EVIDENCE CONTRACTS: VERIFIABLE CONTROL OF EVIDENCE RELIANCE IN MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models can produce correct predictions while relying on evidence they are meant to ignore, or failing to rely on evidence they are asked to use. Existing evidence-control methods can alter routing patterns or reliance scores, but these changes do not establish the stronger property of *behavioral compliance*: whether a requested evidence-use relationship holds under admissible intervention mechanisms excluded from adaptation and model selection. We introduce **Evidence Contracts**, a *specify–enforce–verify* framework that compiles natural-language evidence-use instructions into explicit, behaviorally testable requirements. Evidence Contracts support compositional `IGNORE`, `REQUIRE`, `PREFER`, and `PRESERVE` clauses, respectively specifying invariance to designated evidence, task-consistent dependence on required evidence, relative dependence between evidence sources, and preservation of task performance. We ground each clause in multimodal inputs, enforce the resulting contract through contract-constrained adaptation, and verify compliance using validated intervention families held out from adaptation, model selection, and calibration. Across seven multimodal benchmarks, Evidence Contracts achieve 68.4% held-out contract satisfaction rate (CSR) on the common applicable benchmark–backbone settings, compared with 36.5% for GUIDE, while retaining 98.6% of original task utility. Compliance also transfers to unseen contract targets, operator–target pairings, and clause compositions, reaching 63.9% average CSR. These results show that evidence use can be specified, enforced, and independently verified as a behavioral commitment within declared, validated intervention distributions, rather than treated only as a steering preference.

1 INTRODUCTION

Multimodal models can produce the correct answer for the wrong reason. A vision–language model may recognize an action while relying on the surrounding scene rather than the action itself, or answer a question from textual cues while largely ignoring the visual evidence that should support its prediction. Such behavior is invisible to task accuracy alone: two models can produce the same output while depending on substantially different evidence. As multimodal models are deployed in settings where particular evidence should be trusted, ignored, or prioritized, controlling the output is not enough. We must also control *how a prediction is allowed to depend on evidence*. Recent methods make evidence reliance controllable. FiLoRA (Chung et al., 2026) uses language-conditioned parameter-efficient adaptation to increase or suppress reliance on predefined feature groups, while GUIDE (Han et al., 2026) dynamically modulates multimodal evidence pathways during reasoning and generation. These methods establish an important capability: a model's evidence-use behavior can be steered in response to an instruction. However, *controllability does not imply compliance*. If an instruction asks a model to "ignore the background," a change in an internal gate or reliance score does not establish that the prediction is invariant to admissible background changes. Likewise, increasing the apparent contribution of action evidence does not establish that the prediction functionally depends on that evidence. A model may appear to follow an instruction under familiar transformations while continuing to use prohibited evidence under unseen ones. This gap is related to work on appropriate feature dependence, counterfactual invariance, verification, and faithful reasoning. Existing approaches train models to avoid annotated irrelevant features (Ross et al., 2017), remain invariant under spurious counterfactual changes (Veitch et al., 2021), certify robustness to

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.

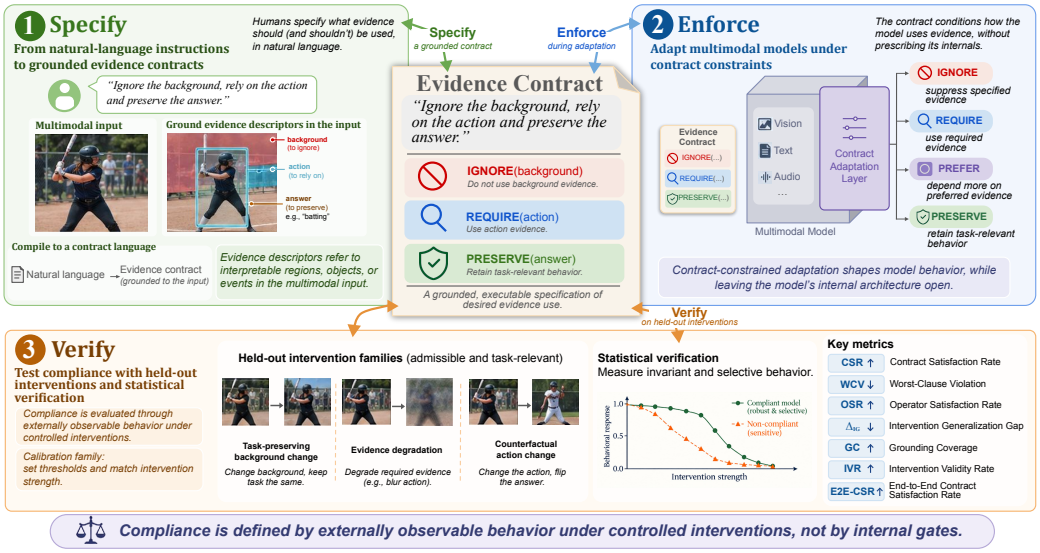

Figure 1: **Evidence Contracts: specify, enforce, and verify.** A user-specified evidence-use instruction is compiled into grounded behavioral clauses and enforced using adaptation-family interventions. Thresholds are calibrated using a disjoint intervention family, and compliance is independently verified using held-out intervention families. Compliance is defined by the model's behavioral response to interventions rather than by an internal routing or reliance score.

semantic variation (Yang et al., 2026), or ground reasoning in task-relevant and causally motivated evidence (Tian et al., 2026; Huynh et al., 2026; Kancheti et al., 2026; Li et al., 2026). These directions typically target a fixed robustness property, a task-defined notion of faithfulness, or externally supplied causal structure. They do not directly address the complementary question: *can a user specify how a multimodal prediction should depend on evidence, and can that requirement be both enforced and independently tested?*

We introduce **Evidence Contracts**, a *specify–enforce–verify* framework that treats user-specified evidence-use instructions as explicit behavioral requirements rather than soft steering preferences. A contract specifies how a model's prediction should respond to admissible interventions on designated evidence: `IGNORE(background)`, for example, requires invariance to task-preserving background changes, whereas `REQUIRE(action)` requires task-consistent dependence on action evidence under a designated admissible intervention. Compliance is therefore defined through observable model behavior, rather than through a particular attention, gate, or routing value. We instantiate Evidence Contracts with four compositional operators: `IGNORE`, `REQUIRE`, `PREFER`, and `PRESERVE`, covering invariance, functional dependence, relative dependence, and task preservation. We enforce these requirements through **contract-constrained adaptation**, an architecture-agnostic objective that jointly optimizes task performance and contract compliance. Crucially, enforcement and verification are separated: contracts are optimized using adaptation-family interventions, but evaluated using *held-out intervention families* excluded from adaptation, model selection, and calibration. This tests whether a model has learned the intended evidence-use relationship rather than merely overfitting to a familiar intervention mechanism. We evaluate Evidence Contracts across seven benchmarks spanning controlled counterfactual evaluation, multimodal reasoning, audiovisual understanding, classification, and generation. Evidence Contracts achieve 68.4% held-out CSR, compared with 36.5% for GUIDE, while retaining 98.6% of original task utility; compliance further transfers to unseen contract targets, operator–target pairings, and clause compositions. Our experiments therefore test whether evidence requirements can be compiled into testable constraints, enforced while preserving task utility, and generalized both across held-out intervention families and beyond seen contract specifications. **Our contributions** are: (1) a grounded, compositional contract language for specifying behavioral evidence-use requirements; (2) contract-constrained adaptation for enforcing these requirements while preserving task-relevant behavior; and (3) an independent verification protocol that evaluates compliance using validated intervention families excluded from adaptation, model selection, and calibration, across seven multimodal benchmarks, multiple backbones, and alternative adaptation strategies. This moves multimodal evidence control from asking ***"Can we steer the model toward the evidence we want?"*** to the stronger question: ***"Can we specify, enforce, and verify how a model's predictions depend on evidence?"***

2 Related Work[1]

Controlling Evidence Reliance in Multimodal Models. Recent work moves beyond controlling model outputs toward controlling the evidence on which multimodal predictions depend. FiLoRA (Chung et al., 2026) uses language-conditioned adaptation to alter reliance on task-relevant and shortcut-associated feature groups, while GUIDE (Han et al., 2026) uses instructions to modulate multimodal evidence pathways during reasoning and generation. However, a change in a reliance score or routing pattern does not establish that the requested behavioral relationship holds under admissible intervention mechanisms excluded from adaptation and model selection. Evidence Contracts therefore treats evidence-use instructions as behavioral requirements and verifies their compliance using held-out intervention families.

Invariance, Specification, and Verification. Right for the Right Reasons (Ross et al., 2017) constrains differentiable explanations, while Counterfactual Invariance (Veitch et al., 2021) studies invariance to spurious correlations under counterfactual changes. Related work also certifies semantic robustness for vision–language models (Yang et al., 2026) and translates high-level semantic properties into verifiable specifications (Elboher et al., 2026). These directions address fixed invariance or robustness properties. In contrast, Evidence Contracts specifies *user-defined evidence-use policies*, including which evidence should be ignored, required, or preferred, and evaluates them through grounded behavioral interventions.

Faithful Multimodal Reasoning. Faithful multimodal reasoning methods encourage predictions to be grounded in task-relevant or causally motivated evidence. Faithful-MR1 (Tian et al., 2026) anchors and reinforces visual attention, FACR (Huynh et al., 2026) uses action-unit causal reasoning for counterfactually faithful emotion explanations, and recent work studies constrained optimization for spatial reasoning (Kancheti et al., 2026) and faithfulness in multimodal reasoning, planning, and acting (Li et al., 2026). These approaches typically enforce or evaluate a task-defined notion of faithfulness based on task supervision or supplied causal structure. Evidence Contracts instead allows the desired evidence-use behavior to vary across user-specified contracts, provided that each remains compatible with the underlying task semantics.

Our Perspective. Prior work separately studies evidence control, fixed robustness or verification properties, and task-defined multimodal faithfulness. Evidence Contracts connects these concerns through a *user-specified, compositional evidence-use policy* whose compliance is enforced during adaptation and independently tested using held-out intervention families.

3 Method

Evidence Contracts turn natural-language evidence-use instructions into behavioral requirements enforced during adaptation and independently verified afterward. The framework *specifies* a structured contract, grounds its evidence descriptors, *enforces* the resulting constraints, and *verifies* compliance only on intervention families excluded from adaptation, model selection, and calibration. Unlike approaches that assess evidence use through internal gates, attention weights, or routing values, we define compliance through externally observable responses to controlled interventions.

