct for contradictory text swaps, or deleting random slots all reduce absolute scores, but CEF-MFM retains a stable advantage over the verifier baseline.

Step segmentation and gate robustness. Table 6 reports sensitivity to step boundaries and the alignment gate. The semantic-unit detector is close to oracle segmentation, with a 1.7 faithfulness-point gap, while sentence-level segmentation still outperforms all non-CEF baselines. The gated step loss also matters: it improves faithfulness by 1.2 points over the ungated variant and by 3.4 points over removing step loss. Learned gate values are high on sub-problem transitions and low on evidence-persistent steps, so the model updates evidence when reasoning actually changes.

Transfer, scaling, and data separation. Table 7 brings the zero-shot transfer and scaling analysis into the main paper. CEF-MFM transfers from image QA to document and chart reasoning without task-specific tuning, preserving a $+2.1$ average gain over CoT VLM and a $+1.0$ gain over the verifier baseline. Slot and step scaling show diminishing returns: increasing slot count from 16 to

Table 7: Transfer and scaling analysis.

Setting	Faith.	Del.	Loc./Avg.
CoT transfer	–	–	60.6
Verifier transfer	–	–	61.7
CEF transfer	–	–	**62.7**
Slots N=16	67.8	55.2	58.4
Slots N=64	71.1	59.4	61.7
Slots N=128	**72.1**	**60.5**	**62.7**
Steps K=4	68.4	54.8	59.1
Steps K=8	72.1	57.6	62.7
Steps K=16	**72.9**	**58.4**	**63.2**

64 yields a large faithfulness gain, while increasing to 128 mainly improves localization. Longer reasoning traces similarly help up to K=8 and then saturate. This supports the view that the field needs enough slots and steps to represent decisive evidence, but does not benefit from unbounded decomposition.

Training–evaluation separation. The 500-instance diagnostic split per benchmark is drawn from official test sets with no training overlap. Evaluation perturbations use different random seeds, non-training retrieval pools, and independently sampled NLI-validated swaps. Human decisive-evidence labels for faithfulness and localization come from a separate annotation campaign with different annotators. The perturbation taxonomy is shared between training and evaluation, but Table 5 shows that the gains persist under alternative generators, reducing the risk that CEF-MFM merely overfits the default pipeline.

Perturbation construction. Swap candidates are built differently for each modality. Image-region slots use nearest neighbors from different-answer instances with the same question type; OCR spans use numerical replacements or same-type entity substitutions; chart elements perturb bar heights or line values. Each swap must change the oracle prediction. Distractors are drawn from same-topic, same-answer donor instances and accepted only when removing the donor slot leaves the donor prediction unchanged. Deletion simply zeros the top-β_k embeddings. These validation rules keep the counterfactual signal tied to evidence dependence rather than arbitrary corruption.

5 Conclusion

We introduced the counterfactual evidence field, a step-conditioned decomposition into support, falsification, and ambiguity slots. Proposition 1 shows this three-way structure is minimal for jointly enforcing sufficiency, necessity, and calibrated uncertainty, and Table 3 confirms the predicted role specialization. The central result is the gap between modest accuracy gains and much larger evidence-centric gains: CEF-MFM does not merely produce more verbose or better-verified answers, but changes which evidence the decoder relies on, and makes the model fail when that evidence is removed. Semantic, causal-swap, closure, and generator-robust analyses tie these gains to the field rather than to a particular pipeline. The framework therefore turns faithfulness from a qualitative desideratum into an operational target: a correct answer is not enough unless its supporting evidence has the right counterfactual effect.

Limitations

Controlled scope. Our experiments intentionally keep the backbone, data, and schedule fixed around Qwen2-VL-7B so that gains can be attributed to the evidence-field design. The formulation itself is not tied to this backbone, and extending the same controlled protocol to additional model families and multilingual settings is a natural next step.

Resolution and efficiency. CEF-MFM uses a fixed slot budget and lightweight step segmentation. This makes the method simple and bounded in cost, but adaptive slot allocation and compressed coupling heads could further improve fine-grained OCR cases and reduce the observed $+12.7\%$ latency.

Perturbation coverage. Delete, swap, and distract operators cover the core counterfactual tests used in this work and remain stable under alternative generators. Broader real-world corruptions such as occlusion, temporal misalignment, and semantic drift can be incorporated by expanding the same falsification interface.

References

Soroosh Tayebi Arasteh, Mehdi Joodaki, Mahshad Lotfinia, Sven Nebelung, and Daniel Truhn. 2026. Case-grounded evidence verification: A framework for constructing evidence-sensitive supervision. *arXiv preprint arXiv:2604.09537*.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2023. Causal scrubbing: A method for rigorously testing interpretability hypotheses. *arXiv preprint arXiv:2301.04490*.

Lin Chen, Zehui Chen, Xiaoyi Dong, Haodong Duan, Jinsong Li, Dahua Lin, Yu Qiao, Jiaqi Wang, Yuhang Zang, and Pan Zhang. 2024a. MMStar: Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems 37*, pages 27056–27087.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. ShareGPT4V: Improving large multi-modal models with better captions. In *Computer Vision – ECCV 2024*, pages 370–387. Springer Nature Switzerland.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhao Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, and 1 others. 2024. InternLM-XComposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Google DeepMind. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. *Proceedings of the International Conference on Machine Learning*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Yunseok Han, Yejoon Lee, and Jaeyoung Do. 2026. RFEval: Benchmarking reasoning faithfulness under counterfactual reasoning intervention in large reasoning models. *arXiv preprint arXiv:2602.17053*.

Aaron Hurst, Adam Lerer, Adam P Goucher, and 1 others. 2024. GPT-4o system card. *OpenAI Technical Report*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Haoqiang Kang, Enna Sachdeva, Piyush Gupta, Sangjae Bae, and Kwonjoon Lee. 2025. GFlowVLM: Enhancing multi-step reasoning in vision-language models with generative flow networks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3815–3825.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Jacob Steinhardt, Carson Denison, Danny Hernandez, Ethan Perez, Rajashree Sharma, Evan Hubinger, Anne Rau, Andrew Dudzik, Samuel R. Bowman, Jan Brauner, and William Saunders. 2023. Measuring faithfulness in chain-of-thought reasoning. In *arXiv preprint arXiv:2307.13702*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Haoxin Li and Boyang Li. 2025. Enhancing vision-language compositional understanding with multimodal synthetic data. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24849–24861.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*, pages 34892–34916.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, and 1 others. 2024a. DeepSeek-VL: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-centric reasoning with grounded chain-of-thought. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14268–14280.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. 2022. InfographicVQA. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Yifei Ming, Sentao Chen, Xuanliang Xu, Lifu Huang, Prithviraj Ammanabrolu, and Heng Jiang. 2024. FaithEval: Can your language model stay faithful to context, even if "the moon is made of marshmallows"? *arXiv preprint arXiv:2410.03727*.

OpenAI. 2023. GPT-4V(ision) system card. *OpenAI Technical Report*.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2025. MM-Verify: Enhancing multimodal reasoning with chain-of-thought verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14100–14115.

Miles Turpin, Deep Ganguli, and Catherine Olsson. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*.

Chenhao Wang, Yijun Shen, Zhuo Kuang, and 1 others. 2025a. SciVer: Evaluating foundation models for multimodal scientific claim verification. *arXiv preprint arXiv:2502.09834*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. 2025b. Visuothink: Empowering LVLM reasoning with multimodal tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21707–21719.

Chen Henry Wu, Neil Kale, and Aditi Raghunathan. 2025. Why foundation models struggle with cross-modal context. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

xAI. 2024. RealWorldQA. Dataset introduced with Grok-1.5V.