3.1 Evidence Contracts and Grounded Intervention Families

Let f_θ map a multimodal input $x \in \mathcal{X}$ to a predictive distribution $p_\theta(\cdot \mid x)$. We use e to denote a natural-language evidence descriptor, such as an object, action, background, OCR text, acoustic cue, or temporal event. An *Evidence Contract* is a conjunction of clauses: $\mathcal{C} = c_1 \wedge c_2 \wedge \cdots \wedge c_K$, where

$$c ::= \texttt{IGNORE}(e) \mid \texttt{REQUIRE}(e) \mid \texttt{PREFER}(e_a, e_b) \mid \texttt{PRESERVE}(y). \tag{1}$$

The operators encode negative dependence, positive dependence, relative dependence, and task preservation, respectively. A frozen compiler g_ψ maps an instruction I into its canonical representation:

$$g_\psi(I) = \mathcal{C}. \tag{2}$$

For example, "Ignore the scenery, rely on what the person is doing, and preserve the task answer" becomes

$$\texttt{IGNORE(background)} \wedge \texttt{REQUIRE(action)} \wedge \texttt{PRESERVE(answer)}. \tag{3}$$

Malformed or unresolved compiler outputs count as failures in end-to-end evaluation. We report compiler fidelity separately from grounding and behavioral compliance.

[1] We provide comparison table (Table 2) to prior approaches in Appendix A.

Grounding and admissibility. We ground each descriptor e to its input support:

$$(M_e, q_e) = G(x, e), \tag{4}$$

where M_e is an image region, text span, temporal segment, acoustic interval, or other structured support, and $q_e \in [0, 1]$ is grounding confidence. The frozen grounder G is external to the adapted model. Groundings below a fixed threshold $q_{\min}$ are unresolved and count as failures in end-to-end evaluation. Given grounded evidence e, an intervention is $\tilde{x} = T_{e,\xi}(x)$. We use task-preserving interventions $\mathcal{A}_e^0$, for which $y_T = y$; evidence-degrading interventions $\mathcal{A}_e^-$, which reduce access to e while preserving y; and diagnostic counterfactual interventions $\mathcal{A}_e^1$, for which $y_T \neq y$ is known. An intervention is admissible only when it changes the designated evidence, preserves non-target content within a fixed tolerance, and has valid task semantics:

$$\mathrm{Valid}(T, x, e) = V_{\text{target}} \wedge V_{\text{non-target}} \wedge V_{\text{task}}. \tag{5}$$

Only interventions with $\mathrm{Valid}(T, x, e) = 1$ are used. Dataset-specific grounders, validators, thresholds, and intervention protocols are provided in the Appendix B.

Mechanism-level family separation. For each evidence source and intervention type, we construct disjoint collections of intervention *families* $\mathfrak{F}_e^{\text{train}}$, $\mathfrak{F}_e^{\text{cal}}$, and $\mathfrak{F}_e^{\text{ver}}$. These collections differ in their intervention-generating mechanisms, rather than merely containing disjoint instances from a common transformation procedure. Let

$$\mathcal{A}_e^r = \bigcup_{\mathcal{F} \in \mathfrak{F}_e^r} \mathcal{F}, \qquad r \in \{\text{train}, \text{cal}, \text{ver}\}. \tag{6}$$

Adaptation uses $\mathcal{A}_e^{\text{train}}$; $\mathcal{A}_e^{\text{cal}}$ fixes contract thresholds and strength-matching rules; and $\mathcal{A}_e^{\text{ver}}$ remains untouched until final verification. For example, background masking and blurring may be used during adaptation, controlled replacement during calibration, and scene substitution or contextual swaps only for verification.

3.2 Behavioral Contract Semantics

For intervention T, let

$$R_\theta(x, T) = D(p_\theta(\cdot \mid x), p_\theta(\cdot \mid T(x))), \tag{7}$$

where $D \in [0, 1]$ is a bounded behavioral divergence. The dependence on evidence e under family $\mathcal{A}_e$ is

$$S_\theta(e; x, \mathcal{A}_e) = \mathbb{E}_{T \sim \mathcal{A}_e}\left[R_\theta(x, T)\right]. \tag{8}$$

We also use a bounded target-support function $Q_\theta(x, y) \in [0, 1]$, instantiated by target probability or a task-appropriate differentiable surrogate. These quantities define operational behavioral dependence within declared intervention spaces; they do not claim to identify a unique causal mechanism.

`IGNORE`: negative dependence.

$$S_\theta(e; x, \mathcal{A}_e^0) \leq \epsilon_{\text{ign}}, \qquad v_{\text{ign}}(x, e) = \left[S_\theta(e; x, \mathcal{A}_e^0) - \epsilon_{\text{ign}}\right]_+. \tag{9}$$

`REQUIRE`: positive dependence. We measure necessity using evidence-degrading interventions:

$$N_\theta(e; x) = \mathbb{E}_{T \sim \mathcal{A}_e^-}\left[\left[Q_\theta(x, y) - Q_\theta(T(x), y)\right]_+\right]. \tag{10}$$

The requirement is $N_\theta(e; x) \geq \delta_{\text{req}}$. When a valid counterfactual target is available, we additionally require

$$C_\theta(e; x) = \mathbb{E}_{T \sim \mathcal{A}_e^1}\left[Q_\theta(T(x), y_T)\right] \geq \tau_{\text{cf}}. \tag{11}$$

The violation is

$$\begin{aligned} v_{\text{req}}(x, e) = {} & \left[\delta_{\text{req}} - N_\theta(e; x)\right]_+ \\ & + \mathbf{1}_{\text{cf}}\left[\tau_{\text{cf}} - C_\theta(e; x)\right]_+, \end{aligned} \tag{12}$$

where $\mathbf{1}_{\text{cf}} = 1$ only when a valid counterfactual target is available.

`PREFER`: relative dependence. We use a frozen localized evidence-change score over M_e to measure intervention strength. The calibration families are used only to fix strength quantile boundaries and matching rules. During adaptation and final verification, dependence is computed using the corresponding adaptation or verification family, respectively, after matching interventions within these fixed strength quantiles. Let $\widetilde{S}_\theta(e; x)$ denote the resulting strength-matched dependence. We require

$$\widetilde{S}_\theta(e_a; x) - \widetilde{S}_\theta(e_b; x) \geq \gamma_{\text{pref}}, \tag{13}$$

with violation

$$v_{\text{pref}}(x, e_a, e_b) = \left[\gamma_{\text{pref}} - \left(\widetilde{S}_\theta(e_a; x) - \widetilde{S}_\theta(e_b; x)\right)\right]_+. \tag{14}$$

PRESERVE: task preservation. Let $\mathcal{A}_{\mathcal{C}}^{0}$ contain the task-preserving interventions relevant to $\mathcal{C}$, including the identity transformation. We require

$$P_\theta(x, y, \mathcal{C}) = \mathbb{E}_{T \sim \mathcal{A}_{\mathcal{C}}^{0}} \left[Q_\theta(T(x), y) \right] \geq \tau_{\text{pres}}, \tag{15}$$

with violation

$$v_{\text{pres}}(x, y) = \left[\tau_{\text{pres}} - P_\theta(x, y, \mathcal{C}) \right]_+ . \tag{16}$$

PRESERVE applies only to task-preserving interventions and therefore does not require the original answer to remain fixed under diagnostic counterfactuals.

A contract is satisfied only when every clause is satisfied:

$$\text{Sat}(x, \mathcal{C}) = \mathbf{1} \left[v_k(x) = 0 \quad \forall c_k \in \mathcal{C} \right]. \tag{17}$$

3.3 Contract-Constrained Adaptation

Let θ_0 be frozen pretrained parameters and ϕ trainable adaptation parameters. Given (x, y, I), we compile I into $\mathcal{C}$ and instantiate its clauses using $\mathcal{A}^{\text{train}}$. The framework is architecture-agnostic: ϕ may represent LoRA parameters or an explicit evidence-routing module. Because clause violations can have different empirical scales across operators, evidence types, and benchmarks, we apply *clause-wise normalization* only during adaptation. For each clause c_k, we assign a fixed calibration stratum $\omega(c_k)$ and estimate a positive scale $s_{\omega(c_k)}$ from calibration-family violations only. The scale is fixed before adaptation and is used only to normalize the adaptation objective; final verification uses the original behavioral quantities and thresholds. The normalized violation is

$$\bar{v}_k(x) = \frac{v_k(x)}{s_{\omega(c_k)} + \varepsilon}, \tag{18}$$

where $\varepsilon > 0$ ensures numerical stability. This prevents higher-scale clause types from dominating the adaptation objective solely because of their measurement range. The calibration strata, scale estimator, and unseen-contract back-off rule are specified in Appendix B.5. Let $\mathcal{O} = \{\text{ign}, \text{req}, \text{pref}, \text{pres}\}$ denote operator types and Z_o the corresponding distribution of normalized training-time clause violations $\bar{v}_k$ for operator o. We solve

$$\begin{aligned} \min_{\phi} \quad & \mathbb{E}_{(x,y)} \left[\ell_{\text{task}}\big(f_{\theta_0,\phi}(x), y\big) \right] \\ \text{s.t.} \quad & \text{CVaR}_\beta(Z_o) \leq \kappa_o, \qquad \forall o \in \mathcal{O}, \end{aligned} \tag{19}$$

where

$$\text{CVaR}_\beta(Z_o) = \min_{\zeta_o} \left[\zeta_o + \frac{1}{1-\beta} \mathbb{E}_{z \sim Z_o} \left[(z - \zeta_o)_+ \right] \right]. \tag{20}$$

We use an augmented primal–dual implementation, detailed in the Appendix B, which increases optimization pressure on operator types with persistently large tail violations without manually fixing static operator weights.

3.4 Independent Verification and Metrics

Final verification uses only $\mathcal{A}^{\text{ver}}$, after adaptation and calibration are complete. For contract $\mathcal{C}$, let $J(\mathcal{C})$ be the number of scalar statistics required for verification. For each statistic j, we estimate $\hat{\mu}_j \in [0, 1]$ from m_j i.i.d. held-out admissible interventions and use

$$r_j = \sqrt{\frac{\log(2J(\mathcal{C})/\alpha)}{2m_j}}. \tag{21}$$

Let $U_j(\hat{\mu}_j) = \min(1, \hat{\mu}_j + r_j)$ and $L_j(\hat{\mu}_j) = \max(0, \hat{\mu}_j - r_j)$. A clause is verified only when

$$\begin{aligned} \texttt{IGNORE}(e): \quad & U(\widehat{S}_\theta(e)) \leq \epsilon_{\text{ign}}, \\ \texttt{REQUIRE}(e): \quad & L(\widehat{N}_\theta(e)) \geq \delta_{\text{req}}, \\ \texttt{PREFER}(e_a, e_b): \quad & L(\widehat{\widetilde{S}}_\theta(e_a)) - U(\widehat{\widetilde{S}}_\theta(e_b)) \geq \gamma_{\text{pref}}, \\ \texttt{PRESERVE}(y): \quad & L(\widehat{P}_\theta) \geq \tau_{\text{pres}}. \end{aligned} \tag{22}$$

When a valid counterfactual is available, REQUIRE additionally requires $L(\widehat{C}_\theta(e)) \geq \tau_{\text{cf}}$. Conditioned on the frozen compiler, grounding procedure, admissibility validators, and calibration parameters, Hoeffding's inequality and a union bound guarantee that all confidence intervals for a contract simultaneously contain their expected behavioral quantities with probability at least $1 - \alpha$. Thus, a verified contract satisfies its expected constraints within the declared held-out intervention distributions; this is not formal certification beyond those distributions. The complete theorem and proof are provided in the Appendix B. For each clause, let ρ^{ver} denote its confidence-adjusted verification margin:

$$\rho_{\text{ign}} = \epsilon_{\text{ign}} - U(\widehat{S}), \qquad \rho_{\text{req}} = L(\widehat{N}) - \delta_{\text{req}},$$
$$\rho_{\text{pref}} = L(\widehat{\widetilde{S}}_\theta(e_a)) - U(\widehat{\widetilde{S}}_\theta(e_b)) - \gamma_{\text{pref}}, \qquad \rho_{\text{pres}} = L(\widehat{P}_\theta) - \tau_{\text{pres}}. \tag{23}$$

For REQUIRE clauses with a diagnostic counterfactual, the margin is the minimum of the necessity margin and $L(\widehat{C}_\theta(e)) - \tau_{\text{cf}}$. We report **Contract Satisfaction Rate** (CSR), the fraction of examples for which every clause has non-negative verification margin; and **End-to-End CSR** (E2E-CSR), which additionally counts malformed compiler outputs, unresolved groundings, and unavailable valid interventions as failures. We report **Operator Satisfaction Rate** (OSR), the clause-level satisfaction rate for each operator, to identify which requirements limit complete-contract performance. Finally, we report the **Intervention Generalization Gap**,

$$\Delta_{\text{IG}} = \text{CSR}_{\text{adaptation-family}} - \text{CSR}_{\text{held-out}}, \tag{24}$$

using fresh matched examples and changing only the intervention-generating family. The mean verification margin distinguishes robust satisfaction from clauses that narrowly pass a binary threshold. We report compliance metrics alongside task utility and retention relative to the unadapted model.

4 Experimental Setup

Tasks and datasets. We evaluate Evidence Contracts on seven benchmarks spanning counterfactual evaluation, visual question answering and reasoning, audiovisual understanding, multimodal classification, and image–text retrieval: Visual Reality Check (Visual R.C.), M^{2}CQA, GQA, TextVQA, EmoReAlM, MM-IMDb, and Flickr30K. Dataset sources, preprocessing, and split construction are in Appendix C.1.

Models and adaptation. We evaluate frozen backbones, prompt steering, task-only parameter-efficient adaptation, counterfactual augmentation, fixed-weight contract regularization, FiLoRA-style evidence control, GUIDE-style routing, and our contract-constrained adaptation. *Fixed-weight Contract (F Contract)* uses the same grounded clauses and adaptation-family interventions as Evidence Contracts, but minimizes a fixed globally tuned sum of clause violations rather than adaptive operator-wise tail-risk constraints. Unless a baseline requires a different architecture, methods are matched in backbone, trainable-parameter budget, task-training examples, minibatch schedule, and optimization steps. Inapplicable method–benchmark pairs are marked N/A and excluded from dataset averages; deviations are in Appendix B.8; backbones are provided in Appendix C.2.

Contract suite and generalization splits. For each benchmark, we instantiate atomic IGNORE, REQUIRE, PREFER, and PRESERVE clauses and compose compatible clauses into multi-clause contracts. Natural-language realizations are paraphrased independently of the evaluated model. Beyond observed contracts, we construct four disjoint generalization settings: unseen paraphrases, evidence targets, operator–target pairings, and clause compositions. These splits test reusable evidence-use specification rather than memorization of a fixed contract set.

Grounding and intervention protocol. For each benchmark, evidence descriptors are grounded to visual, textual, or auditory supports using task-appropriate annotations and validators. Adaptation, calibration, and verification use disjoint *intervention-generating mechanisms*, rather than random instance splits from one mechanism. Interventions are validated to ensure designated-evidence change, bounded non-target modification, and valid task semantics where required. Details are in Appendix B.2.

Evaluation protocol and metrics. Contracts are optimized using adaptation-family interventions. Calibration families are used only for checkpoint selection, hyperparameter choice, threshold calibration, and strength matching; verification families are excluded throughout. Our primary metric is held-out intervention-family CSR, which requires every clause in a contract to be satisfied. We additionally report E2E-CSR, OSR, Δ_{IG}, and mean verification margin; task-utility retention, WCV, GC, and IVR are in Appendix C.3. We report mean and standard deviation over three seeds, using $\alpha = 0.05$ and $m_j = 100$ held-out admissible interventions per scalar statistic.

5 RESULTS

5.1 EVIDENCE CONTROL DOES NOT GUARANTEE VERIFIED CONTRACT COMPLIANCE

We first test the premise underlying Evidence Contracts: evidence-use behavior observed under adaptation-family interventions does not by itself establish contract compliance under held-out intervention mechanisms. Table 1 therefore separates **Adapt.-Family CSR**, measured on fresh instances from families available during adaptation, from **H.O.-Family CSR**, measured using disjoint mechanisms excluded from parameter updates, model selection, threshold calibration, and strength matching. Their difference, Δ_{IG}, measures the loss under independent verification; a small gap is meaningful only together with high absolute held-out compliance. Prompting and LoRA provide limited held-out compliance, reaching only 17.9% and 23.6% H.O.-Family CSR. Counterfactual augmentation produces stronger adaptation-family performance (61.7% CSR), but loses 27.5 points under held-out verification. FiLoRA and GUIDE likewise achieve substantial adaptation-family control (58.4% and 63.1%) while retaining only 31.8% and 36.5% H.O.-Family CSR, respectively. Thus, control under familiar transformations does not reliably transfer to independently generated interventions. Direct contract optimization narrows this gap. F Contract improves H.O.-Family CSR to 49.6%, but still loses 19.9 points across evaluation families. Evidence Contracts achieve 68.4% H.O.-Family CSR with a 5.7-point gap, while retaining 74.1% Adapt.-Family CSR. Its positive mean verification margin (+0.11) further indicates that held-out clauses satisfy their confidence-adjusted constraints *on average* with slack, rather than narrowly passing a binary threshold. The base model's small gap reflects uniformly low compliance, not generalization; all adapted methods retain at least 98.6% of unadapted task retention. Overall, adaptation-family evidence control and verified contract compliance are distinct. By optimizing behavioral constraints while reserving distinct intervention mechanisms for verification, Evidence Contracts achieve high held-out compliance with a small generalization gap. Dataset-level results and seed variation for Evidence Contracts are provided in Appendix D.1; baseline applicability is reported in Appendix B.8.

Table 1: **Evidence control does not necessarily imply verified contract compliance. Task Ret.** denotes task-utility retention relative to the unadapted model. **Ada. CSR** and **H.O.-Family CSR** measure contract satisfaction under adaptation-family and disjoint verification-family interventions. Δ_{IG} denotes their difference (equation 24). **Ver. Margin** is the mean confidence-adjusted margin over held-out clauses; positive mean values indicate satisfaction on average. Results are macro-averaged across datasets using the common evaluation subset for which all methods are applicable.

	Method*	Task Ret. (%) ↑	Ada. CSR(%) ↑	H.O CSR(%) ↑	Δ_{IG}† ↓	Ver. Margin ↑
No explicit contract	Base	100.0	18.6	16.9	1.7	-0.24
	Prompt	99.4	24.7	17.9	6.8	-0.21
	LoRA	99.1	31.5	23.6	7.9	-0.16
	CF Aug.	98.9	61.7	34.2	27.5	-0.09
Explicit control	FiLoRA	99.0	58.4	31.8	26.6	-0.07
	GUIDE	98.8	63.1	36.5	26.6	-0.04
Explicit contract	F Contract	98.7	69.5	49.6	19.9	+0.01
	Evidence Contracts	98.6	**74.1**	**68.4**	**5.7**	**+0.11**

* Abbreviations: CF Aug., counterfactual augmentation; F Contract, fixed-weight contract optimization; Prompt, prompt-only control.

† A small Δ_{IG} is meaningful only together with high absolute compliance; the base model has a small gap because its CSR is low under both adaptation and held-out interventions.

5.2 COMPLIANCE GENERALIZES ACROSS HELD-OUT INTERVENTION FAMILIES

Table 1 establishes aggregate improvements under held-out verification. We next test whether those gains persist across distinct intervention-generating mechanisms, rather than being driven by a favorable subset of transformations. The held-out families comprise semantic scene substitution, localized object–attribute and relation–action counterfactuals, OCR replacement and re-rendering, and facial and acoustic evidence interventions. Each displayed family is excluded from parameter updates, model selection, threshold calibration, and strength matching. Figure 2 reports H.O.-Family CSR for each mechanism, macro-averaged over applicable datasets and contracts.

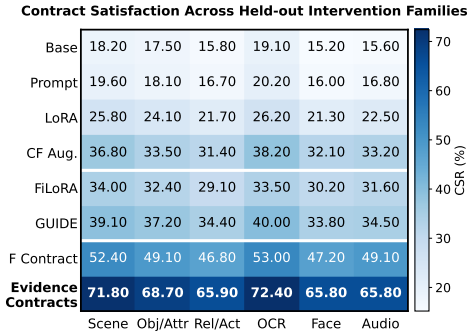

Figure 2: **Contract satisfaction across held-out intervention families.**

Scene denotes semantic scene substitution; Obj./Attr. denotes object or attribute replacement; Rel./Act. denotes relation or action counterfactuals; and OCR denotes text replacement or re-rendering. Full family-level results are in Appendix D.2 (Table 11). Evidence Contracts achieve the highest CSR among applicable methods for every held-out family, ranging from 65.8% to 72.4%. In contrast, the base, prompt-only, and LoRA baselines remain below 27% across families, and counterfactual augmentation remains below 39%. F Contract is stronger (46.8–53.0%), but Evidence Contracts improve on it in every family by 16.7–19.6 points. Thus, the aggregate result is not attributable to a favorable modality or transformation type: compliance generalizes to intervention mechanisms excluded from adaptation, rather than merely to new instances generated by a familiar mechanism. We next test the distinct question of generalization to unseen contract specifications.