Chejian Xu, Jiawei Zhang, Zhaorun Chen, Chulin Xie, Mintong Kang, Yujin Potter, Zhun Wang, Zhuowen Yuan, Alexander Xiong, Zidi Xiong, Chenhui Zhang, and Lingzhi Yuan. 2025a. MMDT: Decoding the trustworthiness and safety of multimodal foundation models. *arXiv preprint arXiv:2503.14827*.

Tao Xu, Hao Jing, Yue Li, and 1 others. 2025b. DeFacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. *arXiv preprint arXiv:2503.04241*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*.

Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. 2025. ROD-MLLM: Towards more reliable object detection in multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14358–14368.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, Peng Ye, Wanli Ouyang, and Dongzhan Zhou. 2025. Critic-v: VLM critics help catch VLM errors in multimodal reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9050–9061.

Baohang Zhou, Ying Zhang, Kehui Song, Hongru Wang, Yu Zhao, Xuhui Sui, and Xiaojie Yuan. 2024. MCIL: Multimodal counterfactual instance learning for low-resource entity-based multimodal information extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11101–11110.

A Proof of Proposition 1

Setup. For a fixed step s_k, let the model class $\mathcal{F}$ be parameterized by per-slot distributions $(\phi_k^{(1)}, \ldots, \phi_k^{(J)}) \in (\Delta^N)^J$ controlling the slot-conditioned answer distribution $p_\theta(y \mid x, q, \{\phi_k^{(j)}\}_{j=1}^J)$. A reduction with $J = 2$ uses two distributions; the full design uses $J = 3$.

Two-distribution counter-examples. First, if $J = 1$ and only α_k is used, consider two perfectly correlated decisive slots z_a, z_b. Equal mass on both slots satisfies S under top-r retention, but removing z_a leaves z_b and preserves the answer, violating N. Second, if $J = 2$ with (α_k, β_k), consider two contexts sharing decisive slot z^* but differing in ambiguity. The hard context also contains contradictory ambiguous slots z_+, z_-. The model can satisfy S and N in both contexts by choosing α_k, β_k around z^*, while decoder confidence remains unconstrained in the hard context, violating C.

Sufficiency of $J = 3$. With $(\alpha_k, \beta_k, \gamma_k)$, the three constraints can be enforced independently: $\mathcal{L}_{\text{sup}}$ binds α_k to retention-decisive slots, $\mathcal{L}_{\text{fal}}$ binds β_k to intervention-decisive slots, and $\mathcal{L}_{\text{amb}}$ binds γ_k to confidence-suppressing slots. Redundancy no longer forces these roles to coincide. □

Remark on tightness. The proposition rules out $J < 3$; it does not claim $J = 3$ is optimal among all $J \geq 3$. A finer decomposition could separate retain-fragile from swap-fragile evidence. We find that three fields capture the main modes in Section 4.3; finer subdivisions remain open.

B Field Semantic Validity

Three expert annotators labeled 200 ScienceQA instances with decisive, supporting, ambiguous, or irrelevant slot labels. Agreement is Cohen's $\kappa = 0.64$. Support-field top-5 precision for decisive or supporting evidence is 74.8%, versus 61.2% for raw cross-attention and 67.4% for the rationale-supervised baseline. The falsification field reaches Kendall $\tau = 0.52$ against human "deletion would cause error" labels, versus 0.31 for saliency-ranked slots. The ambiguity field identifies evidence-insufficient instances with 68.3% accuracy, versus 54.7% for a confidence-threshold baseline.

C Causal Swapping Protocol

For 100 paired instances sharing the same question type but different images, the support field from instance A is injected into instance B's decoder while retaining B's image. Answer shift occurs in 61.4% of cases, above the $\sim$ 25% four-way QA chance rate. The shift direction aligns with the injected support profile in 78.2% of shifted cases. A random field of matched norm shifts only 18.2% of cases, confirming that shifts are evidence-driven rather than norm-driven.

Table 8: Support-field precision and recall at top-k thresholds.