5.3 Generalization to Unseen Contracts

The preceding experiments evaluate familiar contract specifications using held-out intervention mechanisms. We next ask whether verified compliance transfers when the contract itself is excluded from adaptation. We consider four forms of contract-level generalization: held-out linguistic realizations (*Para.*); familiar operators applied to unseen evidence-target categories (*New Target*); unseen combinations of individually observed operators and targets (*New Pairing*); and novel conjunctions of individually familiar clauses (*New Comp.*). All splits are matched to the seen-contract evaluation with respect to dataset, operator distribution, clause count where applicable, and held-out verification-family allocation. Figure 3 reports verified CSR. Evidence Contracts achieve the highest CSR among applicable methods in every unseen-contract setting. They retain 67.5% CSR under held-out paraphrases, close to their 68.4% seen-contract result, and achieve 65.8% on new targets and 63.1% on unseen operator–target pairings. Thus, the learned constraints transfer beyond the linguistic templates, evidence categories, and operator–target associations observed during adaptation. Unseen clause composition is the lowest-scoring setting for Evidence Contracts, as a contract passes only when all constituent clauses satisfy verification. Evidence Contracts retain 59.2% CSR, exceeding F Contract by 27.3 points and GUIDE by 33.8 points. Across all unseen settings, they achieve 63.9% CSR, compared with 40.3% for F Contract and 31.0% for GUIDE. Their 4.5-point decrease from seen to unseen contracts is also smaller than the 9.3-point decrease for F Contract optimization. Together, the high absolute unseen CSR and limited degradation show that Evidence Contracts generalize to held-out contract specifications, rather than only to familiar requirements expressed through new interventions.

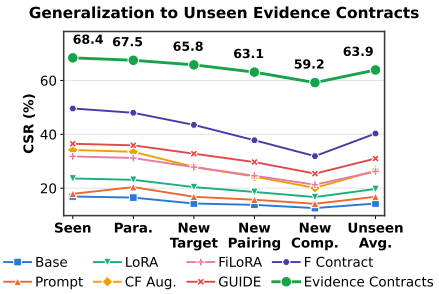

Figure 3: **Generalization to unseen contracts.** Full results appear in Table 12.

5.4 What Enables Verified Compliance?

We ablate adaptive enforcement, clause-wise normalization, intervention validity and family separation, and explicit task preservation. Figure 4 summarizes the resulting generalization and utility trade-offs; full ablations are provided in Appendix D.5 and Table 14.

1) Adaptive enforcement and normalization. Replacing adaptive clause weighting with fixed clause weights reduces H.O.-Family CSR from 68.4% to 49.6% and worst-operator OSR from 77.1% to 59.3%, despite retaining 69.5% Adapt.-Family CSR and 98.7% task retention. Static weighting therefore leaves difficult operator types as bottlenecks. Removing clause-wise normalization likewise lowers H.O.-Family CSR to 53.8% and worst-operator OSR to 55.7%, showing that adaptive weighting must compare heterogeneous clause violations on a common scale.

2) Valid and independent intervention protocols. Removing admissibility filtering lowers H.O.-Family CSR to 51.4%, indicating that interventions which fail to isolate the designated evidence or preserve task semantics provide unreliable optimization signals. Selecting checkpoints on distinct instances generated by the adaptation-family mechanism increases Adapt.-Family CSR from 74.1% to 76.1%, but lowers H.O.-Family CSR from 68.4% to 54.9%, increasing Δ_{IG} from 5.7 to 21.2 points. The effect arises from reusing an intervention mechanism.

3) Non-degenerate compliance. Without `PRESERVE` constraints, task retention falls from 98.6% to 87.6% and worst-operator OSR from 77.1% to 49.5%. Some clause violations can therefore be reduced through broad prediction changes rather than selective evidence control. Verified compliance requires adaptive, scale-aware enforcement; valid and independent intervention protocols; and explicit preservation of task-relevant behavior.

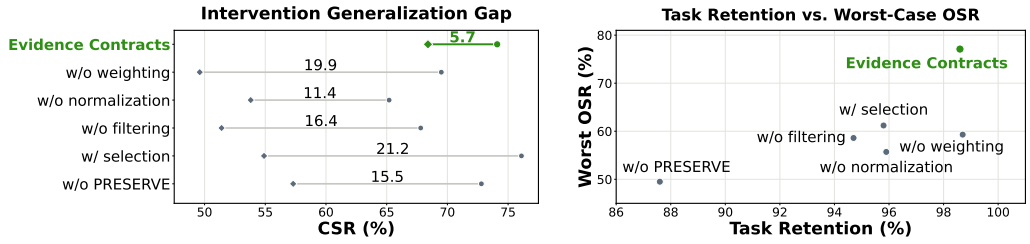

Figure 4: Ablation of Evidence Contract. In Intervention Generalization Gap, • denotes Adapt.-Family CSR and ⧫ denotes H.O.-Family CSR.[2]

5.5 Qualitative Verification and Failure Diagnosis[3]

We illustrate held-out verification with three satisfied contracts and one rejected contract spanning Visual R.C., TextVQA, M^{2}CQA, and GQA (Figure 5). The examples are selected to illustrate distinct contract operators and a localized verification failure; complete case specifications and verification details are provided in Appendix D.6.1. Each example reports operator-specific interventions, model responses, and the resulting verification decision. Contract satisfaction is determined by their joint pattern, rather than agreement with a single desired prediction. The successful cases demonstrate selective rather than indiscriminate robustness. In (a), changing the kitchen background has a small effect (0.08), whereas replacing the pizza with a salad changes the prediction (effect 0.71). In (b), changing *STOP* to *STEP* produces an OCR effect of 0.74, compared with 0.06 for a matched control. In (c), visual evidence has a larger effect than the conflicting prior cue (0.74 versus 0.21), yielding a `PREFER` margin of 0.53. Example (d) shows why clause-level verification is necessary. The background intervention satisfies `IGNORE` (effect 0.05), but modifying the relevant knee-pad attribute produces insufficient response (effect 0.16) for `REQUIRE`. The contract is therefore rejected despite partial success. Rather than collapsing these responses into an average sensitivity score, Evidence Contracts localize the remaining violation to the failed clause. This distinction between complete and superficially plausible partial compliance is central to the framework.



Figure 5: **Qualitative held-out verification.** (a)–(c) satisfy `IGNORE`/`REQUIRE`, `REQUIRE`, and `PREFER`, respectively. (d) passes `IGNORE` but fails `REQUIRE`; the complete contract is rejected. Values are intervention effects; PASS/FAIL uses confidence-adjusted verification.

6 Conclusion

We introduced **Evidence Contracts**, which compile natural-language evidence-use instructions into grounded behavioral constraints, enforce them during adaptation, and verify complete compliance using intervention families held out from adaptation and model selection. Across multimodal benchmarks, Evidence Contracts improves held-out contract satisfaction while preserving task utility, and generalizes across unseen intervention mechanisms and contract specifications. More broadly, the framework shifts evaluation from whether a model responds as intended under familiar interventions to whether independently grounded interventions verify user-specified behavioral requirements.

[2]Compliance across contract operators is provided in Appendix D.4

[3]We provide additional qualitative example in Appendix D.6.

AI USE STATEMENT

In this work, generative AI tools were used only to edit the manuscript for grammar and readability. We have not used generative AI tools for any of the required-disclosure tasks. All such work is the authors' own. All AI-assisted edits were reviewed by authors. We take full responsibility for the final content of this work, including text, claims or artifacts.

REPRODUCIBILITY STATEMENT

Formal definitions, prrofs, and optimization details for our methods are provided in Section 3 and Appendix B. Dataset sources, preprocessing, and experimental configuration are provided in Appendix C.1 and B.8. All results are averaged over three random seeds.

ETHICS STATEMENT

All dataset used in this work (Visual R.C., M^2CQA, GQA, TextVQA, EmoReAlM, MM-IMDb, and Flickr30K) are publicly available benchmarks used under their original licenses and intended research purposes.

All human annotation including complier fidelity audit, intervention validation and region-caption alignment on Flickr30K are is done by authors. Human annotation details are provided in Appendix D.

REFERENCES

Hyunsuk Chung, Soyeon Caren Han, Seungyeon Ji, Jinwoo Kim, Eun-Jung Holden, and Kyungreem Han. Filora: Focus-and-ignore lora for controllable feature reliance, 2026. URL https://arxiv.org/abs/2602.02060.

Yizhak Y Elboher, Reuven Peleg, Zhouxing Shi, Guy Katz, and Jan Křetínskỳ. Talking with verifiers: Automatic specification generation for neural network verification. *arXiv preprint arXiv:2603.02235*, 2026.

Soyeon Caren Han, Hyunsuk Chung, Jinwoo Kim, Seungyeon Ji, and Kyungreem Han. Guide: Guiding internal evidence with language instructions. *arXiv preprint arXiv:2608.30712*, 2026.

Van Thong Huynh, Hong Hai Nguyen, Thuy Pham, Trong Nghia Nguyen, and Soo-Hyung Kim. Faithful action-unit causal reasoning for counterfactually faithful emotion explanations. *ArXiv*, abs/2606.15779, 2026. URL https://api.semanticscholar.org/CorpusID:289299898.

Sai Srinivas Kancheti, Aditya Sanjiv Kanade, Rohit Sinha, Vineeth N. Balasubramanian, and Tanuja Ganu. Faithful GRPO: Improving visual spatial reasoning in multimodal language models via constrained policy optimization. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=Sdh1YOpaGF.

Junxian Li, Xinyue Xu, Sai Ma, Di Zhang, and Sichao Li. Faithful-first reasoning, planning, and acting for multimodal LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6777–6793, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.336. URL https://aclanthology.org/2026.findings-acl.336/.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL https://doi.org/10.24963/ijcai.2017/371.

Changyuan Tian, Zhicong Lu, Huaxing Liu, Xiang Wang, Shuai Li, Yu Chen, Wenqian Lv, Zichuan Lin, Juncheng Diao, and Deheng Ye. Faithful-mr1: Faithful multimodal reasoning via anchoring and reinforcing visual attention. *arXiv preprint arXiv:2605.22072*, 2026.

Victor Veitch, Alexander D' Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. 34:16196–16208, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/8710ef761bbb29a6f9d12e4ef8e4379c-Paper.pdf.

Peiyu Yang, Paul Montague, Feng Liu, Andrew Craig Cullen, Amardeep Kaur, Christopher Leckie, and Sarah Monazam Erfani. Semantic robustness certification for vision-language models. In *Forty-third International Conference on Machine Learning*, 2026.

A Additional Related Works

Table 2: **Positioning Evidence Contracts.** Unlike prior approaches, Evidence Contracts combine user-specified compositional evidence-use policies, behavioral enforcement, and verification using intervention families held out from adaptation, model selection, and calibration.