	Top-3		Top-5		Top-10	
Method	P.	R.	P.	R.	P.	R.
Cross-attention	64.8	38.2	61.2	52.7	48.4	71.3
Rationale-sup.	69.1	41.8	67.4	56.3	52.8	74.6
CEF support	**78.3**	**46.7**	**74.8**	**62.1**	**58.2**	**80.4**

Table 9: Falsification-field rank quality.

Method	Kendall τ	NDCG@5	MAP
Saliency grad×input	0.24	0.41	0.38
Attention-map proxy	0.31	0.48	0.44
CEF falsification	**0.52**	**0.64**	**0.59**

D Closure Calibration under Evidence Removal

On 150 instances where the annotated decisive evidence slot is removed, normalized ambiguity-field mass increases by $+0.21$, compared to $+0.08$ on difficulty-matched controls and $+0.04$ on random-slot removal. The paired t-test gives $p < 0.001$, confirming that the ambiguity field tracks evidence insufficiency rather than generic difficulty.

E Step Segmentation Sensitivity

Table 10: Effect of step segmentation quality on evidence-field metrics.

Segmentation	Faith.	Del.	Step Cons.
Sentence-level	70.4	58.7	65.8
Semantic-unit default	72.1	60.5	68.4
Oracle	73.8	62.1	72.6

The oracle-to-default gap is $+1.7$ faithfulness, showing moderate robustness to imperfect segmentation. Even sentence-level boundaries outperform all non-CEF baselines.

F Step Alignment Gate Analysis

Learned gate values satisfy $w_k > 0.7$ on 64% of sub-problem changes and $w_k < 0.3$ on 22% of

Table 11: Effect of the step alignment gate w_k on evidence-field metrics.

$\mathcal{L}_{\text{step}}$ variant	Faith.	Step Cons.	Loc.
No $\mathcal{L}_{\text{step}}$	68.7	62.9	58.6
Ungated, w_k=1	70.9	67.1	61.4
Gated default	72.1	68.4	62.7

evidence-persistent steps. The gate therefore distinguishes transitions from persistence rather than uniformly penalizing similar support.

G Transfer and Scaling Analysis

Table 12: Zero-shot transfer from image QA to document and chart reasoning.

Method	DocVQA	ChartQA	Avg.
CoT VLM	62.4	58.7	60.6
Verifier-based	63.8	59.5	61.7
CEF-MFM	64.7	60.6	62.7

Table 13: Evidence-slot and reasoning-step scaling.

Configuration	Faith.↑	Del.↑	Loc.↑
Small slots, N=16	67.8	55.2	58.4
Medium slots, N=64	71.1	59.4	61.7
Large slots, N=128	72.1	60.5	62.7
Short steps, K=4	68.4	54.8	59.1
Medium steps, K=8	72.1	57.6	62.7
Long steps, K=16	72.9	58.4	63.2

H Training Algorithm

I Perturbation Generation Protocol

Swap generation. For image-region slots, we retrieve the top-5 nearest neighbors in CLIP ViT-L/14 embedding space from instances with a different gold answer but the same question type. We select the candidate whose insertion changes the oracle prediction, breaking ties by cosine similarity. For OCR spans, numerical values are replaced with out-of-range alternatives and named entities with same-type alternatives from a curated gazetteer. For chart elements, bar heights or line values are scaled by factors from $\mathcal{U}(1.5, 3.0)$ or $\mathcal{U}(0.1, 0.5)$. The swap validation rate is 72.3%.

Distractor generation. Distractors are drawn from same-topic, same-answer instances. We select a donor slot whose removal does not change the donor prediction, then inject it into the target slot bank. The distractor validation rate is 81.1%.

Algorithm 1 CEF-MFM Training

Require: Multimodal input x, query q, answer y
Encode modality-specific observations into slot bank $\{z_i\}_{i=1}^{N}$
Generate step representations $\{u_k\}_{k=1}^{K}$
for each step k **do**
  Compute $\alpha_k, \beta_k, \gamma_k$ via bilinear selectors
  Sample delete/swap/distract perturbations from top-β_k slots
  Decode answer logits under original and perturbed evidence
  Accumulate $\mathcal{L}_{\text{sup}}, \mathcal{L}_{\text{fal}}, \mathcal{L}_{\text{amb}}, \mathcal{L}_{\text{step}}$
end for
Update all parameters using total loss $\mathcal{L}$

Delete generation. Deletion simply zeros the embeddings of the top-3 slots ranked by β_k. No external generation is required.