Approach	Specification	Enforcement	Verification / evaluation	User-specified instruction
Feature constraints & invariance (Ross et al., 2017; Veitch et al., 2021)	Fixed feature property	✓	Empirical counterfactual eval.	–
Semantic verification & certification (Elboher et al., 2026; Yang et al., 2026)	Semantic / robustness property	–	Formal / certified	Limited
Faithful multimodal reasoning (Tian et al., 2026; Huynh et al., 2026; Kancheti et al., 2026; Li et al., 2026)	Task-defined / causal faithfulness	✓	Task-specific evaluation	–
FiLoRA (Chung et al., 2026)	Feature-reliance instruction	✓	Reliance diagnostics	✓
GUIDE (Han et al., 2026)	Evidence-routing instruction	✓	Functional evaluation	✓
Evidence Contracts (Ours)	Compositional contract	✓	Held-out behavioral verification	✓

Table 2 compares Evidence Contracts with representative approaches along four dimensions: the form of the specified property, whether the property is enforced during learning or adaptation, the type of verification or evaluation, and whether the desired behavior can be directed by a user-specified instruction. The table summarizes the primary capability and evaluation protocol of each line of work; it does not imply that an existing method could not be combined with an additional intervention-based evaluation procedure.

Feature constraints and counterfactual invariance. Right-for-the-Right-Reasons methods constrain a model using annotated features or differentiable explanation signals, while counterfactual-invariance methods seek invariance to specified spurious changes. These approaches provide important mechanisms for reducing reliance on known unwanted signals. Their target property, however, is typically fixed by feature annotations, counterfactual construction, or a task-level invariance assumption, rather than compiled from a compositional instruction specifying several desired evidence-use relationships simultaneously.

Semantic specification and verification. Semantic robustness and specification-based approaches formalize properties such as robustness to a class of semantic changes and may verify or certify them under a defined threat model. Evidence Contracts is complementary: its specification concerns how a prediction should depend on designated evidence sources, including invariance, required dependence, relative preference, and task preservation. Its verification is based on grounded behavioral responses to admissible intervention families rather than solely on a fixed semantic property.

Faithful reasoning and evidence control. Faithful multimodal reasoning methods encourage grounding in task-relevant or causally motivated evidence, often under task-specific supervision or supplied causal structure. FiLoRA and GUIDE further enable instruction-conditioned control of feature reliance or evidence routing. Evidence Contracts does not replace these mechanisms: contract-constrained adaptation can operate with standard parameter-efficient adaptation or explicit routing mechanisms. Its distinct contribution is to require that the resulting behavior satisfy a complete contract under intervention mechanisms excluded from adaptation, model selection, and calibration.

B Additional Method Details

B.1 Contract Compiler and Fidelity Audit

The frozen contract compiler uses constrained decoding under the grammar in Eq. equation 1. It maps each instruction to a conjunction of canonical `IGNORE`, `REQUIRE`, `PREFER`, and `PRESERVE` clauses. Outputs that are malformed, omit a required argument, or contain an unresolved descriptor are retained as unresolved rather than repaired post hoc.

Table 3: **Dataset-level grounding and intervention protocol.** (a) Evidence types and grounding support per dataset; $q_{\min}$ is the grounding-confidence threshold. (b) Intervention families and validators / thresholds per dataset; τ_{tar} and τ_{non} are the target-efficacy and non-target-preservation thresholds, respectively.

(a) Grounding

Dataset	Evidence	Grounding support	$q_{\min}$
VLM R. C.	Obj., action, bg.	Detector + segmentation mask	0.80
M^{2}CQA	Visual evidence, prior cue	Image region; textual prior span	0.80
GQA	Obj., attr., rel., action, bg.	Scene-graph region support	0.85
TextVQA	OCR text, obj., bg.	OCR polygon; object region	0.85
EmoReAlM	Face, vocal cue, prosody, motion	Face track; aligned audio segment	0.80
MM-IMDb	Poster appearance, text semantics	Poster region; plot-text span	0.80
Flickr30K	Obj., action, appearance, scene	Region grounding; caption span	0.80

(b) Interventions

	Intervention family			
Dataset	**Train**	**Calibration**	**Verification**	**Validators / thresholds**
VLM R. C.	Local masking; Gaussian blur	Controlled region replacement	Semantic scene substitution; contextual swap	Frozen visual validator; task metadata; τ_{tar}=0.20, τ_{non}=0.10
M^{2}CQA	Region masking; prior-cue deletion	Partial replacement	Visual–prior conflict swap	Frozen VLM validator; answer annotation; τ_{tar}=0.20, τ_{non}=0.10
GQA	Masking; localized blur	Attribute or object replacement	Relation/action counterfactual	Scene graph + target answer; τ_{tar}=0.20, τ_{non}=0.10
TextVQA	OCR masking; partial deletion	Token deletion; re-rendering control	Text-content replacement; re-rendering	OCR agreement + answer annotation; τ_{tar}=0.20, τ_{non}=0.10
EmoReAlM	Face masking; audio attenuation	Partial temporal occlusion	Facial or acoustic emotion edit	Frozen AV validator; emotion label; τ_{tar}=0.15, τ_{non}=0.10
MM-IMDb	Image blur; text masking	Partial text deletion	Poster-style or text-semantic replacement	Metadata label + semantic validator; τ_{tar}=0.15, τ_{non}=0.10
Flickr30K	Region masking; localized blur	Controlled appearance edit	Object/action/scene substitution	Region–caption alignment; human audit; τ_{tar}=0.15, τ_{non}=0.10

We separately audit compiler fidelity on held-out instructions. Human annotators independently construct canonical contracts from each instruction and adjudicate disagreements at the operator, evidence-target, and clause-composition levels. We report exact-contract match, clause-level precision and recall, and the frequency of each error type. This audit distinguishes compiler errors from grounding and behavioral-compliance failures.

B.2 Grounding, Validators, and Intervention Construction

For each dataset and evidence descriptor, we fix a grounder, input support representation, intervention generators, and external validity checks before adaptation. Table 3 reports the fixed grounding, validation, and intervention protocol used in our experiments. These choices were specified before adaptation and shared across all applicable methods.

All candidate interventions are filtered by the validity indicator in Eq. equation 5. The target-efficacy validator confirms that the intervention changes evidence within the grounded support by at least τ_{tar}. The non-target validator requires changes outside the support to remain below τ_{non}. Task-semantic validity is checked against the original target for task-preserving and evidence-degrading interventions, and against the declared counterfactual target for diagnostic counterfactuals. Frozen external validators, task annotations, or human verification are used where applicable.

B.3 Mechanism-Level Family Separation

The separation in Eq. equation 6 is performed at the level of intervention-generating mechanisms rather than through a random split of instances produced by one mechanism. For example, background masking and blurring are used during adaptation, controlled replacement is reserved for

calibration, and semantic scene substitution or contextual swaps are held out for verification. Likewise, OCR-region masking is used for adaptation, partial deletion for calibration, and text-content replacement or re-rendering for verification.

All family assignments are fixed before adaptation and remain identical across compared methods. Verification-family interventions are excluded from parameter updates, early stopping, hyperparameter selection, threshold calibration, and strength matching. We report the number of candidate and validated interventions in each family and the resulting Intervention Validity Rate.

B.4 Optimization Details

We optimize Eq. equation 19 using an augmented primal–dual objective. For minibatch $\mathcal{B}$, let

$$g_o(\mathcal{B}) = \widehat{\mathrm{CVaR}}_\beta(Z_o; \mathcal{B}) - \kappa_o. \tag{25}$$

The adaptation objective is

$$\mathcal{L}_{\mathrm{CCA}} = \mathcal{L}_{\mathrm{task}} + \sum_{o \in \mathcal{O}} \left[\lambda_o g_o(\mathcal{B}) + \frac{\rho}{2} [g_o(\mathcal{B})]_+^2 \right], \tag{26}$$

where $\lambda_o \geq 0$ is the dual variable for operator o and ρ controls the quadratic penalty. Dual variables are updated as

$$\lambda_o \leftarrow [\lambda_o + \eta_\lambda g_o(\mathcal{B})]_+ . \tag{27}$$

The update is applied only when $\mathcal{B}$ contains at least one clause of operator type o. This increases optimization pressure on operator types with persistently large tail violations without manually fixing static operator weights.

B.5 Clause-Wise Normalization

Clause-wise normalization is used only in the adaptation objective; final verification always applies the original, unnormalized behavioral quantities and contract thresholds. We assign each clause c to a fixed calibration stratum

$$\omega(c) = \big(d(c),\, o(c),\, t(c)\big), \tag{28}$$

where $d(c)$ denotes the benchmark, $o(c) \in \{\mathrm{ign}, \mathrm{req}, \mathrm{pref}, \mathrm{pres}\}$ denotes the operator, and $t(c)$ denotes the grounded evidence-type family. For `PREFER`, $t(c)$ is the ordered pair $(t(e_a), t(e_b))$; for `PRESERVE`, it is the task family of the corresponding benchmark.

For each observed calibration stratum ω, we collect raw clause violations from calibration-family interventions only and define the scale using the root mean-square violation:

$$s_\omega = \max\left(\sqrt{\frac{1}{|\mathcal{V}_\omega^{\mathrm{cal}}|} \sum_{v \in \mathcal{V}_\omega^{\mathrm{cal}}} v^2},\, 10^{-3} \right), \tag{29}$$

where $\mathcal{V}_\omega^{\mathrm{cal}}$ is the set of raw calibration violations in stratum ω. We use $\varepsilon = 10^{-6}$ in Eq. equation 18. All scales are fixed before adaptation and are not updated using adaptation-family or verification-family outcomes.

For a contract excluded from adaptation and calibration, including unseen evidence targets, operator–target pairings, and clause compositions, each clause inherits the scale of its corresponding observed calibration stratum $\omega(c)$. Thus, unseen contracts do not introduce a new normalization parameter or use their own outcomes for calibration. If an exact benchmark–operator–evidence-type stratum is unavailable, we deterministically back off to the benchmark–operator stratum. This back-off rule is fixed before adaptation and is shared by all compared methods.

B.6 Finite-Sample Verification Guarantee

Theorem. Fix a contract $\mathcal{C}$ and condition on the frozen compiler, grounding procedure, admissibility validators, calibration parameters, and adapted model. Suppose each scalar verification statistic is bounded in $[0, 1]$ and estimated from m_j i.i.d. samples from its declared held-out intervention

distribution. Then, with probability at least $1-\alpha$, every confidence interval used in Eq. equation 22 simultaneously contains its corresponding expected behavioral quantity. Consequently, every contract declared verified satisfies all of its expected behavioral constraints within the declared held-out intervention distributions.

Proof. For scalar statistic j, Hoeffding's inequality gives

$$\Pr\left(|\hat{\mu}_j - \mu_j| > r_j\right) \leq \frac{\alpha}{J(\mathcal{C})}, \tag{30}$$

for the radius r_j in Eq. equation 21. A union bound over the $J(\mathcal{C})$ scalar statistics implies that all confidence intervals simultaneously contain their expectations with probability at least $1-\alpha$. On this event, the upper confidence bound for `IGNORE` upper-bounds expected dependence; the lower confidence bounds for `REQUIRE` and `PRESERVE` lower-bound expected target support; and the lower-minus-upper bound for `PREFER` lower-bounds the expected dependence difference. Therefore every verified clause satisfies its expected constraint, and conjunctive verification implies satisfaction of the complete contract.

This guarantee concerns statistical behavioral testing within declared, validated held-out intervention distributions. It does not certify robustness to arbitrary transformations outside those distributions.

B.7 Additional Diagnostic Metrics

Worst-Clause Contract Violation. For example i and clause c_k, let ρ_{ik}^{ver} be the confidence-adjusted verification margin. We define

$$\mathrm{WCV} = \frac{1}{N}\sum_{i=1}^{N} \max_{c_k \in \mathcal{C}_i} \left[-\rho_{ik}^{\mathrm{ver}}\right]_+, \tag{31}$$

which captures the magnitude of the strongest remaining violation in each contract.

Grounding Coverage and Intervention Validity Rate. Grounding Coverage (GC) is the proportion of requested evidence descriptors resolved above the threshold $q_{\min}$. Intervention Validity Rate (IVR) is the proportion of candidate interventions satisfying Eq. equation 5. We report GC and IVR by dataset, operator, and intervention family to distinguish failures of compilation, grounding, intervention construction, and behavioral compliance.

B.8 Reproducibility and Evaluation Protocol

Table 4 reports the benchmark-level intervention protocol. For each benchmark, we list candidate and validated interventions in the adaptation, calibration, and verification splits; the number of independent training seeds; the verification confidence level α; and the number m_j of held-out admissible interventions used for each scalar verification statistic. Family assignments, validators, thresholds, and strength-matching rules are fixed before adaptation. Verification-family interventions are excluded from parameter updates, early stopping, hyperparameter selection, threshold calibration, and strength matching.

Table 5 reports adaptation budgets and optimization settings by backbone family. Results are aggregated over three independent seeds and reported as mean and standard deviation. A baseline is marked N/A when its architecture, modality, or required evidence representation is unavailable for a benchmark–backbone setting; such pairs are excluded from dataset averages.

C Evaluation Setup Details

C.1 Dataset Details and Split Construction

C.1.1 Dataset Sources

Table 6 lists the source task and construction used for each benchmark. Visual R.C., M^{2}CQA, and EmoReAlM use their corresponding controlled benchmark protocols; GQA, TextVQA, MM-IMDb,

Table 4: **Benchmark-level evaluation protocol.** $N_{\text{cand}}/N_{\text{valid}}$ denotes the total number of candidate and validated interventions in each split. m_j is the number of held-out admissible interventions sampled from the verification split to estimate each scalar statistic. All family assignments, validators, thresholds, and strength-matching rules are fixed before adaptation.

Benchmark	Interventions $N_{\text{cand}}/N_{\text{valid}}$			Seeds	α	m_j
	Adaptation	Calibration	Verification			
Visual R.C.	1500/1200	375/300	750/600	3	0.05	100
M^{2}CQA	1250/1000	313/250	625/500	3	0.05	100
GQA	2500/2000	625/500	1250/1000	3	0.05	100
TextVQA	1875/1500	469/375	938/750	3	0.05	100
EmoReAlM	1500/1200	375/300	750/600	3	0.05	100
MM-IMDb	1250/1000	313/250	625/500	3	0.05	100
Flickr30K	1875/1500	469/375	938/750	3	0.05	100

Table 5: **Adaptation budgets and optimization settings.** β is the CVaR tail level and ρ is the augmented-Lagrangian penalty coefficient. LoRA modules are inserted into attention and MLP projections. Trainable parameters exclude frozen backbone parameters. For comparable adaptation baselines, we match the backbone, trainable-parameter budget, task-training examples, minibatch schedule, and optimization steps.

Backbone family	LoRA rank	Trainable params.	Steps	Batch	Learning rate	Weight decay	β	ρ
Vision–language	16	$\leq 0.20\%$	1,000	16	2×10^{-4}	0.01	0.90	1.0
Audio–visual	16	$\leq 0.20\%$	1,000	16	2×10^{-4}	0.01	0.90	1.0
Image–text	16	$\leq 0.20\%$	1,000	16	2×10^{-4}	0.01	0.90	1.0

and Flickr30K follow their public task definitions. We preserve the original task target for every retained example and use only the annotations required to ground the designated evidence source.

C.1.2 Data Source Link

We provide the source links for the evaluation datasets used in our experiments:

- **Visual R. C.**: `https://huggingface.co/datasets/anonymous9457/vlm-reality-check`
- **M^{2}CQA**: `https://huggingface.co/datasets/QCRI/M2CQA`
- **GQA**: `https://cs.stanford.edu/people/dorarad/gqa/about.html`
- **TextVQA**: `https://huggingface.co/datasets/facebook/textvqa`
- **EmoReAlM**: `https://avere-iclr.github.io/`
- **MM-IMDb**: `https://github.com/johnarevalo/gmu-mmimdb`
- **Flickr30K**: `https://huggingface.co/datasets/nlphuji/flickr30k`

C.1.3 Split Construction

We first partition base examples into disjoint adaptation, calibration, and final-verification sets. We then construct contracts and admissible interventions within the corresponding partition. Thus, no base example, contract instance, or intervention instance is shared across these splits. Adaptation families are used for parameter updates only; calibration families are used for checkpoint selection, hyperparameter choice, threshold calibration, and strength matching; verification families are not accessed until final evaluation.

For unseen-contract evaluation, we additionally construct four disjoint generalization settings: held-out linguistic realizations (unseen paraphrases), evidence targets, operator–target pairings, and multi-clause compositions. The held-out contract specifications are absent from adaptation and

Table 6: **Dataset sources and task construction.** "Filtered" means that examples without a valid task target, resolved grounding, or admissible intervention are excluded.

Benchmark	Source task	Retained input	Preprocessing
Visual R.C.	Controlled counterfactual recognition	image, question, answer	Filtered for valid target evidence
M^2CQA	Multimodal causal question answering	image, question, answer	Filtered for grounded visual/textual cues
GQA	Visual question answering	image, question, answer	Original task target; valid scene-graph support
TextVQA	Text-based visual question answering	image, question, answer	OCR normalization and valid text-region support
EmoReAlM	Audiovisual emotion understanding	video, audio, label	Aligned clips with valid face/audio support
MM-IMDb	Multimodal genre classification	poster, plot, genre label	Valid poster and plot-text support
Flickr30K	Image–text retrieval	image, caption pair	Valid image-region and caption-span support

Table 7: **Reproducibility configuration.** "Families" denotes distinct intervention-generating mechanisms. Train / Cal. / Ver. instances are validated interventions available in the corresponding split, and m_j is the number used to estimate each scalar verification statistic. All family assignments are fixed before adaptation.

Benchmark	Train families	Cal. families	Ver. families	Train / Cal. / Ver. instances	m_j	Seeds
Visual R.C.	2	1	2	1,200 / 300 / 600	100	3
M^2CQA	2	1	2	1,000 / 250 / 500	100	3
GQA	2	1	2	2,000 / 500 / 1,000	100	3
TextVQA	2	1	2	1,500 / 375 / 750	100	3
EmoReAlM	2	1	2	1,200 / 300 / 600	100	3
MM-IMDb	2	1	2	1,000 / 250 / 500	100	3
Flickr30K	2	1	2	1,500 / 375 / 750	100	3

calibration, while their task inputs remain drawn from the same benchmark-specific base partition. These splits test generalization of contract semantics separately from generalization across intervention-generating mechanisms. Table 4 reports the intervention allocation; Table 3 specifies the grounding, validation, and intervention mechanisms.

C.1.4 Reproducibility protocol.

Table 7 summarizes the intervention-family allocation and validated-instance counts used by our protocol. The complete reproducibility configuration, including candidate intervention counts, the verification confidence level α, and the number m_j of held-out admissible interventions used for each scalar statistic, is provided in Table 4. To avoid optimistic evaluation, verification interventions are generated by mechanisms excluded from adaptation, threshold calibration, model selection, and early stopping. We train each method with three random seeds and report the mean and standard deviation across seeds. Complete benchmark-specific generator, validator, and threshold details are provided in Appendix B.2.

Adaptation configuration and baseline coverage. Unless a baseline requires a different architecture, all applicable methods use the same backbone, task-training examples, minibatch schedule, and number of optimization steps. In the configuration, we adapt rank-16 LoRA modules inserted into attention and MLP projections, corresponding to at most 0.20% trainable parameters of the backbone, for $1{,}000$ optimization steps. We use AdamW with learning rate 2×10^{-4}, weight decay 0.01, batch size 16, and CVaR level $\beta = 0.90$. Early stopping, when used, is selected only using

Table 8: **Baseline applicability matrix.** A dash indicates that a method is not applicable because its required modality, architecture, or implementation is unavailable; it is neither treated as a zero nor included in aggregate rankings.

Method	Vision–language	Audio–visual	Retrieval	Budget
Frozen backbone	✓	✓	✓	0
Task-only LoRA	✓	✓	✓	Matched
Prompt steering	✓	✓	✓	0
Counterfactual augmentation	✓	✓	✓	Matched
F Contract	✓	✓	✓	Matched
FiLoRA-style control	✓	✓	✓	Matched
GUIDE-style routing	✓	✓	✓	Matched
Evidence Contracts	✓	✓	✓	Matched

Table 9: **Benchmark–backbone assignments.** All checkpoints are used without full-model fine-tuning. "Adapter" denotes the method-specific trainable components, including the rank-16 LoRA modules used by Evidence Contracts and matched PEFT baselines.

Benchmark	**Task family**	**Frozen backbone**	**Trainable Comp.**
Visual R.C.	Vision–language	`Qwen3-VL-8B-Instruct`	Adapter
M^2CQA	Vision–language	`Qwen3-VL-8B-Instruct`	Adapter
GQA	Vision–language	`Qwen3-VL-8B-Instruct`	Adapter
TextVQA	Vision–language	`Qwen3-VL-8B-Instruct`	Adapter
MM-IMDb	Image–text classification	`Qwen3-VL-8B-Instruct`	Adapter
Flickr30K	Image–text retrieval	`Qwen3-VL-8B-Instruct`	Adapter
EmoReAlM	Audio–visual understanding	`Qwen3-Omni-30B-A3B-Instruct`	Adapter

calibration-family CSR. Contract thresholds are calibrated on $\mathcal{A}^{\mathrm{cal}}$ separately for each benchmark and remain fixed during held-out verification.

For every reported comparison, we match the backbone and trainable-parameter budget whenever the baseline architecture permits it. Results are reported only on datasets and modalities for which each baseline is applicable; unavailable method–benchmark pairs are marked "–" and excluded from dataset averages.