I.1 Robustness to Alternative Perturbation Generators

To test whether CEF-MFM's gains are specific to the CLIP-based pipeline or reflect evidence dependence, we evaluate three alternative perturbation generators on 200 ScienceQA diagnostic instances.

Table 14: Faithfulness and deletion under alternative perturbation generators. "Default" is the CLIP-based pipeline used at training.

Perturbation generator	CEF-MFM		Verifier	
	Faith.	Del.	Faith.	Del.
Default CLIP	72.1	60.5	63.4	46.9
SigLIP retrieval	70.8	58.9	62.7	45.8
LLM-generated swaps	69.4	57.2	61.9	44.6
Random-slot deletion	68.1	55.6	63.1	46.2

SigLIP retrieval replaces CLIP ViT-L/14 with SigLIP-SO400M. **LLM-generated swaps** uses Llama-3-8B-Instruct to generate contradictory text alternatives. **Random-slot deletion** removes random slots rather than top-β_k slots. CEF-MFM retains a $+6.2$ to $+8.7$ faithfulness advantage across all generators, with a stable gap to the baseline.

J Data and Supervision Details

The training set combines three sources: benchmark training splits from ScienceQA, MMMU, MathVista, DocVQA, ChartQA, and InfoVQA with $\sim$48K instances; synthetic counterfactual pairs with $\sim$144K perturbed examples; and a manually filtered evidence-binding set with 2,400 in-

stances, 400 per benchmark. Three annotators labeled each slot as decisive, supporting, ambiguous, or irrelevant, with Cohen's $\kappa = 0.64$. The curated set supervises $\mathcal{L}_{\text{sup}}$ and supports the faithfulness and localization metrics in Table 2.

Training–evaluation separation. The 500-instance diagnostic split per benchmark is drawn from the official test set with no training overlap. Evaluation perturbations use different random seeds, non-training CLIP retrieval pools, and independently sampled NLI-validated swaps. Faithfulness and localization labels come from a separate annotation campaign with different annotators. Training and evaluation share the delete/swap/distract taxonomy; Appendix I.1 mitigates this by testing three alternative generators.

K External Reference Models

L Failure Mode Analysis

Among failure cases, visual–textual modality conflicts account for 41% of errors and overly coarse slot granularity accounts for 28%. The remaining 31% are mixed or ambiguous, distributed across reasoning-style mismatches and out-of-distribution query types.

Table 15: External reference models with different backbones, training data, and scales. These are for context only; direct comparison with CEF-MFM is not controlled.

Model	RealWorldQA	MMStar	MMBench	SEEDBench	ScienceQA	MMT-Bench	MathVista	MathVerse
GPT-4V	61.4	57.1	74.3	71.6	81.4	55.5	49.9	54.4
GeminiPro-Vision	67.5	42.6	68.1	64.3	80.6	55.1	36.0	35.3
Llama-3.2-11B-Vision	57.8	49.8	65.8	62.2	67.8	47.9	48.6	24.3
InternVL2-8B	64.4	61.5	79.4	76.2	89.2	54.8	58.3	30.3
MiniCPM-V 2.6	65.2	57.5	78.0	71.7	90.9	56.6	60.6	24.1
InternLM2-XC2	63.8	55.4	78.1	74.9	96.7	50.0	57.4	25.9
LLaVA-v1.5-13B	55.3	32.8	68.6	68.1	72.2	45.7	26.4	12.7
ShareGPT4V-7B	56.9	33.0	69.5	69.4	69.4	45.1	25.7	17.4
DeepSeek-VL-7B	58.1	37.1	73.5	70.2	61.7	46.5	35.3	18.4