C.2 Backbones and Implementation Configuration

Table 9 specifies the pretrained backbone used for each benchmark. Unless a baseline intrinsically requires a different architecture, methods within a benchmark use the same backbone, task interface, and trainable-parameter budget. All backbone parameters are frozen; only method-specific adaptation modules are trained.

For Flickr30K, the Qwen3-VL backbone ranks candidate captions by their conditional answer likelihood; contracts are compiled into constraints on these retrieval scores rather than on a free-form generated response.

C.3 Additional Diagnostic Metrics

The primary CSR metric requires every clause in a contract to satisfy its confidence-adjusted verification criterion. We additionally report diagnostic metrics that localize residual failures in behavioral compliance, grounding, and intervention construction.

Task-utility Retention Task-utility Retention reports performance on the original task relative to the unadapted model, where 100 denotes full retention. Let $U_\theta(\mathcal{D})$ be the task utility on evaluation set $\mathcal{D}$, with all task metrics oriented so that larger values are better. We define

$$\text{Task-Utility Retention} = 100\,\frac{U_\theta(\mathcal{D})}{U_{\theta_0}(\mathcal{D})}, \tag{32}$$

where θ_0 denotes the unadapted backbone. Values above 100 indicate an improvement over the unadapted model, whereas values below 100 indicate a utility cost of adaptation. We report the macro-average across datasets and random seeds.

Worst-Clause Contract Violation. For example i and clause $c_k \in \mathcal{C}_i$, let ρ_{ik}^{ver} be the confidence-adjusted verification margin defined in Eq. equation 23. We define

$$\mathrm{WCV} = \frac{1}{N} \sum_{i=1}^{N} \max_{c_k \in \mathcal{C}_i} \left[-\rho_{ik}^{\mathrm{ver}} \right]_+ . \tag{33}$$

WCV measures the magnitude of the strongest remaining violation in each contract. Unlike CSR, it distinguishes contracts that narrowly fail a threshold from those with substantial non-compliance.

Grounding Coverage. Grounding Coverage (GC) measures whether requested evidence descriptors can be resolved before behavioral evaluation:

$$\mathrm{GC} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left[q_{e_i} \geq q_{\min} \right] . \tag{34}$$

We report GC by dataset and evidence type. Unresolved descriptors count as failures for E2E-CSR, but are separated from behavioral-compliance failures in the diagnostic breakdown.

Intervention Validity Rate. For a candidate intervention set $\mathcal{T}$, Intervention Validity Rate (IVR) is

$$\mathrm{IVR}(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} \mathbf{1} \left[\mathrm{Valid}(T, x, e) = 1 \right] . \tag{35}$$

We report IVR by dataset, intervention type, and family assignment. This distinguishes failures of intervention construction from failures to satisfy a valid grounded contract.

D Results Details

D.1 Dataset-Level Control-Gap Results

Table 10 reports the dataset-level results for Evidence Contracts underlying its macro-average in Table 1. Values are mean ± standard deviation across three seeds. Baseline applicability and the construction of the common evaluation subset are reported in Table 8.

Table 10: **Dataset-level verified-compliance results for Evidence Contracts.** Task Ret. is task-utility retention relative to the unadapted backbone. Ada. CSR and H.O.-Family CSR are measured using fresh adaptation-family and disjoint verification-family interventions, respectively. All values except Δ_{IG} and Ver. Margin are percentages.

Dataset	Task Ret.	Ada. CSR	H.O.-Family CSR	Δ_{IG}	Ver. Margin
Visual R.C.	98.6 ± 0.3	69.8 ± 1.6	63.4 ± 1.8	6.4 ± 0.5	+0.06 ± 0.02
M^{2}CQA	98.7 ± 0.2	74.1 ± 1.2	68.3 ± 1.4	5.8 ± 0.4	+0.11 ± 0.02
GQA	98.5 ± 0.4	74.7 ± 1.4	69.1 ± 1.5	5.6 ± 0.5	+0.12 ± 0.03
TextVQA	98.7 ± 0.3	76.2 ± 1.1	70.4 ± 1.2	5.8 ± 0.4	+0.15 ± 0.02
EmoReAlM	98.8 ± 0.2	73.3 ± 1.5	67.8 ± 1.6	5.5 ± 0.5	+0.10 ± 0.03
MM-IMDb	98.9 ± 0.2	74.3 ± 1.3	69.5 ± 1.4	4.8 ± 0.4	+0.12 ± 0.02
Flickr30K	98.0 ± 0.4	76.3 ± 1.2	70.3 ± 1.3	6.0 ± 0.5	+0.11 ± 0.03

Evidence Contracts is somewhat lower on Visual R.C. than on the other benchmarks, but remains within the same overall performance range. This benchmark emphasizes counterfactual answer validity under tightly constrained edits, rather than the broader multimodal evidence-use patterns represented by the remaining tasks; importantly, it does not exhibit an anomalously large adaptation–verification gap.

D.2 Family-Level Held-Out Compliance Results

Table 11 presents the full results for each held-out intervention family, with performance reported as mean ± standard deviation over three independent seeds. These results are summarized in Figure 2 to highlight the overall pattern across intervention-generating mechanisms. Family-level CSR is macro-averaged over applicable datasets and contracts on the same common evaluation subset used in Table 1.

Table 11: **Contract satisfaction across held-out intervention families.** Values are mean ± standard deviation over three seeds and are percentages. Scene denotes semantic scene substitution; Obj./Attr., object or attribute replacement; Rel./Act., relation or action counterfactuals; OCR, text-content replacement or re-rendering; and Face and Audio, interventions that modify facial and acoustic emotion evidence, respectively. Each family-level CSR is macro-averaged over applicable datasets and contracts. **Average** is the unweighted macro-average across the six family-level cells, evaluated on the same common applicable subset as Table 1.

Method	Scene	Obj./Attr.	Rel./Act.	OCR	Face	Audio	Average
No explicit contract enforcement							
Base	18.2	17.5	15.8	19.1	15.2	15.6	16.9
Prompt	19.6	18.1	16.7	20.2	16.0	16.8	17.9
LoRA	25.8	24.1	21.7	26.2	21.3	22.5	23.6
CF Aug.	36.8	33.5	31.4	38.2	32.1	33.2	34.2
Explicit evidence control							
FiLoRA	34.0	32.4	29.1	33.5	30.2	31.6	31.8
GUIDE	39.1	37.2	34.4	40.0	33.8	34.5	36.5
Explicit contract optimization							
F Contract	52.4	49.1	46.8	53.0	47.2	49.1	49.6
Evidence Contracts	**71.8** ± 1.2	**68.7** ± 1.4	**65.9** ± 1.5	**72.4** ± 1.1	**65.8** ± 1.6	**65.8** ± 1.5	**68.4** ± 0.8

D.2.1 Family-Level Verification Validity

The family-level comparison uses only interventions that pass the same benchmark-specific admissibility checks used in the primary verification protocol. In particular, validators confirm the designated evidence change, bounded non-target modification, and preservation of valid task semantics where required. Candidate interventions rejected by these checks are excluded before CSR is computed; no method-specific filtering is applied. The generator, validator, and strength-matching specifications for each benchmark–family pair are provided in Appendix B.2.

The displayed family cells aggregate only benchmark–contract pairs to which the corresponding intervention mechanism applies. Thus, cross-family differences in absolute CSR are descriptive rather than evidence that one family is intrinsically easier; the relevant comparison is between methods within the same held-out family. All displayed verification families remain unseen during adaptation, checkpoint selection, threshold calibration, and strength matching.

D.3 Generalization to Unseen Contracts

Table 12 reports the full results for generalization to unseen evidence contracts across three independent seeds. It provides detailed CSR results for the seen-contract setting and four forms of contract-level generalization across Evidence Contracts and the comparison methods, including FiLoRA, GUIDE, and F Contract. Figure 3 summarizes the corresponding performance trends across these unseen-contract settings.

D.4 Compliance Across Contract Operators

The previous analyses evaluate complete contracts, which are satisfied only when every constituent clause passes verification. We now disaggregate compliance by operator to determine whether the overall improvement is shared across `IGNORE`, `REQUIRE`, `PREFER`, and `PRESERVE`, or driven primarily by a single type of constraint.

Table 12: **Generalization to unseen evidence contracts.** All CSR values are means over three random seeds and are reported as percentages. **Seen** reports held-out-intervention CSR for contract structures observed during adaptation. **Paraphrase (Para.)** uses held-out linguistic realizations of familiar contracts. **New Target** introduces evidence-target categories absent from adaptation-time contracts. **New Pairing** combines individually observed operators and targets that were never paired during adaptation. **New Composition (New comp.)** evaluates novel conjunctions of individually familiar clauses. **Unseen Avg.** is the unweighted macro-average over the four unseen-contract settings. All splits are matched by dataset, operator distribution, clause count where applicable, and verification-family intervention. **F Contract (Fixed-weight)** denotes contract optimization with static clause weights.

Method	Seen	Para.	New Target	New Pairing	New Comp.	Unseen Avg.
No explicit contract enforcement						
Base	16.9	16.5	14.3	13.8	12.6	14.3
Prompt	17.9	20.4	16.8	15.7	14.2	16.8
LoRA	23.6	23.1	20.4	18.6	16.7	19.7
CF Aug.	34.2	33.5	27.9	24.3	20.1	26.5
Explicit evidence control						
FiLoRA	31.8	31.2	27.8	24.6	21.3	26.2
GUIDE	36.5	35.9	32.8	29.7	25.4	31.0
Explicit contract optimization						
F Contract	49.6	48.0	43.5	37.8	31.9	40.3
Evidence Contracts	**68.4** $\pm$ 0.8	**67.5** $\pm$ 1.0	**65.8** $\pm$ 1.1	**63.1** $\pm$ 1.3	**59.2** $\pm$ 1.5	**63.9** $\pm$ 1.1

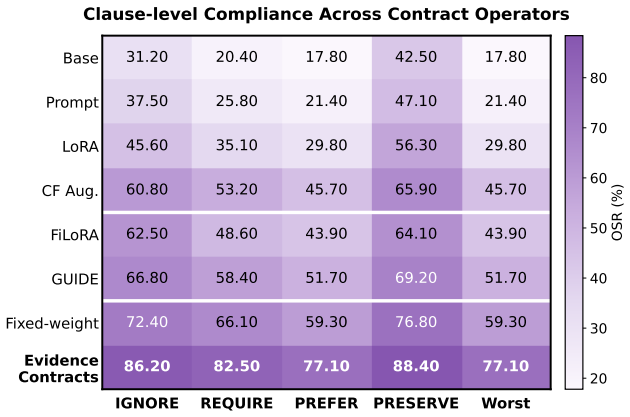

Figure 6: Clause-level compliance across Contract Operators.

For this analysis, we report Operator Satisfaction Rate (OSR). For each dataset and random seed, OSR is computed as the fraction of held-out clauses of a given operator type that satisfy their confidence-adjusted criterion under held-out verification-family interventions, and is then macro-averaged across datasets and seeds. Unlike contract-level CSR, OSR evaluates clauses individually and therefore does not require the remaining clauses in the same contract to be satisfied. The two metrics are complementary: CSR measures complete contract satisfaction, whereas OSR identifies which behavioral requirements remain difficult.

Table 13 shows that Evidence Contracts improve compliance across all four operator types. They achieve 86.2% OSR for `IGNORE`, 82.5% for `REQUIRE`, 77.1% for `PREFER`, and 88.4% for `PRESERVE`. The improvement is therefore not explained solely by learning invariance to designated evidence. Evidence Contracts also increase responsiveness to required evidence, enforce relative preferences between evidence sources, and maintain task-relevant behavior under task-preserving interventions.

Table 13: **Clause-level compliance across contract operators.** All OSR values are percentages. Each operator column reports the percentage of held-out clauses satisfying their confidence-adjusted criterion under held-out verification-family interventions. **Worst** is the lowest OSR across the four operators. OSR is measured at the clause level and is therefore not directly comparable to the conjunctive contract-level CSR. Results are macro-averaged across datasets and random seeds.

Method	IGNORE	REQUIRE	PREFER	PRESERVE	Worst
No explicit contract enforcement					
Base	31.2	20.4	17.8	42.5	17.8
Prompt	37.5	25.8	21.4	47.1	21.4
LoRA	45.6	35.1	29.8	56.3	29.8
CF Aug.	60.8	53.2	45.7	65.9	45.7
Explicit evidence control					
FiLoRA	62.5	48.6	43.9	64.1	43.9
GUIDE	66.8	58.4	51.7	69.2	51.7
Explicit contract optimization					
F Contract	72.4	66.1	59.3	76.8	59.3
Evidence Contracts	**86.2** $\pm$ **0.8**	**82.5** $\pm$ **1.0**	**77.1** $\pm$ **1.2**	**88.4** $\pm$ **0.7**	**77.1** $\pm$ **1.2**

IGNORE and PRESERVE are the easiest operators for most methods. Prompting, standard LoRA, and counterfactual augmentation can reduce dependence on a designated source or maintain task-relevant behavior under task-preserving interventions. However, satisfying IGNORE without violating PRESERVE requires selective invariance: the model must attenuate behavioral dependence on the designated evidence while retaining information needed for the task. Evidence Contracts reach 86.2% and 88.4% OSR on these operators, respectively, compared with 72.4% and 76.8% for F Contract. This result suggests that the compliance gain is not obtained through indiscriminate suppression of evidence sensitivity or degenerate prediction stability.

REQUIRE is more difficult because it demands positive behavioral dependence on the designated evidence. A model cannot satisfy this operator merely by producing a stable prediction; its predictive behavior must respond sufficiently to validated diagnostic counterfactuals. Evidence Contracts achieve 82.5% REQUIRE satisfaction, improving over F Contract by 16.4 percentage points, GUIDE by 24.1 points, and FiLoRA by 33.9 points. This improvement indicates that the method controls both evidence invariance and evidence sensitivity.

PREFER remains the most challenging operator across all methods. Unlike IGNORE and REQUIRE, it imposes a relative ordering between evidence sources and therefore requires the interventions associated with both sources to be grounded, validated, and matched in strength. The base model achieves only 17.8% OSR, while GUIDE and F Contract reach 51.7% and 59.3%, respectively. Evidence Contracts raise PREFER satisfaction to 77.1%, a 17.8-point improvement over F Contract. The gain on the most difficult operator shows that the proposed optimization does more than independently increase or decrease sensitivity to a single source; it can enforce comparative evidence-use requirements.

The Worst column further distinguishes methods that achieve high compliance for selected operators from those that perform consistently across all four. For every baseline, PREFER determines the worst-operator result, leaving even F Contract below 60%. Evidence Contracts raise worst-operator satisfaction to 77.1%, reducing the extent to which complete contracts are limited by a single weak clause type. This balance is particularly important under conjunctive contract semantics, where failure of any constituent clause causes the entire contract to fail.

Overall, Evidence Contracts improve both absolute and worst-operator compliance across invariance, sensitivity, relative preference, and task preservation requirements. The results support the use of operator-specific constraints rather than a single undifferentiated evidence-control objective. We next ablate the components responsible for these gains.

D.5 Ablation Studies

This section provides the full ablation protocol and results for Section 5.4. All variants use the same backbone, trainable-parameter budget, task-training data, contract distribution, adaptation-

Table 14: **Full ablation results for Evidence Contracts.** Values are mean ± standard deviation over three random seeds and are percentages, macro-averaged across datasets. **Ada. CSR** measures contract satisfaction on fresh instances generated by the intervention family available during adaptation. **H.O. CSR** measures compliance on a distinct verification family excluded from adaptation, threshold calibration, and model selection. **Worst OSR** is the lowest held-out clause-level satisfaction rate across `IGNORE`, `REQUIRE`, `PREFER`, and `PRESERVE`. **Task Ret.** is original-task performance relative to the unadapted model, where 100 denotes full retention.

Method	Ada.-Family CSR	H.O.-Family CSR	Worst OSR	Task Ret.
Evidence Contracts	**74.1** ± **0.9**	**68.4** ± **0.8**	**77.1** ± **1.0**	**98.6** ± **0.3**
w/o adaptive clause **weighting**	69.5 ± 1.1	49.6 ± 1.4	59.3 ± 1.7	98.7 ± 0.2
w/o clause-wise **normalization**	65.2 ± 1.4	53.8 ± 1.6	55.7 ± 1.8	95.9 ± 0.5
w/o admissibility **filtering**	67.8 ± 1.3	51.4 ± 1.7	58.6 ± 1.9	94.7 ± 0.6
w/ same-family model **selection**	76.1 ± 1.0	54.9 ± 1.5	61.2 ± 1.6	95.8 ± 0.5
w/o **`PRESERVE`** constraints	72.8 ± 1.1	57.3 ± 1.5	49.5 ± 2.0	87.6 ± 0.8

step budget, and optimization schedule as Evidence Contracts. Unless explicitly varied below, they also use the same grounding procedure, intervention generators, validators, calibration protocol, and held-out verification families.

Ablation variants. **Fixed clause weighting** replaces the adaptive operator-wise weighting rule with static clause weights, while retaining the same grounded clauses, interventions, and adaptation family. **No clause-wise normalization** applies the optimization objective to raw clause violations, without rescaling violations across clause strata. **No admissibility filtering** retains candidate interventions before the benchmark-specific validity checks; the intervention and optimization-step budgets are held fixed by sampling from this unfiltered pool. **Same-family model selection** selects checkpoints and hyperparameters using validation instances generated by the same intervention mechanism used for adaptation. These validation instances are disjoint from adaptation instances, and final evaluation remains on the original held-out verification families. This variant therefore isolates reuse of the intervention mechanism, rather than reuse of adaptation examples. Finally, **no `PRESERVE` constraints** removes `PRESERVE` clauses during adaptation. All variants are evaluated using the original full evaluation protocol, including held-out intervention-family verification and task-utility evaluation.

Adaptive enforcement and normalization. Fixed clause weighting retains moderately high adaptation-family CSR but substantially reduces held-out CSR and worst-operator OSR. This indicates that static weighting does not adequately prioritize persistently unsatisfied operator types under conjunctive contract semantics. Clause-wise normalization is separately necessary: raw violation magnitudes differ across operators and evidence types, so optimizing unnormalized violations overemphasizes clauses with numerically larger scales rather than those with the greatest standardized non-compliance.

Intervention validity and family separation. Without admissibility filtering, H.O.-Family CSR falls to 51.4%, despite remaining adaptation-family performance of 67.8%. Candidate interventions that fail to isolate designated evidence, preserve task semantics, or satisfy the prescribed non-target modification bound can therefore provide misleading optimization signals. Reusing the adaptation intervention mechanism for model selection similarly raises Adapt.-Family CSR but produces a 21.2-point intervention generalization gap. The comparison uses disjoint adaptation and selection instances, showing that the degradation is attributable to mechanism-level reuse rather than instance-level leakage.

Preservation constraints. Removing `PRESERVE` constraints reduces task retention from 98.6% to 87.6% and worst-operator OSR from 77.1% to 49.5%. Thus, reductions in some clause violations can be achieved by broad changes in prediction behavior rather than selective control of the designated evidence. Explicit `PRESERVE` requirements are needed to rule out this degenerate route to apparent compliance.



Figure 7: **Qualitative examples of Evidence Contract verification** across datasets and operators: (a) Visual R.C. (IGNORE + REQUIRE), (b) TextVQA (REQUIRE), (c) M²CQA (PREFER), and (d) GQA (IGNORE + REQUIRE).

Table 15: **Specifications of qualitative verification cases in Figure 5.** Each column lists one case with its datset, contract clauses, designated evidence, held-out interventions, and clause-level verification outcome.

Case	(a) Visual R.C.	(b) TextVQA	(c) M^{2}CQA	(d) GQA
Contract clauses	`IGNORE` background; `REQUIRE` object	`REQUIRE` OCR content	`PREFER` visual over prior cue	`IGNORE` background; `REQUIRE` knee-pad attribute
Evidence	kitchen background; pizza	displayed text	visual evidence; conflicting prior	background; knee-pad attribute
Held-out interventions	background substitution; pizza-to-salad replacement	*STOP*-to-*STEP*; matched control	visual intervention; prior-cue intervention	background substitution; attribute modification
Outcome	**PASS**	**PASS**	**PASS**	**FAIL**: `REQUIRE`

D.6 Additional Qualitative Examples

Figure 7 shows the additional qualitative examples of evidence contract verification. The successful cases demonstrate selective rather than indiscriminate robustness. In (a), changing the study room background has a small effect (0.08), whereas replacing the pipette with a pen changes the prediction (effect 0.71). In (b), changing *Coca-Cola* to *Pepsi* produces an OCR effect of 0.72, compared with 0.09 for a matched control. In (c), visual evidence has a larger effect than the conflicting prior cue (0.76 versus 0.18), yielding a `PREFER` margin of $+0.58$.

D.6.1 Qualitative Case Specifications

This section provides the full specifications for the qualitative cases in Figure 5. Each case uses an intervention family held out from adaptation, threshold calibration, and model selection. All displayed interventions pass the benchmark-specific admissibility checks.

For each case, the displayed PASS/FAIL result is computed from the same confidence-adjusted clause criterion used for aggregate CSR. The examples are diagnostic illustrations and are not used for threshold selection, model selection, or hyperparameter choice.